Dokumentation
Abschlussprojekt
Data Visualization
2015 Flight Analysis

Christine Arnoldt

Abstract

Das im Folgenden dokumentierte Projekt ist Ergebnis einer Abschlussaufgabe im Rahmen der Data Science – Data Visualization Vorlesung an der DHBW Stuttgart. Gegenstand der Aufgabe war es ein Daten-Set mit Flügen aus dem Jahr 2015 in drei bis fünf Grafiken zu visualisieren. Für die Bearbeitung der Aufgabe wurde an dieser Stelle jedoch statt des bereitgestellten, bereit bereinigtem Datensets das Original verwendet und selbst bereinigt. Es enthält unter anderem Informationen zu Datum, Abflugs- und Landezeiten, Verspätungen, Distanz, sowie die Koordinaten der Flughäfen.

Der Code für dieses Projekt (ohne die csv-Dateien), sowie Screenshots der Visualisierungen und der HTML-Code finden Sie hier: https://github.com/ChristineArnoldt/data-visualization
Den gesamten Code für dieses Projekt inkl. Der csv-Dateien, sowie die Screenshots der Grafiken und HTML-Files finden Sie hier: https://owncloud.dhbw-stuttgart.de/index.php/s/pYsf6G6o2iY76k5

Zielsetzung und Stakeholder

Ziel war es, Visualisierungen zu erstellen, die Flugreisenden die Möglichkeit gibt, eine Reihe an Informationen zu Airlines, Flughäfen und Flügen zu bekommen, die bei der Reiseplanung und Fragen nach der Fluganbieterwahl, dem Reisezeitpunkt und den Abflug- und Ankunftsorten unterstützen können. Fokus lag hierbei auf Langsteckenflügen mit einer Distanz von über 1400 Meilen (was ca. 2.253 km entspricht). Damit sind Flugreisende, die eine längere Flugreise innerhalb der USA anstreben auch in erster Linie die Zielgruppe, weshalb sämtliche Visualisierungen in englischer Sprache beschriftet sind.

Visualisierungsarten

Die erste Visualisierung ist ein Bar Chart, der dem Betrachter die Flughäfen mit den höchsten durchschnittlichen Verspätungen aufzeigt, wobei Nutzer interaktiv über einen Slider auswählen können, wie viele Flughäfen gezeigt werden. Ein Hovertext zeigt zusätzliche Informationen. Bei der zweiten Grafik handelt es sich um eine Heatmap. Diese zeigt die Anzahl der Flüge pro Tag, wodurch Rückschlüsse auf das Reiseaufkommen möglich sind. Zudem zeigt eine Textbox zusätzliche Informationen zu dem Tag, z.B. wie viele der Flüge an dem entsprechenden Tag verspätet waren und wie viele Meilen an Distanz an dem Tag zurückgelegt wurden. Eine weitere Visualisierung ist ein Set an Radar Charts, die die Performance nach Airline aufzeigen soll. Dabei werden die Dimensionen der prozentualen Verspätung, vergangenen Flugzeit, Anzahl der annullierten Flüge, Start- und Landeflughäfen (jeweils gemessen an der Gesamtzeit bzw. Gesamtmenge) dargestellt. Die Flughäfen sind nach Größe (gemessen an der Anzahl der Flüge) auf die vier Charts verteilt. Der Nutzer kann hier zudem in der Legende einzelne Airlines ein- oder ausblenden. Die letzten beiden Visualisierungen sind Landkarten der USA. Auf der einen Karte sind die US-Bundesstaaten farbig anhand der durchschnittlichen Ankunftsverspätung visualisiert. Auf der anderen sind die Flughäfen in Form einer Bubble-Map abgebildet, wobei die Größe der Bubbles die Größe des Flughafens (gemessen an der Anzahl der eintreffenden Flüge) widerspiegelt und die Farbe die durchschnittliche Verspätung in Minuten anzeigt.

Data-Cleaning

Bereits am Anfang des Projektes war klar, dass das bereitgestellte Daten-Set in seinem Umfang für die geplanten Visualisierungen (u.A. eine Heatmap, die das gesamte Jahr abbildet) nicht ausreichend war und damit ein eigenes Daten-Cleaning vonnöten. Die im Rahmen des Projektes bereinigten Daten unterscheiden sich von dem gegebenen Daten-Set in erster Linie von einigen zusätzlichen Spalten, die Auskunft über annullierte und umgeleitete Flüge bieten, die Gründe für Flug-Annullierungen und Verspätungen enthalten, sowie Spalten mit dem ausgeschriebenen Namen der Flughäfen und den Bundesstaaten in denen die Flughäfen liegen. Bei dem Cleaning wurden mehrere Dateien erstellt, die als Grundlage für die verschiedenen Visualisierungen dienen. Dabei wurde insbesondere zwischen annullierten und nicht annullierten Flügen sowie dem Bereinigungsgrad unterschieden.

