Data Visualization

Vorgehen bei Daten Visualisierungen

Dozent: Herr Florian Eichin

Eingereicht am: 07.05.21 | Abgabetermin: 16.05.21

Lea Marie Mayer

Kurs: WWI2020F | Matrikelnummer: 6813665

Email: wi20185@lehre.dhbw-stuttgart.de

Inhaltsverzeichnis

Vorbereitung

Heutzutage geht die Informations-Technologie über das einfache Speichern und Anzeigen von Daten hinaus. Die Teildisziplin Data Science, Daten zu akquirieren, auszuwerten und daraus Wissen zu generieren, gewinnt immer mehr an Bedeutung. Vor allem um im wirtschaftlichen Umfeld die Unternehmenssteuerung zu optimieren, die Entscheidungsfindung zu unterstützen oder im Kontext von Marketing sogar markante Sachlagen ausfindig zu machen und zu präsentieren. Dabei ist es nicht nur wichtig, zu welchen Aussagen die Daten führen, sondern auch, wie diese visualisiert werden. Entscheidend ist dabei, die akquirierten Daten für verschiedene Interessensgruppen individuell adäquat aufzubereiten.
Im Folgenden geht es um drei Visualisierungsbeispiele aus der Logistikbranche. Hierzu werden Flugdaten auf verschiedene Weisen ausgewertet und aufbereitet.

Herangehensweise

Der erste Schritt ist es, die Rahmenbedingungen für alle folgenden Visualisierungen festzulegen.
Die visuellen Auswertungen der Flugdaten können verschiedene Interessensgruppen adressieren, beispielweise die Interessenten oder Kunden der Airline, die sich für die angebotenen Flüge und insbesondere deren Verspätung und Zuverlässigkeit interessieren. Auf der anderen Seite sind auch den Airlines oder Flughäfen selbst ein Vergleich mit ihrer Konkurrenz wichtig. Zudem kann die Airline über eine Darstellung ihrer Flugdaten auf eigene Schwächen und Verbesserungen schließen. Im Folgenden werden die genannten Interessensvertreter adressiert.
Bei den Darstellungen muss auf gewissere Grundcharakteristiken geachtet werden, wie sie beispielweise bei Dieter Rams Design Regeln aufgeführt sind (Rams, 2021). Davon sind für die folgenden Visualisierungen besonders die Verständlichkeit, die Ästhetik und die Innovation der Darstellung, sowie die Ehrlichkeit der Daten wichtig.
Eine weitere Bedingung für die Visualisierungen ist das Medium, über das sie präsentiert werden. Bei den folgenden Visualisierungen handelt es sich um Web-Plots, die man zum Beispiel auf einer Website finden könnte. Es besteht also nicht die Möglichkeit dem Publikum zu der Darstellung eine mündliche Erklärung zu liefern oder direkte Rückfragen zu ermöglichen. Daher muss insbesondere darauf geachtet werden, alle benötigten Informationen mitzuliefern, damit der Betrachter eigene Schlüsse ziehen kann.
Ein Vorteil der digitalen Darstellung ist die Möglichkeit, Animationen oder Interaktionsmöglichkeiten einzubauen, mit denen der Betrachter selbst durch die Darstellung navigieren und somit seinen eigenen Fokus setzen kann. Damit bekommt der Betrachter ein sogenanntes Exploratory-Erlebnis, bei dem er mit der Visualisierung interagieren kann.
Da es sich bei den Daten ausschließlich um Flugdaten aus den USA handelt, sind die Plots in internationaler Sprache beschriftet, da die meisten der Adressierten entweder aus den USA stammen oder an Internationalen Reisen interessiert sind.

