29.4.2023

**Dokumentation**

Eine Dokumentation zur Prüfungsleistung im Rahmen der Vorlesung Data Visualization

**Laura Krone**

**Kurs**: WWI2022F

**Matrikelnummer**: 1130769

**Vorlesung**: Data Visualization

**Dozent**: Florian Eichin

Inhaltsverzeichnis

**1. Formulating your brief**1

**2. Working with Data1**

2.1 Erhebung der Datensätze1

2.2 Untersuchung der Datensätze1

2.3 Transformation der Datensätze2

2.3.1 OrigDelaySummary.csv & DestDelaySummary.csv 2

2.3.2 Airlines.csv 2

2.3.3 Connections.csv 2

2.3.4 AirlineOrgDelay.csv & AirlineDestDelay.csv2

2.3.5 BestConnect.csv 2

2.3.5 SortedAirports.csv 2

**3. Establish your editorial thinking2**

3.1 Visualisierung 1 - Flughäfen 2

3.2 Visualisierung 2 - Fluglinien 2

3.3 Visualisierung 3 – State Verbindungen 2

3.4 Visualisierung 4 -Verspätungsvergleich 2

3.5 Visualisierung 5 – Beste Verbindungen 2

**4. Develop the design solution2**

4.1 Visualisierung 1 - Flughäfen 2

4.2 Visualisierung 2 - Fluglinien 2

4.3 Visualisierung 3 – State Verbindungen 2

4.4 Visualisierung 4 - Verspätungsvergleich 2

4.5 Visualisierung 5 - Beste Verbindungen 2

**1 Formulating your brief**

Als eine Prüfungsleistung für die Vorlesung Data Visualization wurden am 28. März 2023 seitens des Dozenten, Florian Eichin, 5 Visualisierungen für Erkenntnisse aus einem frei wählbaren Datensatz, sowie eine Dokumentation dieser gefordert. Stakeholder dieses Projektes sind der Dozent Florian Eichin sowie der Kurs WWI2022F, insbesondere Laura Krone. Die Zielgruppe der von Laura Krone erstellten Visualisierungen beschränkt sich einzig und allein auf Herr Florian Eichin. Da nicht bekannt ist welche Erkenntnisse für die Zielgruppe für Interesse sind, wird im Folgenden darauf geachtet, dass der Betrachter von der Interaktivität des Mediums einer HTML-Seite profitieren kann und so nach Belieben die für ihn relevanten Aspekte entdecken kann (Exploratory). Da weiterhin nichts über die Vorkenntnisse der Zielgruppe bekannt ist und aufgrund der asynchronen Präsentation der Darstellungen keinerlei Rückfragen möglich sind, werden zusätzlich auch Aspekte eines Explanatory – Ansatzes benötigt. Anfängliche Problemstellung aller Visualisierungen ist es also beider Ansätze zu verknüpfen. Da es sich bei der Prüfungsleistung um ein einmaliges Event handelt könnte die Wiederverwertbarkeit des Visualisierungsprozesses vernachlässigt werden, doch wird darauf im Folgenden, aufgrund eines besseren Stils und einer nachhaltigeren Verarbeitung der Daten, dennoch geachtet.

**2 Working with data**

**2.1 Erhebung der Datensätze**

Durch den Auftraggeber, Herr Florian Eichin, wurde im Rahmen der Vorlesung ein Datensatz über Flüge am 01.01.2015 zur Verfügung gestellt. Dieser Datensatz ist ein Teilausschnitt eines größeren Datensatzes, welcher auf Kaggle zu finden ist. Um die Datendichte zu erhöhen, die Analyse repräsentativer zu machen und dennoch eine vertretbare Performance gewährleisten zu können, fiel die Entscheidung darauf für 4 von 5 Grafiken einen etwas größeren Teilausschnitt der auf Kaggle zu finden Daten, welcher nun den ganzen Januar enthält, zu verwenden. Bei der 5. Grafik wäre die Visualisierung (Vis. 2: Airlines) des ganzen Januars nicht sinnvoll gewesen, weshalb dort nur ein Tag visualisiert und mit zusätzlichen Daten aus dem ganzen Januar ergänzt wurde.

