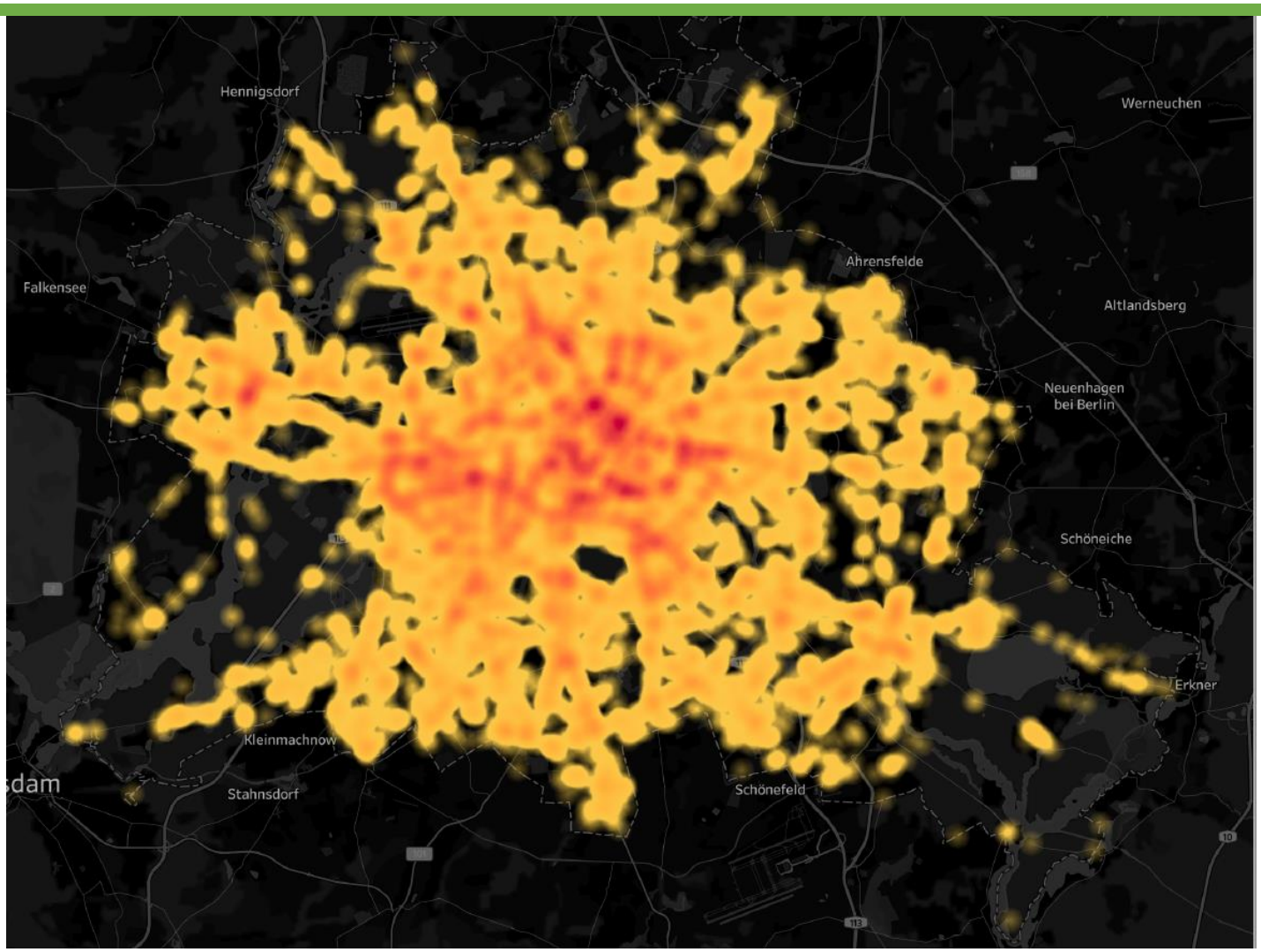
Hochschule für Technik
und Wirtschaft Berlin
University of Applied Sciences

Wissensmanagement Projekt

Nutzenpotentiale von Open Data am Beispiel des Unfalldatensatzes Berlin 2018



Team: Joshua Hammesfahr, Benjamin Wuthe

Inhaltsverzeichnis

Einführung ... 4
Zielsetzung ... 4
Vorgehen ... 4
Sichtung und Auswahl der Datenbasen ... 5
Ausgangsdatenbasis ... 5
Evaluation weitere Datenquellen im Fis-Broker ... 6
Evaluation weiterer externer Datenquellen ... 7
Definition Fragenkatalog ... 7
Konsolidierung und Aufbereitung der Datenbasis ... 7
Deskriptive Betrachtung der Datenbasis ... 9
Orts Dashboard ... 9
Zeitpunkt Dashboard ... 10
Beteiligten Dashboard ... 11
Betrachtung ausgewählter Analyseansätze ... 12
Assoziationsanalyse zur Sichtung von Regeln in der Datenbasis ... 12
Auswirkung von Ampeln auf Abbiege- und Kreuzungsunfälle ... 13
Identifikation potenzieller Faktoren für die Unfallschwere ... 15
Abschließende Betrachtungen ... 17
Literatur ... 19

Abkürzungsverzeichnis

TSB	Technologiestiftung Berlin
ODIS	Open Data Informationsstelle Berlin
JSON	Javascript Object Notation
PKW	Personenkraftwagen
BVG	Berliner Verkehrsbetriebe

Abbildungsverzeichnis

Abbildung 1: Projektablauf 5
Abbildung 2: Verwendete Datenquellen 6
Abbildung 3: Verknüpfung mit Google Geo Informationen 8
Abbildung 4: Dashboard Ort 9
Abbildung 5: Dashboard Zeitpunkt 10
Abbildung 6: Dashboard Beteiligte 11
Abbildung 7: Assoziationsanalyse Explorer 12
Abbildung 8: Assoziationsanalyse Grafen 13
Abbildung 9: Geclusterte Unfälle 14

Einführung

Das vorliegende Dokument beinhaltet die Dokumentation eines studentischen Datenanalyseprojektes auf Basis eines Unfalldatensatzes von Berlin aus 2018, welches in Kooperation mit der Technologiestiftung Berlin (im Folgenden TSB) im Rahmen der Veranstaltung Wissensmanagement – Projekt an der HTW Berlin durchgeführt wurde. Die Ausarbeitungen sind in einem GitHub Verzeichnis (https://github.com/s0554849/unfallanalyse_berlin) und unter Tableau Public (https://public.tableau.com/profile/benjamin.wuthe#!/vizhome/UnfallanalyseBerlin/UnfallanalysefrBerlin) öffentlich zugänglich.

Die TSB ist eine staatlich geförderte Einrichtung, welche vor allem öffentlichen Einrichtungen Informationen, Software und Infrastruktur sowie praktische Anwendungspotentiale dieser im Kontext der Digitalisierung aufzeigt[1]. Innerhalb der TSB gibt es die Open Data Informationsstelle Berlin (im Folgenden ODIS), welche ein kooperatives Projekt darstellt. Das Ziel von ODIS ist hierbei das Aufzeigen von Nutzenpotentialen von offenen Daten im öffentlichen Sektor, welche von jedem verwendet werden könnten (Open Data). Im Zuge einer zunehmenden Bedeutung von Open Data im öffentlichen Sektor stellt die Stadt Berlin mit dem fachübergreifenden Informationssystem Fis-Broker ein Geoportal bereit, über das Geodaten für die Stadt Berlin zur Verfügung gestellt werden. In diesem wurde kürzlich ein Datensatz für Verkehrsunfälle in Berlin aus dem Jahre 2018 veröffentlicht.

Zielsetzung

Zielsetzung des Projektes ist die Analyse dieses Unfalldatensatzes sowie die potenzielle Ableitung von Erkenntnissen, welche einen Mehrwert bieten können. Hierbei handelt es sich um eine Abbildung des generellen Zieles der ODIS Nutzenpotentiale von Open Data aufzuzeigen mit einem Fokus auf den Unfalldatensatz. Um den Mehrwert von offenen Daten aufzeigen und das potenzielle Erkenntnisspektrum zu erweitern, ist es weiterhin im Zuge der Aufgabenstellung explizit gewünscht, den Datensatz mit weiteren offenen Datensätzen zu verlinken.