Datenauswahl & visuelle Ansätze

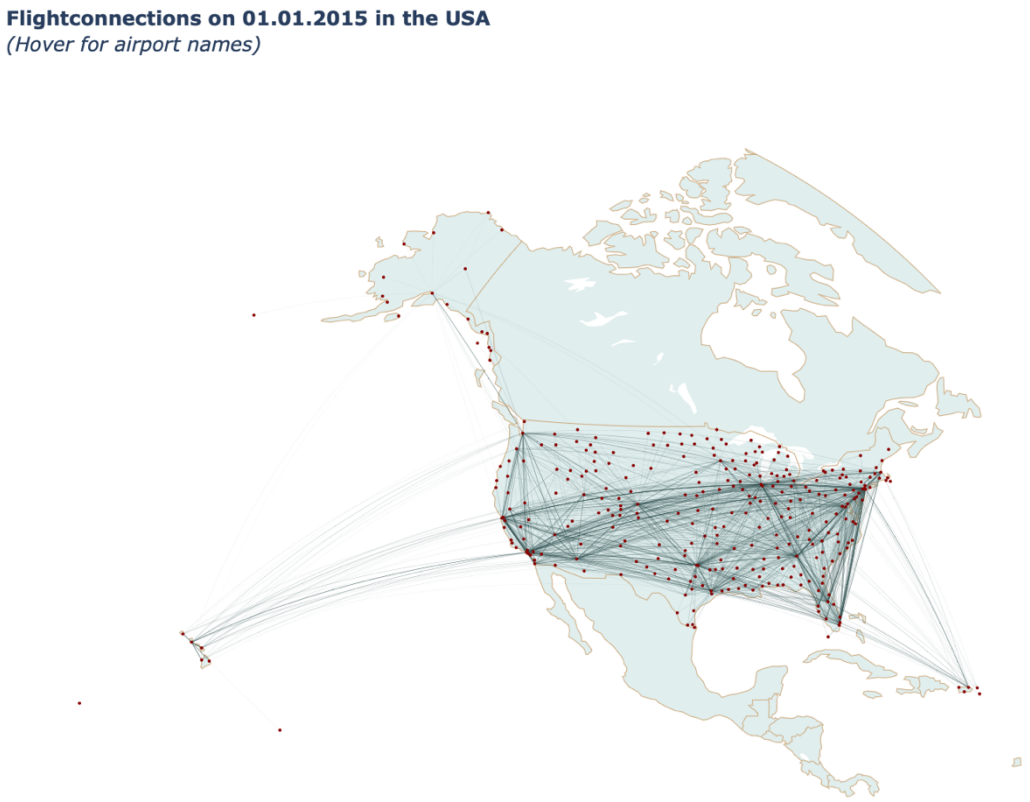
Ein Großteil der Daten ist bereits durch die Aufgabenstellung vorgegeben. Bei einigen Visualisierungen können aber noch externe Daten mit einbezogen werden, um die Dimensionen der Darstellungen zu erweitern. Im Folgenden wird für jede Visualisierung beschrieben, um welchen Ausschnitt der Daten es sich handelt und warum dieser ausgewählt wurde.
Um einen ersten Eindruck der Daten zu bekommen, wurden zu Beginn mit den Python Libraries Pandas und Numpy die Durchschnitte, Minima, Maxima, sowie der Meridian einiger Werte gewonnen (siehe firstview_inspect .py). Hieraus entstehen dann drei Aspekte, die im Folgenden näher betrachtet werden.
Im Folgenden ist die Dokumentation über die Umsetzung in Kommentaren in den angegebenen Python-Datei angefügt.

Visualisierungen

Flight Connections (connections.py)

Thema der Darstellung & Auswahl der Daten

Die erste Visualisierung präsentiert die Flugverbindungen zwischen den verschiedenen Flughäfen. Die Visualisierung ermöglicht einen ersten Eindruck in die Vernetzung der amerikanischen Luftfahrtinfrastruktur und soll unter anderem ermöglichen, Aussagen darüber treffen zu können, welche Verbindungen häufiger genutzt werden, an welchen Flughäfen viel Verkehr stattfindet und ob ein Zusammenhang zur geographischen Lage besteht.
Jede Flugverbindung wird als Linie dargestellt. Um die Darstellung übersichtlich zu halten, sind nur die Daten eines einzelnen Tages einbezogen.



