Dokumentation: Data Visualization Florian Eichin

Formulating your brief:

Um das Wirtschaftsinformatik-Data Science Studium erfolgreich abzuschließen, wird in dem Modul Data Visualization eine Prüfungsleistung vom Dozenten Herrn Florian Eichin eingefordert. Die Anforderungen enthalten fünf visuell dargestellte Grafiken eines Datensatzes mit anschließender schriftlicher Ausarbeitung. Als beteiligte Stakeholder werden sowohl Herrn Florian Eichin als auch der Kurs WWI2022F identifiziert. Zielgruppe der Visualisierung ist der Auftraggeber Herr Florian Eichin. Die Visualisierung beinhaltet fünf Grafiken, welche auf einer Webseite aufgelistet sind und nach der Komplexität absteigend sortiert wurden. Da es sich bei der Visualisierung um eine Abgabe handelt und somit keine Fragen beantwortet werden können, wie es im Falle einer Präsentation wäre, wurde darauf geachtet, die Grafiken nicht zu kompliziert und zu überladen zu gestalten. Durch eine genaue Beschriftung der Grafiken und der Legenden, soll die Zielgruppe ohne viel Aufwand direkt verstehen, was diese Grafik darstellt. („Explanatory“). Mit eingebauten Dropdown Menüs ist der User auch in der Lage, die Visualisierungen in kleinem Maße zu verändern und somit Interessante Informationen selbst zu suchen. („Exploratory“). Die weitere Verwendbarkeit der Visualisierungen ist nicht geplant, da es sich hierbei um eine einmalige visuelle Darstellung der Daten handelt, es wurde auf eine saubere und ehrliche Darstellung der Daten geachtet.

Working with data:

1. Erhebung der Daten:

Für die Datenvisualisierung wurde durch Herrn Florian Eichin ein bereits gesäubertes Daten Set mitgeliefert. In diesem sind Flüge, welche am 01.01.2015 stattfanden, verschiedener Fluggesellschaften aus den USA enthalten. Dieses Subdatenset ist ein kleiner Ausschnitt eines großen Datensets welches auf Kaggle.com zu finden ist. In dem großen Datenset sind Flüge vom 01.01.2015 bis zum 31.12.2015 der USA enthalten. Um eine qualitativ hochwertige Auswertung abliefern zu können wurde ein neues Subdatenset erstellt, welches aus dem großen Datenset besteht und den kompletten Januar mit seinen Flugdaten enthält. Da ich nicht so schnell mit der Reinigung von Datensätzen bin und uns vom Auftraggeber ein bereits gereinigtes Datenset überlassen wurde, habe ich das komplette gereinigte Datenset von einem Kommilitonen (Lucas Michael) nehmen dürfen und mir selbst die Januardaten extrahiert. (Ich hoffe das ist für Sie in Ordnung, dass das Datenset von einem Kommilitonen gereinigt wurde).

2. Untersuchung der Daten:

Es gab ein großes und zwei kleine Datensets, die untersucht werden konnten. Das kleinste der Datensets (airlines.csv) beinhaltet 14 Zeilen. In der ersten Spalte befinden sich die IATA\_CODES (dies ist ein Code, der von International Air Transport Association entwickelt wurde, um Fluggesellschaften eindeutig identifizieren zu können). In der zweiten Spalte befindet sich unter AIRLINE der Name der Fluggesellschaft.

Das mittleren Datenset (airports.csv) enthält 322 einzigartige Zeilen. In der ersten Spalte befindet sich wie im ersten Datenset der IATA\_CODE (auch dieser wurde von der International Air Transport Association entwickelt, steht hier jedoch für das eindeutige Identifizieren der Flughäfen). Zudem sind die Spalten: AIRPORT (Name des Flughafens), CITY (Stadt, in der sich der Flughafen befindet), STATE (Staat, in dem der Flughafen liegt), COUNTRY (Land, in dem sich der Flughafen befindet. In diesem Fall sind alle Flughäfen in den USA), sowie LATITUDE (Breitengrad, auf dem sich der Flughafen befindet) und LONGITUTE (Längengrad, auf dem sich der Flughafen befindet) enthalten.

