Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Fußball, Städte, Daten

- Wechselwirkungen von dem Erfolg von Fußballmannschaften und Städten-

Max Bahne, Sarah Bock, Xenia Sataev

Ausarbeitung im Rahmen des Aufbauprojektes im Wintersemester 2015/2016

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Fußball, Städte, Daten

- Wechselwirkungen von dem Erfolg von Fußballmannschaften und Städten-

Ausarbeitung im Rahmen des Aufbauprojektes WS 2015/16

im Studiengang Next Media (M.A.)
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Ausgangslage

Wir leben heute in einer rundum vermessenen Welt. Sensoren in unseren Smartphones zeigen, wo wir sind, woher wir kommen und wohin wir gehen. Cookies zeichnen auf, wie wir uns auf einer Website bewegen und welche Inhalte uns gefallen. Die Smartwatch überwacht unseren Puls und misst jeden unserer Schritte.
Produkt dieser digitalen Vermessung der Welt ist ein riesiger Datenberg, der aufgeteilt in Bytes auf unzähligen Speichermedien rund um die Welt verteilt liegt. Unausgewertet sind die Daten nahezu wertlos. Erst durch Data-Mining lässt sich das Potential dieser erkennen, verwertbares Wissen generieren und in Mehrwert umsetzen.

Im zweiten Fachsemester des Studiengangs Next Media begeben wir uns auf die Suche nach spannenden Zusammenhängen, nach unentdecktem Wissen, das wir aus von uns gesammelten Daten extrahieren können. Unser Interesse gilt dabei dem Fußball und den Städten, die die Fußballvereine beherbergen. Gibt es Zusammenhänge zwischen Fußball- und den Städtedaten? Haben sie einen Einfluss aufeinander? Ändert sich die Stadt, wenn eine Fußballmannschaft Erfolg hat? Im Fokus steht allerdings auch die Frage nach der Umsetzbarkeit der Beantwortung dieser Fragen. Welche Daten bekommen wir überhaupt für unsere Analyse und wie können wir diese analysieren? Wird es möglich sein, die Daten in einem auswertbaren Format zu bekommen beziehungsweise sie soweit aufzuarbeiten, dass eine Analyse möglich wird? Welche Schritte müssen berücksichtigt werden? Welche Schwierigkeiten erwarten uns?

Im Mittelpunkt des Projekts steht demzufolge insbesondere auch die Frage, wie ein klassischer KDD-Prozess (**K**nowledge **D**iscovery in **D**atabases) funktioniert und verläuft. Das Ziel ist es, den KDD Prozess vollständig zu durchlaufen und zu dokumentieren.

Der Ablauf eines klassischen KDD-Prozesses gibt in Grundzügen die Gliederung unseres Projekts vor. Dabei halten wir uns an die Phasen des CRISP-DM Modells, das einen standardisierten KDD-Prozess beschreibt.

CRISP-DM[1] steht für Cross Industry Standard Process for Data Mining. Dieses branchenübergreifende Prozess-Modell wurde bereits Mitte der 1990er Jahre von Daimler Chrysler und SPSS entwickelt und wird heutzutage in der Praxis weit verbreitet genutzt. In dem CRISP-DM-Modell werden sechs Phasen unterschieden. Diese werden in der entsprechenden Reihenfolge sowie mit ihren Wechselwirkungen in Abbildung 1 veranschaulicht.

Abbildung 1 : Das Prozessmodell als Data Mining-Zyklus
(Quelle: an Anlehnung an Shearer[2])

Die Pfeile stellen die Abhängigkeiten und Wechselwirkungen zwischen den Phasen dar, wobei das Durchlaufen der Phasen nicht zwangsläufig in der angegebenen Reihenfolge erfolgen muss. Oft muss in der praktischen Folge in Projekten zwischen den Phasen hin- und hergewechselt werden um die Qualität der Untersuchung zu erhöhen.

Startpunkt eines Data-Mining Projektes ist zumeist das Geschäftsverständnis (Business Understanding). In dieser Phase werden die Ziele und Anforderung an das Projekt

[1] IBM SPSS Modeler CRISP-DM-Handbuch, Copyright IBM Corporation 1994,2012 (ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/de/CRISP-DM.pdf
[2] Quelle: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

festgelegt. Die Aufgaben- bzw. Fragestellung wird abgeleitet sowie eine grobe Vorgehensweise festgelegt. In unserer Projektarbeit betrifft diese Phase zunächst die Suche nach einem Thema. In dieser Phase wird beschrieben, wie wir das nun festgelegte und untersuchte Thema gefunden und gegenüber anderen Fragestellungen abgegrenzt haben. Die folgende Phase ist das Datenverständnis (Data Understanding). In diese Phase fällt die Datensammlung, -beschaffung sowie die erste Sichtung dieser. Dabei werden bereits mögliche Probleme ermittelt, die in der Datenqualität oder der Verfügbarkeit von Daten auftreten können. In diesem Abschnitt werden wir vermehrt von Schwierigkeiten berichten, die bei der Beschaffung gut aufbereiteter Daten auftreten. Weiterhin gehen wir auf Problematiken ein, die bei der Bereinigung von Datensätzen entstehen. Aus diesem Grund beschreiben wir in diesem Abschnitt ebenfalls die in dem Prozessmodell genannte dritte Phase; die Datenvorbereitung (Data Preparation). Die Phase beinhaltet die Konstruktion eines finalen Datensatzes für das darauffolgende Modeling, die Entwicklung eines Analysemodells. In diesem Teil wird beschrieben, welche Überlegungen zum Berechnungsmodell erfolgten und welche Punkte wir dabei zu berücksichtigten versuchten. Zum Schluss gehen wir auf unsere Ergebnisse der Analyse ein. Hier werden die Ergebnisse unserer Analyse zunächst erläutert und unter Berücksichtigung aufgetretener Problematiken beschrieben. Abschließend bieten wir in einem Fazit eine Zusammenfassung der Projektarbeit und arbeiten die relevantesten Kritikpunkte heraus.

2 Die Themensuche

Der Umfang von zur Verfügung stehenden Daten ist – wie bereits in der Einleitung angeführt – gigantisch. Eine Studie im Auftrag des IT-Dienstleisters EMC maß, dass im Jahr 2013 weltweit rund 4,4 Zettabytes an Daten erzeugt wurden[3]. Die Zahl klingt zunächst sehr abstrakt und nicht spektakulär. In einer Visualisierung der Studie wird jedoch der Umfang des jährlich erzeugten Datensatzes deutlicher. Würde man die 4,4 Zettabytes auf iPads mit jeweils 128 Gigabyte Speicher und jeweils einer Dicke von 7,5 Millimeter laden, würde der Stapel dieser iPads von der Erdoberfläche bis zum Mond

[3] Vgl. Turner, Vernon u.a.: The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, Framingham, 2014. Hier als PDF verfügbar: http://idcdocserv.com/1678.

reichen[4]. In Anbetracht des Umfangs dieses Datenbergs verwundert es nicht, dass der Senior Director of Search bei Microsoft, Stefan Weitz, Daten als Lingua Franca des 21. Jahrhunderts ausruft.

> „Natural language and analog multimedia like books, photographs, records, and film are no longer the dominant descriptors of the world. The lingua franca of todays physical descriptions are aspects like location, time, associated people, abilities, and visual representations.“[5]

Diese Daten - heruntergebrochen auf ihre Struktur – liegen alle digital vor.
An unsere Vorgehensweise stellten wir zunächst stets den Anspruch ein sinnvolles Ergebnis erarbeiten zu können. Dieses sollte wiederum einen Mehrwert bieten und in der Praxis anwendbar sein. So entstand die Idee einer Stauvorhersage aufgrund von Verkehrsdaten für die Stadt Hamburg. Ziel war eine Predictive Analysis, eine Vorhersageanalyse, durchzuführen und entsprechende Ausweich- und Vermeidungsrouten auszuspielen. Nach der notwendigen Recherche stellte sich allerdings heraus, dass Google nach dem Zukauf vom Kartendienst Waze diese Funktion bereits in den Google Maps-Service integriert hat. Aus diesem Grund aber auch aufgrund mangelnden Zugangs zu den notwendigen Daten verwarfen wir die Idee wieder.
Die darauffolgende Idee war ein Vorhersagemodell für Flug- beziehungsweise Bahnverspätungen. Eine kurze Recherche ergab, dass auch in diesem Bereich bereits mehrere Anbieter existieren. Bei der Vorhersage der Flugverspätungen kam erschwerend hinzu, dass der Flugverkehr von einer Vielzahl von Faktoren, wie etwa Wettervorhersage, Streiks, Zwischenfälle an Flughäfen oder politischen Ereignissen in den verschiedenen Ländern, abhängig ist. Diese große Zahl an potenziell relevanten Variablen würde ein zu komplexes Vorhersagemodell nach sich ziehen, das für den Umfang des Projektes nicht realistisch umsetzbar wäre. Somit entschieden wir uns auch in diesem Fall gegen das Thema.

---

[4] Vgl. ebd. Visualisierte Form als PDF: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf.
[5] S. Weitz, Stefan: Search: How the Data Explosion makes us smarter, Brookline, 2014, S. 39.

Um uns einem umsetzbaren und dennoch spannenden Thema zu nähern, war es notwendig, die eingangs beschriebene Herangehensweise zurückzulassen und mit einer Fragestellung zu beginnen. Es sollte somit nicht das sinnvolle Ergebnis vorangestellt werden, sondern eine interessante Frage, die wir durch eine Datenanalyse beantworten wollten. Nach wiederum unzähligen Themenentwürfen und Recherchen fanden wir das Thema unseres Projektes: Fußball. Das Ziel unseres Projektes war es demnach zu untersuchen, ob sich der Erfolg von Bundesliga-Mannschaften positiv auf die Entwicklung der Heimatstädte der Vereine auswirkt. Im selben Zuge stellte sich die Frage, ob auch umgekehrt eine positive Stadtentwicklung Einfluss auf den Erfolg der jeweiligen Vereine hat. Die Hypothese war dabei, dass der Erfolg von Bundesliga-Mannschaften – und damit in der höchsten deutschen Spielklasse – finanziell der jeweiligen Stadt zugute und somit auch die Attraktivität der Heimatstadt des Vereins gesteigert wird. Mittel- bis langfristig könne sich ein konstant hoher Erfolg eines Teams also positiv für die Stadtentwicklung auswirken.

3 Datenbeschaffung

Mit der oben genannten Fragestellung stiegen wir in die darauffolgende Phase der Analyse ein. Gemäß Colin Shearer, der 2000 mit seinem CRISP-DM-Modell einen wichtigen Grundstein für den Data Mining- bzw. den KDD-Prozess legte, begannen wir mit der zweiten Phase des bereits erläuterten Modells, dem Business und Data Understanding[6].

