**Data Visualization – Documentation – Airlines and their delays**

**I. Formulating your brief**

Das Visualisierungsprojekt ist eine Einzelleistung und stellt eine Prüfungsleistung bzw. einen Leistungsnachweis im Fach Data Visualization dar. Das Projekt lässt sich als zu bewertendes Freizeitprojekt charakterisieren. Es zeigt, wie Studierende einen Datenbestand analysieren, um daraus wertvolle Informationen zu gewinnen und diese zu veranschaulichen. Idealerweise können aus den Ergebnissen Handlungsempfehlungen abgeleitet werden. Die Studierenden arbeiten mit praxisnahen und konkreten Daten. Ziel ist es, dass theoretische Wissen im Bereich der Datenanalyse zu vertiefen. Der Dozent stellt das Datenset zur Verfügung, Der Betrachter erhält anhand von fünf verschiedenen Visualisierungen einen Überblick. Die Visualisierungen werden auf einem interaktiven Dashboard in Python dargestellt. Vorteil von Python ist die Nachvollziehbarkeit des Frameworks. Die Visualisierungen decken sowohl „Explanatory“, also erklärende, als auch untersuchende, sogenannte „Exploratory“ Darstellungen ab. Der Nutzer erhält bekannte wichtige Informationen und gewinnt weitere Erkenntnisse. Darüber hinaus kann er interaktiv mit den Darstellungen agieren. Die genaueren Erläuterungen der einzelnen Visualisierungen im Detail folgen im Abschnitt **IV. Develop Design Solution.**

**II. Working with data**

Dieses Kapitel befasst sich mit den grundlegenden Daten aus dem Datenset und mit

der Vorgehensweise, wie die Daten für die Visualisierungen aufbereitet werden.

Das Datenset bietet Einblicke in die Verspätungen und Flugzeiten aller Flüge eines Tages des Jahres 2015 in den USA inklusive der Daten zu den Abflug- und Zielflughafen, Airlines und Flugnummern. Im Folgenden werden die Werte des Datensets genauer erläutert und kategorisiert. Das Datenset wurde vorab bereinigt zur Verfügung gestellt. Damit kann die Visualisierung zielführend und effizient erstellt werden. Schwerpunkt ist die Analysetiefe um eigene Ideen bereichert. Die Daten umfassen quantitative und qualitative Werte. Zu den qualitativen Werten gehören nominale Werte, wie Bezeichnungen für Start- und Zielflughäfen oder die verschiedenen Fluggesellschaften.

Diese Bezeichnungen besitzen keine natürliche Reihenfolge und lassen sich nicht berechnen. Beispielsweise lässt sich kein Median aus diesen Werten bilden, da sie „Strings“ darstellen. Daneben stellen die Flugnummer (eine „unique number“), die Abflug- und Ankunftsverspätungen (in Minuten), die voraussichtliche und tatsächliche Flugdauer (in Minuten), sowie die Flugstrecke (in Flugmeilen) sogenannte qualitative Werte dar. Da das Datenset schon von fehlenden Werten vom Auftraggeber bereinigt wurde, erübrigen sich Angaben zum Daten Cleaning. Für die Visualisierungen wurden die Abkürzungen der Fluggesellschaften, durch ausgeschriebene vollständige Namen ersetzt, um dem Nutzer die Ergebnisse möglichst schnell greifbar und eindeutig darstellen zu können.

Außerdem war es für bestimmte Grafiken notwendig, ausschließlich mit der „positiven“ Verspätung zu arbeiten. Negative Werte für Verspätungen sind so zu interpretieren, dass das Flugzeug früher losflog oder früher ankam. Somit wurden für einen Teil der Grafiken die Verspätungen auf null Minuten gesetzt. Das Datenset wurde durch berechnete und gezählte Spalten erweitert. Ebenfalls wichtig zu wissen ist die Summe der Flüge je Flughafen am 01.01.2015. Daher wurden die Werte gezählt und in einem neuen Dataframe abgelegt. Des Weiteren gibt das Datenset keine Auskunft über das Delta der planmäßigen und der tatsächlich geflogenen Zeit, weshalb die Differenz in Minuten ebenfalls dem Dataframe hinzugefügt wurde.