Vorgehen

Die zur Erreichung der Zielsetzung gewählte Vorgehensweise ist in Abbildung 1 visualisiert. So wurden zunächst im Rahmen eines Kickoff-Meetings mit den Ansprechpartnern innerhalb der TSB die Stiftung vorgestellt, die Erwartungen an das Projekt abgesteckt und die Links zu den zu nutzenden Ressourcen bereitgestellt. Nach dem generellen Einstieg wurden anschließend die verfügbaren Datenquellen evaluiert. Hierbei wurde zunächst die Ausgangsdatenbasis genauer betrachtet, anschließend wurden potenzielle Datenquellen im Fis-Broker evaluiert und abschließend fand eine Betrachtung weiterer potenziell sinnvoller externer Datenquellen statt. Dieser Schritt wird im Kapitel *Sichtung und Auswahl der Datenbasis* detailliert betrachtet. Nachdem ein generelles Verständnis der verfügbaren Daten besteht, wurden Ideen für potenzielle Fragestellungen, welche im Zuge des Projektes konkret beantwortet werden können, gesammelt und mit der TSB abgestimmt. Das Ergebnis dieser Phase ist ein Fragenkatalog, welcher im Kapitel *Definition Fragenkatalog* vorgestellt wird. Nachdem ein generelles Verständnis für die verwendeten Datenquellen hergestellt wurde und Leitfragen definiert waren, wurden die Daten verknüpft und aufbereitet, sodass eine konsolidierte Datenbasis besteht, welche als Grundlage für eine folgende deskriptive Analyse sowie eine methodische Beantwortung ausgewählter Fragen mithilfe von Data Mining Techniken genutzt werden kann. Im Anschluss an die folgende deskriptive Analyse der Datenbasis in Tableau wurden die Fragen aus dem Fragenpool iterativ abgearbeitet. Diese

[1] TSB, 2020

Bearbeitung wird im Kapitel *Betrachtung ausgewählter Analyseansätze* vorgestellt. Die Bearbeitung des Projektes erfolgte in enger Abstimmung mit der TSB. Aufgrund des explorativen Charakters wurde im Zuge des Projektes eine agile Projektabstimmung in iterativen Zyklen gewählt, in denen der Status Quo, Abstimmungspunkte, Hürden und das weitere Vorgehen mit den Projektpartner abgestimmt wurden. Die Zykluszeit lag hierbei durchschnittlich bei drei Wochen.

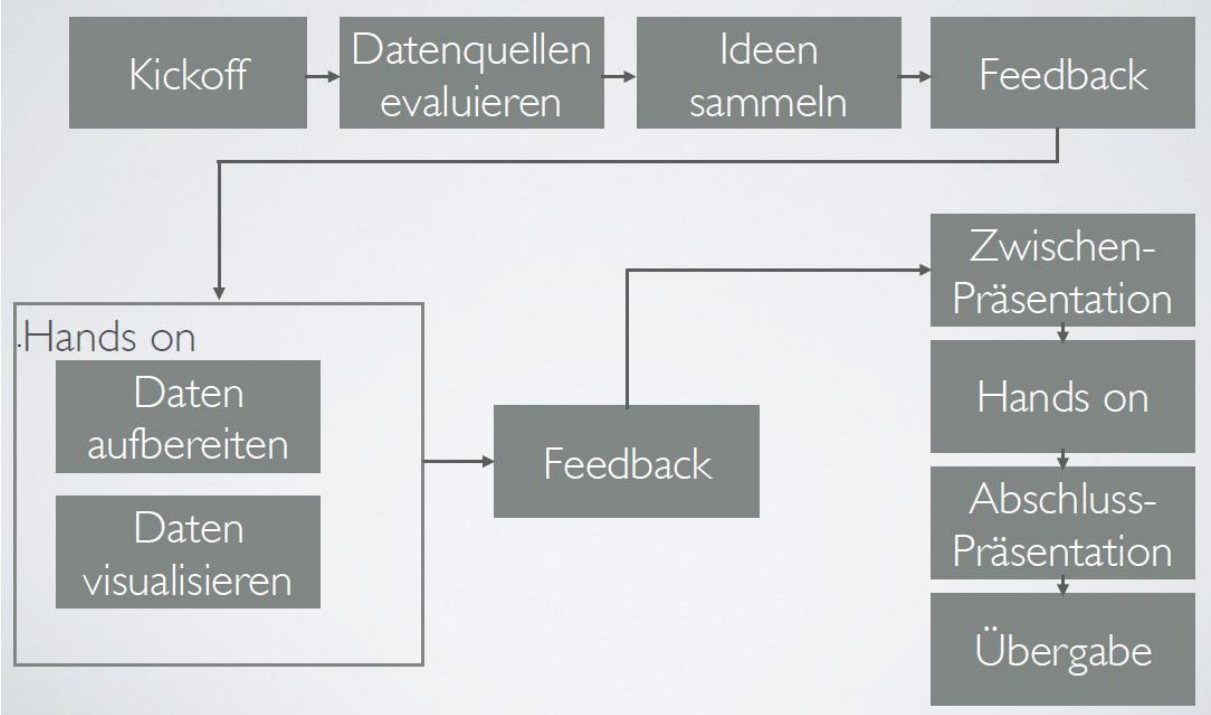


Abbildung 1: Projektablauf

Sichtung und Auswahl der Datenbasen

Ausgangsdatenbasis

In der Ausgangsdatenbasis werden die Unfälle in Berlin auf öffentlichen Wegen und Plätzen mit Personenschaden aus dem Jahre 2018 erfasst. Es handelt sich somit erstmal um eine gefilterte Sicht auf alle Unfalldaten in Berlin in 2018, da die Unfälle mit Sachschaden nicht erfasst sind. Der Datensatz enthält circa 13.000 Unfälle mit 23 Spalten. Relevante Information im Datensatz sind unter anderem, wo der Unfall passiert ist (UKREIS, UGEMEINDE, LINEREFX, LINEREFY), wann der Unfall passiert ist (UMONAT, USTUNDE, UWOCHENTAG), welche Verkehrsteilnehmer beteiligt waren (IstRad, IstPKW etc.), um welche Art von Unfall es sich in Bezug auf den Unfallvorgang und Folgen für Beteiligte handelt (UKATEGORIE, UART, UTYP1) und wie die Umweltbedingungen waren (ULICHTVERH, USTRZUSTAND). Für eine generelle Impression zu dem Datensatz werden dieser und die zugehörigen textuellen Beschreibungen in 02_Datenbereinigung/ OpenDataWMProjekt.ipynb betrachtet.

Evaluation weitere Datenquellen im Fis-Broker

Um den Unfalldatensatz mit weiteren offenen Geodatensätzen der Stadt Berlin anzureichern, wurde anschließend im Geoportal Fis-Broker geschaut, welche Datensätze mit diesem tendenziell sinnvoll verknüpft werden könnten. Im Zuge einer Grobauswahl wurden zunächst sieben Datenquellen in die engere Auswahl aufgenommen:

- Standorte von Ampeln in Berlin
- Lärmbelästigung
- Verkehrsmengen
- Tempolimits
- Straßenlaternen
- Bodenrichtwerte
- Einwohner nach Bezirken

Die Daten wurden zunächst über die Geoinformationssystemsoftware QGIS geladen, in dieser betrachtet und anschließend als GeoJSON, ein Datenformat für Geodaten, welches auf der JavaScript Object Notation (im Folgenden JSON) basiert, exportiert. Anschließend wurden die exportierten Daten mithilfe der Python-Bibiothek GeoPandas zur weiteren Betrachtung in einem Jupiter Notebook (02_Datenbereinigung/OpenDataWMProject.ipynb) und in Tableau evaluiert. Im Zuge der Evaluation wurden der Datensatz der Straßenlaternen sowie der Datensatz der Einwohner nach Bezirken ausgeschlossen. Die Straßenlaternen wurden im Projekt nicht verwendet, da es aufgrund der Allgegenwärtigkeit von Straßenlaternen in Berlin keine Unfallorte gab, bei denen keine Straßenlaterne in direkter Nähe war. Die Einwohner nach Bezirken wurden aus der weiteren Betrachtung ausgeschlossen, da keine logische Kausalität zwischen der Einwohnerzahl eines Bezirkes und der Anzahl an Unfällen identifiziert werden konnte, welche nicht bereits durch die Verkehrsmenge abgebildet worden wäre. Die Bodenrichtwerte wurden zunächst aufgenommen, sind jedoch im Zuge der Folgeverarbeitung nicht mehr von Bedeutung, da sich diese nur mit einem Bruchteil der Unfälle geographisch verknüpfen ließen.

Es verbleiben somit die vier weiteren Datenquellen mit den Ampelstandorten in Berlin, den Tempolimits, der Lärmbelästigung und den Verkehrsmengen, die in Abbildung 2 dargestellt sind.