Es ist zu berücksichtigen, dass das CRISP-DM-Modell entwickelt wurde, um im betrieblichen Umfeld Erkenntnisse aus Datensätzen zu gewinnen, die einen monetären Mehrwert generierten. Dementsprechend hat Shearer in seiner Definition die einzelnen Schritte des Modells dem betrieblichen Nutzen angepasst. Wir stützten unser Projekt dennoch auf die Grundannahmen Shearers, vernachlässigten jedoch die erwähnten betrieblichen Aspekte und konzentrierten uns auf seine Ideen zur Knowledge Discovery from Databases. Laut Shearer beginne das Modell mit einem

[6] Vgl. Shearer, Colin: The CRISP-DM-Model. The New Blueprint for Data Mining, In: Journal of Data Warehousing, Volume 5, Number 4, Seattle, 2000, S. 14.

Data Mining-Problem, das zu einer Fragestellung führt[7]. Die Fragestellung hatten wir bereits erarbeitet und formuliert und somit das Ziel der Analyse festgelegt. Im zweiten Schritt, dem Data Understanding, gilt es zu erläutern, welche Daten für die Hauptfragestellung genutzt werden können und müssen, bzw. welche von diesen Daten vorhanden sind und zur Verfügung stehen. Somit stand gemäß unserer Frage das Verständnis davon an erster Stelle, welchen Zeitraum, welche Vereine und Vereinsdaten sowie welche dazugehörigen Städte und Stadtdaten wir für die Auswertung benötigten.

3.1 Untersuchungszeitraum

Da die Entwicklung einer Stadt ein langfristiger Prozess ist, entschieden wir uns für einen möglichst langen Zeitraum, der mit der Gründung der deutschen Bundesliga beginnen und mit der Saison 2014/2015 enden sollte. Diese wurde bereits 1963 gegründet und diese Tatsache bot uns somit einen Untersuchungszeitraum von über 50 Jahren.

Für den gewählten Untersuchungszeitraum ergaben sich jedoch zwei hauptsächlich zu nennende Probleme. Zunächst war das Saisonprinzip der Bundesliga problematisch, da die Platzierungen stets über einen Jahreswechsel erfolgten. Eine Bundesligasaison dauert demnach nicht vom 1. Januar bis zum 31. Dezember, sondern von Mitte August bis Mitte Mai. Dadurch war es problematisch die Platzierungsdaten direkt mit den Stadtdaten im selben Jahr zu vergleichen, da diese in einem Kalender-Jahreszeitraum erhoben werden.
Aufgrund dessen, dass wir keine kurzfristigen, sondern mittelfristigen Veränderungen untersuchen wollten, vermuteten wir, dass diese Problematik sich nicht zu stark auf die Ergebnisqualität ausüben würde.

Um Dopplungen zu vermeiden soll das erste Problem in diesem Abschnitt nur kurz angeschnitten und vor allem im letzten Abschnitt des dritten Kapitels noch etwas ausführlicher behandelt werden.

[7] Vgl. Shearer, Colin: The CRISP-DM-Model. The New Blueprint for Data Mining, S. 15.

Weiterhin ergab sich die Problematik, dass die für unsere Analyse relevanten Stadtdaten, zwischen 1963 bis 1991 fast durchgehend nicht zur Verfügung standen beziehungsweise zum großen Teil nicht existierten. Viele Daten standen auch aufgrund der deutschen Wiedervereinigung nicht zur Verfügung. So wurden sie entweder nicht erhoben oder aber mit unterschiedlichsten Methoden gemessen und festgehalten und sind daher qualitativ nicht vergleichbar.
Aufgrund der mangelnden Daten mussten wir den Untersuchungszeitraum deutlich verkürzen und wählten deshalb das Jahr 1991 als Startzeitpunkt und die Saison 2014/2015 als Endzeitpunkt der Analyse.

Somit umschließt unser Untersuchungszeitraum knapp die Hälfte der Bundesliga-Historie. Selbstverständlich ist dieser Zeitraum nur begrenzt nützlich um tatsächlich längerfristige Veränderungen in Städten zu erkennen, kurz- beziehungsweise mittelfristige Veränderungen sollten jedoch bereits in diesem Zeitraum zu erkennen sein.

3.2 Vereine und Vereinsdaten

Bei der Wahl der relevanten Fußballvereine entschieden wir uns aus Gründen der geringeren Komplexität dazu, nur Mannschaften mit einzubeziehen, die in dem festgelegten Untersuchungszeitraum mindestens eine Saison der ersten Bundesliga zugehörig waren. Obwohl die Zweite Liga ebenfalls zum Profibereich gezählt wird, beschlossen wir die Anzahl der zu untersuchenden Vereine zu begrenzen. Der ursprünglich gewählte Untersuchungszeitraum von 1963 bis heute umfasste 54 Vereine. Nachdem wir das Startjahr der Untersuchung auf 1991 bis heute reduzierten, blieben 41 relevante Fußballvereine, die in dem genannten Zeitraum mindestens eine Saison in der ersten Bundesliga spielten (vgl. Tabelle 1).

Tabelle 1: Übersicht über die 41 relevanten Vereine

VfB Stuttgart	Borussia Dortmund	Eintracht Frankfurt	1. FC Kaiserslautern	1. FC Nürnberg
Bayer Leverkusen	1. FC Köln	Karlsruher SC	Werder Bremen	Bayern München
Dynamo Dresden	FC Schalke 04	Borussia Mönchengladbach	Hamburger SV	VfL Bochum
Wattenscheid 09	Stuttgarter Kickers	Hansa Rostock	MSV Duisburg	Fortuna Düsseldorf
Bayer Uerdingen	1. FC Saarbrücken	SC Freiburg	VfB Leipzig	1860 München
Fortuna Düsseldorf	FC St. Pauli	KFC Uerdingen	Arminia Bielefeld	Hertha BSC Berlin
VfL Wolfsburg	SpVgg Unterhaching	SSV Ulm	Energie Cottbus	Hannover 96
1. FSV Mainz 05	Alemannia Aachen	1899 Hoffenehim (Sinsheim)	FC Augsburg,	SpVgg Greuther Fürth
Eintracht Braunschweig				

Nachdem die relevanten Vereine festgelegt wurden, folgte die Frage, wie der Erfolg der Vereine in unserer Analyse definiert wird und wie sich dieser messen lässt. Für eine Mannschaft, die in jeder Saison gegen den Abstieg spielt, wäre nicht abzusteigen und die Klasse zu halten bereits ein Erfolg. Ein Verein, der den Anspruch hat, möglichst weit vorne in der Tabelle platziert zu werden, wäre ein um Haaresbreite verhinderter Abstieg hingegen kein Erfolg.

Um eine einheitliche Definition festzulegen, bestimmten wir die Tabellenplatzierung einer Fußballmannschaft am Ende der jeweiligen Saison als Hauptfaktor. Neben diesem bestimmten wir drei weitere Faktoren für den Erfolg: Umsätze und Gewinne der Mannschaft sowie die Top-Spieler des Vereins, gemessen an der Anzahl der Tore, die der jeweilige Spieler erzielte.

Somit wurden folgende vereinsbezogenen Faktoren für die Untersuchung definiert:

- Platzierung
- Umsätze
- Gewinne
- Top-Spieler

Eine weitere Überlegung war, die Marktwerte der Mannschaften in die Auswertung einfließen zu lassen. Diese waren jedoch auf der einen Seite nur bis 2004 zu ermitteln und basierten auf der anderen Seite auf einer Schätzung von transfermarkt.de, deren Schätzungsgrundlage nicht offen zugängig und demzufolge für uns nicht nachvollziehbar war.
Die Schulden der Vereine betrachteten wir zwar auch als wichtige Faktoren, diese standen jedoch ebenfalls nur sehr vereinzelt zur Verfügung, sodass selbst bei der Hochrechnung der fehlenden Werte kein konsistenter Datensatz gebildet werden konnte.

Dieselbe Problematik bestand bei den Umsatz- und Gewinn-Daten, die zu großen Teilen von der Seite fussball-geld.de stammen oder aus Presseberichten herausgefiltert wurden. Grund für die fehlende Verfügbarkeit der Finanzdaten der Bundesligisten ist hauptsächlich, dass es sich nicht bei allen Vereinen um Aktiengesellschaften handelt und diese daher größtenteils nicht verpflichtet sind Geschäftsberichte zu veröffentlichen. In diesen Fällen stammen die Daten aus anderen Quellen wie Presseberichten oder Eigenauskünften der Vereine, die wir manuell gesammelt haben.
Abbildung 2 verdeutlicht, wie viele Vereine der Bundesliga in unserem Untersuchungszeitraum einen Geschäftsbericht aufweisen müssen.

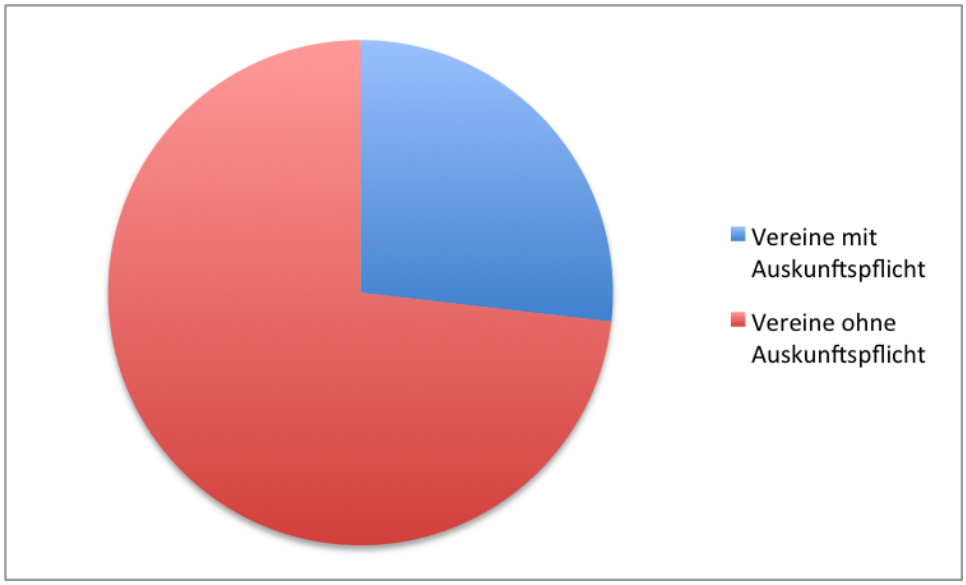

Abbildung 2 : Finanzielle Auskunftspflicht des Bundesligavereine
(Quelle: eigene Darstellung)

Um die benötigten Daten für die relevanten Vereine aufzubereiten, wurden zunächst alle Mannschaften, die von 1991 bis 2014 in der ersten Bundesliga spielten, in einer Excel-Tabelle erfasst. Dabei bildeten die Vereine die Spalten und die Jahre die Zeilen.
Alle Vereine, die nicht unter die Top 18 und somit in die erste Bundesliga des aktuellen Jahres kamen, bekamen den Wert 0 zugeschrieben. Die eigentliche Platzierung, die die Nummer 1 für den besten und Nummer 18 für den schlechtesten Verein vorsieht, wurde in umgekehrter Reihenfolge in Punkte umgewandelt. Somit bekam die erfolgreichste Mannschaft 18 Punkte, der Verein auf dem letzten Tabellenplatz einen Punkt. Auf diese Weise war die Platzierung der einzelnen Vereine mit den restlichen zu untersuchenden Variablen vergleichbar, da wir angenommen haben, dass eine positive Stadt- oder Vereinsentwicklung von steigenden Werten charakterisiert wird. Die jeweiligen Werte wurden manuell Mannschaft für Mannschaft in die Tabelle eingefügt.