Bar-Chart

Datengrundlage und Umsetzung

Grundlage für den Bar-Chart sind nicht-annullierten Flüge, für die auch Informationen zu den Flughäfen bereit standen, wobei nur die Flughäfen, die Anzahl der Flüge sowie die Verspätung an den Startflughäfen betrachtet wurde. Bei den Flughafenkürzeln

bzw. -namen handelt es sich um qualitative Daten. Die Verspätung und Anzahl sind quantitative Daten. Zudem wurde eine Datei mit allen Flughäfen zur IATA-Code Namensauflösung genutzt. Für die Darstellung wurde der Durchschnitt anstelle des Medians als vergleichbarer Wert aus zwei Gründen gewählt: 1. Um das Diagramm für alle Nutzer so leicht verständlichen wie möglich zu gestalten – viele Menschen können mit einem Durchschnitt mehr anfangen als mit einem Median und 2. weil es Flughäfen mit weniger als drei Flügen gibt, bei denen der Median nur schlecht den ‚typischen' Fall abbilden kann. Zudem gibt es bei der Darstellung des Medians viele Flughäfen mit einem Median von Null (s. Abb. 1), die im Graphen eine große Lücke darstellen. Um dennoch den Median-Wert nicht außen vor zu lassen, wird dieser im Hovertext der jeweiligen Balken angezeigt.

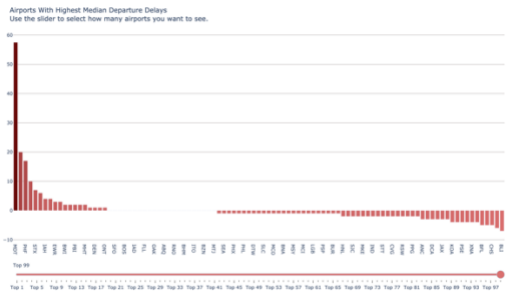

Abb. 1 Median

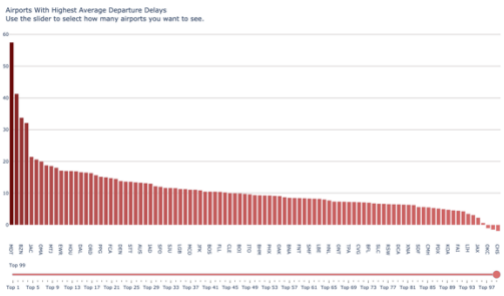

Abb. 2 Durchschnitt

Zur Erstellung des Charts wurden zunächst zwei Dataframes erstellt: eines für Abflüge und eines für Ankünfte, jeweils sortiert nach Lande- bzw. Startflughafen. Beiden wurden zwei Spalten mit den durchschnittlichen Abflugs- bzw. Ankunftsverspätungen und dem Median der Abflugs- bzw. Ankunftsverspätungen hinzugefügt. Zudem wurde das Data Frame um eine Spalte mit dem ausgeschriebenen Namen des Flughafens ergänzt. Dann wurden die beiden Data Frames gemerged. Die Flughäfen wurden für die spätere richtige Reihenfolge im Bar Chart nach der höchsten Abflugsverspätung und der Anzahl der abgeflogenen Flüge sortiert. Dann wurden in einem Loop mit plotly 99 Traces des Bar Charts geschaffen, die jeweils einen Step auf dem im Anschluss implementierten Slider darstellen. Der Slider verändert hierbei die Sichtbarkeit der verschiedenen Traces. Wählt man nun also als Slider-Step die 30, wird einem nur der entsprechende Trace angezeigt, welcher in diesem Fall die Top 30 Flughäfen nach höchster Abflugsverspätung als Bar Chart zeigen würde.