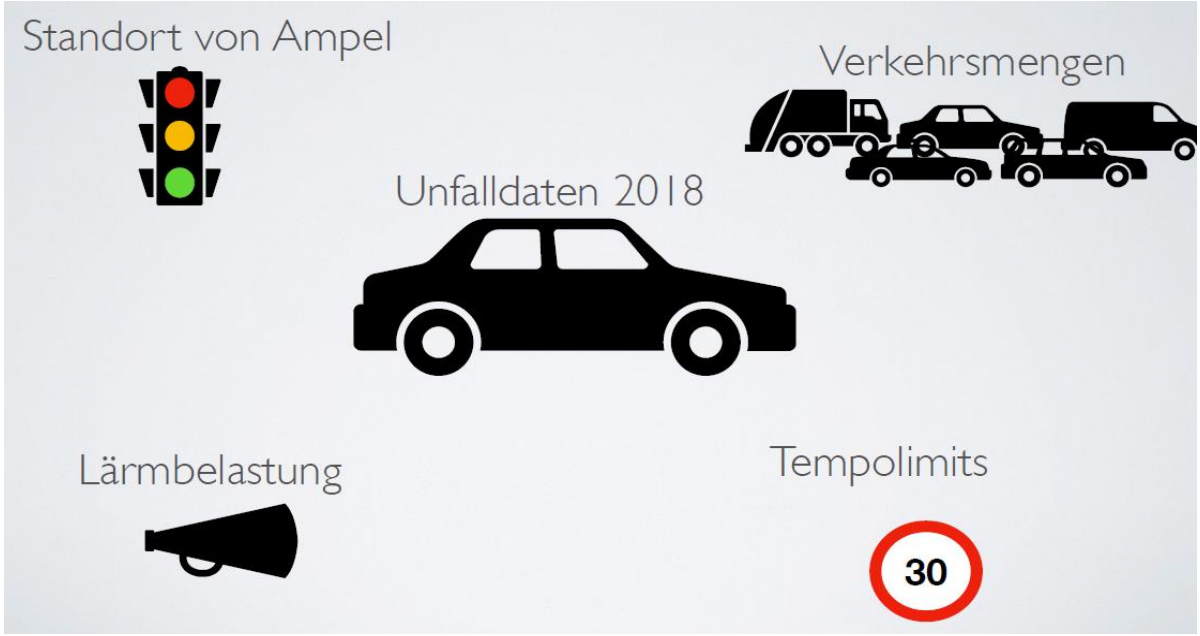


Abbildung 2: Verwendete Datenquellen

Evaluation weiterer externer Datenquellen

Neben den im Fis-Broker bereitgestellten Datenquellen wurde zunächst versucht weitere externe Datenquellen zur Anreichung des Unfalldatensatzes zu erschließen, Als potentielle Datenquellen wurden hierbei initial Wetterinformationen (siehe 02_Datenbereinigung/ OpenDataWMProjekt.ipynb) und erweiterte Geoinformationen über die Google Maps Geocode API (siehe 02_Datenbereinigung/getLocationsFromLatLong.py) identifiziert.
Die Wetterinformationen konnten im Zuge des Projektes nicht weiterverwendet werden, da die Unfalldaten im Hauptdatensatz keine konkrete Datumszuordnung haben und sich diese somit nicht verknüpfen lassen. Von den Geoinformationen der Google Maps Geocode API konnten die Straße, Postleitzahl und der Bezirk, in dem ein Unfall passiert ist, erschlossen werden. Diese Daten konnten vor allem als zusätzliche Geoinformationen genutzt werden, wenn im Ausgangsdatensatz zum Teil keine Geoinformationen für einzelne Unfälle erfasst worden sind.

Definition Fragenkatalog

Nachdem die wesentlichen Datenquellen spezifiziert und deren Inhalte überblicksartig erschlossen wurden, konnten Leitfragen in Form eines Fragenkatalogs für die weitere Ausarbeitung definiert werden, Die Erstellung der Leitfragen wurde zunächst mittels Brainstorming teamintern initiiert, anschließend wurde die Machbarkeit dieser diskutiert und abschließend fand eine Abstimmung mit dem Projektpartner statt. Die folgenden Fragen wurden hierbei spezifiziert:

1. Wie entwickeln sich die Unfallzahlen?
2. Wann und wo ist es als Radfahrer besonders gefährlich zu fahren?
3. Wie wirken sich die Lichtverhältnisse auf Unfälle aus?
4. Gibt es Orte, die besonders gefährlich sind, was macht diese Orte so gefährlich?
5. Wie wirkt sich der Anteil von bestimmten Verkehrsteilnehmergruppen auf einer Straße auf die Häufigkeit sowie Schwere von Verkehrsunfällen aus?
6. Wie wirkt sich die Lautstärke an Orten auf die Unfallzahlen aus?
7. Welche Faktoren begünstigen oder senken die Chance bei einem Unfall schwer verletzt zu werden oder zu sterben?
8. Helfen Ampeln bei der Reduktion von Abbiege und Kreuzungsunfällen?

Die Beantwortung dieser Fragen wird in den Kapiteln Deskriptive Betrachtung der Datenbasis und Betrachtung ausgewählter Analyseansätze vorgestellt. Im Zuge der deskriptiven Betrachtung werden hier zunächst mittels Visualisierungen Zusammenhänge zwischen Eigenschaften gesucht. Anschließend wird diese Suche nach generellen Zusammenhängen mithilfe einer Assoziationsanalyse fortgeführt. Im Zuge dieser Ansätze werden unter anderem die Fragen 1 bis 4 betrachtet. Abschließend werden die Fragen 7 und 8 explizit herausgegriffen und es wird betrachtet, wie diese Fragen beantwortet werden können und was die Ergebnisse auf Basis der Datenbasis sind.

Konsolidierung und Aufbereitung der Datenbasis

Zur Konsolidierung und Aufbereitung der Datenbasis mussten die gewählten Datensätze zunächst verknüpft werden. Die Verknüpfung der Geodatensätze, welche vom Fis-Broker gezogen worden sind, erfolgte hierbei mithilfe der Geoinformationssystemsoftware QGIS. Die Herausforderung bei der Verknüpfung von Geodaten besteht darin, dass die Verknüpfung über räumliche Beziehungen zwischen zwei Objekten stattfindet. So werden die Unfallstandorte beispielsweise über einen Punkt lokalisiert, die Tempolimits über eine Linie und Lärmareale über ein Polygon. QGIS bietet in diesem Kontext die Möglichkeit Datensätze auf Basis einer geographischen Überschneidung zu verbinden. Eine weitere Herausforderung beim geographischen Verbinden der Datensätze ist im linken

Bereich von Abbildung 3 anhand der Verbindung zwischen den Unfallorten (grün) und den Tempolimitbereichen (orange) visualisiert. So überschneiden sich diese zunächst nicht, obwohl sie auf der gleichen Straße liegen. Um dies zu lösen bietet QGIS die Möglichkeit einer Pufferung, durch welche definiert werden kann, dass alles in einem Radius von fünf Metern um den Punkt zu dem Unfallort gehört und alles, was beispielsweise drei Meter von der Linie des Tempolimitbereiches entfernt ist, zu diesem gehört. Durch diese Pufferung und die räumliche Verknüpfung mit einem Left-Join konnte schrittweise vom Unfalldatensatz ausgehend eine verknüpfte Datenbasis für alle Geoinformationen erstellt werden.

Mit dieser verknüpften Geodatenbasis konnte die weitere Verarbeitung innerhalb des Jupiter Notebooks 02_Datenbereinigung/Create_Cosolidated_Dataset.ipynb durchgeführt werden. In dieser wird die Datenbasis zunächst mit den zusätzlichen Straßeninformationen der Google Maps Geocode API verknüpft. Anschließend findet die Aufbereitung der Daten statt. Diese umfasst zum einen das sinnvolle Ersetzen von Null-Werten und unmöglichen Werten. Ein Beispiel für dieses stellt das Tempolimit dar. So ist innerhalb Berlins das Tempolimit generell 50 km/h, wenn es keine weiteren Einschränkungen gibt oder das Tempolimit im Datensatz beispielsweise bei einem Wert von 5923 km/h liegt. Weiterhin wurden die Identifikationsnummern im Datensatz für spezielle Felder durch deren textuelle Werte ersetzt. So wurde zum Beispiel im Feld UTYP1 der Wert 2 durch den Zeichenkette “Abbiegeunfall” ersetzt. Abschließend wurden im Zuge der Aufbereitung aus den bestehenden Features neue Features über Berechnungen abgeleitet. So wurden beispielsweise die neuen Features Anteil von Kraftfahrzeugen am Straßenverkehr (ANTEIL_KRAD), Anzahl an Unfällen je Bezirk (UGEMEINDE_AVG_TARGET) oder ein Feature, welches den Anteil von Verletzten an der Gesamtverkehrsmenge auf einer Straße abbildet (CRASH_BY_DTV), erzeugt.