Bereits an dieser Stelle wird deutlich, dass die Verfügbarkeit beziehungsweise das Fehlen der relevanten Daten eine große Herausforderung bei dieser Analyse darstellt. Gleichzeitig sind die Festlegung von messbaren Parametern und die genaue Definition dieser unabdingbar.

3.3 Stadtdaten

Um die Fußballdaten mit den Stadtdaten vergleichen zu können, sammelten wir die Daten aller Städte, die die 41 für die Analyse relevanten Fußballvereine beherbergen. Insgesamt waren für unsere Untersuchung 36 Städte von Bedeutung.
Eine genaue Übereinstimmung mit der Anzahl der Vereine ergibt sich nicht, da es Städte gibt, zu denen mehrere Vereine zugehörig sind. So teilen sich beispielsweise der HSV und der FC St. Pauli dieselbe Stadt Hamburg sowie der FC Bayern München und 1860 München beide zu der Stadt München gehören.

Als Auswahlkriterien für die stadtrelevanten Daten war es wichtig, Indikatoren zu finden, die für eine Stadtentwicklung ausschlaggebend sind und diese entsprechend

widerspiegeln. Für den festgelegten Untersuchungszeitraum legten wir folgende Daten der jeweiligen Städte fest:

- Einwohnerzahl
- BIP
- Brutto-Durchschnittseinkommen der Einwohner
- verfügbares Einkommen der Einwohner
- Anzahl der Erwerbstätigen
- Wetterdaten (bestehend aus Durchschnittstemperatur, Niederschlag und Sonnenstunden)

Bei der Beschaffung der stadtbezogenen Daten entstanden ähnliche Schwierigkeiten wie bei den Daten zu den Bundesligavereinen. Nach langwieriger Recherche und Anschreiben relevanter Stellen, stellte uns das Bundesamt für Bau-, Stadt- und Raumforschung (BBSR) den Großteil der Daten für den gesamten relevanten Zeitraum 1991 bis 2015 zur Verfügung.

Zwar standen die Daten auch für den Zeitraum von 1963 bis 2015 zur Verfügung, aufgrund bereits beschriebener Schwierigkeiten, war es jedoch nicht möglich, den vollständig zur Verfügung stehenden Zeitraum zu nutzen. Zudem wurden Daten laut BBSR für die Jahre vor 1991 zum Teil mit anderen Messmethoden sowie durch andere zuständige Institutionen erhoben und sind somit nicht gut mit den Jahren ab 1991 vergleichbar. Grund dafür sei die bereits erwähnte Zweiteilung Deutschlands in BRD und DDR, denn erst nach der Wiedervereinigung 1990 seien verlässliche und vergleichbare Daten produziert beziehungsweise einheitlich erhoben worden.

Wie bereits erwähnt umfasst der reduzierte Untersuchungszeitraum 36 relevante Städte, denen jeweils mindestens ein relevanter Bundesligaverein zugehörig ist.
Uns lagen Daten über die Anzahl der Erwerbstätigen als auch die Einwohnerzahlen vor. Da man jedoch davon ausgehen kann, dass diese aufgrund der steigenden Bevölkerungszahlen von Jahr zu Jahr ebenfalls ansteigen, betrachteten wir den Quotienten der beiden Werte, in dem wir die Zahl der Erwerbstätigen pro Einwohner

berechneten. Auf diese Weise ließ sich zum einen die tatsächliche Entwicklung darstellen, zum anderen der Vergleich der verschiedenen Städte untereinander bewerkstelligen.

Die Wetterdaten sind zwar nicht relevant hinsichtlich der Frage nach dem Einfluss der Fußballmannschaften auf die Stadtentwicklung, könnten jedoch andersrum einen Einfluss auf die Ergebnisse einer Fußballmannschaft haben. Zudem wollten wir Faktoren mit in die Analyse aufnehmen, die auf den ersten Blick keinen Einfluss haben. Aus diesem Grund entschieden wir uns dafür diese aufzunehmen.

Der Bezug der Wetterdaten war nach zeitintensiver Suche zwar möglich, jedoch ebenfalls unvollständig. Die Website wetterkontor.de, von der die meist angegebenen Wetterdaten stammen, zieht seine angegebenen und veröffentlichen Daten nach eigenen Angaben vom Deutschen Wetterdienst. Die verwendeten Tabellen zeigen Jahreswerte für Temperatur (Grad), Niederschlag (Liter pro Quadratmeter) und Sonnenschein (Stunden) als Mittelwert für das gesamte Jahr (vgl. Abbildung 3).

Jahreswerte

Zeitraum	Temperatur		Niederschlag		Sonnenschein	
	Mittel	Abw.	Summe	Abw.	Summe	Abw.
1996	7,5	-1,9	491,1	62%	1483,9	94%

Abbildung 3: Tabellenausschnitt der Wetterdaten der Stadt Hamburg im Jahre 1996 (Quelle: wetterkontor.de)

Die fehlenden Werte waren insbesondere im Ruhrgebiet und Umgebung zu verzeichnen und waren dort nicht oder nur sehr vereinzelt verfügbar. Wir vermuten, dass dieser Sachverhalt aus der Tatsache der verschlechterten Luftverschmutzung, die aus der starken Industrie in dem Gebiet herrührt, zurückzuführen ist. Dies blieb jedoch weitestgehend eine Vermutung, da die zeitliche Verfügbarkeit für eine Nachverfolgung dieser Hypothese leider ausblieb.

Die nicht vorhandenen Werte in bestimmten Jahren sowie lückenhafte Daten in einigen Städten wurden von uns manuell durch Mittelwerte ersetzt. Dabei wurde für eine Stadt der Mittelwert über die Wetterdaten aller vorhandenen Jahre gebildet und für den fehlenden Wert eingesetzt. Fehlten hingegen alle Jahreswerte zu einer bestimmten Stadt, wurden die vorhandenen Werte der nächstgelegenen Stadt eingesetzt, so z.B. die Wetterdaten von Hannover für die fehlenden Daten von Braunschweig.

Ein weiterer interessanter Faktor waren die Mietpreise in den unterschiedlichen Städten. Doch auch dieser stand uns trotz langwieriger Recherche nicht zur Verfügung.
Die Seite wohnungsboerse.net beispielsweise verfügt zwar über einen Mietspiegel der meisten Städte in den letzten vier Jahren, es ist jedoch – abgesehen von dem Fehlen der restlichen Jahre – nicht ersichtlich, woher diese Daten stammen, wie diese erhoben wurden und ob diese valide sind. Die Auskunft der Städte selbst ist zwar teilweise vorhanden, variiert aber hinsichtlich der zur Berechnung herangezogenen Faktoren stark untereinander. So stellte ein Großteil der Städte ein PDF-Dokument mit dem Mietspiegel des vergangenen Jahres zur Verfügung, einige Städte sogar nur in gedruckter Form. Lediglich zwei Städte (Bochum und Berlin) boten ein Archiv der Mietspiegel an. Aufgrund der Menge der von uns benötigten Daten ist ein manuelles Rauskopieren beziehungsweise Abschreiben der Daten nicht realisierbar. Aus diesem Grund entschieden wird uns, auch die Mietpreisdaten zu ignorieren.

Da auch der Mietpreis einer Stadt ein bedeutender Indikator für die Stadtentwicklung ist, ist das Fehlen dieser Werte durchaus erheblich. Es besteht bei diesen Daten eine ähnliche Problematik wie bei den Stadtdaten vor 1991. Es gibt in Deutschland keine einheitliche Stelle, die eine verlässliche Aufnahme, Dokumentation sowie Aufbereitung der Daten in der Bundesrepublik zur Aufgabe hat.

4 Entwicklung eines Analysemodells

Mit den teilweise problematischen Ergebnissen der ersten beiden Phasen der Analyse folgten wir der Reihenfolge des CRISP-DM-Modells in die kommenden beiden Phasen, der Datenbereinigung (Data Preparation) und dem Modell (Modeling).
Bei der Datenbereinigung wird der Datensatz in ein analysefähiges Format transferiert. Zu diesem Zweck müssen fehlende Daten erkannt und gegebenenfalls durch andere ersetzt werden. Dies kann beispielsweise durch den Einsatz von Mittelwerten oder Schätzwerten durchgeführt werden.
Weiterhin müssen fehlerhafte Daten und Ausreißer erkannt und auch hier gegebenenfalls ergänzt, ersetzt oder eliminiert werden. Nicht zuletzt müssen die zur Verfügung stehenden Daten laut Shearer an einem Ort zusammengebracht werden um diese dann schlussendlich auch analysieren und gegebenenfalls direkt vergleichen zu können[8].
Im Zentrum der darauffolgenden Phase, dem Modeling, stand der Entwurf eines Modells, das unseren Datensatz entsprechend auffassen und verwerten kann. Das Ziel des Modells ist eine Korrelation zwischen Stadt- und Bundesligadaten zu ermitteln. Für die Überprüfung eines potenziellen Zusammenhangs dieser Daten, setzen wir zwei verschiedene Methoden ein. Zum einen die deskriptive Analyse und zum anderen die logistische Regressionsanalyse mit der Statistik-Software R.