Im dritten und somit dem größten Datenset ( flights.csv) sind wichtige Daten aller vorhergesehenen Flüge der USA über das komplette Jahr 2015 enthalten. Das Datenset beinhaltet 31 Spalten und 1.048.576 Zeilen. Um einen besseren Überblick bereitzustellen, werden die Spaltennamen hier einmal aufgelistet: YEAR (das Jahr), MONTH (der Monat), DAY\_OF\_WEEK ( Wochentage von „1“ (Montag) – „7“ (Sonntag)), AIRLINE ( IATA-CODE der Fluggesellschaften), FLIGHT\_NUMBER ( Flugnummer), TAIL\_NUMBER ( Flugzeugnummer), ORIGIN\_AIRPORT ( Startflughafen als IATA\_CODE), DESTINATION\_AIRPORT ( Zielflughafen als IATA\_CODE), SCHEDULED\_DEPARTURE ( geplante Abreise), DEPARTURE\_TIME ( tatsächliche Abflugzeit), DEPARTURE\_DELAY ( Verspätung beim Abflug), TAXI\_OUT (Zeit auf der Rollbahn), WHEELS\_OFF ( Zeit an dem die Räder den Boden verlassen), SCHEDULED\_TIME ( geplante Reisezeit), ELAPSED\_TIME (tatsächliche Reisezeit), AIR\_TIME ( Flugzeit), DISTANCE ( Distanz des Fluges), WHEELS\_ON (wann die Räder ausgefahren wurden), TAXI\_IN (Rollbahnzeit zurück ans Gate), SCHEDULED\_ARRIVAL (geplante Ankunft), ARRIVAL\_TIME ( tatsächliche Ankunftszeit), ARRIVAL\_DELAY ( Ankunftsverspätung), DIVERTED ( umgeleitet), CANCELLED ( gestrichen „0“ – nein, „1“ -ja), CANNCELLATION\_REASON ( der Grund, weswegen der Flug gestrichen wurde). In den letzten fünf Spalten stehen, falls Verspätung angefallen, die Verspätungsdauer, die auf Belegung des Luftraum, aus Sicherheitsgründen, auf die Fluggesellschaft, andere verspätete Fluggesellschaften oder das Wetter zurückzuführen sind. Bei der Untersuchung der Daten stachen einige Spalten mit sehr vielen Nullwerten hervor, so z.B. die Spalten mit Gründen der Verspätung. Dies ist vermutlich darauf zurückzuführen, dass bei einer Verspätung eines Fluges nicht immer genau gesagt werden kann, was die genaue Ursache der Verspätung ausmacht. Vereinzelt kam es auch vor, dass nur die ersten Spalten eines Daten Tupels ausgefüllt wurden. Ich vermute, dass hier bei der automatischen Datensammlung der Daten Fehler unterlaufen sind und Werte verloren gingen.

3. Transformation der Daten:

Damit die Analyse und die daraus folgende Visualisierung mit den Schullaptops vernünftig ausgeführt werden kann, wurde das Datenset im Zuge des Cleaning-Prozesses verkürzt. Es wurden lediglich die Daten des Januars 2015 angeschaut und für Grafiken verwendet. Spalten mit vielen leeren Zeilen oder Spalten deren Inhalt weniger von Bedeutung waren wurden, ebenfalls entfernt. Für spätere Visualisierungen wurden die 3 Datensets zu einem Datenset vereint (Datensets.py), damit sowohl der ausgeschriebene Name der Fluggesellschaft, der ausgeschriebene Staat in dem sich der Flughafen befindet, das Kürzel des Staates und der Name des Flughafens enthalten sind. Das für die Visualisierung fertige Datenset ( FinalesDatenset.csv) beinhaltet nun folgende Spalten: YEAR, MONTH, DAY\_OF\_WEEK, AIRLINE, FLIGHT\_NUMBER, ORIGIN\_AIRPORT, DESTINATION\_AIRPORT, SCHEDULED\_DEPARTURE, DEPARTURE\_TIME, DEPARTURE\_DELAY, ELAPSED\_TIME, AIR\_TIME, DISTANCE, SCHEDULED\_ARRIVAL, ARRIVAL\_TIME, ARRIVAL\_DELAY, ORIGIN\_AIRPORT\_LAT, ORIGIN\_AIRPORT\_ON, DESTINATION\_AIRPORT\_LAT, DESTINATION\_AIRPORT\_LON, STATE, AIRPORT, AIRLINE\_NAMES, State (State wurde in diesem Fall klein geschrieben, das es bereits eine Spalte gibt, die jedoch nur die Kürzel der Staaten enthält und in dieser Spalte der vollständige Name enthalten ist).