Chartauswahl und Design

Für die Darstellung der höchsten durchschnittlichen Ablufgsverspätungen wurde ein Bar-Chart gewählt, da dieser die notwendigen Informationen sehr klar, unkompliziert und vor allem gut vergleichbar vermitteln kann. Als Startpunkt für den Slider wurde die 10 gewählt, da es generell als üblich gilt die Top 3 oder Top 10 darzustellen. Da es sich allerdings um ein Balkendiagramm handelt und 3 Balken weniger Vergleichspunkte bieten und weniger aussagekräftig sind, fiel die Wahl auf die Top 10. Als Farbe wurde für die Balken keine feste Farbe gewählt, sondern ein Farbverlauf um die höchsten Verspätungen mit einem dunkleren Rot-Ton zusätzlich hervorzuheben. Für den Farbverlauf wurde lediglich der Start und Endwert festgelegt. So kalkuliert plotly selbst den Farbverlauf, was auf Grund benötigten Flexibilität durch den Slider und die hohe Anzahl an Traces ideal ist. Um die Informationen genau ablesbar zu machen und Zusatzinformationen zu liefern, wird in dem Hovertext noch einmal der genaue Wert der durchschnittlichen Verspätung, der Median der Verspätung, die Anzahl der gestarteten Flüge, sowie der ausgeschriebene Name des Flughafens angezeigt.

Heatmap

Datengrundlage und Umsetzung

Die Datengrundlage für die Heatmap bilden alle nicht-annullierten Flüge. Bei der Erstellung der Grafik wurden drei quantitative Datentypen herangezogen: das Datum des Fluges, die zurückgelegte Distanz sowie die Verspätungsminuten bei Ankunft. Zunächst wurden die Monats-, Tages- und Jahres-Spalten durch eine Datumsspalte ersetzt. Obwohl in der Spalte der Verspätungsangabe einige NaN Werte waren, wurde die Entscheidung getroffen, diese dennoch abzubilden, da der Fokus der Grafik die Anzahl der Flüge ist. Die Information zur Anzahl der verspäteten Flüge ist sekundär. Anschließend wurde eine neue Spalte erstellt, in der markiert wurde ob ein Flug als verspätet gewertet wird oder nicht – basierend auf persönlichen Erfahrungswerten wurde die Annahme getroffen, dass Fluggästen eine Verspätung von weniger als fünf Minuten egal ist, weshalb nur Flüge mit einer größeren Verspätung als ‚verspätet' gewertet wurden. Es wurde die Anzahl der Flüge pro Tag sowie die zurückgelegte Distanz und die Gesamtverspätung in Minuten pro Tag ermittelt. Anschließend wurde die Anzahl der Flüge, der Tag und der Monat auf eine plotly Heatmap übertragen, bei der die x-Achse die Tage im Monat und die y-Achse die Monate darstellt. Die Farbe zeigt die Anzahl der Flüge an.

Chartauswahl und Design

Die Heatmap wurde gewählt, da mit ihr eine gute Übersicht über das ganze Jahr möglich ist, weil die Heatmap drei Dimensionen abbilden kann. Sie lässt leicht Muster innerhalb des Jahres erkennen. Es wurde ein schlichtes Design gewählt mit der von plotly bereitgestellten Colorscale „OrRd". Die Monatsbezeichnung wurde zur besseren Lesbarkeit in Monatskürzel umbenannt. Ein Hovertext gibt zusätzliche Informationen zu annullierten Flügen und zurückgelegter Distanz. Der Grafik wurde eine Annotation für den Nutzer hinzugefügt, insbesondere um einige Punkte des Vorgehens und der Wertung der Daten kurz zu erklären, wie zum Beispiel welche Flüge gewertet wurden, ab wann der Flug als verspätet gilt und dass die Distanz annullierte Flüge ausschließt. Damit soll die Transparenz gegenüber dem Nutzer gewahrt werden.