Adressat

Adressaten bei dieser Darstellung sind Airline Betreiber und Kunden. Die Betreiber interessieren sich dafür, welche Verbindungen häufig von anderen Airlines angeboten werden. Die Kunden wollen wissen, von welchen Flughäfen aus viele andere Flughäfen angeflogen werden.

Umsetzung & Design

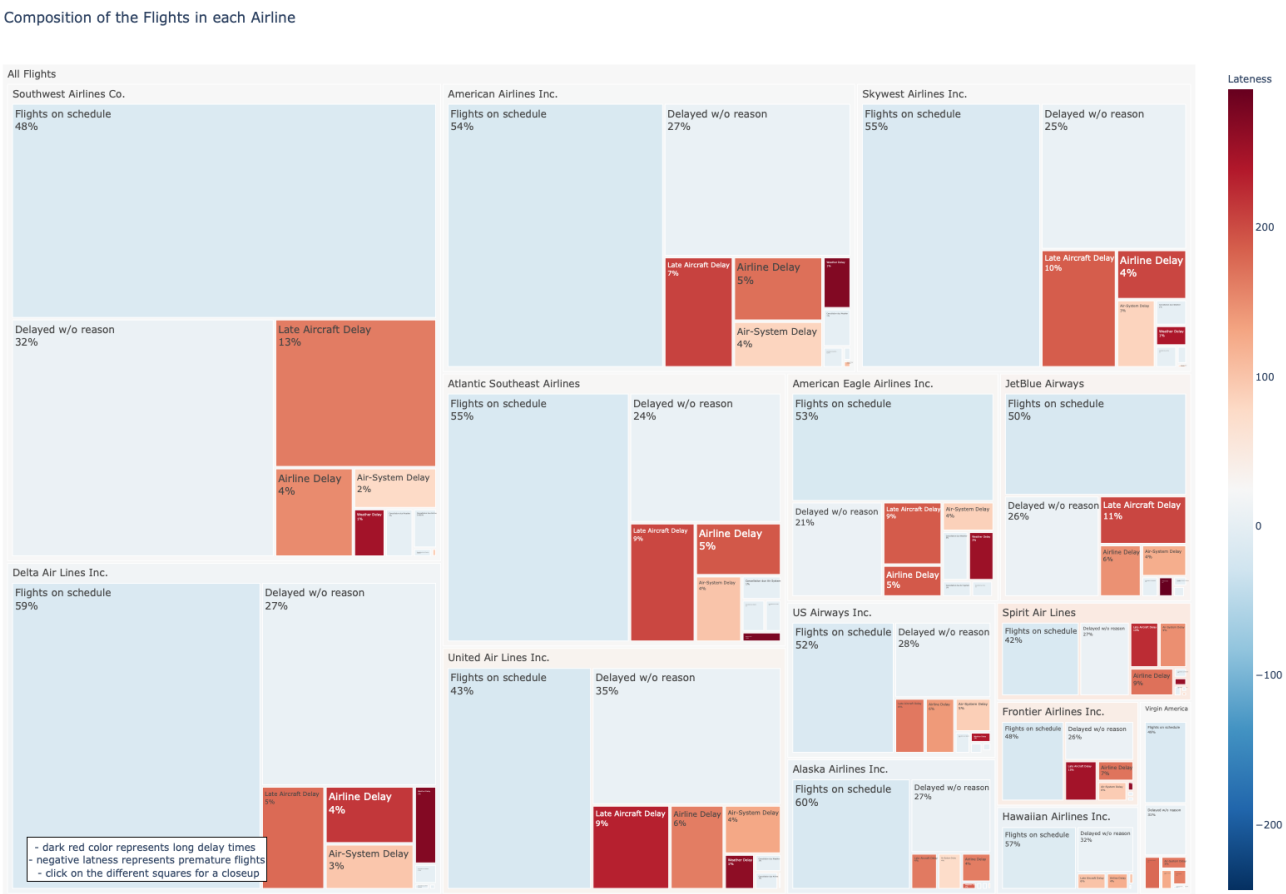
Die Dimensionen, die dargestellt werden sollen, begrenzen sich auf die geografische Lage der Flughäfen und die Verbindungen dieser Punkte. Daher eignet sich am besten die Visualisierung auf einer Landkarte. Mit dem Scattergeo-Objekt aus der Plotly Libraries wurde dazu über die Koordinaten der Flughäfen Standorte Marker-Punkte auf der Karte erzeugt. Für die Verbindungen der entstandenen Punkte wurden Linien gezogen. (siehe connections.py)
Um die vielen entstandenen Linien als Betrachter unterscheiden zu können, wurde noch die Deckkraft der Linien je nach Häufigkeit der Verbindung angepasst. Die häufig geflogenen Strecken sind nun viel deutlicher dargestellt als die selten geflogenen.