Establish your editorial thinking:

Mit der Visualisierung der Daten habe ich versucht, einen möglichst breiten Einblick zu den Daten der Flüge der USA im Januar zu erhalten. Die Grafiken sind nach der Komplexität absteigend sortiert, so dass die Zielgruppe bei den ersten Visualisierungen erst einen groben Überblick erhält, was das Datenset zu bieten hat. Bei der letzten Grafik kann der User dann sogar schon interagieren, in dem er Filter über das Dropdown Menü benutzt oder sich auf der Karte bewegt. Ich habe mich dafür entschieden, alle Grafiken auf eine Seite zu plazieren, um dem User eine möglichst einfache Bedienung zu gewährleisten. So muss nicht für jede Grafik des gleichen Datensets ständig das Tap gewechselt werden und es ermöglicht zudem direkte Vergleiche mit anderen Grafiken. Durch das Untereinanderliegen der ersten zwei Grafiken lässt sich für den User direkt herausfinden, welche Fluggesellschaft und welcher Flughafen der größte ist.

1. *Anzahl der Flüge pro Tag der zehn größten Flughäfen im Januar 2015*

In dieser Grafik werden die Daten der zehn größten Fluggesellschaften betrachtet. Die Rangliste der Fluggesellschaften wurde anhand der Flüge bestimmt. Die Summe der Flüge wird pro Tag dargestellt, um eine Veränderung innerhalb des Januars zu präsentieren. Der Fokus dieser Grafik liegt auf dem direkten Blick der Größe jeder Fluggesellschaft.

1. *Flugzahlen pro Flughafen und pro Tag*

Es werden die gesamten Daten der Flughäfen betrachtet und wieder nach dem Tagesdatum des Januars sortiert. Die Anzahl der Flüge wurde hier durch das Summieren der Ankunft und Abflüge ermittelt. Der Fokus dieser Grafik liegt auf dem direkten Erkennen der Größe eines jeden Flughafens.

1. *Durchschnittliche aufgeholte Zeit pro Fluggesellschaft*

Bei dieser Grafik werden die Daten der Flugart und der durchschnittlichen aufgeholten Zeit pro Fluggesellschaft betrachtet. Mit der aufgeholten Zeit wird hier die Zeit, ab der das Flugzeug losrollt, bis zu dem Zeitpunkt, an dem das Flugzeug am Finger andockt (falls Finger vorhanden). Bei dieser Grafik liegt der Fokus auf der Pünktlichkeit jeder Fluggesellschaft.

1. *Top 15 Flugrouten*

Es werden die gesamten Daten des Januars genommen und die Top 15 Flugrouten, welche sich aus Start- und Zielflughafen zusammensetzen, ermittelt. Diese Grafik stellt dem Benutzer eine direkte und einfache Aufzeichnung der Flugrouten auf der Landkarte der USA dar. Der Fokus liegt hierbei nicht mehr nur auf einem Flughafen, sondern es ist die Kombination aus Start- und Zielflughafen.

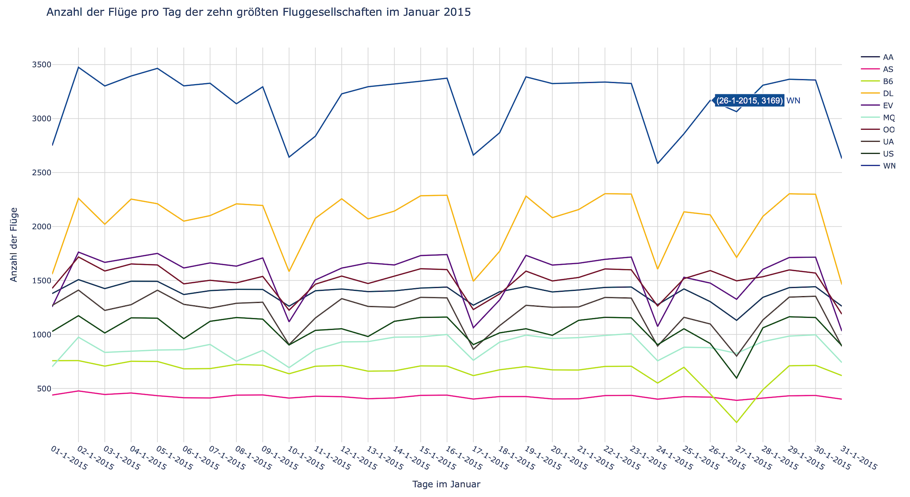
1. *Anzahl der Flüge nach Staaten und Flughäfen*

Die letzte Grafik beinhaltet ein Teil der zweiten Grafik. Hier werden die Staaten nach der Anzahl der Farben gefärbt. Aus der zweiten Grafik kennen wir bereits die Anzahl der Flüge pro Flughafen nach Tagesdatum getrennt. Diese wird hier zu einer gesamten Anzahl der Flüge des Monats Januar summiert und nach der Anzahl eingefärbt. Der Fokus dieser Grafik liegt auf der Kombination aus der Anzahl der Flüge der Staaten und den Top 20 Flughäfen.