4.1 Das Bereinigen der Daten

Einen hohen Anteil der Datenbereinigung nahm die Ergänzung von fehlenden Daten in Anspruch. Bei der Vorbereitung der Daten arbeiteten wir vor allem mit Microsoft Excel.
Explizit in den Fußball-Datensätzen ergaben die Untersuchungen eine Vielzahl an fehlenden Daten. Problematisch waren dabei vor allem die bereits beschriebenen fehlenden Auskünfte der Vereine über die eigene Finanzsituation. Weiterhin ergaben sich verändernde Vereinsstrukturen bei einigen Vereinen. So löste sich

[8] Vgl. Shearer, Colin: The CRISP-DM-Model. The New Blueprint for Data Mining, In: Journal of Data Warehousing, Volume 5, Number 4, Seattle, 2000, S. 16.

beispielsweise der VfB Leipzig zur Saison 2004/2005 auf, aus dem KFC Uerdingen wurde Bayer 05 Uerdingen.
Zudem ließen sich von Vereinen, die zu 100 Prozent einem Konzern angehören, generell nur erschwert oder gar keine Finanzdaten einsehen. Bereits bestehende und ermittelte Finanzdaten standen zum Teil weiterhin lediglich bis in das Jahr 2010 zur Verfügung, der Zeitraum bis 2014 musste daher in den meisten Fällen extrapoliert oder geschätzt werden.

Hilfreich war dabei die Aussage von Liga-Präsident Reinhard Rauball, der sich auf eine Berechnung berief, die ergab, dass 1991 der Gesamtumsatz aller Vereine der Bundesliga bei 190 Millionen Euro lag[9]. Auf dieser Basis war es möglich anhand der bereits gesammelten Daten und der Tabellenkonstellation der Vereine sowie der Beteiligung an internationalen Wettbewerben, (die zusätzliches Geld in die Vereinskassen brachte), die fehlenden Daten zu ergänzen.
Wie bereits eingangs erwähnt, lagen uns die Daten der Platzierung der Mannschaften saisonal vor. Um die Fußball- mit den Städtedaten zu vereinheitlichen, definierten wir deshalb beispielweise den Wert der Saison 2013/2014 als den Wert für das Jahr 2014.

Ähnliche Problematiken wie die soeben beschriebenen, ergaben sich bei den Städtedaten – wenn auch in deutlich geringerem Umfang. Die vom BBSR zur Verfügung gestellten Daten waren im Zeitraum von 1992 bis 1995 für jede Stadt, jedoch nicht für jede Kategorie lückenhaft. Für Rostock beispielweise fehlten jedoch bis 1992 die Daten in allen Kategorien.
Eine weitere Schwierigkeit in den Daten war, dass nur Kreise und kreisfreie Städte aufgeführt waren. Nicht alle Städte, die einen Bundesligaverein beherbergen oder beherbergten, sind jedoch kreisfrei. So fanden sich zum Beispiel Sinsheim (1899 Hoffenheim) oder auch Uerdingen (Bayer 05 Uerdingen) nicht in der uns zugesandten Liste. Einige fehlende Daten mussten wir anhand anderer zur Verfügung stehender Daten schätzen oder berechnen. Die fehlenden Daten des jeweiligen BIP

berechneten wir beispielsweise anhand des Bevölkerungsanteils der Stadt an der Bevölkerung des Kreises, wodurch sich auch der Anteil der Stadt am BIP abschätzen ließ.

Auch die Bereinigung der vorliegenden Wetterdaten brachte, wie bereits erwähnt, Schwierigkeiten hervor. Diese lagen vor allem für das Ruhrgebiet fast ausschließlich nicht vor und mussten extrapoliert werden. Dies geschah anhand der Wetterdaten der nächstgelegenen Stationen.

Die einzeln gesammelten, erhobenen und umgewandelten Daten, die in unterschiedlichen Tabellen und in verschiedenen Formen vorlagen, mussten im letzten Schritt in eine einheitliche Form gebracht werden, um sie vergleichbar zu machen und in eine große Gesamttabelle umzuwandeln. Bei diesem Prozessschritt bemerkten wir an einigen Stellen, dass wir auch da – trotz vorher festgelegter einheitlicher Struktur – unterschiedliche Ausgangstabellen erstellt haben. Dies lag insbesondere daran, dass wir die Struktur offenbar nicht klar genug definiert haben und somit jeder Projektteilnehmer ein unterschiedliches Verständnis davon hatte. Dies betraf allerdings nur einige Tabellen und konnte mit verhältnismäßig geringem Aufwand behoben werden.

Umsatz und Gewinn der Bundesligavereine in Mio €

Umsatz	1991-1992	1992-1993	1993-1994	1994-1995	1995-1996	1996-1997	1997-1998	1998-1999	1999-2000	2000-2001	2001-2002	2002-2003	2003-2004	2004-2005	2005-2006
1.FC Kaiserslautern	3,2	4,3	3,8	4,2	5,7	6,5	24,5	36,5	50,3	63,291	47,195	40	39,8	39,4	35,1
1.FC Köln	12,3	15,6	17,6	20,4	24,3	22,2	21,6	10,3	12,4	23,8	29,6	15,6	23,9	37,6	40,5
1.FC Nürnberg	9,5	12,5	14,3	6,7	5,4	2,3	7,6	23,5	13,2	14,5	34,5	33,2	16,5	37,8	38,7
1.FC Saarbrücken	0	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.
1.FSV Mainz 05	3,2	2,6	2,8	3	3,1	4,2	4,3	3,9	4,5	4,6	4,2	4,7	5,6	24,2	27,6
1860 München	2,5	3,1	2,3	12,4	15,4	13,6	15,6	20,4	19,7	21,6	25,4	34,3	27,5	20,4	23,4
1899 Hoffenheim (Sinsheim)	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,9	1	1,1	1,2	2,3	4,5
Alemannia Aachen	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	1,5	1,7	2,1	1,6	1,7	1,4	0,45
Arminia Bielefeld	0,1	0,1	0,1	0,1	3,4	10,3	9,7	12,3	31,2	20,3	21,3	23,4	24,5	27,4	31,3
Bayer Leverkusen	0	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.
Bayer Uerdingen	0	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.	k.a.
Bayern München	34,5	43,2	54,3	67,8	76,8	75,4	83,9	84,5	110,2	123,5	142,8	154,09	160,99	189,5	204,7
Borussia Mönchengladbach	13,4	15,4	14,5	34,5	23,5	32,5	31,4	40,4	41,2	37,6	38,6	45,3	47,5	51,2	55,6
BorussiaDortmund	23,6	24,5	31,2	44,5	54,3	65,7	64,5	77,4	90,5	104,5	110,3	124,42	94,95	73,95	83,26
Dynamo Dresden	3,45	4,5	5,4	6,7	2,1	1,3	1,6	2,8	3,4	2,7	2,4	2,3	3,6	3,6	3,9
Eintracht Braunschweig	1,2	1,45	1,1	1,6	2,4	3,4	4,4	3,9	4,1	3,9	4,7	10,5	7,5	12,3	14,3
EintrachtFrankfurt	8,9	9,6	10,4	11,3	9,8	11,2	10,5	12,3	14,5	16,4	17,6	18,7	24,3	26,8	32,8
Energie Cottbus	1,1	2,3	2,1	1,8	2,3	3,4	4,5	4,7	6,7	8,5	10,6	11,5	16,4	20,3	21,5
FC Augsburg	1,1	1,4	1,5	1,3	2,4	2,1	5,6	6,4	7,4	9,5	10,2	9,4	11,2	13,2	11,2
FC Schalke 04	24,5	31,2	44,3	54,2	56,2	67,2	72,8	79,1	87,5	85,4	95,4	98,5	104	102	142
FC St.Pauli	5,6	3,2	4,5	6,5	7,6	23,6	15,6	8,6	7,9	9,5	11,3	12,3	13,4	16,4	20,4
Fortuna Düsseldorf	9,6	4,3	2,2	5,6	16,7	19,8	20,3	12,6	11,5	9,8	10,2	11,8	10,3	9,6	8,7
Hamburger SV	12	13,4	14,3	16,7	25,6	36	37,8	44,7	55,2	61,3	72,6	75,6	82,3	83	102
Hannover 96	2,3	4,5	6,7	8,7	7,4	10,3	11,3	10,7	12,4	15,4	25,6	27,5	31,4	36,4	42,3
Hansa Rostock	8,7	7,6	9,8	12,5	13,4	17,6	19,8	23,4	27,8	29,7	31,2	32,3	35,7	34,2	33,2
Hertha BSCBerlin	2,3	2,6	3,1	2,9	3,2	4,5	7	19,6	23,4	27,6	34,2	41,8	44,6	51,4	54,3
Karlsruher SC	8,7	9,6	10,4	11,5	12,3	15,6	13,5	6,7	15,8	3,4	5,6	7,6	12,3	13,4	14,6
MSV Duisburg	3,4	4,3	5,6	6,4	10,3	12,5	14,3	12,3	9,6	9,4	10,2	11,2	12,5	13,8	21,3
SC Freiburg	4,3	4,8	5,1	4,9	7,5	13,2	11,3	16,5	15,7	17,8	17,3	16,5	19,5	19,1	20,5
SpVgg Greuther Fürth	1.1	1.2	0.9	1	1.43	2.4	2.5	9.5	10.4	12.4	14.5	16.7	17.5	16.7	18.9

Abbildung 4: Umsätze der Fußballvereine
(Quelle: Eigene Darstellung)

[9] Vgl. http://www.focus.de/sport/fussball/bundesliga1/tid-26910/50-jahre-bundesliga-die-bundesliga-in-dekaden-die-90er-jahre-_aid_799402.html, zuletzt eingesehen am 17.01.16 um 14:22.