Radar-Charts

Datengrundlage und Umsetzung

Bei den Radar-Charts ist die Datengrundlage eine csv-Datei mit allen Airlines zur IATA-Code Namensauflösung und eine csv-Datei mit allen Flügen, in welcher einige Flüge keinen Ankunfts- oder Startflughafen haben. Bei der Bereinigung wurden alle Reihen entfernt, in denen nicht-annullierte Flügen in der Spalte der vergangenen Zeit NaN-Werte hatten: bei annullierten Flügen wären dieser Werte egal (ein annullierter Flug hat auch keine Flugzeit). Es stellte sich raus, dass damit bereits alle Reihen mit nicht-annullierten Flügen, bis auf einige fehlende Flughäfen, bereinigt waren. Die NaN Werte in der Spalte der Flughäfen wurden nicht entfernt, da die Flüge trotzdem stattgefunden haben und so ggf. nur ein paar wenige Flughäfen in der Anzahl fehlen (da die Flughäfen quantifiziert werden) verglichen mit ganzen Data Frame Reihen, hätte man sie entfernt. Für die Visualisierung fiel die Dimensionsauswahl auf Anzahl der annullierten Flüge, die gesamte Verspätungszeit, die gesamte Flugzeit (quantitative Datentypen) und die Anzahl angeflogener Start- und Landeflughäfen (hier wurden qualitative Daten, also die Namen der Flughäfen, zur Vergleichbarkeit quantifiziert), da diese Werte für Flugreisende besonders wichtig und gleichzeitig durch gut vergleichbar sind. Nun wurden alle numerischen Werte sowie die Flughäfen für die annullieren Flüge auf NaN gesetzt: Flüge die nicht geflogen wurden, werden bei Distanz, vergangener Zeit und der Anzahl an Flughäfen nicht gezählt, da die Airline diese Flugzeit und -distanz nicht zurückgelegt hat und die Flughäfen nicht bedient wurden. Anschließend wurde für alle Spalten eine

Summe oder die Anzahl pro Airline bestimmt und als neue Spalten eingefügt Zusätzlich wurden die prozentualen Werte dieser Spalten, gemessen am Gesamtwert, ermittelt, um alle Radar-Dimensionen im Chart in eine ähnlichen, vergleichbare Dimension zu bringen. Nun wurden mit plotly vier Subplots erstellt, auf die jeweils (sortiert nach Größe) drei Airlines geplottet werden. Zudem wurde ein Hovertext hinzugefügt, der neben den Dimensionswerten und dem Airline-Namen die Anzahl der von der Fluglinie geplanten Flüge anzeigt.

Chartauswahl und Design

Um die Airline-Performance vergleichbar zu machen, eignet sich ein Radar-Chart am besten, im Radar-Chart viele Dimensionen dargestellt werden können. Die Airlines wurden auf vier Charts verteilt um alle Traces gut lesbar zu gestalten. Die beiden Charts, die zusammen die sechs größten Airlines (gemessen an der Anzahl der Flüge) abbilden, haben die gleiche Range und auch die beiden Charts mit den kleinsten Airlines haben die gleiche Range, um jeweils die größten und die kleinsten Airlines besser vergleichbar zu machen. Es handelt sich jedoch um einen Trade-Off: die kleinsten Airlines sind noch nicht ideal ablesbar, allerdings könnten verschiedene Chart-Ranges für jedes der Diagramme den Betrachter in die Irre führen und die Werte schlechter vergleichbar machen. Es konnte keine gemeinsame Chart-Range für alle Plots gewählt werden, da sonst nicht alle Airline-Traces gut erkennbar abgebildet gewesen wären. Da man in der Legende einzelne Traces aus- oder einblenden kann, wurde sich für die bestmögliche Vergleichbarkeit, ohne die Traces unlesbar zu machen, entschieden. Die Farben wurden einzeln gesetzt. Dabei wurden mit coolers Farben gesucht, die visuell zusammenpassen, gleichzeitig aber die Differenzierbarkeit der Traces erhöhen. Die Trace-Flächen sind transparent, damit auch die darunter liegenden Traces gut identifizierbar sind. Die Linien sind nicht durchsichtig, damit die Abgrenzungen klarer sind. Als Plot-Hintergrund wurde weiß gewählt, damit sich die Traces gut abheben. Es wurden jeweils Subplot-Überschriften gesetzt, die kurz wiedergeben, welche Airlines abgebildet sind. Die Legende wurde unter die Charts gesetzt, damit der Titel mittig ist und da der Legende selbst eine geringere Signifikanz zugeschrieben wird (die Hovertexte und Subplot-Überschriften zeigen Airline-Namen und die Legende dient hauptsächlich dem ein- und ausblenden der Traces).