Die Datei „short\_look\_at\_the\_data.py” soll einen groben Überblick über das Datenset geben. Sie liefert einen Einblick über die Größe des Datensets, Maximal- und Minimalwerte, Anzahl von verschiedenen Werten, Korrelationen, Durchschnitte und Mediane und unterstützt den Entwickler zum einen bei der Überprüfung der Daten und präsentiert zum anderen Ideen für erste Zusammenhänge und Visualisierungen. Diese Datei soll lediglich einen Datenüberblick geben und soll dem Nutzer der Visualisierungen nicht ausgehändigt und gezeigt werden.

**III. Establish our editorial thinking**

Der Nutzer soll durch die Visualisierungen Einblick für verschiedene Beziehung und Kausalitäten zu den Verspätungen bekommen. Die Verspätungen werden sowohl mit den Airlines als auch mit den Flughäfen und Distanzen verglichen, um Gründe für die Verspätungen zu finden. Der Betrachter erhält eine einmalige Tagesübersicht aller Flüge vom 01.01.2015 und bekommt die wichtigsten Informationen auf einem Dashboard zusammengefasst mit Fokus auf den Verspätungen.

**IV. Develop Design Solution**

Da es für den Betrachter wicb htig ist, möglichst viel auf einen Blick zu sehen, stellt die Datei „dashboard\_flights“ eine Gesamtübersicht mit allen Visualisierungen dar. Daher bietet ein interaktives Dashboard eine gute Alternative die Grafiken beliebig zu ordnen und zu gestalten. Um die Dash-App auszuführen, genügt es auf den Link zu klicken, der im Terminal in Visual Studio Code ausgegeben wird. Das Dashboard ist in schwarz gehalten, damit die Farben, Linien, Kreise oder Balken auf den Visualisierungen besser hervorkommen und ausdrucksstarker wirken. Auf die Visualisierungen wird in diesem Kapitel genauer eingegangen.

**IV.1 Indikatoren**

Das Dashboard zeigt zuallererst drei wichtige Kennzahlen, um auf das Thema visuell hinzuführen und das Kernthema vorzustellen. Die Anzeige aller Flüge, die am 01.01.2015 in den USA stattfanden ist zwar eine signifikante, unübersichtliche Zahl, gibt dem Betrachter ein besseres Verständnis über die Größe der Daten. Da sich das Dashboard viel mit Verspätungen beschäftigt, zeigt der zweite Indikator mit Hilfe eines Tachometers die durchschnittlichen Verspätungen aller Flüge an. Die Farbe Orange soll dabei eine Warnfarbe sein und die Verspätung symbolisieren. Diese Kennzahl ist mittig platziert, um dem Nutzer ein Gespür für die kommenden Visualisierungen zu geben. Um dem Betrachter nicht nur Negatives über die Flüge zu vermitteln, sorgt der dritte Indikator für eine positive Nachricht. Im Durchschnitt sind die Flüge schneller am Zielflughafen angekommen als geplant. Daher ist die Kennzahl auch in grün und als reine signifikante Zahl dargestellt.

**IV.2 Map: Durchschnittliche Verspätungen in Minuten nach Region**

Da das Datenset die Koordinaten der einzelnen Flughäfen enthält, ist es wichtig dem Nutzer als Erstes nahezubringen, wo überhaupt die Flughäfen liegen und wie viele Verspätungen diese im Schnitt haben. Mit Mapbox bietet es sich an, eine Karte zu visualisieren und mit Werten (Verspätungen) zu füllen. Die Karte bietet mehr Vorteile als ein Bar Chart, in dem die Flughäfen und ihre Verspätung aufgelistet sind, da der Nutzer eine Vorstellung des Standorts der Flughafen bekommt und nicht durch eigene Recherche die Lage ausfindig machen muss. Wie auch bei den Indikatoren wird hier Orange als Signalfarbe für die Verspätungen an den Flughäfen verwendet. Da der Nutzer auch häufig eine differenzierte Sicht auf die Verspätungen bekommen möchte, zeigen die dunkleren Kreise größere Verspätungen als die helleren Kreise. Darüber hinaus wird dem Betrachter verdeutlicht, wo besonders viele Flughafen sind und kann sich durch Zoomen frei auf der Karte bewegen.