	A	B	C	D	E	F	G	H	I
		1. FC Kaiserslautern	1. FC Köln	1. FC Lokomotive/VfB Leipzig	1. FC Nürnberg	1. FC Saarbrücken	1. FSV Mainz 05	1899 Hoffenheim (Sinsheim)	Alemannia Aachen
2	1991	18	8	0	2	0	0	0	0
3	1992	13	11	0	14	0	0	0	0
4	1993	9	6	0	4	2	0	0	0
5	1994	15	10	1	4	0	0	0	0
6	1995	12	6	0	0	0	0	0	0
7	1996	3	4	0	0	0	0	0	0
8	1997	0	9	0	0	0	0	0	0
9	1998	18	1	0	0	0	0	0	0
10	1999	11	0	0	6	0	0	0	0
11	2000	11	0	0	0	0	0	0	0
12	2001	6	14	0	0	0	0	0	0
13	2002	12	1	0	5	0	0	0	0
14	2003	4	0	0	2	0	0	0	0
15	2004	1	6	0	0	0	0	0	0
16	2005	1	6	0	0	0	0	0	0
17	2006	1	2	0	13	0	9	0	0
18	2007	0	0	0	15	0	3	0	4
19	2008	0	0	0	7	0	0	0	0
20	2009	0	8	0	0	0	0	17	0
21	2010	0	4	0	3	0	9	10	0
22	2011	10	7	0	13	0	17	11	0
23	2012	1	3	0	6	0	12	8	0
24	2013	0	0	0	6	0	12	2	0
25	2014	0	0	0	4	0	11	13	0

Abbildung 5: Platzierung der Fußballvereine (Quelle: Eigene Darstellung)

Auch die Umwandlung der Städte- und Fußballdaten in ein einheitliches Format stellte es sich als zeitintensiv heraus. So musste manuell und teils durch Verweise jeder einzelnen Mannschaft eine Stadt zugeordnet werden. Die gesammelten Daten wandelten wir so um, dass wir für jedes Jahr und jede Kategorie die Differenz zum Vorjahr berechneten. Somit wurden für das Jahr 2014 die Differenz zum Vorjahr hinsichtlich der Platzierung, der Topspieler, des Umsatzes und des Gewinns einer Mannschaft zugeschrieben. Dasselbe wurde mit den Städtedaten durchgeführt und die Differenz dieser zum Vorjahr errechnet. Auf diese Weise war ein Vergleich der Daten über die Steigung beziehungsweise Stagnation der Werte möglich.

Am Ende der Datenbereinigungsphase stand eine Gesamttabelle, die über 920 Zeilen verfügte. Jede Zeile bildete eine Fußballmannschaft in einem bestimmten Jahr mit den dazugehörigen Werten aus jeder Kategorie.

	A	B	C	D	E	F	G	H	I	J	K
1	Stadt	Platzierung	Topspieler	Umsatz	Gewinn	BIP	durchschn_Einkommen	Verfügb_Einkommen	Erwerbst_EW	Niederschlag	Temperatur
2	Kaiserslautern_1992	0	-5	1,10	-0,22	44,09	214,30	-299,58	-0,01	152,90	0,80
3	Kaiserslautern_1993	0	-4	-0,50	0,13	0,91	214,30	299,58	-0,01	41,90	-0,50
4	Kaiserslautern_1994	1	6	0,40	0,09	0,76	214,30	-299,58	-0,01	28,10	1,20
5	Kaiserslautern_1995	0	-3	1,50	0,86	-11,73	214,30	299,58	0,03	-126,00	-0,80
6	Kaiserslautern_1996	0	-9	0,80	3,5	539,63	214,30	-299,58	-0,20	-66,60	-1,60
7	Kaiserslautern_1997	0	-3	18,00	5,5	-292,98	214,30	299,58	0,20	151,30	1,50
8	Kaiserslautern_1998	1	18	12,00	-3,8	43,23	214,30	-299,58	-0,04	192,00	0,10
9	Kaiserslautern_1999	0	-7	13,80	-1	162,30	214,30	299,58	-0,02	-203,10	0,50
10	Kaiserslautern_2000	0	0	12,99	1,1	-140,42	214,30	-388,00	-0,02	109,30	0,30
11	Kaiserslautern_2001	0	-5	-16,10	-15,9	-67,11	184,06	364,12	0,00	-8,90	-0,80
12	Kaiserslautern_2002	1	6	-7,20	12,13	47,74	332,00	-310,00	-0,02	-66,20	0,40
13	Kaiserslautern_2003	0	-8	-0,20	-9,63	185,95	413,00	168,00	-0,02	-262,50	0,20
14	Kaiserslautern_2004	0	-3	-0,40	8,7	65,50	281,00	-223,00	-0,02	150,60	-0,70
15	Kaiserslautern_2005	1	0	-4,30	-1,39	-176,18	-524,00	219,00	0,00	36,80	0,30
16	Kaiserslautern_2006	0	0	-7,70	-0,66	116,22	392,00	-48,00	-0,02	9,30	0,20
17	Kaiserslautern_2007	0	-1	-7,60	-1,53	79,52	232,00	453,00	-0,02	164,80	0,20
18	Kaiserslautern_2008	0	0	0,90	-0,02	20,67	426,00	-428,00	-0,02	-114,20	-0,40
19	Kaiserslautern_2009	0	0	0,60	-0,02	-44,43	-194,00	327,00	0,00	-1,60	-0,10
20	Kaiserslautern_2010	0	0	21,00	4,12	110,29	502,00	-562,00	0,01	-74,60	-1,10
21	Kaiserslautern_2011	1	10	8,60	2,2	80,75	414,00	362,00	-0,02	-26,30	1,60
22	Kaiserslautern_2012	0	-9	-18,30	-6,4	-69,45	621,00	-188,96	0,01	9,10	-0,50
23	Kaiserslautern_2013	0	-1	6,20	2,06	82,03	-307,04	153,00	-0,03	-112,50	-0,30
24	Kaiserslautern_2014	0	0	-2,30	0,08	33,10	209,09	-299,58	-0,01	166,20	1,50
25	Koeln_1992	1	3	3,30	0,8	933,73	351,01	-292,03	-0,02	289,60	0,90
26	Koeln_1993	0	-5	2,00	0,3	1377,27	351,01	292,03	-0,02	-19,00	-0,70
27	Koeln_1994	1	4	2,80	0,4	982,01	351,01	-292,03	-0,01	-137,50	1,30
28	Koeln_1995	1	-4	3,90	0,3	1479,35	351,01	292,03	0,02	24,20	-0,60
29	Koeln_1996	0	-2	-2,10	0,2	-261,81	351,01	-292,03	-0,01	-189,70	-1,80
30	Koeln_1997	1	5	-0,60	-1,7	1371,91	351,01	292,03	-0,01	180,70	1,60
31	Koeln_1998	0	-8	-11,30	-2,7	778,26	351,01	-292,03	-0,04	161,90	0,10
32	Koeln_1999	0	-1	2,10	0,8	332,36	351,01	292,03	-0,05	-158,60	0,60
33	Koeln_2000	0	0	11,40	1,9	-814,68	351,01	-305,00	-0,06	116,10	0,10
34	Koeln_2001	1	14	5,80	0,1	2071,56	537,74	10,00	-0,02	-22,90	-0,70
35	Koeln_2002	0	-13	-14,00	-0,2	-518,60	564,00	-650,00	-0,01	-8,50	0,60

Abbildung 6: Gesamttabelle aller Vereine mit den jeweiligen Vereins- und Städtedaten (Quelle: Eigene Darstellung)

Insgesamt betrachtet waren die beiden Phasen der Datenbeschaffung und der Datenbereinigung die mit Abstand zeitaufwändigsten Phasen der Analyse und des gesamten Projektes. Hinsichtlich der Tatsache, dass das Projekt eine Projektpräsentation nach sich zieht und bei der Zeit und Arbeit, die in dieses reingesteckt wurden, war die Frustration zeitweise hoch, da für die eigentliche Analyse und die Ergebnisse und somit die Beantwortung der Fragestellung kaum noch Zeit blieb.

4.2 Angewandte Analysemethoden

Gemäß Shearer steht zu Beginn der Analyse der Daten die Wahl der Analysemethoden[10]. Im Fall der Fußballdaten haben wir uns für die oben genannten zwei Methoden entschieden. Wir begannen mit der deskriptiven Analyse, die wir in Excel durchführten.

Für die durchzuführende Analyse und Vergleichsrechnungen war es unerlässlich unsere vorliegenden Daten zu vereinheitlichen. Der Vergleich von Daten

[10] Vgl. Shearer, Colin: The CRISP-DM-Model. The New Blueprint for Data Mining, S. 17.

unterschiedlichster Metriken führt zu keinem brauchbaren Ergebnis, sofern diese nicht vergleichbar gemacht werden.
Im ersten Schritt berechneten wir für alle Fußballvereine und alle Städte die Differenz der Daten des aktuellen Jahres zum Vorjahr. Im zweiten Schritt berechneten wir eine Wachstumsrate der Differenz-Tabelle mittels der Formel:

„=(([berechnete Differenz]*100)/MAX([erste Zelle der Differenz-Tabelle der Stadt/des Vereins]:[letzte Zelle der Differenz-Tabelle der Stadt/des Vereins]))“.

Dank dieser zwei Schritte ließen sich die vorliegenden Daten miteinander vergleichen.

Bei der darauffolgenden logistischen Regression mussten die Daten ebenfalls entsprechend aufbereitet und vergleichbar gemacht werden. Wie im obigen Transformationsprozess berechneten wir dafür die Differenz der verschiedenen Datensätze.

4.2.1 Deskriptive Analyse

Am Ende der Datenbereinigung der ersten Methode lag eine Tabelle mit den Wachstumswerten der Fußball- und Stadtdaten vor, die nun miteinander vergleichbar waren.
Problematisch beim Vergleich dieser Daten war allerdings, dass mögliche Auswirkungen des Erfolgs des Bundesliga-Teams auf die Stadtentwicklung nicht im selben Jahr beziehungsweise in der derselben Saison sichtbar werden würden, in der das Team den Erfolg erringt, sondern in den Folgejahren. Weil sich dieser Zeitraum nicht eindeutig bemessen lässt, konnte die KORREL-Funktion, die Excel anbietet, um Werte zu vergleichen, nicht angewendet werden. Stattdessen fanden wir eine Möglichkeit, die Daten anderweitig zu vergleichen. Aus den vorliegenden Daten erstellten wir Liniendiagramme, die auf den ersten Blick recht unübersichtlich waren (vgl. Abbildung 7).