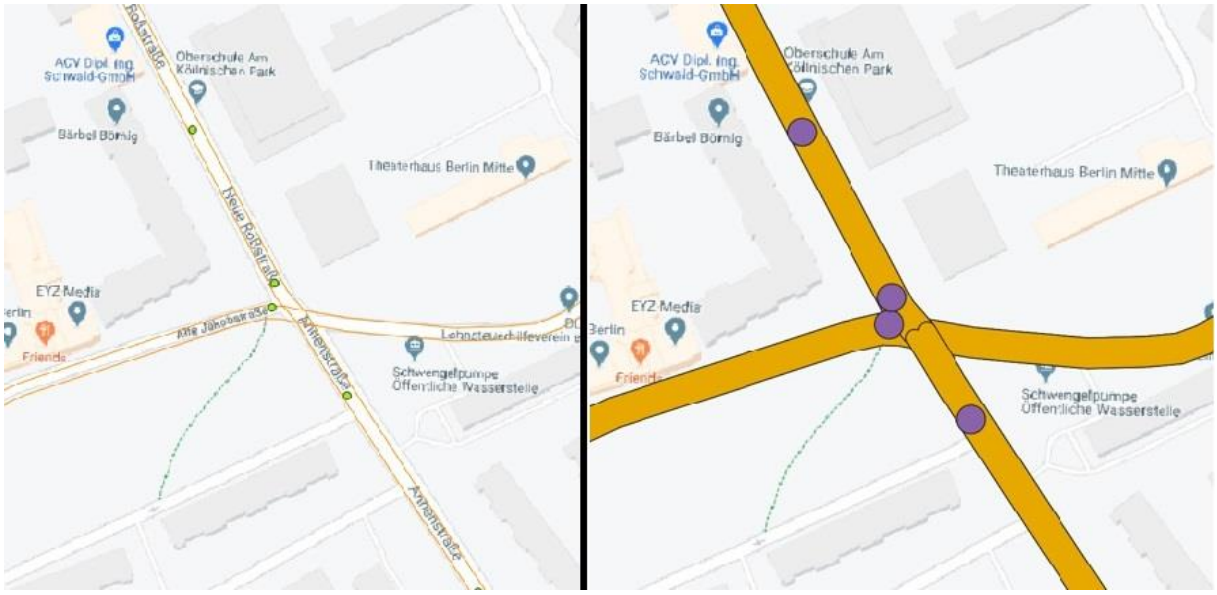


Abbildung 3: Verknüpfung mit Google Geo Informationen

Deskriptive Betrachtung der Datenbasis

Aus der vorliegenden Datengrundlage bestehend aus dem Unfalldatensatz 2018 wurde mit der Self Service Business Intelligence Software Tableau die deskriptive Analyse durchgeführt. Ziel ist es einen Überblick über die Daten zu bekommen und was im Speziellen Unfälle in Berlin ausmachen, und die Fragen 1 bis 3 aus dem Fragenkatalog zu beantworten. Dazu wurden Dashboards erstellt, welche aus diversen Diagrammen und Visualisierungen bestehen. Der Fokus besteht in der Analyse von drei Pfeilern. Thematisch aufgeteilt sind sie nach dem Ort, dem Zeitpunkt und nach den Beteiligten der Unfälle.

Das Dashboard ist online verfügbar auf der Tableau public gallery unter folgendem Link:

https://public.tableau.com/profile/benjamin.wuthe#!/vizhome/UnfallanalyseBerlin/UnfallanalysefrBerlin.

Dadurch ist die Möglichkeit geboten, eine Analyse der Daten interaktiv in den Dashboards durchzuführen.

Orts Dashboard

Das Dashboard der ortsabhängigen Auswertung besteht aus vier Visualisierungen (vgl. Abbildung 4: Dashboard Ort). Bestehend aus einer Karte, die die Unfallorte anhand der Geoinformationen darstellt. Die Unfalldichte wird hierbei dargestellt, indem die Farbskala von Gelb (wenige Unfalle) bis Rot (viele Unfalle) verwendet wird. Eine erhöhte Unfalldichte lässt sich auf zwei Ebenen erkennen. Zum einen im Stadtzentrum bzw. innerhalb des S-Bahn Rings und zum anderen auf den Hauptverkehrsstraßen.
Weiterhin ist die Verteilung der UART in einem Balkendiagramm dargestellt und zeigt auf, wie viele bzw. welche Unfälle Opfer mit leichten Verletzungen, schweren Verletzungen oder mit Todesfolge hatten.
Die Darstellungen *Unfälle in Bezirk* und *Unfälle auf Straßen* listen die jeweiligen Entitäten der Bezirke bzw. Straßen, mit der Anzahl der entsprechend in Zusammenhang gebrachten Unfälle.

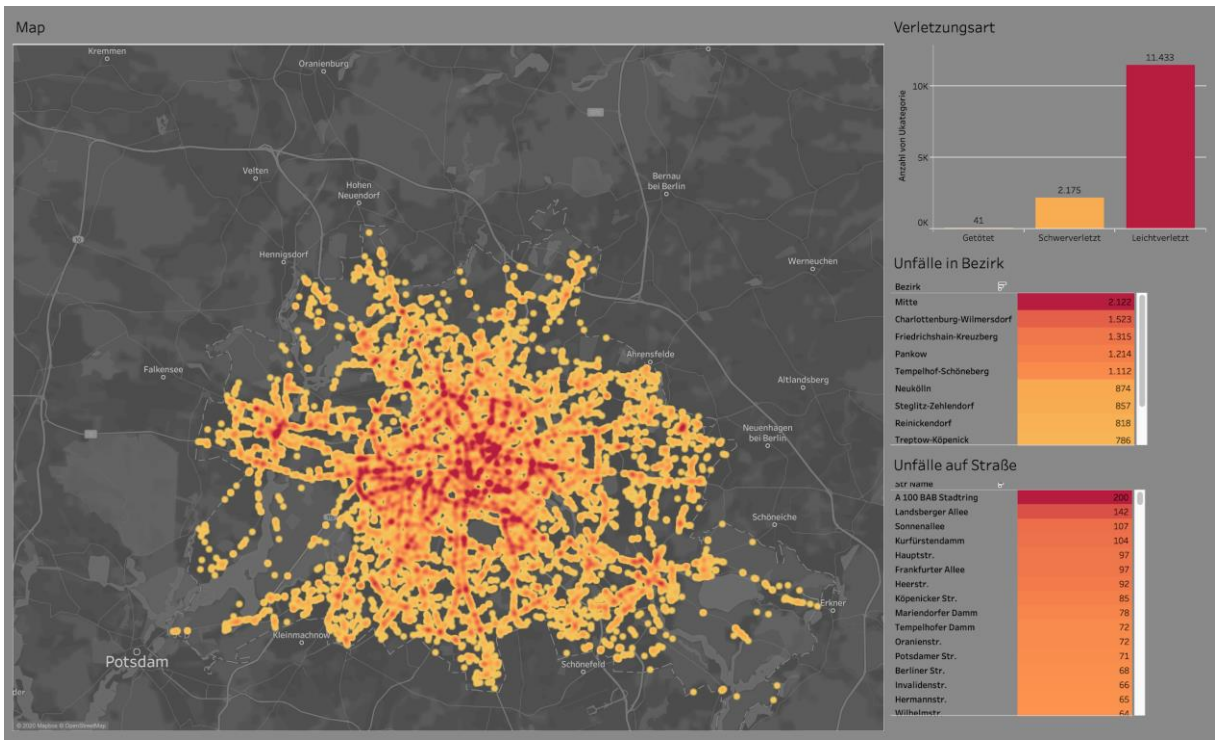


Abbildung 4: Dashboard Ort

Zeitpunkt Dashboard

Auch das Dashboard zur Auswertung der Unfallzeitpunkte besteht aus vier Visualisierungen (vgl. Abbildung 5: Dashboard Zeitpunkt). Der Fokus lag hierbei bei der Auswertung des Monats und des Wochentags, da die Datenbasis eine Auswertung auf Tagesebene nicht zuließ. In der Heatmap *Unfallszeit Monat* sind an der Einfärbung bereits zwei wesentliche Aussagen zu erkennen.

1. Am Wochenende geschehen weniger Unfälle als Werktags und
2. In den Sommermonaten geschehen mehr Unfälle als in den Wintermonaten

Unterstützt wird dies durch die Tatsachen, dass an Sonntagen im Januar die wenigsten Unfälle (49) und an Freitagen im Juni die meisten Unfälle (304) geschehen. Auch die Darstellung *Unfälle zu Lichtverhältnis Monat* unterstützt diese Annahme. Es ist ein deutlicher Anstieg der Unfälle von März bis Mai und ein Senken von Oktober bis Dezember zu erkennen. Unter Betrachtung der dritten Fragestellung aus dem Fragepool ist erkenntlich, dass im Sommer weitaus mehr Unfälle mit Tageslicht geschehen als im Winter. Das klingt zunächst trivial, da es im Sommer länger hell ist, aber das Verhältnis von Sonnenstunden zur Verteilung der Lichtverhältnisse stimmt nicht überein. Im Juni fanden mehr als 90% der Unfälle mit Tageslicht statt, während es im Dezember lediglich 38% waren. Das lässt darauf schließen, dass die Menschen in den Sommermonaten aktiver sind und dadurch auch mehr am Verkehr teilnehmen und mögliche Opfer sind.
Auch die Auswertung der Wochentage in Zusammenhang mit der Uhrzeit führt zu interessanten Erkenntnissen, wie in der Heatmap *Unfallszeit Woche* zu sehen ist. An der Färbung lässt sich nachvollziehen, wann maßgeblich Unfälle verursacht werden, was sich zweifelsfrei mit dem Berufsverkehr überschneidet. Zu sehen ist, dass die Unfallrate Montag bis Freitag von 7 bis 9 Uhr und 15 bis 18 Uhr beträchtlich ist. Anzumerken ist, dass der zweite Block noch wesentlich höher ist als der erste, was unter Umständen aus der Müdigkeit bzw. geringeren Aufmerksamkeit nach der Arbeit resultiert.