**2.2 Untersuchung der Datensätze**

Zu untersuchen gab es im Summe 4 Datensätze, die Flugdaten des ganzen Januars 2015, sowie 3 Hilfsdatensätze, welche die Airline-Namen und die Flughäfen-Namen und die States, in welchen die Flughäfen liegen enthalten. Das Hauptdatenset beinhaltet Eckdaten über Inlandsflüge in den USA im Januar 2015. Sie enthielten das IATA-Kürzel des Start- und des Zielflughafens, sowie der Airline. Darüber hinaus umfasste er auch Eckdaten, wie unteranderem geplante Startzeit, Startverspätung, geplante Zeit, tatsächliche Zeit, geplante Ladezeit, Landeverspätung sowie die Koordinaten des Start- und des Zielflughafen. Insgesamt umfasst der große Datensatz 16 Spalten und 457013 Zeilen. Zusätzlich zu diesem Hauptdatensatz wurden noch 2 Hilfsdatensets verwendet. Wichtig für die Visualisierungen waren im Airline Datensatzes sind die Spalten IATA, welche das zweistellige IATA-Kürzel der Airline enthielt, sowie Name, welche den gängigen Namen der Airline enthielt. Die Spalten Alias, IACO, Country und Active wurden nicht benötigt. In Summe hatte dieses Datenset 8 Spalten und 6162 Zeilen.

Der dritte Datensatz enthielt zusätzliche Informationen zu den Flughäfen. Er bestand aus nur 7 Spalten und 341 Zeilen. Die Spalten beinhaltet Auskünfte zu IATA, Flughafen Name (AIRPORT), Stadt, Staat, Land sowie über die Koordinaten.

Der letzte und kleinste Datensatz beinhaltete die Kürzel der US States in der Spalte „abbreviation“ und den ausgeschriebenen Namen der US States. Es hat somit 2 Spalten und 50 Zeilen. In diesem Datenset fehlen allerdings die nicht inkorporierten Außengebiete der Vereinigten Staaten, weshalb diese händisch ergänzt werden mussten, da Flüge dorthin dennoch im Hauptdatenset vorhanden sind.[[1]](#footnote-1)

**2.3 Transformation der Datensätze**

Bei der Erstellung des Hauptdatensatzes aus den Originaldaten, wurden alle nicht benötigten Spalten entfernt. Zudem wurde ausgefallene Flüge und Zeilen die NAN-Werte enthielten gelöscht. Damit die Daten auch geographisch eingeordnet werden konnten wurde der Datensatz zudem auch noch mit den Koordinaten der Flughäfen, welche ebenfalls in den Originaldaten zu finden waren, angereichert. Danach galt es die Daten einem gut in JavaScript verwertbaren Format für die Visualisierungen vorzubereiten. [[2]](#footnote-2)

**2.3.1 OrigDelaySummary.csv & DestDelaySummary.csv**

Diese beiden Datensätze wurden für die Visualisierungen Eins und Vier benötigt. Die Erstellung fand einmal für den Start- und einmal für den Zielflughafen statt. Als erstes wurden die benötigten Spalten mit den Informationen zum Flughafen, dessen Position sowie der Verspätung aus dem Hauptdatensatz entnommen und eine neue Spalte „hour“ erzeugt, welche die Abflugs- bzw. Ankunftsstunde enthält. Danach wurden die neuen Datensets nach Flughafen bzw. nach Flughafen und Stunde gegroupt. Mithilfe dieses Objekts wurde dann ein neues Datenset befüllt. Dieses enthält das Flughafen IATA Kürzel, die Stunde, wobei in der Stunde -1 die Durchschnittswerte des kompletten Tages stehen, die Flughafen Position, die Höchste, niedrigste und durchschnittliche Verspätung pro Stunde an einem Flughafen, sowie die durchschnittliche Anzahl an Flügen pro Tag und Stunde.