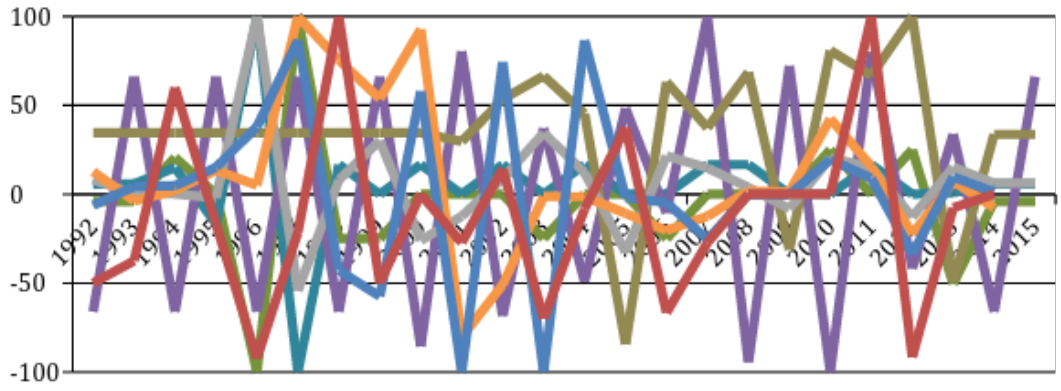

Abbildung 7: Zusammenfassung aller Daten in einem Liniendiagramm am Beispiel Kaiserslautern (Quelle: Eigene Darstellung)

Durch das Ein- und Ausblenden von Linien ließen sich aber so über einen gewissen Zeitverlauf Zusammenhänge sichtbar machbar. Großer Nachteil dieser Methode ist allerdings – wie es Cukier und Mayer-Schönberger in ihrem Buch zu Big Data an vielen Stellen formulieren, [...]dass nur gezeigt wird, *dass* ein Zusammenhang besteht, aber nicht, *warum* ein Zusammenhang besteht[11]. Ein möglicher Zusammenhang zeigte sich zum Beispiel in Wolfsburg:

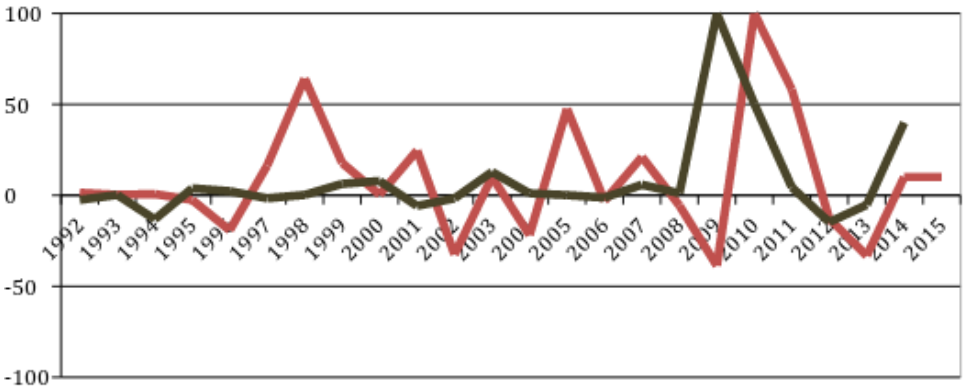

Abbildung 8: Scheinbarer Zusammenhang zwischen BIP der Stadt und dem Gewinn des VfL Wolfsburg (Quelle: Eigene Darstellung)

Vorzugsweise in den vergangenen zehn Jahren wird deutlich, dass womöglich auf ein Jahr mit gestiegenem oder gefallenem Gewinn (dunkelgrüne Linie) beim VfL Wolfsburg auch eine vergleichbare Veränderung im BIP der Stadt Wolfsburg (rote

[11] Vgl. Mayer-Schönberger, Viktor; Cukier, Kenneth: Big Data . Die Revolution, die unser Leben verändern wird, München, 2013, S. 13 f.

Linie) erfolgte. Mit dieser Methode erarbeiteten wir für 36 Städte auf insgesamt 85 mögliche Korrelationen. Das entspricht 2,4 Korrelationen pro Stadt. Auf diese wird im Folgenden genauer eingegangen.

Von der Gesamtzahl der Städte war es weiterhin notwendig, diejenigen Städte abzuziehen, die für eine deskriptive Analyse zu wenige Daten hatten. Schließlich konnten für diese Städte zwangsläufig keine Zusammenhänge festgestellt werde.

Grund für die geringe Datendichte war entweder die zu kurze Zeit in der Bundesliga oder schlichtweg fehlende Finanz-Daten wie zum Beispiel bei Bayer Leverkusen oder Bayer Uerdingen. In solchen Fällen sah das erstellte Liniendiagramm[12] zum Beispiel wie folgt aus:

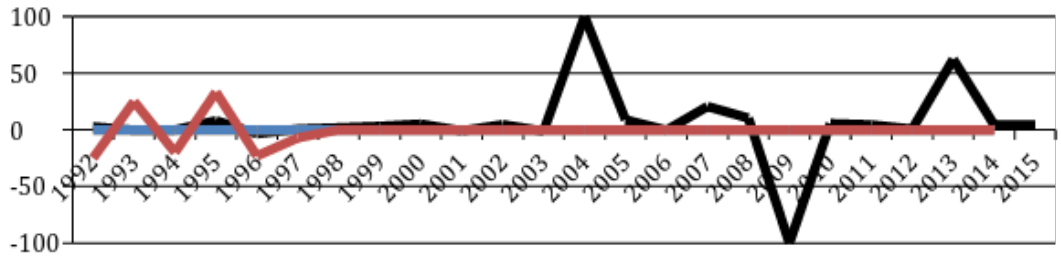

Abbildung 9: Fußball-Daten von Bayer Uerdingen in Zusammenhang mit dem BIP der Stadt Uerdingen (Quelle: Eigene Darstellung)

Im oben gezeigten Diagramm sind die Fußball-Daten von Bayer Uerdingen in Verbindung zu dem BIP der Stadt Uerdingen gesetzt. Es ist deutlich zu erkennen, dass Zusammenhänge sowohl aufgrund der kurzen Verweildauer des Vereins in der Bundesliga, als auch wegen der fehlenden Daten zu Umsatz und Gewinn des Vereins nicht nachweisbar sind. Diese Problematik umfasst zehn weitere Städte[13].
Lediglich bei der Stadt Kaiserslautern bzw. dem Verein 1. FC Kaiserslautern fanden wir für uns keine ersichtlichen Zusammenhänge trotz vorhandener Datensätze. Nach Abzug dieser Städte von allen relevanten und untersuchten Städten verblieben 25

[12] Auf der x-Achse sind die Jahre, auf der Y-Achse das Wachstum in Prozent. Die schwarze Linie ist das BIP, die rote der Umsatz und die blaue der Gewinn des Vereins.

Städte in unserer Auswertung. Die Quote für Korrelationen pro Stadt erhöhte sich auf 3,4.

Weitergehend haben wir untersucht, welche Verbindungen zwischen Fußball- und Stadtdaten in unserer deskriptiven Analyse am stärksten ausgeprägt waren. Zu diesem Zwecke haben wir die Verbindungen in drei verschiedene Kategorien aufgeteilt: leichte Verbindungen, mittelstarke Verbindungen und starke Verbindungen.

Je nachdem, wie stark die Übereinstimmungen zwischen zwei Linien waren, wurde sie eine der Kategorien zugeordnet. Gemessen an der Ausprägungsstärke der aufgetretenen Peaks in Diagrammen wurde eine Abstufung der Stärke dieser Ausprägungen vorgenommen. Dies hatte einen Vergleich der Daten erleichtert. Für die insgesamt 85 gefundenen Zusammenhänge ergibt sich folgende Zuordnung:
46 Zusammenhänge der Kategorie „leicht“, 37 der Kategorie „mittelstark“ und 2 Verbindungen der Kategorie „stark“.
Auf die jeweiligen Verbindungen aufgeteilt ergeben sich folgende Top Vier bzw. Top Zwei für jede Kategorie:

[13] Neben Uerdingen betrifft dieses Problem auch Leverkusen, Saarbrücken, Sinsheim, Cottbus, Augsburg, Aachen, Wattenscheid, Ulm, Fürth und Leipzig.

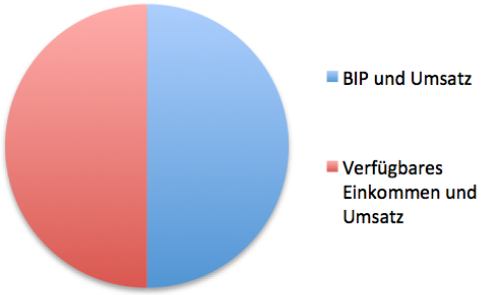

Abbildung 10: Verteilung der Verbindungen auf die Kategorien leicht (links oben), mittelstark (rechts oben) und schwer (unten) (Quelle: Eigene Darstellung)

Neben der Verteilung konnten wir auch die Häufigkeit der Verbindungen bestimmen, die in den Liniendiagrammen vertreten sind.
Um der Kategorisierung gerecht zu werden, multiplizierten wir die Anzahl der als leicht kategorisierten Verbindungen mit 1, die der mittelstarken mit 2 und die der starken mit 3. Durch die entsprechende Bewertung der Verbindungen lassen sich diese in eine neu geordnete Reihenfolge anordnen. Folgende Rangfolge legten wir im Zuge der Bewertung fest:

1. BIP und Platzierung
2. Erwerbstätige und Platzierung
3. BIP und Umsatz
4. Erwerbstätige und Gewinn
5. Durchschnittseinkommen und Umsatz

Des Weiteren gingen wir der Frage nach, in welchen Städten die Verbindungen zwischen Fußball- und Stadtdaten dem Anschein nach am stärksten ausgeprägt sind.
Eine starke Ausprägung bedeutet in diesem Fall die Anzahl der Verbindungen zwischen Stadt- und Fußballdaten. Wolfsburg belegte in diesem Ranking mit Abstand den ersten Platz, gefolgt von München, Hamburg und Gelsenkirchen.

4.2.2 Logistische Regressionsanalyse mit R

Neben der deskriptiven Analyse, die wir manuell durchgeführt haben, analysierten wir die gesammelten Daten mittels logistischer Regression im Analysetool *R*[14].

Für die logistische Regression entschieden wir uns weil wir eine dichotome abhängige Variable sowie verschiedene, unabhängige Variablen mit unterschiedlichem Skalenniveau vorliegen hatten.

Ziel war es dabei, die Abhängigkeit der Zielvariable „Platzierung" von den als relevant vermuteten Faktoren zu untersuchen. Die Platzierungsvariable wurde 0/1 kodiert – 1, falls die Platzierung im Vergleich zum Vorjahr anstieg und 0 für den Abstieg des Vereins oder das Beibehalten der Position. Bei allen unabhängigen Variablen flossen die errechneten Differenzen zum Vorjahr in das Modell ein, ohne umkodiert zu werden. Die unabhängigen Variablen setzten sich aus Daten zusammen, die zum Verein gesammelt wurden sowie aus Daten zu den jeweiligen Städten, die diese Vereine beherbergen. Somit ergab sich eine Tabelle mit folgenden Variablen:

- Platzierung (abhängige Zielvariable)
- Topspieler
- Umsatz
- Gewinn
- BIP
- durchschnittliches Einkommen
- Verfügbares Einkommen
- Erwerbstätige pro Einwohner
- Niederschlag
- Temperatur

An jeden Verein wurde jedes Jahr rangespielt, sodass pro Verein und Jahr eine Zeile entstand, die mit den dazugehörigen Werten eine Datenreihe. Insgesamt umfasste die Tabelle 920 Zeilen (vgl. Abbildung 6, Kapitel 4.1).