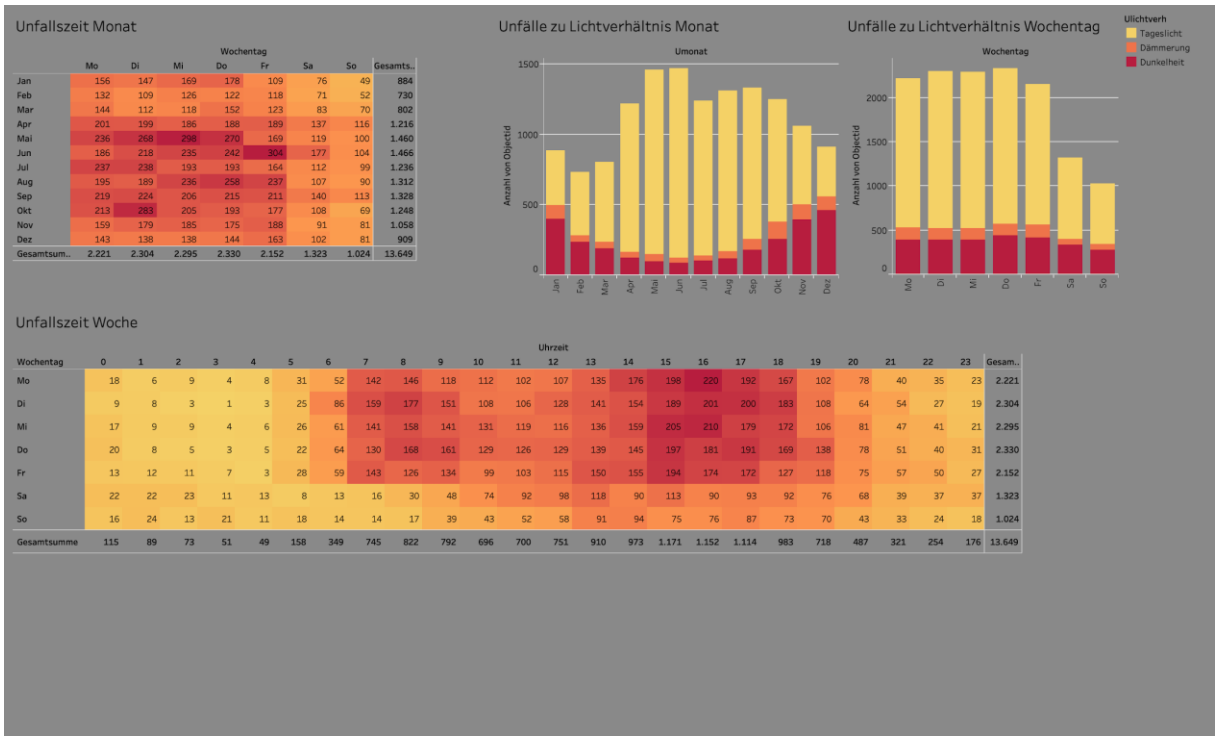


Abbildung 5: Dashboard Zeitpunkt

Beteiligten Dashboard

Welche Verkehrsteilnehmer involviert waren, wird im abschließenden Dashboard behandelt (vgl. Abbildung 6: Dashboard Beteiligte). Mehr als die vorherigen Dashboards setzt dieses auf gezieltes Suchen und Finden von Erkenntnissen mit Hilfe der Filter innerhalb dieses. Wie zuvor ist auch hier eine Karte vorhanden, die die Orte der Unfälle darstellt und eine Auflistung der Bezirke mit der entsprechenden Anzahl der Unfälle. Zudem sind zwei Balkendiagramme vorhanden, die die jeweiligen Summen darstellen. Zum einen handelt es sich hierbei um die Anzahl der Unfälle nach Verkehrsteilnehmer in Gruppen bzw. Fortbewegungsmittel (*Unfallsummen nach Fortbewegung*) und zum anderen um die Anzahl der Unfälle nach Monaten (*Unfälle im Monat*). Im Diagramm der Verkehrsteilnehmer nach Gruppen ist zu beachten, dass ein Unfall mehrere Beteiligte innehaben kann. So ergibt sich, dass von den 13.649 betrachteten Unfällen bei 11.152 mindestens ein PKW beteiligt war, bei 5.191 mindestens ein Fahrradfahrer usw.
Auf der rechten Seite des Dashboards befindet sich eine Gruppierung der jeweiligen Unfallbeteiligten mit der Anzahl des Vorkommens in einer Hervorhebungstabelle. Diese Tabelle lässt sich dabei mit einer Art Baumstruktur vergleichen und zeigt jede mögliche Kombination von Verkehrsteilnehmer auf. Über ihr sind ihre entsprechenden Filter, die ein “stutzen” des Baumes ermöglichen. Zu lesen ist die Tabelle beispielsweise wie folgt: Unter Betrachtung der ersten Zeile lässt sich ablesen, dass folgende Verkehrsteilnehmergruppen beteiligt waren: Radfahrer, Fußgänger, PKW Fahrer und Kraftradfahrer. Diese Konstellation an Verkehrsteilnehmer kam im Datensatz zwei Mal vor. Die Tabelle lässt sich nicht nur ausgehend von den Verkehrsteilnehmern lesen, sondern auch ausgehend von der Anzahl. Unter Betrachtung der größten Menge (3.633) liest sich die entsprechende Zeile, dass bis auf PKW alle anderen Felder auf “Nein” gesetzt sind. Der zweite höchste Wert wiederum (3.470) beinhaltet erneut die Gruppe PKW und zudem die Radfahrer. Diese Werte sind auf gesamt Berlin bezogen. Eine Erweiterung der Filter auf Bezirksebene ergibt, dass im Zentrum Berlins wesentlich mehr Radfahrer in Unfälle verwickelt werden als in Bezirken außerhalb unter Betrachtung des Verhältnisses PKW - PKW und Rad. Werden die Radfahrer gesondert betrachtet, ist erkenntlich, dass diese insbesondere in den Sommermonaten die Zahl der Unfälle steigen lassen. Werden sie exkludiert, zeigt sich der wellenartige Trend weniger stark ausgeprägt. An dieser Stelle soll nochmals auf Tableau Public verwiesen werden. Hier kann die Aussage mit entsprechenden Filtersetzung nachvollzogen werden.

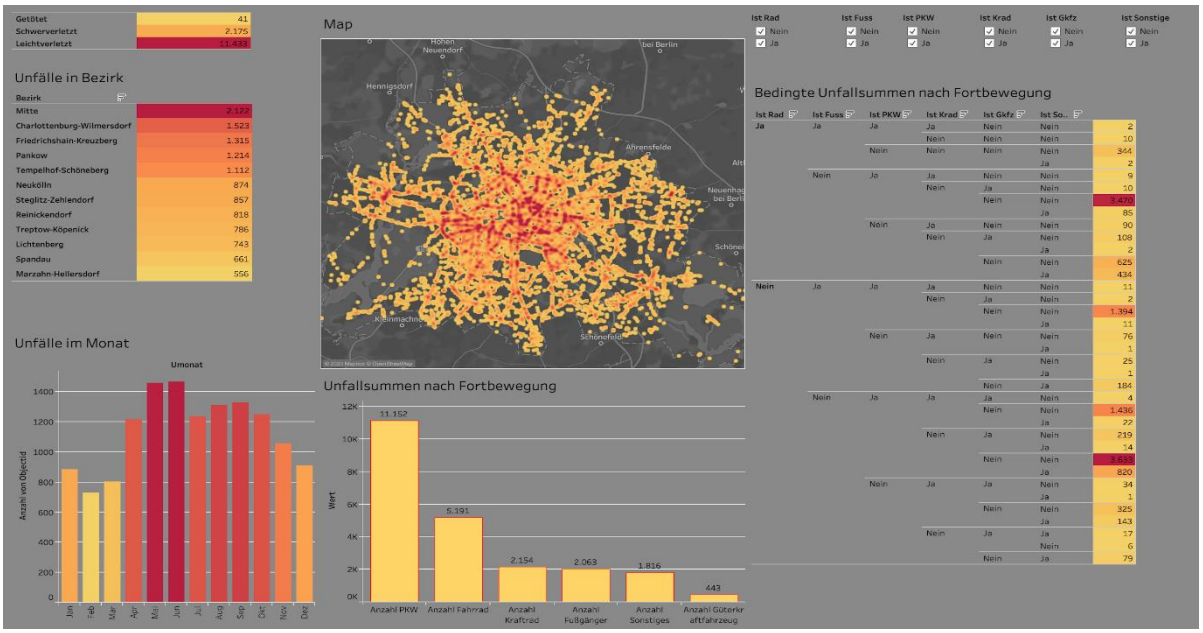


Abbildung 6: Dashboard Beteiligte

Betrachtung ausgewählter Analyseansätze

Nachdem die Unfälle deskriptiv beschrieben wurden, soll im weiteren Verlauf die explorative Analyse beschrieben werden. Folglich wurden Techniken des Data Minings verwendet. Im Detail ist das zum einen eine Assoziationsanalyse zur Regelfindung und zum anderen die Verwendung eines Random Forrests zur Beantwortung der Fragestellung, was potentielle Faktoren für schwere Unfälle sind. Weiterhin wurde zur Betrachtung der Fragestellung, welchen Einfluss Ampeln auf Abbiege- und Kreuzungsunfälle haben eine im Tool QGIS integrierte Clusterfunktionalität verwendet, welche auf DBSCAN-Clustering basiert.