Develop the design solution:

In den folgenden Visualisierungen wurden vier Vergleichsdiagramme (Comparison) und ein Distribution-Diagramm benutzt. Alle Diagramme wurden mit Graph Objects programmiert. Es wurde stehts darauf geachtet etwas Farbe in die Visualisierungen zu bringen, um dem Betrachtenden etwas mehr Abwechslung zu bieten. Jede Grafik enthält eine passende Legende mit farblicher Kennzeichnung. Des weiteren ist jede Visualisierung mit einer Überschrift versehen, damit direkt ersichtlich ist, welche Relationen dargestellt wurden. Wie oben bereits erwähnt, enthalten manche Visualisierungen Dropdown Menüs, entweder zur spezifischeren Auswahl von Daten oder zur farblichen Darstellung der Grafiken.

1. Anzahl der Flüge pro Tag der zehn größten Flughäfen im Januar 2015



Um eine gute Übersicht zu haben, welche Fluggesellschaft an jedem Tag wie viele Flüge fliegt, wurde hier ein Linien Diagramm verwendet. Um besser zwischen den Fluggesellschaften unterscheiden zu können, wurde jeder Fluggesellschaft eine andere Farbe zugewiesen -siehe Legende- . Es wurde zudem ein weißer Hintergrund mit einem Rastermuster gewählt, um dem Betrachtenden das Ablesen der Werte zu erleichtern. In dieser Grafik wurden bewusst nur die zehn größten Fluggesellschaften verwendet, um das Liniendiagramm nicht zu überladen. Wer sich weniger Fluggesellschaften anzeigen lassen möchte, um diese besser vergleichen zu können, kann durch Klicken auf die IATA\_CODES (rechte Seite) manuell die jeweilige Fluggesellschaft ab- oder anwählen. Für eine vereinfachte Ablesemöglichkeit ist der Hovereffekt zuständig. Wie hier im Beispiel zu sehen, ist der Mauszeiger auf den 26.1.2015 und auf die Fluggesellschaft WN gerichtet. Zudem wird die Anzahl der Flüge direkt rechts daneben angezeigt, wie im Beispiel zu sehen, 3.169 Flüge.

Aus der Grafik lässt sich gut ablesen, dass die Fluggesellschaften, bis auf kleine Abweichungen, einen sehr ähnlichen Verlauf haben. So ist bei fast allen ein Einbruch der Anzahl der Flüge am 10.1.2015/ 17.1.2015 / 24.1.2015/ 27.1.2015 und am 31.1.2015 zu identifizieren.

2. Flugzahlen pro Flughafen und pro Tag



Da nun die größte Fluggesellschaft bekannt ist, wird in dieser Visualisierung dem Betrachtenden die Größe des Flughafens nahegebracht. Bei dieser Grafik wurde ein Streudiagramm gewählt, um dem Betrachtenden direkt mit der Größe des Kreises zu verdeutlichen, welches der größte Flughafen an welchem Tag ist. Um wieder besser unterscheiden zu können, wurde pro Flughafen eine andere Farbe gewählt. Der vollständige Name des Flughafens ist der Legende (rechts) zu entnehmen. Bei der Skalierung wurde wieder auf die tagesweise Skalierung zugegriffen, da eine komplette für den Monat in Grafik 5 enthalten ist. Der Hintergrund des Streudiagramms wurde dieses mal in einem bläulichen Ton gehalten, da die Anzahl der Flüge mit dem „über den Kreis hovern“ angezeigt wird. Beim Hovereffekt sind sowohl die Anzahl der Flüge, Tag des Januars als auch der Flughafenname zu sehen. Die Skalierung der Y-Achse wurde bewusst nicht gröber skaliert, da die Anzahl der Flüge durch die Größe des Kreises dargestellt wird und die Y-Achse somit nur zur groben Einordnung dient. Die Kreise haben eine durchsichtige Farbe, um Überschneidungen gut sichtbar zu machen. Für eine bessere Sichtbarkeit der Kreise haben diese auch eine schwarze Umrandung erhalten. Hier in diesem Beispiel zeigt der Hovereffekt den 11.01.2015 mit 722 Flügen des Flughafens Dallas/Fort Worth International Airport.

3. Durchschnittlich aufgeholte Zeit pro Fluggesellschaft

Ein Bild, das Diagramm enthält.