[14] „R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues." (www.r-project.org/about)

Vereine, die sich eine Stadt teilten wurden gekennzeichnet. So wurden der HSV und der FC St. Pauli als „Hamburg_H“ bzw. „Hamburg_P“ bezeichnet und besaßen zwar unterschiedliche Vereinsvariablen (Platzierung, Topspieler, Umsatz und Gewinn), jedoch identische Stadtvariablen.

Die Daten wurden als CSV-Datei in das Programm *R* geladen und die logistische Regression automatisch durchgeführt (vgl. Abbildung 11)

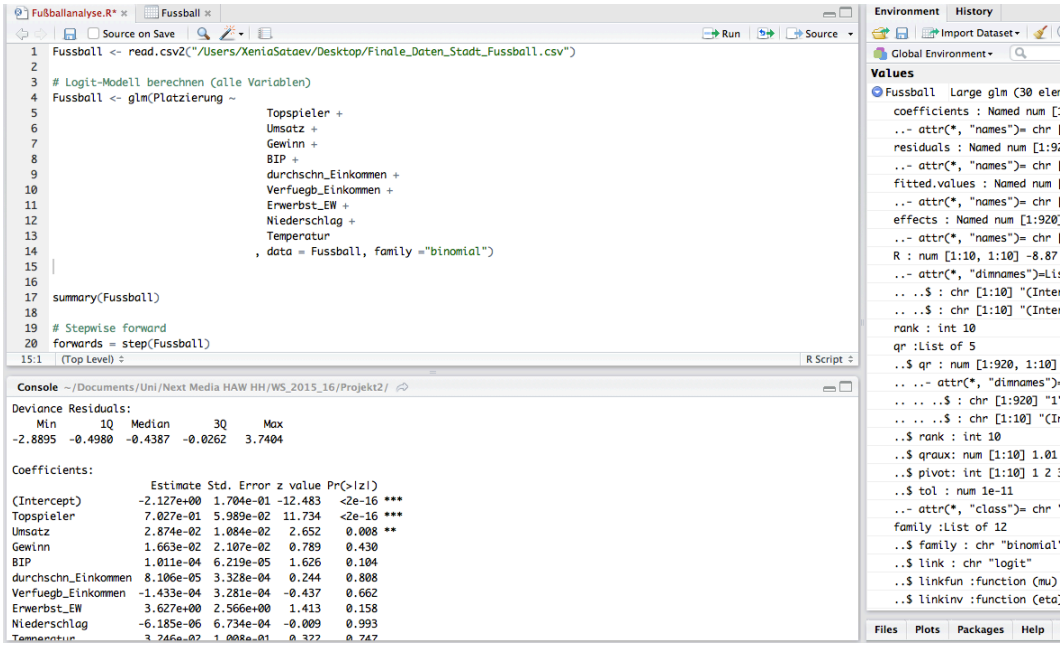

Abbildung 11: Datenspiegelung in R (Quelle: Eigene Darstellung)

5 Ergebnisse der Analyse

Laut Shearer gehört die Ergebnisauswertung der Analyse immer noch zur vierten Phase des CRISP-DM-Models[15]. Hier werden die Ergebnisse unserer verwendeten Analysemethoden zusammengeführt und miteinander verglichen.

5.1 Ergebnisse der deskriptiven Analyse

Die deskriptive Analyse der Fußball- und Stadtdaten hat folgende Ergebnisse hervorgebracht:

1. Die weiter oben bereits erwähnten 3,4 Verbindungen zwischen Fußball- und Stadtdaten sind nicht aussagekräftig genug, um zweifelsfrei eine Verbindung zwischen den korrelierten Daten nachzuweisen. Der Grund dafür ist unter Anderem der vergleichsweise kurze Untersuchungszeitraum von lediglich 24 Jahren, der nicht ausreicht, um einen Einfluss auf einen langfristigen und komplexen Prozess wie Stadtentwicklung sichtbar zu machen. Dennoch konnten vereinzelnd Verbindungen in den Daten aufgezeigt werden, die zumindest Auffälligkeiten aufwiesen. Die am stärksten ausgeprägten Korrelationen fanden sich hier zwischen dem BIP der Stadt und der Platzierung des Vereins. An zweiter Stelle folgen Erwerbstätige und die Platzierung des Vereins in der saisonalen Ergebnistabelle.

Für die sichtbaren Auffälligkeiten gilt hingegen, dass keine dieser Verbindungen durchgängig für die Fußball- und Stadtdaten aller Mannschaften beziehungsweise durchgehend für alle Jahre einer einzelnen Mannschaft nachgewiesen werden konnten. Aus diesem Grund kann ein eindeutiger Zusammenhang zwischen Fußball- und Stadtdaten nicht abschließend nachgewiesen und somit nicht angenommen werden.

2. Das Ranking der Städte zeigte, dass die meisten Verbindungen zwischen Fußball- und Stadtdaten in Wolfsburg ersichtlich waren - mit Abstand belegt die niedersächsische Stadt den ersten Platz der auffälligen Korrelationen. Das könnte auf die hohe wirtschaftliche Einflussnahme von VW in der Stadt zurückzuführen sein, die der Konzern sowohl in der Stadt, als auch im Verein spielt. Erwähnenswert ist die Tatsache, dass VW als Hauptsponsor des VfL Wolfsburg agiert und der größte Arbeitgeber in Wolfsburg ist[16].

An dieser Stelle wäre es interessant gewesen, diese Situation mit anderen Städten zu vergleichen, in denen die Situation ähnlich ist – wie zum Beispiel in Leverkusen mit

---

[15] Vgl. Shearer, Colin: The CRISP-DM-Model. The New Blueprint for Data Mining, S. 17.

[16] Vgl. http://www.manager-magazin.de/unternehmen/industrie/ranking-in-diesen-staedten-dominiert-eine-einzige-firma-a-1081886-6.html, zuletzt besucht am 15.03.2016 u 15:35 Uhr. Laut dem Manager-Magazin leben in Wolfsburg rund 123.000 Menschen, VW beschäftigt in der Stadt rund 60.000 Mitarbeiter.

dem Chemie-Konzern Bayer. Aufgrund fehlender Daten war uns dieser Vergleich leider nicht möglich.

3. Auffällig ist des Weiteren, dass bei vielen Liniendiagrammen die Verbindungen zwischen Fußball- und Stadtdaten erst nach dem Jahr 2000 deutlich sichtbar werden. Ein Grund dafür könnte sein, dass wir viele Stadtdaten extrapolieren mussten und somit auf Mittelwerten und Schätzungen basieren. Dadurch könnte eine mögliche Korrelation mit den Fußball-Daten verhindert worden sein. Ein weiterer Grund ist, dass die Schwankungen in den Umsätzen und Gewinnen mit zunehmender Zeit, seit 1991 nahezu explodierten. Das bereits erwähnte Zitat von Liga-Präsident Rauball, der berichtete, die Umsätze der Vereine seien von 190 Millionen Euro 1991 auf circa 2,5 Milliarden Euro im Jahr 2015 gestiegen, verdeutlicht das. Das entspricht einer Steigerung von über 1000 Prozent. Aufgrund unseres zweistufigen Rechenmodells werden zudem die geringeren Differenzen in den Daten nicht stark genug berücksichtigt. Die größeren Differenzen in den Daten nach dem Jahr 2000 sind daher für die großen Peaks in den Liniendiagrammen und somit auch für die aufgezeichneten und aufgefallenen Verbindungen verantwortlich.

4. Ein weiteres Ergebnis der Analyse ist, dass trotz vorhandener Daten eine Analyse nicht immer möglich ist. Das betrifft vor allem die Vereine, die nur ein oder zwei Jahre in der Bundesliga verweilten und danach wieder in die Zweite Bundesliga abstiegen. In dieser Zeit konnten die Vereine vermutlich nicht genügend „Erfolg“ erwirtschaften, der sich positiv auf die Stadtentwicklung hätte auswirken können.

5. Schade war zudem, dass die deskriptive Analyse fast ausnahmelos bei den kleineren Städten wie Saarbrücken, Wattenscheid oder auch Sinsheim (noch) nicht möglich war, weil der Datensatz nicht aussagekräftig genug war. Vorzugsweise in diesen Städten wäre es jedoch sehr interessant gewesen, den Einfluss der Fußballvereine auf die Stadtdaten und die Stadtentwicklung zu untersuchen. Nicht zuletzt, weil die oben beschriebene Komplexität der Faktoren, die die Stadtentwicklung bestimmen, hier etwas leichter zu durchdringen gewesen wären und ein Einfluss so leichter festzustellen gewesen wäre.

5.2 Ergebnisse der logistischen Regression

Auch die Ergebnisse der logistischen Regression wiesen keine eindeutigen Zusammenhänge auf.
Eine starke Signifikanz wiesen die beiden Variablen Topspieler sowie Umsatz auf. Alle weiteren Variablen gingen als nicht signifikant in das Modell ein. Dieses Ergebnis ist jedoch nicht besonders überraschend.
Die Abhängigkeit der Vereinsplatzierung von ihren Topspielern und dem Umsatz wurde bereits vor der Analyse stark angenommen und die Variablen somit als Kontrollvariablen berücksichtigt. Das Ergebnis der Analyse lässt darauf schließen, dass die Stadtentwicklung keinen Einfluss auf die Fußballergebnisse hat.

6 Fazit

Das Ziel unseres einsemestrigen Projektes war die Beantwortung der zentralen Fragestellung nach einem möglichen Zusammenhang zwischen dem Erfolg von Bundesligamannschaften und den sie beherbergenden Städten, um auf diese Weise den gesamten Datenanalyseprozess zu durchlaufen. Zusammenfassend kann man sagen, dass die Datenbeschaffung und die Datenbereinigung die zeitintensivsten Phasen des Projektes waren und fast die gesamte Bearbeitungszeit, die wir für das Projekt hatten, in Anspruch genommen haben. Dies ist auf diverse Schwierigkeiten und Herausforderungen zurückzuführen, die bereits näher erläutert wurden.
Somit blieb für die eigentlich Analyse der Daten und damit das Erhalten von wertvollen Ergebnissen kaum noch Zeit. Die Ergebnisse, die wir mittels der von uns gewählten Methoden erhalten haben, sind zudem leider nicht sehr zufriedenstellend, auch wenn nicht unterwartet. Ein Zusammenhang zwischen Fußball und Städten konnte folglich nicht festgestellt beziehungsweise hinreichend bewiesen werden. Die nachfolgenden Absätze fassen das Vorgehen und die Schwierigkeiten, die dabei auftraten zusammen, und beleuchten Verbesserungsmöglichkeiten für zukünftige Analysen.