Damit sich die einzelnen Punkte der Flughäfen gut vom Festland-Hintergrund abheben, der angelehnt an die Weltkugel als dezent grün dargestellt wird, sind die Standorte in dunkelrot verzeichnet. So erinnert die Darstellung an eine Weltkugel und ist mit dieser als Vergleich leicht zu verstehen. Der Fokus der Darstellung liegt auf dem Kontinent Nordamerika, da alle Flugdaten aus den USA bezogen werden.
Um die Karte nun für den Betrachter verständlich und interaktiv zu machen, ist der Hovertext so eingestellt, dass die einzelnen vollen Namen der Flughäfen bei Berührung mit dem Cursor angezeigt werden. Um den Betrachter auf diese Funktion hinzuweisen ist der entsprechende Hinweis in der Überschrift angehängt.

Composition (compositon.py)

Thema der Darstellung & Auswahl der Daten

Diese Visualisierung bietet einen Überblick über die Zusammensetzung von verspäteten, ausgefallenen und pünktlichen Flügen nach Airline. Zusätzlich ist noch der Grund der Verspätung angegeben. Um die Daten besser vergleichen zu können, sind in einer Ansicht alle Airlines mit ihrem Verspätungstyp sichtbar.
In dieser Visualisierung geht es in erster Linie nicht darum, die Airlines direkt zu vergleichen, sondern eher darum, einen groben Überblick zu bekommen über die Zusammensetzung, der Verspätungs- und Ausfallgründe der Airline zu bekommen und bei Interaktion mit der Darstellung genauere Informationen zu erlangen
Die Aussagen der Visualisierung sollen eine möglichste hohe Genauigkeit haben. Ausreißende Extremwerte oder von Jahreszeiten abhängige Schwankungen sollten daher umgangen werden. Deshalb wurde das große Datensetz verwendet das die Flugdaten über ein ganzes Jahr bereitstellt.



Adressat

Hierbei sind Kunden sowie Betreiber der Airlines die Adressaten. Durch die vielen Informationen in einer Darstellung können unterschiedlichste Aussagen, je nach Interesse des Betrachters, betrachtet/näher analysiert werden.

Daten zusammenfassen

Es sollen verschiedene Dimensionen des Datensatzes und deren Zusammenwirkung auf einmal dargestellt werden. Daher ist die Plotly Express Treemap die ideale Wahl. Um die Visualisierung zu erstellen, müssen als erstes alle Informationen, die dargestellt werden sollen, in einer Spalte kombiniert werden, damit diese in der Darstellung angezeigt werden können. Ansonsten wird die verschachtelte Darstellung nicht aussagekräftig, es sollen vergleichbare Vierecke dargestellt werden und keine endlose Verschachtelung. Durch die Kombination wird außerdem die durch die große Datenmenge lange Laufzeit verkürzt (siehe composition_treemap.py).
Nun kann mit der Treemap der Anteil, den die Kategoriebei allen Flügen einer Airline visualisiert werden.