2.3.2 Airlines.csv

Zur Erstellung dieses Datensatzes, welcher in den Visualisierungen Zwei, Vier und Fünf verwendet wird, kam eines Hilfsdatensets zum Einsatz. Dieses wurde zu Beginn gefiltert, sodass es nur noch die Airlines enthielt, welche auch wirklich innerhalb des großen Flugdatensets vertreten waren. Im großen Datensatz wurde für diesen neuen Datensatz auch eine neue Spalte names „AirTimeDiff“ erstellt, welche von der tatsächlich gebrauchten Zeit die geplante Zeit abzieht und somit dann die aufgeholte oder die für den Flug zusätzlich gebrauchte Zeit enthält. Für jede Airline, sowie für die Summe aller Airlines, wurden dann die Maximal-, Minimal-, Durchschnitts- und Quantils-Werte ermittelt und in einen Datensatz eingepflegt.

2.3.3 Connections.csv

Dieser Datensatz bedurfte ebenfalls eines Hilfsdatensatzes, da zu jedem Flughafen der US State benötigt wurde in dem er liegt. Als erstes wurde der Hauptdatensatz durch die States der Start- und Zielflughäfen ergänzt und nach diesen gegrouped. Die Anzahl der Verbindungen wurden dann einem Dictionary mit dem Key „Startflughafen,Zielflughafen“ zusammen mit der durchschnittlichen Start und Landeverspätung abgelegt. Danach wurde eine Liste mit allen States erstellt in welche oder aus welchen Flüge im Januar 2015 geflogen sind. Da die Daten für Visualisierung Drei in einem speziellen Format benötigt wurden, um ein Chord-Diagramm visualisieren zu können, wurde in diesem Fall der CSV-String von durch eine Schleife selbst zusammengesetzt.

2.3.4 AirlineOrgDelay.csv & AirlineDestDelay.csv

Diese beiden CSV-Dateien wurden auf eine sehr ähnliche Art erstellt. Der erste Datensatz enthält dabei die Werte für die Airlines beim Start und der zweite für die Airlines bei der Ankunft. Beide enthalten die Werte durchschnittliche, minimale und maximale Verspätung sowie die durchschnittliche Anzahl der Flüge pro Stunde. Beide enthalten ebenfalls die Start- bzw. Landestunde bzw. den Wert -1, wenn es um die Werte für den ganzen Tag geht. Genutzt werden beide Datensätze in der Visualisierung Vier.

2.3.5 BestConnect.csv

Die BestConnect CSV beinhaltet Informationen zu einer festgelegten Kombination von Startflughafen, Zielflughafen und Airline, welche im folgenden „Verbindung“ genannt wird. Für jede Verbindung werden die Werte minimale, maximale und durchschnittliche Startverspätung, Zielverspätung sowie die gebrauchte Zeit in den Datensatz notiert. Zusätzlich enthält der Datensatz auch die von der Airline geplante Zeit.

2.3.5 SortedAirports.csv

Die letzte CSV- Datei fungiert selbst als Hilfsdatensatz. Sie enthält alle im Hauptdatensatz vorkommenden Flughäfen, sowie ihren Namen nach IATA-Kürzel sortiert.

**3 Establish your editorial thinking**

Grundsätzliches Ziel der Visualisierungsreihe ist, dass der Betrachter am Ende die Möglichkeit hat, für ihn relevante Einsichten über Verspätungen des Monats Januars zu erkennen und somit zukünftig in der Lage ist Verspätungen möglichst aus dem Weg zu gehen und seine Chancen auf ein pünktliches Ankommen zu erhöhen. Zum einen soll die Rolle der Start- und Zielflughäfen für die Verspätung näher betrachtet werden, darüber hinaus ist aber auch interessant wie es generell bei der Verbindung zwischen States mit Verspätungen aussieht, letzter einzeln betrachteter Faktor für die Verspätungen sind die Airlines. Zum Schluss wird auch noch die Kombination aus Startflughafen, Zielflughafen und Airline evaluiert.

**3.1 Visualisierung 1 – Flughäfen**

Es werden alle Flughäfen betrachtet und deren minimale, maximale und durchschnittliche Verspätung in Kombination mit der Anzahl der abgeflogenen bzw. gelandeten Flüge betrachtet. Die Visualisierung soll so auch helfen zu verstehen, wo im Land die größten Flughäfen sind und wann im Tagesverlauf an diesen die meiste Aktivität stattfindet. Dabei wird zusätzlich für die Verspätung sowohl der Durchschnitt einer Stunde des Tages im Laufe des Monats als auch ein Durchschnitt des ganzen Monats errechnet, um so generell den Verlauf eines durchschnittlichen Tages darstellen zu können.