Assoziationsanalyse zur Sichtung von Regeln in der Datenbasis

Als eine Disziplin des Data Minings wurde im Rahmen des Projekts eine Assoziationsanalyse, auch Warenkorbanalyse genannt, durchgeführt. Ziel dieser Technik war es Regeln und Abhängigkeiten innerhalb des Datensatzes zu identifizieren. Eine Assoziationsanalyse benötigt Item Sets, Datensätze in denen Items hinterlegt sind. Ein Item Set besteht im Unfalldatensatz beispielsweise aus den Items {PKW, Montag, Schwerverletzt, ...}.
Mithilfe dieser Item Sets werden Regeln erstellt, nach dem Schema *Wenn Item X, dann Item Y.* Das Ergebnis der Assoziationsanalyse sind für jede entstandene Regel drei Kennzahlen: *Support*, *Confidence* und *Lift.*
Der Support gibt die relative Häufigkeit an, an der die Regel angewendet werden kann. Die Confidence die relative Häufigkeit unter der diese Regel richtig ist.
Der Lift gibt an, wie sehr der Konfidenzwert für die Regel den Erwartungswert übertrifft. Die Vorbereitung des verwendeten Datensatzes in eine für die Assoziationsanalyse verwendbare Form, wurde in Python durchgeführt (*03_Analytics/apriori_prep.py*). Es wurde das One Hot Encoding auf qualitative Merkmale durchgeführt, um die benötigte Struktur für die Items zu erhalten, die in der Assoziationsanalyse benötigt werden. Anschließend erfolgte die Zusammenfassung der Spalten der 24 Stunden in Tageszeiten (Morgen, Mittag, Abend, Nacht) sowie der Monate in Jahreszeiten (Frühling, Sommer, Herbst, Winter). In diesem Projekt wurde hauptsächlich die Programmiersprache Python verwendet, um die Assoziationsanalyse durchzuführen wurde dennoch auf die Programmiersprache R zurückgegriffen. Nachdem nach intensiven Recherchen zu Paketen, die eine Assoziationsanalyse in Python durchführen können kein adäquates Paket gefunden werden konnte, fiel die Entscheidung die Assoziationsanalyse in R durchzuführen. *Aarules* und *arulesViz* sind Pakete, die entsprechende Berechnungen durchführen können, bzw. die Visualisierung und interaktive Handhabung der Ergebnisse ermöglichen. Anwendung fand dies in der Markdown Datei *03_Analytics/association_rules.Rmd*, deren Ausführung die Analyse interaktiv betrachten lässt, wie in Abbildung 7 dargestellt.

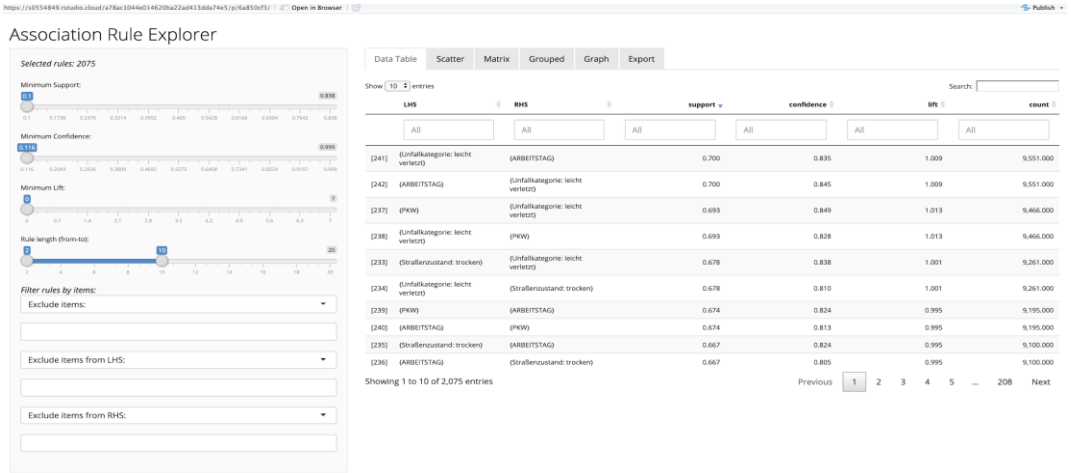


Abbildung 7: Assoziationsanalyse Explorer

In der interaktiven Ansicht können diverse Filter gesetzt werden. So können Minimalwerte für Support, Confidence und Lift gesetzt, Anzahl der Items in einer Regeln definiert oder bestimmte Items ausgeschlossen werden. Für die Auswertung der Analyse wurden die Filter auf folgende Werte gesetzt:

Minimum Support 0.12
Minimum Confidence 0.6
Minimum Lift 1.2

Die Einstellung der Filter erfolgte, um eine gewisse Aussagekraft der Regeln vorauszusetzen wodurch ca. 100 Regeln übrig blieben. Die gefundenen Regeln ergeben sich teilweise aus dem Kontext heraus, sodass die bei der Unfallart "*Zusammenstoß zwischen Fahrzeug und Fußgänger*" und das Vorhandensein von Fußgängern einen hohen Wert sowohl beim Confidence als auch beim Lift verursacht. Ebenso nachvollziehbar ist es, dass beispielsweise die Items *Tageszeit Mittag* und *Lichtverhältnisse Tageslicht* einen hohen Zusammenhang vorweisen wie in Abbildung 8 zu sehen. Nach Durchsicht aller Regeln, die innerhalb der Filter geblieben sind, fließen sich keine Regeln finden, die neue Erkenntnis hervorbrachten und nicht aus der Sache per se erkennbar sind.

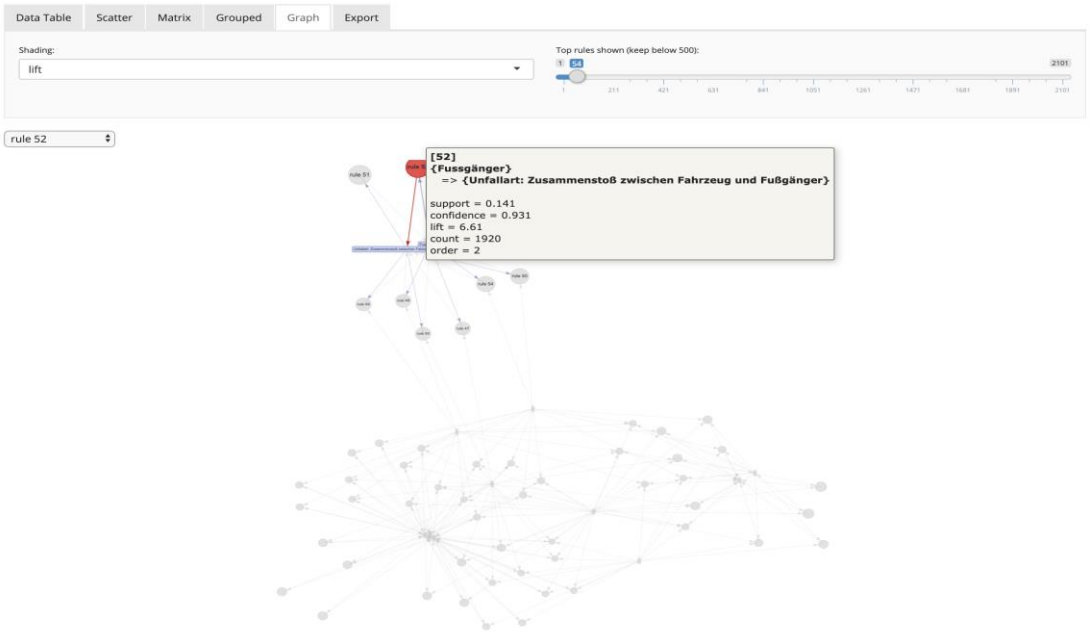


Abbildung 8: Assoziationsanalyse Grafen

Auswirkung von Ampeln auf Abbiege- und Kreuzungsunfälle

Der Fragestellung, welche Auswirkung Ampeln auf Abbiege- und Kreuzungsunfälle haben, wird im Jupiter Notebook 03_Analytics/Frage1_Ampeln_Abbiegeunfälle.ipynb nachgegangen. Zunächst wird zur Unterscheidung, ob der Unfall in der Nähe einer Ampel passiert ist, das Feld NAHE_AMPEL mit 1 für true und 0 für false erzeugt. Der Wert wird hierbei auf Basis der Fragestellung, ob der Identifikator des Ampeldatensatzes im konsolidierten Datensatz einen Null-Wert beinhaltet, gebildet. Weiterhin werden alle nicht Abbiege- oder Kreuzungsunfälle aus dem Datensatz herausgefiltert.