Design

Die Textinformationen sind so erstellt, dass in jedem Viereck der prozentuale Fluganteil, den die Kategorie pro Airline hat, angezeigt wird, damit dem Betrachter der Vergleich leichtert fällt.
Doch ist damit nur die Anzahl der Flüge und nicht die verursachte Verspätungszeit der Flüge miteinbezogen. Die Möglichkeit, die entstandenen Kasten einzufärben, wurde genutzt, um noch eine weitere Dimension darzustellen, die durchschnittliche Dauer der Verspätung. Dazu wird im DataFrame eine neue Spalte erzeugt, welche die Verspätung enthält. Auf diese wird eine Farbskale angewendet. Rot ist eine Signalfarbe, die bei vielen Anzeigen von Maschinen oder anderen Skalen den negativ verstandenen Werten zugeordnet wird. Dieses Verständnis wird hier auch angewendet. Die Farbskala ist also so gewählt, dass lange Verspätungszeiten dunkelrot angezeigt werden und so gleich auf sich aufmerksam machen und einen negativen Eindruck auslösen.
Zum besseren Verständnis der recht komplizierten Darstellung sind die Tooltips mit einem eigenen Hovertemplate überarbeitet. Damit werden noch weitere Informationen dargestellt, um dem Betrachter beim Verständnis der Darstellung zu helfen.
Zusätzlich werden noch Anmerkungen eingefügt, die dem Betrachter beim ersten Kontakt mit der Darstellung das Verständnis erleichtern sollen. Damit diese hervorstechen, zeichnet sich der Text durch weißen Hintergrund und schwarzen Rand aus. Außerdem werden alle Namen der Airlines ausgeschrieben, um eine bessere Aussagekraft zu haben, als bei Abkürzungen.