**3.2 Visualisierung 2 – Fluglinien**

Es werden repräsentativ alle Flüge einer Airline an einem Tag betrachtet, um so Einblicke über Verspätungen und örtliche Vertretung zu erlangen. Ergänzt werden diese Erkenntnisse mit generellen für den Monat durchschnittlichen Verspätungswerten.

**3.3 Visualisierung 3 – State Verbindungen**

Die Idee hinter der dritten Visualisierung ist es, zu erkennen welche Verbindungen es zwischen den verschiedenen US-States gibt und zusätzliche Informationen über die Häufigkeit, sowie die zu erwartende Pünktlichkeit der Verbindungen zu geben.

**3.4 Visualisierung 4 – Verspätungsvergleich**

Bei dieser Visualisierung sollen die Verspätung an Startflughäfen, Zielflughäfen, Airlines beim Abflug und Airlines bei der Landung untereinander verglichen, um so jeweils die 10 pünktlichsten und die 10 unpünktlichsten zu ermitteln. Dies soll ebenfalls wieder im Laufe eines Tages visualisierbar sein.

**3.5 Visualisierung 5 – Beste Verbindung**

Die letzte Visualisierung soll dem Benutzer die Möglichkeit geben eine für ihn relevante Verbindung zwischen 2 Flughäfen auszuwählen und alle Airlines angezeigt bekommen, die diese Verbindung anbieten. Es werden auch zusätzliche Informationen, wie die Verbindungsdichte, Dauer der Flüge und zu erwartende Verspätung für den Benutzer verfügbar werden.

**4 Develop the design solution**

Das Design der Visualisierungen orienteierte sich in allen Fällen an dem Exploratory Ansatz. Dem Betrachter sollten möglichst viele Optionen zur persönlichen Konfiguration der Visualisierung geboten werden. Dabei sollte die Bedienung möglichst intuitiv sein. Um dennoch auftretenden Unklarheiten vorzubeugen, wurden in allen Fällen Tooltips verwendet. Für alle Visualisierungen wurde dabei oben eine einheitlich designte Überschrift und ein Menu mit Dropdownbutton verwendet. Die Visualisierungen haben alle ein schlichtes Design mit einem kohärenten Farbschema. Für alle Grafiken wird eine klare Legende angezeigt und es werden Hover-Effekte verwendet, um die Werte klarer ablesen zu können und zudem zusätzliche Informationen zu bereitzustellen.

**4.1 Visualisierung 1 – Flughäfen**

Die Anzahl der abgeflogenen bzw. gelandeten Flüge wird durch die Größe der Kreise für die Flughäfen dargestellt. So erhält wird die Aufmerksamkeit des Betrachters sofort auf die größten und somit relevantesten Flughäfen gelenkt, ohne die Informationen über die kleineren zu unterschlagen. Um die Verspätung an den Flughäfen visuell darzustellen, wird die Farbe der Kreise verwendet. Um den Betrachter nicht durch verschiedene Skalen zu verwirren, bleibt die Farbskala unabhängig von der gewählten Verspätungsoption gleich und wurde mit einem möglichst breiten Farbschema gewählt. Durch die Verwendung einer Karte hat der Betrachter in dieser Visualisierung die größten Flughäfen der USA auch geografisch einordnen zu können. Dabei wurde für das Farbschema Karte der Darkmode gewählt, um bei der Visualisierung die Möglichkeit zu haben einen pünktlichen Flughafen mit der neutralen Farbe Weiß darstellen zu können. Neben der statischen Darstellung der Werte gibt es noch eine weitere Option, bei der die Werte pro Stunde animiert werden, um so die Veränderungen im Laufe eines typischen Tages darstellen zu können.

Es wird darüber hinaus auch ein Tooltip verwendet, damit der Betrachter die genauen Werte ebenfalls ansehen kann.