Um die Auswirkung einer Ampel ermitteln zu können, wird weiterhin festgelegt, dass dies geprüft werden kann, indem die durchschnittliche Anzahl an Abbiege- und Kreuzungs- unfällen an Kreuzungen oder Abbiegungen mit einer Ampel mit denen ohne Ampel verglichen wird. Während diese Berechnung für Kreuzungen oder Abbiegungen mit Ampel noch relativ einfach erscheint, da die Gruppierung dieser anhand des Identifikators der jeweiligen Ampel erfolgen kann, ist dies für Kreuzungen und Abbiegungen ohne Ampel schwieriger, da es keinen direkten

Bezugspunkt gibt, zu dem gruppiert werden kann. Um dieses Problem zu lösen kam die Idee auf die Bezugspunkte mithilfe von Clustern zu realisieren, welche über ein Clusteringverfahren wie beispielsweise DBSCAN gebildet werden. Nachdem dies zunächst im Jupiter Notebook getestet wurde und eine Reihe von Problemen auftraten, konnte ermittelt werden, dass im Werkzeug QGIS diese Funktionalität mittels vorgegebener Analysewerkzeuge verwendet und ein Mapping zwischen Cluster Identifikator und Unfall Identifikator exportiert werden kann. Das Nutzen der Funktionalität in QGIS bietet weiterhin den Vorteil, dass die entstandenen Cluster in unterschiedlichen Farben auf der Karte visualisiert werden und es somit ermöglicht wird visuell zu prüfen, ob die Cluster so wie erwartet gebildet wurden oder, ob gewisse Parameter angepasst werden müssen. Exemplarisch ist die Sicht in QGIS in Abbildung 9 visualisiert.

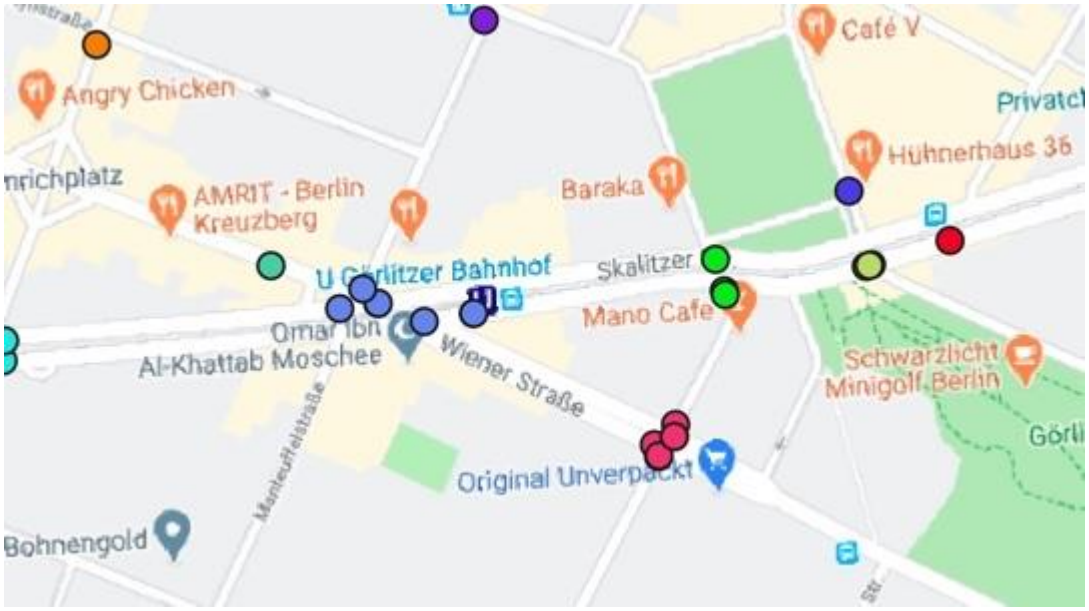


Abbildung 9: Geclusterte Unfälle

Mit diesen Clustern als Bezugspunkt konnte anschließend die durchschnittliche Anzahl an Unfällen je Kreuzung und Abbiegung mit und ohne Ampeln gebildet werden. Als Ergebnis kam heraus, dass die durchschnittliche Anzahl an Kreuzungen und Abbiegungen mit Ampel um circa 50% höher ist als an Kreuzungen und Abbiegungen ohne Ampel. Aus dieser Erkenntnis lässt sich allerdings noch nicht ableiten, dass Ampeln für mehr Kreuzungs- und Abbiegeunfälle sorgen, da hier nicht beachtet wird, dass Ampeln eher an Orten, an denen es in der Vergangenheit viele Unfälle gab oder die aufgrund ihrer Lage eher gefährlich sind, positioniert sein könnten. Um sich diesem anzunähern wurde im nächsten Schritt anstatt der absoluten Unfallzahlen die Summe der relativen Anzahl an Unfällen im Verhältnis zur Verkehrsmenge auf einer Straße betrachtet. Hierbei ergibt sich eine ähnliche Relation wie zuvor. Ein weiterer Kritikpunkt, welchen man bei diesem Vergleich bei Betrachtung der jeweiligen Verteilung der Unfallzahlen an Kreuzungen und Abbiegungen äußern könnte, wäre dass es deutlich mehr Kreuzungen und Abbiegungen mit einem Unfall ohne Ampel als mit Ampel in der Nähe gibt und dies den Durchschnitt stark beeinflusst. Um dies zu prüfen wurden abschließend alle Kreuzungen und Abbiegungen mit einem Kreuzungs- und Abbiegeunfall entfernt und die absolute und relative Zahl wieder verglichen. Dabei ergibt sich ein ähnliches Bild. Insgesamt scheint es also ein Indiz zu geben, dass Ampeln einen negativen Einfluss auf die Zahl von Abbiege- und Kreuzungsunfällen haben. Um dies zu validieren müssten jedoch weitere Einflussfaktoren für die Gefährlichkeit einer Kreuzung oder Abbiegung ermittelt werden und in die Berechnung einfließen. Diese scheinen jedoch im vorliegenden Datensatz nicht eindeutig erkennbar zu sein.

Identifikation potenzieller Faktoren für die Unfallschwere

Der Fragestellung, welche Merkmale eines Unfalls einen Einfluss auf die Schwere der Verletzungen der Unfallopfer haben könnten, wird im Jupiter Notebook 03_Analytics/ Frage2_Vorhersage_Verletzungsgrad.ipynb nachgegangen. Der Ansatz, der hier zur Beantwortung der Frage verfolgt wird, ist, dass mittels einer Klassifikation versucht wird anhand der Features im Datensatz vorherzusagen, ob es bei einem Unfall Schwerverletzte oder Tote gab (1) oder nur Leichtverletzte (0). Unfälle mit Toten werden im Zuge der Analyse mit dem gleichen Label versehen wie Unfälle mit Schwerletzten. Diese Entscheidung wurde zu Beginn in Abstimmung mit den Projektpartnern getroffen, da der Anteil der Unfälle mit Toten weniger als 0,5% der gesamten Unfälle ausmacht und eine gute Klassifikation auf Basis der bestehenden Merkmale fraglich erscheint. Als Klassifikationsverfahren wurde ein Random Forest verwendet, da sich dieser einfach und stabil umsetzen lässt. Weiterhin konnte in der Literatur in ähnlichen Projekten aufgezeigt werden, dass dieser die Besten Ergebnisse liefert[2]**.** Um die Güte der Klassifikation festzustellen und zu optimieren wurde zunächst die Accuracy (Anteil der richtig vorhergesagten Klassen) verwendet. Als Referenzwert für diese wurde die Verteilung der Labels zugrunde gelegt. So werden bei 80,6% der Unfälle Unfallopfer leicht verletzt. Es ließe sich somit eine Accuracy von 80,6% erreichen, wenn man immer vorhersagt, dass es nur Leichtverletzte gibt. Nach diversen Optimierungen des Random Forests konnte eine Accuracy von circa 84% erreicht werden, welche im Vergleich zum einfachen Vorhersagemodell keinen besonderen Mehrwert liefert.
Aufgrund des geringen Mehrwertes und der Überlegung, dass es für die Identifikation der Merkmale, welche den Verletzungsgrad vorhersagen, sinnvoller scheint möglichst viele schwere Verletzungen richtig vorherzusagen, wurde der Random Forest anschließend anhand des Gütekriteriums F1-Score optimiert. In diesem fließt zum einen der Anteil der richtig klassifizierten Unfälle, bei denen es Schwerverletzte oder Tote gabe, an allen die als solche klassifiziert wurden (Precision) und zum anderen der Anteil der richtig klassifizierten Unfälle mit Leichtverletzten an allen, die als solche klassifiziert wurde (Recall). Abschließend konnte ein F1-Score von 42% mit einer Precision von 32,8% und einem Recall von 56,9% erreicht werden. Die fünf relevantesten Features in dem Kontext waren der relative Anteil an Unfällen mit Schwerverletzten an der Verkehrsmenge auf der jeweiligen Straße (29,5%), der Anteil der Unfälle auf einer Straße im Verhältnis zur Verkehrsmenge (5,5%), die Unfallart Zusammenstoß mit einem vorausfahrendem/ wartendem Fahrzeug (4,6%), der Anteil an Krafträder auf einer Straße (3,6%) und der Anteil an Lieferwagen auf einer Straße (3,4%). Insgesamt scheinen die Features zur Vorhersage im Modell sehr vielfältig zu sein, da 20 weitere Faktoren noch jeweils ein bis drei Prozent der Vorhersage ausmachen. Es können somit keine eindeutigen Merkmale im Datensatz identifiziert werden, welche einen gravierenden Einfluss auf die Unfallschwere haben. Weiterhin bleibt kritisch zu betrachten, dass die Güte des Modells zu gering ist, sodass sich kaum sinnvolle Ableitungen ergeben. Ein Schluss, der aus diesen Ergebnissen gezogen werden kann, ist, dass der Datensatz nicht die passenden Features beinhaltet, um eine aussagekräftige Vorhersage der Schwere eines Unfalls treffen zu können. Im Zuge einer anschließenden Literaturrecherche wurde ermittelt, welche Features in ähnlichen Projekten verwendet wurden und welche Erkenntnisse in diesen gesammelt werden konnten. Eine wichtige Erkenntnis, welche in diesen erlangt wurde ist, dass die Schwere von Unfällen vor allem auf menschliche Faktoren zurückzuführen ist und kaum mit Umweltfaktoren erklärt werden kann [3]. Zu den Features, die in weiteren Projekten als aussagekräftig eingestuft wurden, zählen unter anderem die Lichtverhältnisse und die Geschwindigkeitsbegrenzungen[4]. Die Geschwindigkeitsbegrenzungen sind im Datensatz vorhanden jedoch insofern problematisch, dass es in Berlin vorrangig nur zwei