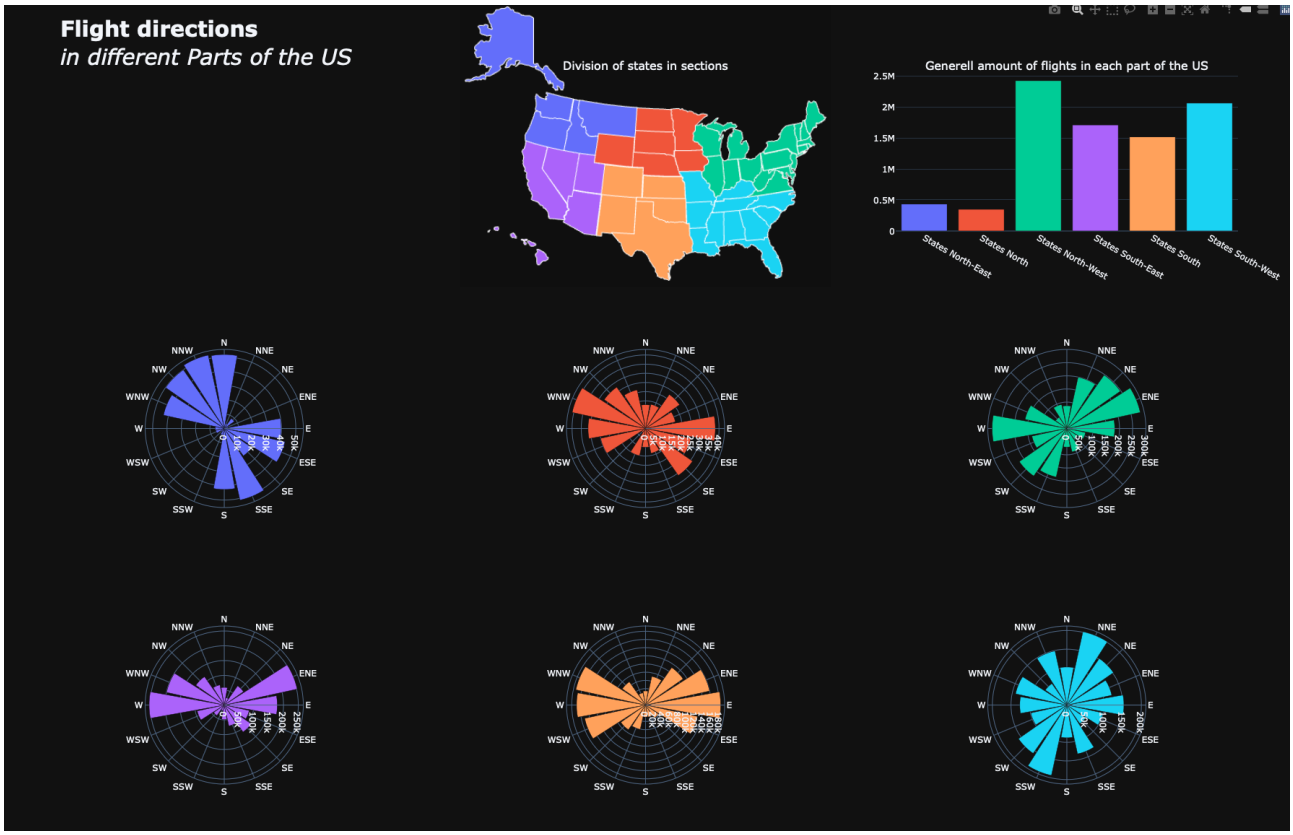
Fazit

Die Visualisierung enthält eine Menge an Informationen und ist daher nicht für schnelle Aussagen geeignet, sondern gehört in die Kategorie des Exploratory-Erlebnis. Der Betrachter kann sich Informationen anzeigen lassen, die ihn interessieren und zum Beispiel mit dem Klicken auf die Vierecke seinen eigenen Fokus setzen. Auch kann durch die Einfärbung optisch nach Airlines gesucht werden, die wenig schwerwiegende Verspätungen haben. Auch kann beispielsweise leicht die Airline mit der größten Pünktlichkeit durch die Annotation herausgelesen werden. Die Darstellung bietet einen guten Überblick und gleichzeitig Möglichkeiten, verschiedenste Informationen zu bekommen.

Flightdirections (windrose.py)

Thema der Darstellung & Auswahl der Daten

Die Visualisierung zeigt auf, in welche Himmelsrichtung die Flüge vermehrt fliegen und wie sich die Flugrichtungen zwischen verschiedenen US-Staaten unterscheiden.
Daher wurde das vollumfassende Daten-Set ausgewählt, in dem alle Flüge eines Jahres verzeichnet sind. Damit werden jahreszeitliche Schwankungen oder einzelne Extremwerte in den Daten umgangen, welche die Visualisierung, die allgemeine Aussagen treffen soll, verfälschen könnten. Um die Laufzeit des Programms trotzdem relativ gering zu halten, werden später verschiedene Maßnahmen getroffen.



Adressat

Adressaten der Visualisierung, sind hierbei Flughafen- oder Airlinebetreiber, die aus den Informationen Wissen generieren können, in welche Himmelsrichtung am meisten Flugverkehr stattfindet. Außerdem kann die gesamte Reise und Tourismusbranche losgelöst von den Flügen herausfinden, wohin die Kunden von einem bestimmten Ort aus reisen wollen und dementsprechend Angebote oder Werbung machen.