**4.2 Visualisierung 2 – Fluglinien**

Diese Visualisierung arbeitet ebenfalls mit einer Karte, die auch im Darkmode angezeigt wird. Auf dieser Karte werden alle Flüge eines Tages visualisiert. Über einen Dropdown Button hat der User zudem auch die Möglichkeit auszuwählen, welchen Tag er visualisiert haben will und so verschiedenen Tage vergleichen. Standardmäßig werden Flüge aller Airlines angezeigt, doch sind die Flüge auch über einen Dropdown - Button filterbar, sodass nur noch Flüge einer Airline angezeigt werden. So erhält der Betrachter einen Überblick darüber, welche Airlines du welchen Zeiten und in welchen Gebieten fliegen. Für jeden Flug wird zudem auch die Verspätung durch die Farbe visualisiert. Diese verändert sich im Laufe des Fluges, wenn Verspätung auf- oder abgebaut wird. Die Farbskala ist dabei die gleiche, wie die der ersten Visualisierung für den Betrachter einheitlich und möglichst nicht verwirrend zu sein.

Neben der Visualisierung auf der Karte hat der Benutzer zudem auch die Möglichkeit sich zusätzliche Informationen über die angezeigte Airline abzurufen. Dafür steht ihm eine Checkbox zur Verfügung, die eine Zeile mit Zusatzinformationen einblendet. In dieser Zusatzzeile befinden sich 3 Boxplots die für den ganzen Monat die Startverspätung, die Aufgeholte Zeit in der Luft und die Verspätung bei Ankunft visualisieren. Zudem wird auch der Anteil der Airline an allen Flügen in einem Donut-Diagramm sowie die absolute Anzahl aller von der Airline geflogenen Flüge im ganzen Januar angezeigt. Das Donut–Diagramm besitz dabei ebenfalls einen Tooltip, welcher den Namen der Airline anzeigt sowie den Prozentsatz an allen Flügen. Die verschiedenen Abteile des Diagramms können auch angeklickt werden, um so zwischen den Airlines zu wechseln.

Da die Boxplots klein sind, um die Karte nicht zu sehr zu überdecken, gibt es für die Boxplots auch noch einen Hover Effekt, der dann über eine Tooltip die visualisierten Werte anzeigt.

**4.3 Visualisierung 3 – State Verbindungen**

Bei dieser Visualisierung bot sich ein Chord-Diagramm an, um die Verbindungen zwischen den States darzustellen, die Bereite ist dabei repräsentativ für die Anzahl. Da nicht bekannt ist, ob dem Betrachter die Kürzel der States bekannt sind, befindet sich rechts neben der Visualisierung eine Legende mit der Farbe, dem Kürzel des States und dem vollständigen State-Namen. Über einen Klick auf einen bestimmten Pfad kann dieser hervorgehoben werden. Zudem ist auch das Hervoreben einer ganzen Gruppe über einen Klick auf die Legende oder die Gruppe in der Visualisierung möglich. So kann der Benutzer der Visualisierung genau die Verbindungen hervorheben, die für ihn interessant sind. Über einen Button oben im Menu können alle Verbindungen wieder unselektiert werden.

Ein Hover Effekt über den Verbindungen gibt dem Betrachter darüber hinaus auch die Möglichkeit, genaueres über die Verbindungen zu erfahren. Diese zusätzlichen Informationen beinhalten die Anzahl der Flüge auf dieser Verbindung in beide Richtungen im ganzen Januar 2015 als auch die durchschnittlichen Start- und Ankunftsverspätungen.