**IV.3 Airline Verspätungen**

Neben den Verspätungen der einzelnen Flughäfen, welche durch die Karte visualisiert wurden, zeigt das Histogramm dem Betrachter die Verspätungen der jeweiligen Fluggesellschaften. Die Verspätungen wurden in 20 Minuten Zeiträume unterteilt. Somit erhält der Nutzer eine Differenzierung zwischen den einzelnen Verspätungen und kann so besser für sich erkennen, welche Fluggesellschaften auch im 40-60 min Verspätungsbereich liegen. Durch die Legende hat der Betrachter die Möglichkeit, spezifische Airlines auszuwählen, um diese gegenüberzustellen und zu vergleichen. Durch das Histogramm erkennt man die Verspätungsverteilung besser und kann beliebige Zeitspannen (im Code) festlegen. Um die Fluggesellschaften visuell voneinander abzutrennen, verwendet das Histogramm unterschiedliche Farben für die einzelnen Airlines. Diese Grafik dient auch beispielsweise Nutzern, die für ihren nächsten Urlaub zwei Airlines miteinander vergleichen wollen, diese mit Hilfe der Legende zu selektieren und Rückschlüsse ziehen können, welche Fluggesellschaften pünktlicher und somit vertraulicher sind.

**IV.4 Heatmap Distanzen**

Um den Blickwinkel von Airlines und Airports ein wenig wegzulenken, zeigt die Heatmap ob größere Flugdistanzen (Langstreckenflüge ab 2000 Flugmeilen) mit größerer Verspätung in Verbindung stehen. Gründe dafür könnten zum Beispiel Zwischenstopps, längere Tankfüllungen oder Wartungen des Flugzeugs sein. Die Heatmap zeigt in umgedrehter Farbskala die Verspätungen der einzelnen Airlines in Verbindung mit den Distanzen an, da sich das weiß auf dem schwarzen Hintergrund besser abhebt und der Nutzer die Visualisierung leichter lesen kann. Die Farbe für die Verspätungen wurde, wie bei der Map und dem Indikator „Orange“ gewählt, um im Farbschema zu bleiben und dem Nutzer unterbewusst zu zeigen, dass die Visualisierung Verspätungen anzeigt. Die Leitfrage, ob größere Distanzen zu mehr Verspätungen führen, wird durch die Grafik widerlegt. Außerdem zeigt die Grafik gut, dass viel mehr Kurzstrecke als Langstrecke geflogen wurde. Eine Heatmap ist hier sehr passend, da einerseits durch die Farbskala die Länge der Verspätung gezeigt wird und andererseits die Häufigkeit der Flüge pro Airline mit ihrer Distanz dargestellt.

**IV.5 Korreliert hoher Flugbetrieb mit großen Verspätungen?**

Die letzte Grafik soll dem Nutzer ein Verständnis zwischen der Anzahl der Flüge und den Verspätungen vermitteln. Hierbei erfährt der Betrachter die Anzahl der Flüge der jeweiligen Flughäfen und darüber hinaus, ob die hohe Fluganzahl Grund für größere Verspätungen ist. Die Visualisierung setzt sich aus zwei einzelnen Charts zusammen. Sie besteht aus einem Bar Chart, welches für jeden Flughäfen die jeweilige Anzahl der Flüge am 01.0 1.2015 zeigt und aus einem Liniendiagramm, das in Abhängigkeit die Verspätungen darstellt. Auch hier wurde für die Verspätungen die Farbe Orange gewählt, damit der Nutzer die Symbolfarbe für diese Kategorie im Gedächtnis behält und die Grafik sich leichter erschließen lässt. Man kann deutlich erkennen, dass die Anzahl der Flüge pro Flughafen nicht ausschlaggebend für größere Verspätungen ist. Flughafen mit kleinerer Anzahl weißen deutlich größere Verspätungen auf und schneiden im Durchschnitt schlechter ab als die großen Flughäfen. Der hohe Durchschnitt bei kleinerer Anzahl lässt sich aber damit begründen, dass bei weniger Flügen, einzelne Anomalien viel deutlicher ausschlagen und nicht durch eine große Anzahl an kleineren Verspätungen „herunterkorrigiert“ werden.