Daten Erweiterung

Zur Darstellung eignet sich hier insbesondere das Polarbar/Windrosen Diagramm. Ähnlich wie beim Balkendiagramm können hier verschiedene Werte in Balken dargestellt werden, die sich im Kreis anordnen und damit einen optischen Vergleich zu einem Kompass herstellen. Dadurch wird die Darstellung einfacher zu verstehen sein.
Der Datensatz liefert für diese Analyse nicht alle benötigten Daten. Die noch benötigten Daten werden, wie im Code dokumentiert, aus den Stammdaten berechnet. In einer ausgelagerten Cleaning-Datei wird der Dataframe um die Koordinaten und Bundesstaaten der Flughäfen ergänzt (Cleaning_windrose_1.py). In einer weiteren wird die Himmelsrichtung der Flügeberechnet (Cleaning_windrose_2.py).
Die beiden Dateien sind ausgelagert und werden nicht bei jedem Durchlauf neu geschrieben, dadurch wird die Laufzeit des Programms extrem gekürzt. Jetzt sind alle benötigten Daten vorhanden und die Datenbeschaffung ist abgeschlossen.

Daten sortieren & zusammenfassen

Es besteht nun die Möglichkeit, Flüge zusammenzufassen, um damit eine Aussage über die Flugrichtung zu erzielen. An dieser Stelle kommt auch die Frage auf, ob die Flugrichtung etwas mit der Verspätung zu tun hat, die im Weather-Delay verzeichnet ist. Ob bestimmte Richtungen beispielsweise durch wiederkehrende Winde häufiger Verspätungen aufweisen als andere. Nach Überprüfung stellte sich jedoch heraus, dass sie Flugrichtung für den Weather-Delay keine Rolle spielt und es wird weiterhin das Ziel verfolgt, generell Informationen über alle Flüge zu erfassen.
Es gibt nun mehrere Möglichkeiten, die Daten einzuteilen, um diese zu vergleichen. Um die Übersichtlichkeit und Aussagekraft beizubehalten, werden die Bundesstatten in sechs geografische Gruppen eingeteilt, die jeweils einen zusammenhängenden Teil der USA zusammenzufassen. Nach der Aufteilung werden die Flüge

verschiedener Himmelsrichtungen zusammengefasst und für die Darstellung im Windrosen-Diagramm angepasst.

Subplots erstellen

Für jede Zusammenfassung von Staaten wird eine solche Datei erstellt, diese kann nun als Barpolar-Diagramm dargestellt werden. Das Ziel ist es nun die entstandenen Teildatensätze nebeneinander als Diagramme darzustellen.

Darstellung designen & erklären

Die entstandene Darstellung erweist sich nun als unverständlich und bedarf weiterer Erklärungen. Welcher Subplot nun welche Daten über die verschiedenen Staaten enthält kann nun schwer mit einer Legende dargestellt werden. Daher eignet sich eine visuelle Darstellung einer USA Karte am besten, um die Staaten den einzelnen Plots zuordnen zu können. Diese werden hierfür in den gleichen Farben dargestellt, wie die Subplots selbst.
Damit die Verbindung schneller hervorsticht, wird die Hintergrundfarbe durch das Anpassen des Layouts auf schwarz gesetzt. Nun treten die grellen Farben stark hervor und die Zusammenhänge sind deutlicher zu erkennen.
Die Subplots werden unter dem Erklärungsbild so angeordnet, dass sie der geografischen Lage entsprechen, der der Subplot zugeordnet wird. Damit sind die Daten der blau eingefärbten Staaten oben links auf der Karte ebenfalls oben links in den Plots zu finden, hierdurch soll wieder schneller der Zusammenhang für den Betrachter hergestellt werden.
Um die Aussagekraft der Darstellung zu erhöhen wird noch ein weiterer Subplot hinzugefügt. Jeder der Windrosen-Subplots hat eine eigene Skala, damit der Vergleich zwischen den unterschiedlichen Mengen von Flügen pro Flugrichtungen deutlicher sichtbar ist. Damit gerät aber der Vergleich über die gesamte Menge der Flüge in den Hintergrund, ein voller Ausschlag auf der einen Skale ist mit der des nächsten Plots nicht zu vergleichen und die unterschiedlichen Zahlen auf den Skalen fallen oft erst spät auf. Um dieses Problem zu lösen und nicht noch viel Beschreibungstext hinzuzufügen oder durch eine Angleichung der Skalen die visuelle Aussagekräftigkeit zu verlieren, wird ein weiterer Plot hinzugefügt. Ein einfaches Barchart zeigt in passender Farbe die kumulierten Flüge für jede Region an.
Dieser Chart steht zusammen mit dem Bild der Staatenaufteilung über den restlichen Plots, damit der Betrachter zuerst der visuellen Erklärung Aufmerksamkeit schenkt, bevor er sich den erklärungsbedürftigen Subplots widmet.