[2] Ramya et al., S.533, 2019
[3] Li et al., S.370, 2017
[4] Rana et al. S.3496, 2019

Klassenausprägungen gibt (30km/h und 50km/h), welche als solche noch nicht sehr hoch sind. Die Lichtverhältnisse lassen sich insofern nicht gut bestimmen, da in Berlin nahezu überall Straßenlaternen sind (Ergibt sich aus der Betrachtung des im Fis-Broker bereitgestellten Datensatzes zu Straßenlaternen in Berlin) und sich diese Information somit kaum aus den Unfalldaten entnehmen lässt.

Abschließende Betrachtungen

Die Auswertung des Unfalldatensatzes lässt resümierend folgende Annahmen zu: Die Analyse förderte keine bahnbrechenden Erkenntnisse zutage und es sind keine direkten Zusammenhänge der Features erkennbar. Das bedeutet nicht unbedingt, dass es keine Zusammenhänge gibt, sondern lediglich, dass die Daten die Informationen nicht hergeben. Wichtige Informationen zu Unfällen sind nicht vorhanden, die Aussagekräftig wären. So wären Informationen, die die Verkehrsteilnehmer betreffen weitaus aussagekräftiger, als Umweltbedingungen wie Tageslicht oder Fahrbahnbeschaffenheit. Wie schon in der Assoziationsanalyse festgestellt gibt es keine Regeln, durch die abzusehen ist, warum schwere Unfälle geschehen. Im Umkehrschluss bedeutet das wiederum, jeder Unfall hat einen menschlichen Fehler als Ursache. Daher wurden die folgenden Handlungsempfehlungen erarbeitet welche jedoch lediglich als Denkanstoß dienen sollen und deren weitreichenden Folgen nicht im Detail evaluiert wurden. Die Handlungsempfehlungen sind von der Politik abzuwägen und durchzusetzen

- **Vorsicht und Aufmerksamkeit im Straßenverkehr**
  Es scheint trivial, dennoch sollte es jedem Bürger bewusst sein, dass er sein Leben gefährdet, nimmt er am Straßenverkehr teil. Es sollte in größerem Umfang die Aufmerksamkeit auf die Gefahren im Straßenverkehr hingewiesen werden, auch um die Leichtsinnigkeit von Verkehrsteilnehmer anzuprangern.
- **Ausbau und Schutz von Fahrradwegen**
  Wie in der deskriptiven Analyse festgestellt, sind insbesondere in der Innenstadt Radfahrer die gefährdetsten Verkehrsteilnehmer. Sie sollten durch den Ausbau und Schutz von Fahrradwegen unterstützt werden.
- **Regelmäßige Prüfung der Fahrtüchtigkeit**
  Die Debatte um erneute Führerscheinprüfung ist nicht neu und es lässt sich nicht von der Hand weisen, dass die Reaktionsfähigkeit im höheren Alter abnimmt. Das Argument der Diskrimination wiegt nach Ansicht der Autoren weniger, als ein Menschenleben.
- **Verschärfung der Strafen für Verkehrssünder**
  Höhere Geldstrafen und Führerscheinentzug sollten bei vorsätzlicher Gefährdung des Straßenverkehrs durchgesetzt werden, um potentielle risikoreiche Fahrverhalten zu unterbinden.
- **Bodenampeln**
  Unaufmerksame Smartphone User sind mehr denn je im Straßenverkehr gefährdet. Die Einführung von Ampeln soll die Gefährdungsgruppe schützen, da die Signale peripher wahrgenommen werden können, während sie auf ihr Smartphone sehen.
- **Variable Tempolimits**
  Die zeitliche Auswertung hat gezeigt, dass im Feierabendverkehr die meisten Unfälle geschehen. Stress und Übermüdung tragen dazu bei die Reaktionszeit zu verlängern, daher ist eine Erweiterung der variablen Höchstgeschwindigkeit denkbar. Insbesondere während der Werktage von 7 bis 9 Uhr und 15 bis 18 Uhr wäre generell eine Senkung der Tempolimits sinnvoll.
- **Reduzierung des Individualverkehrs und Ausbau ÖPNV**
  Mit 11.152 von 13.649 der ausgewerteten Unfälle, sind PKWs in mehr als 80% der Fälle beteiligt. Aus diesem Grund sollten den Bürgern attraktive Angebote gemacht werden, auf ein privaten PKW zu verzichten. Beispielsweise ist ein Jahresticket der BVG für 365 € im Gespräch. Entgegengesetzt kann auch die Attraktivität eines PKW durch Maut in der Innenstadt oder erhöhte Parkpreise gesenkt werden
- **Förderung des autonomen Fahrens**
  Wie festgestellt wurde ist der größte Risikofaktor menschliches Versagen. Um das zu

umgehen sollte, ausgiebiges Testen der Funktionsfähigkeit vorausgesetzt, autonomes Fahren gefördert werden und auf den Straßen weiter etabliert werden. Ebenso ist es denkbar, dass künftig bei Neuzulassungen von PKWs das Vorhandensein eines Fahrassistenten Pflicht wird.

Jeder Mensch, der verletzt wird oder umkommt, ist ein Mensch zu viel. Nichtsdestotrotz steht Berlin im Bundesvergleich bei den Getöteten pro eine Million gut dar. Nur Bremen hat mit neun getöteten pro Millionen Einwohner weniger als Berlin mit 12. Grund dafür ist die Siedlungsstruktur der Stadtstaaten. Im Kontrast dazu haben Brandenburg mit 57 und Sachsen-Anhalt 63 Getötete pro Millionen Einwohner [5]. Naheliegend ist daher in diesen Gebieten Ursachenforschung zu betreiben und zunächst an dieser Stelle Optimierungen vorzunehmen, um die Anzahl an Verunglückten zu senken.

[5] Destatis, 2018

Literatur

Destatis: https://www.destatis.de/DE/Presse/Pressemitteilungen/2019/02/PD19_069_46241.html (abgerufen am 18.02.2020)

Li, L. & Shrestha, S. & Hu, G. (2017). Analysis of road traffic fatal accidents using data mining techniques. IEEE 15th Int. Conf. on Software Engineering Research, Management and Applications (SERA).363-370. Abrufbar unterhttps://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7965753 (abgerufen am 18.02.2020)

Ramya, S. & Reshma, SK & Manogna, V. & Saroja, Y. & Gandhi, G. (2019). Accident Severity Prediction Using Data Mining Methods. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.528-536. Abrufbar unter: https://www.researchgate.net/publication/332408305_Accident_Severity_Prediction_Using_Data_Mining_Methods (abgerufen am 18.02.2020)

Rana, V. & Joshi, H. & Parmar, D. & Jadhav, P. & Kanojiya, M (2019) Road Accident Prediction using Machine Learning Algorithm. International Research Journal of Engineering and Technology (IRJET).3494-3496. Abrufbar unter https://mail.irjet.net/archives/V6/i3/IRJET-V6I3657.pdf (abgerufen am 18.02.2020)

TSB: https://www.technologiestiftung-berlin.de/de/stiftung/ (abgerufen am 19.02.2020)