Nachdem wir zu Beginn der Projektarbeit nach einigen verworfenen Ideen die Herangehensweise an die Themenfindung ändern mussten und unseren Fokus auf eine Fragestellung statt auf ein Ergebnis legten, nahm die Datenbeschaffung eine

sehr lange Zeit der Projektarbeit in Anspruch. Das Suchen und Sammeln der Daten war somit von vielen frustrierenden Momenten geprägt.
In dieser Zeit war große Geduld gefragt, da viele Daten entweder gar nicht oder sehr schwer zugänglich sind. Die Daten, die scheinbar von qualitativen Quellen zur Verfügung gestellt wurden, stehen zudem meist in unzureichender Form wie PDF-Formaten oder nur in Papierform zur Verfügung – in Zeiten der Digitalisierung und Open Data Cities ein sehr irritierendes Ergebnis. Die Daten müssen daher in sehr aufwendiger manueller Arbeit in andere Formate und Datenverarbeitungsprogramme übertragen werden. Weiterhin bestehen bei der Qualität der Daten erhebliche Schwächen, da die notwendige Erläuterung der Datenerhebung und Messmethoden unzureichend ist.

Ähnliche Schwierigkeiten sind bei der Datenbereinigung entstanden, die bereits in Kapitel 4.1. erläutert wurden. Datenfehler, Ausreißer, Messfehler und leere Werte müssen nicht nur behoben, sondern im ersten Schritt auch entdeckt werden, was ebenfalls mit einem hohen Zeitaufwand verbunden ist. Nachdem wir den Großteil der Zeit mit den beiden Phasen Datenbeschaffung und –bereinigung zugebracht haben, haben wir mit dem erarbeiteten Datensatz zwei Analysemodelle erarbeitet, um die Zusammenhänge der Entwicklung von Städten und Erfolgsentwicklungen von Bundesligavereinen zu untersuchen. Genutzt wurde dabei die deskriptive Analyse und die logistische Regression mittels R.

Beide Analysen zeigen in den Ergebnissen keine unerwarteten signifikanten Zusammenhänge und Abhängigkeiten beziehungsweise Wechselwirkungen zwischen dem Erfolg von Bundesligavereinen und der Entwicklung von Städten. Die deskriptive Analyse ergab zwar einige Auffälligkeiten, diese könnten aber ebenfalls Zufallsergebnisse sein und konnten nicht bei allen Vereine beziehungsweise Städten beobachtet werden. Die häufigsten Faktoren, die einen möglichen Zusammenhang zum Erfolg von Bundesligamannschaften aufwiesen, waren das BIP und die Anzahl an Erwerbstätigen. Am deutlichsten wurden diese Zusammenhänge und Wechselwirkungen in Wolfsburg sichtbar. Wobei der Einfluss der VW Konzerns dabei nicht vernachlässigt werden kann. Weiterhin konnten bei der Untersuchung

viele Vergleiche aufgrund von fehlenden Daten nicht abschließend durchgeführt werden. Ein Einfluss konnte daher nicht untersucht werden.
Das Ergebnis der logistischen Regression wies zwei stark signifikante Faktoren für den Erfolg eines Vereins auf, den Einfluss der Topspieler und des Umsatzes. Diese wurden jedoch bereits von uns vermutet. Ein Zusammenhang mit den Städtedaten konnte hingegen nicht beobachtet werden. Zwar waren die Temperatur und das BIP schwach signifikant, konnten in der fortschreitenden Analyse jedoch nicht explizit nachgewiesen werden.

Es wird in der abschließenden Betrachtung der Ergebnisse deutlich, dass diese sehr gering aussagekräftig sind. Die Notwendigkeit von vollständigen und deutlich umfangreicheren Datensätzen als die, die uns zur Verfügung standen, zeigt sich als unerlässlich. Denn die aus fehlenden und unvollständigen Datensätzen resultierenden Problematiken haben eine große Bedeutung für die Ergebnisse. Zusätzlich nimmt der Umgang mit diesen Problematiken einen Zeitaufwand in Anspruch, der schwer in ein gesundes Verhältnis zum abschließenden Ergebnis zu setzen ist. Festzuhalten ist daher die Notwendigkeit von großen, umfangreichen und vollständigen Datensätzen in nutzbarer Form.
Die Phase der Datenbeschaffung, -aufbereitung und –bereinigung nimmt indes den größten Teil der Arbeitszeit ein und sollte daher keinesfalls unterschätzt werden. Es ist daher notwendig ein Analyseprojekt sehr detailliert vorzubereiten und möglichst früh einen Zeitplan zu erstellen. Dabei sollte festgelegt werden wie realistisch die Beantwortung einer Fragestellung und die damit einhergehende Analyse in einer bestimmten Zeit ist. Zudem sollte das Ziel genau definiert und im Verlauf des Projektes regelmäßig überprüft und gegebenenfalls überarbeitet werden. Von besonderer Bedeutung ist zudem das Festlegen der Form, in der alle Daten vorliegen müssen um eine Analyse durchführen zu können. Das beinhaltet ein gleiches Format, ein einheitliches Skalenniveau und das Bestimmen der Aufteilung und Zuordnung von Zeilen und Spalten.
In diesem Zuge muss die Datenebereinigung äußerst gewissenhaft durchgeführt werden. Auffallende fehlende Daten oder Fehler in den Datensätzen müssen bereits vor Beginn der Analyse aufgedeckt, eliminiert und bearbeitet werden, da es

andernfalls während der Analyse zu erheblichen Schwierigkeiten und Fehlermeldungen kommen kann. Das Verbessern dieser fehlerhaften Daten in der Phase der Analyse ist verhältnismäßig aufwändig und setzt die Projektphase in die Phase der Data Preparation zurück, was wiederum zur Verlängerung der Bearbeitungszeit führt. Deutlich wird der enorme Zeitanspruch, der sowohl durch die Suche nach geeigneten Daten, das Sammeln und Übertragen dieser als auch das Aufbereiten und Bereinigen dieser entsteht.

Um die Qualität der Ergebnisse zu erhöhen, ist zu empfehlen einen deutlich größeren Datensatz zu wählen. In unserem Fallbeispiel wäre daher der Untersuchungszeitraum größer zu wählen und gegebenenfalls die fußballbezogenen Daten auch auf die 2. Bundesliga auszuweiten. Selbstverständlich sind die dabei erwähnten Probleme nicht außer Acht zu lassen, die sich mit einer Vergrößerung der Datenmenge ebenfalls vermehren. Ein größerer Untersuchungszeitraum würde jedoch aus unserer Sicht einer Erhöhung der Ergebnisqualität zuträglich sein.
In diesem Zusammenhang ist zu nennen, dass sich die Einflüsse langfristig entwickeln und entstehen und sich demnach erst zeitversetzt zeigen beziehungsweise nachweisen lassen. Daher müssen die Entwicklungen nicht nur auf die Jahreszahlen bezogen verglichen werden, sondern auch auf einen bestimmten zeitversetzten Zeitraum. Das bedeutet, dass Entwicklungen beispielsweise aus dem Jahre 1992 ebenfalls mit Daten aus den Folgejahren 1993, ‘94 und ‘95 verglichen werden sollten. Erweitert kann die Untersuchung außerdem mit weiteren Faktoren, die deutliche Indikatoren für die Stadtentwicklung beziehungsweise für die Erfolgsentwicklung von Bundesligavereinen sein könnten.

Die wichtigsten zu berücksichtigen Punkte für die Verbesserung der Ergebnisqualität sind somit umfangreichere Datensätze, verbesserte Datenqualität, Einbeziehen von weiteren Faktoren sowie eine genauere Zielsetzung und Ausarbeitung der notwendigen Datenerfassungskriterien. Die akribische Planung und Vorarbeit ist daher bei einem solchen Datenanalyseprojekt unerlässlich. Weiterhin darf der enorme Zeitaufwand in der Datenbeschaffungs-, Datenbereinigungs- und Datenaufbereitungsphase keinesfalls unterschätzt werden. Nichtsdestotrotz ist es

möglich, dass auch bei der besten Planung und der umfangreichsten Analyse einfach kein zufriedenstellendes Ergebnis erzielt wird, da kein Zusammenhang besteht, obwohl dieser anfangs unterstellt beziehungsweise angenommen wird. Auch das ist jedoch ein Ergebnis.

Abschließend möchten wir diese Arbeit mit einem Apell an die zuständigen Behörden zur Veröffentlichung und Bereitstellung von Daten sämtlicher Art schließen. Diese Daten sollten im Gegensatz zur aktuellen Situation einheitlich und in einer Form bereitgestellt werden, die weiterverarbeitet werden kann. Dabei sollten in die Webseiten integrierte Tabellen, PDF-Formate oder Papierdokumente nach Möglichkeit vermieden werden. Es sollten Formate genutzt werden, die für alle zugänglich, einheitlich herunterzuladen und weiterzuverarbeiten sind, um die Daten zumindest mit den üblichen Standardprogrammen verwenden zu können. Das Sprechen von Transparenz, öffentlich zugänglichen Daten und Open Data sind mit dem alleinigen Benutzen der Begriffe nicht getan. Das Vorliegen einer großen Menge von Daten ist für mögliche Analyseprojekte nicht ausreichend sofern die Zugänglichkeit und die mögliche Verwertung dieser nicht gewährleistet werden. Den Namen Open Data City hat aus unserer Sicht bislang keine Stadt in Deutschland verdient.

Literaturverzeichnis

Cross Industry Standard Process for Data Mining; Wikipedia. Online: Quelle: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Eckl-Dorna, Wilfried, (2016). http://www.manager-magazin.de/unternehmen/industrie/ranking-in-diesen-staedten-dominiert-eine-einzige-firma-a-1081886-6.html, zuletzt besucht am 15.03.2016 u 15:35 Uhr.

IBM SPSS Modeler CRISP-DM-Handbuch, Copyright IBM Corporation 1994, 2012; online: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/de/CRISP-DM.pdf

Mayer-Schönberger, Viktor; Cukier, Kenneth (2013): Big Data . Die Revolution, die unser Leben verändern wird, München.

Shearer, Colin (2000). The CRISP-DM-Model. The New Blueprint for Data Mining, In: Journal of Data Warehousing, Volume 5, Number 4, Seattle.

Turner, Vernon (2014) u.a.: The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, Framingham. Hier als PDF verfügbar: http://idcdocserv.com/1678
Visualisierte Form: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf.

Vogt, Martin. Fokus. http://www.focus.de/sport/fussball/bundesliga1/tid-26910/50-jahre-bundesliga-die-bundesliga-in-dekaden-die-90er-jahre-_aid_799402.html, zuletzt eingesehen am 17.01.16 um 14:22.

Weitz, Stefan (2014). Search: Hot the Data Explosion makes us smarter; Brookline.