Fazit

Somit beinhaltet die Darstellung viele interessante Informationen, die aber nicht auf den ersten Blick ersichtlich sind. Der Betreiber einer Airline / Flughafens oder ein Arbeiter aus der Tourismusbranche kann sich mit der Darstellung intensiv beschäftigen, um so aus der Darstellung wichtige Schlüsse zu ziehen, die sein Handeln beeinflussen könnten.

Allgemeines Fazit

Allgemein lässt sich sagen, dass die Art und Weise wie Daten dargestellt werden können, einen großen Einfluss auf den späteren Betrachter hat und daher auf diesen abgestimmt werden muss. Wie zuvor in den aufgeführten Darstellungen, sollten sich Visualisierungen für Kunden auf einfache und schnell erkennbare Aussagen fokussieren. Ein Beispiel wäre hierfür, die an eine Weltkugel angelehnte Flugverbindungsdarstellung (connection.py), die durch die Assoziation einfach zu verstehen ist und der Betrachter schnell erkennen kann, um was es sich handelt. Für Business-Interessierte, wie den Betreiber einer Airline oder eines Flughafens, sollte die Darstellung viele wichtige und technisch relevante Informationen enthalten und muss sich nicht auf eine einzelne Aussage konzentrieren, sondern einen tieferen Einblick in die Daten ermöglichen. Beispiele dafür sind die Treemap-Darstellung und die Visualisierung der Flugrichtungen (composition.py und windrose.py). Doch sollten bei allen Darstellungen, ungeachtet des Adressaten, die Grundsätze des Designs, wie Verständlichkeit, Ästhetik und Innovation umgesetzt werden.

Literaturverzeichnis

Attribute, T. (15. 04 2021). Von https://plotly.com/python/reference/treemap/ abgerufen

Bearing. (25. 04 2021). Von https://stackoverflow.com/questions/54873868/python-calculate-bearing-between-two-lat-long abgerufen

Bearing, C. (10. 04 2021). Von https://gist.github.com/jeromer/2005586 abgerufen

Chart, B. (15. 04 2021). Von https://plotly.com/python/bar-charts/ abgerufen

Doku, M. (05. 04 2021). Von https://plotly.com/python-api-reference/generated/plotly.graph_objects.Scatter.html. abgerufen

Eichin, F. (2021). Data Visualization Skript.

Loc-Funktion. (06. 04 2021). Von https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html abgerufen

Map, L. (06. 04 2021). Von https://plotly.com/python/lines-on-maps/ abgerufen

Map-Funktion. (15. 04 2021). Von https://pandas.pydata.org/docs/reference/api/pandas.Series.map.html abgerufen

Markers, S. (01. 05 2021). Von https://plotly.com/python/v3/marker-style/ abgerufen

Rams, D. (06. 04 2021). *Design Dieter Rams*. Von https://www.vitsoe.com/de/ueber-vitsoe/gutes-design abgerufen

Treemap. (10. 04 2021). Von https://plotly.com/python/treemaps/. abgerufen

Windrose. (30. 03 2021). Von https://plotly.com/python/wind-rose-charts/. abgerufen