Automatisch generierte BeschreibungDie dritte Grafik ist die erste Grafik, in der vom Betrachtenden mitbestimmt werden kann, was gezeigt werden soll. Diese Grafik stellt die aufgeholte Zeit vom Losfahren am Finger bis zum Andocken am Gate im Zielflughafen dar. Da hier zwischen 3 Arten des Fluges unterschieden wird (Kurzstrecke bis 1.000 km, Mittelstrecke 1.001 km bis 3.000 km, Langstrecke ab 3.001 km), wurde die Darstellung eines Balkendiagramms gewählt, da hier alle drei Flugarten pro Fluggesellschaft angezeigt werden können und der Betrachtende somit einen direkten Vergleicht hat. Um die Grafik für den User übersichtlicher zu halten sind alle Flugarten einheitlich in einer Farbe gekennzeichnet. Wird über einen der angezeigten Balken gehovert, so sind Informationen über den IATA\_CODE der Fluggesellschaft und die genaue Anzahl der aufgeholten Minuten erhältlich. Wer nur die einzelnen Flugarten aufgezeigt bekommen möchte, kann mit dem Dropdown-Menü die gewünschte Flugart auswählen. Hier funktioniert das „darüber hovern“ wie auf der Startgrafik.

Auffällig in dieser Grafik ist die vermehrt aufgeholte Zeit in der Luft von den Fluggesellschaften. Dagegen verspäten sich nur 3 Fluggesellschaften in der Luft. Es lässt sich gut erkennen, dass es in den USA im Januar 2015 nur 4 Fluggesellschaften gibt, die auch Langestreckenflüge anbieten.

4. Top 15 Flugrouten

Ein Bild, das Karte enthält.

Automatisch generierte Beschreibung

Durch die Visualisierung mit einer Karte der USA, auf der der Betrachtende die Top 15 Flugrouten angezeigt bekommt, wird dem Nutzer die Informationen bildlich und grafisch dargestellt. Um die Grafik übersichtlich zu gestalten, wurde die Auswahl auf die Top 15 Flugrouten beschränkt. Damit die Flugrouten besser voneinander zu unterscheiden sind, wurde jeder Flugroute eine eigene Farbe zugewiesen. In der Legende (rechts im Bild) wird die Linienfarbe angezeigt sowie der Start- und Zielflughafen. Durch das Drücken der einzelnen Flugrouten können diese ausgeblendet werden. Als Flugroute wird der Flug von X nach Y gewertet. Der Flug Y nach X wird nicht als Rückflug, sondern als separater Flug betrachtet. Der Nutzer kann sich durch schieben und zoomen auf der Karte frei bewegen. Zur vereinfachten Anzeige erhält der Betrachtende beim „darüber hovern“ die Breiten- und Längengrade des Startflughafens, den IATA\_CODE des Start- sowie des Zielflughafens.

In dieser Auswertung ist besonders auffällig, dass vermehrt Flüge zwischen der West- und Ostküste der USA stattfinden. Zwischen Norden und Süden finden nur wenige Flüge statt. Eine zusätzliche überraschende Information ist die hohe Platzierung des Fluges zwischen Kahului und Honolulu.

5. Anzahl der Flüge nach Staaten und Top 20 Flughäfen

Ein Bild, das Karte enthält.

Automatisch generierte BeschreibungFür die letzte Visualisierung wurde eine interaktive Map und ein Dropdown-Menü vereint. Um die Anzahl der Flüge pro Staat je Monat grafisch darzustellen, wurden die Staaten nach der Anzahl der Flüge eingefärbt. Hierfür wurde eine Colorbar eingefügt, die von hell (wenig Flüge) bis dunkel (viele Flüge) die Werte liefert. Da in der Grafik 2 bereits die Flüge pro Flughafen je Tag dargestellt wurde, wird hier die kompletten Fluganzahl im Januar dargestellt. Wem die Grafik farblich zu überlastet ist, hat die Wahl über das Dropdownmenü, die Staaten ohne Füllfarbe (weiß) auszuwählen. Da ich davon ausgehe, dass der Betrachtende nicht auswendig weis, welcher Punkt welchen Flughafen darstellt, wird der vollständige Name des Flughafens beim „darüber hovern“ angezeigt. Beim „hovern“ über die Staaten werden zusätzlich die genaue Anzahl der Flüge angezeigt. Um nicht zu viele Punkte auf der Karte zu platzieren, wurden hier die 20 größten Flughafen ausgewählt.

Eine interessante Erkenntnis dieser Grafik ist, dass der Flughafen(Hartsfield-Jackson Atalanta International Airport) mit den meisten Flügen zwar in einem Vielflieger Staat (Georgia)liegt, jedoch nicht in dem Staat mit der höchsten Anzahl an Flügen(Texas).