**4.4 Visualisierung 4 – Verspätungsvergleich**

Für die vierte Visualisierung wurde eine Balken Diagramm gewählt, da dieses sehr gut einen direkten Vergleich ermöglicht. Welche Daten in dem Balkendiagram dargestellt werden kann der Benutzer selbst entscheiden. Er kann zwischen den Optionen Flughafen bei Abflug, Flughafen bei Ankunft, Airline bei Abflug und Airline bei Ankunft wählen. Außerdem hat der Benutzer der Möglichkeit zu wählen, ob die höchste, niedrigste oder durchschnittliche Verspätung visualisiert werden soll. Zudem kann der Betrachter auch noch über ein Dropdown Button auswählen, ob die Balken aufsteigend oder absteigend sortiert sind. Der Benutzer der Visualisierung hat darüber hinaus auch die Möglichkeit Flughäfen oder Airlines mit für ihn zu wenigen Flügen pro Stunde herauszufiltern, da diese durch ihre geringe Anzahl wesentlich extremere Werte aufweisen können. So werden dem Benutzer nur noch die großen Flughäfen bzw. Airlines angezeigt und der ihm ermöglichte Vergleich ist repräsentativer. Um den User optisch nicht zu überfordern wurde das Design schlicht gehalten und es werden maximal nur die Top 10 Ergebnisse angezeigt. Standardmäßig werden die Ergebnisse des ganzen Januars verwendet. Neben dieser Standardoption gibt es zudem auch eine Animationsmöglichkeit, die dieselben Werte pro Stunde visualisiert.

Durch diese Visualisierung hat der Benutzer die Möglichkeit die Verspätungen der verschiedenen Optionen direkt miteinander zu vergleichen.

Die Balken sind zur besseren Verdeutlichung mit der gleichen Farbskala, wie die in der Visualisierungen 1 und 2 eingefärbt. Darüber hinaus gibt es oben eine Skala für die Verspätung in Minuten und einen Tooltip der den genauen Wert sowie die durchschnittliche Anzahl der Flüge pro Stunde. Durch diese Zusatzinformation kann der Benutzer für sich selbst einschätzen, wie stark er Repräsentativität dieses Balkens bewerten soll.

Um auch bei einer starken Verspätung die Enden der Balken sehen zu können, ohne dass die niedrigen unleserlich klein werden, verändert sich die Skala des Balkendiagramms, wenn bei Auswahl der maximalen Verspätung absteigend sortiert werden soll.

**4.5 Visualisierung 5 – Beste Verbindung**

Die letzte Visualisierung stellt entweder die Verspätung oder die Zeit für eine Verbindung dar. Für diese Visualisierung wurde in der Verspätungsvariante ein Lollipop-Graph gewählt, um die wichtigsten Kennzahlen (minimale, durchschnittliche und Maximale Verspätung) einer Airline Verbindung zwischen zwei Flughäfen darzustellen. Oben im Menu kann der User einen Startflughafen und einen Zielflughafen seiner Wahl auswählen. Zudem hat er die Möglichkeit die Dropdownbuttons mithilfe eines Radiobutton Menus zu filtern um nur noch Flughäfen angezeigt zu bekommen mit denen eine Verbindung existiert.

Da es bei den Filtern eine Option gibt, die es dem User ermöglicht z.B. für alle Startflughäfen, die einen Zielflughafen die Verbindungen anzuzeigen, werden es zu Teil sehr viele Verbindungen. Um dennoch stets die Skala sehen zu können befindest sich diese fixiert am unteren Rand und weitere Zeilen werden erst durch Scrollen sichtbar.

Für die Variante, welche die gebrauchte Zeit, visualisiert, wird der Lollipop-Graph durch ein zusätzliches Rechteck ergänzt, welche die für eine Verbindung durchschnittlich geplante Zeit angibt. Eine neue Legende erklärt die benutzen Symbole in der unteren rechten Ecke.

Zusätzlich dazu gibt es bei dieser Visualisierungsvariante eine farbliche Hinterlegung die Anzeigt, bei welcher zeitlichen Grenze Kurz-, Mittel- und Langstreckenflüge liegen

Mithilfe dieser Grafik kann der Benutzer vergleichen, welche Airline für ihn die beste Verbindung zwischen zwei Flughäfen anhand der Verspätung und anhand der gebrauchten Zeit anbietet. Zudem kann er mithilfe des Tooltips ebenfalls sehen, wie häufig die Airline pro Monat diese Verbindung fliegt und damit Rückschlüsse auf die Verbindungsdichte schließen.

1. Siehe Anhang („connections.html“) [↑](#footnote-ref-1)
2. Vgl. datenvorbereitung.py [↑](#footnote-ref-2)