Map – US-States

Datengrundlage und Umsetzung

Datengrundlage bildet eine bereinigte csv-Datei, mit allen nicht-annullierten Flügen und eine Datei mit der Abkürzungsauflösung der Bundesstaatenkürzel. Es wurden die Bundesstaaten (qualitativ) und die durchschnittliche Ankunftsverspätung (quantitativ) pro Bundesstaat visualisiert und zusätzlich die Summe an Verspätungsminuten (quantitativ) angegeben. Dabei wurde zunächst eine Spalte mit dem ausgeschriebenen Namen des Bundesstaates hinzugefügt, indem das Data Frame mit dem Data Frame der US-Staaten gemerged wurde. Dann wurden die Summe der Verspätungsminuten (Ankunftsverspätung) pro Bundesstaat und die durchschnittliche Ankunftsverspätung ermittelt und als Spalten hinzugefügt. Die Wahl fiel dabei aus Gründen der Einheitlichkeit und leichteren Verständlichkeit für alle Nutzer, wie auch bei dem Balkendiagramm, auf den Durchschnitt anstelle des Medians. Anschließend wurde die durchschnittliche Ankunftsverspätung mit plotly auf eine Karte mit US-Bundesstaaten geplottet, wobei die Farbe der Staaten die Höhe der Verspätung angibt. Ein Hovertext verrät zudem zusätzlich die summierte Verspätung aller Flüge, die in dem entsprechenden Bundesstaat landen.

Chartauswahl und Design

Diese Darstellung ermöglicht einen sehr schnellen und einfachen Überblick darüber, welche Staaten unter der größten Ankunftsverspätung leiden. Die Darstellung ist simpel und für jeden zu verstehen. Als Farbskala wurde die von plotly implementierte Skala „OrRd" analog zur Heatmap genutzt. Diese Farbgebung geht zum einen mit der Heatmap und dem Barchart

einher, zum anderen werden Verspätungen als negativ wahrgenommen, weshalb ein stärkerer Rotton intuitiv eine größere Verspätung suggeriert. Staaten, für die keine Ankunftsflüge vorliegen sind ausgegraut. Ein Hovertext zeigt die dargestellte Information noch einmal in Satzform, wobei auch der Name der Bundesstaaten mit angezeigt wird, weil anzunehmen ist, dass nicht alle User sämtliche Bundesstaaten auf einer Karte zuordnen könnten. Zudem zeigt der Hovertext die Summe der Verspätungsminuten gelandeter Flüge.

Map - Airports

Datengrundlage und Umsetzung

Datengrundlage für die Flughafen-Karte ist eine csv-Datei mit nicht-annullierten Flügen, denen ein Flughafen zugeordnet werden kann und eine weitere, unbereinigte Datei mit allen nicht-annullierten Flügen um später zusätzliche Flughäfen zu markieren. Es wurden Flughäfen (qualitativ), die Anzahl der Flüge pro Flughafen und die Ankunftsverspätung (quantitativ) abgebildet. Die Entscheidung, die Verspätung an den Landeflughäfen und nicht die Verspätung an Startflughäfen darzustellen, wurde getroffen, da die Abflugsverspätung bereits im Balken-Diagramm dargestellt ist. Zunächst werden alle Zeilen im Data Frame ohne Ankunftsflughafen entfernt, da diese für die Darstellung irrelevant sind. Anschließend wird ein Data Frame erstellt, dass nach Landeflughäfen gruppiert ist. Es wird die Anzahl der gelandeten Flüge sowie die durchschnittliche Ankunftsverspätung pro Flughafen bestimmt. Die Wahl fiel hier aus den gleichen Gründen wie bei dem Bar Chart auf den Durchschnitt anstelle des Medians. Zusätzlich wird das zweite Data Frame einmal nach Startflughafen und einmal nach Landeflughafen gruppiert, wobei nur Flughäfen rausgefiltert werden, die nicht bereits in dem anderen Data Frame gelistet sind. Dieses dient später dazu, auch Flughäfen abzubilden, in denen keine Flüge landen. Da in der Legende die Option bestehen soll, Traces ein- und auszublenden, müssen die Flughäfen gruppiert werden (da nicht jeder Flughafen einzeln aus- und eingeblendet werden soll, da die Legende sonst zu lang und unübersichtlich wäre). Die Gruppierung erfolgte nach Verspätungszeit. Dazu wurden zunächst Minimum und Maximum als Randwerte ermittelt. Anschließend wurden eine Einteilung der Schritte zwischen diesen Randwerten ermittelt (genauere Vorgehensweise s. Chartauswahl und Design). Anhand der Limits wurden nun vier Traces mit den zugehörigen Flughäfen erstellt. Dabei gibt die Marker-Größe die Anzahl der angekommenen Flüge und die Farbe den durchschnittlichen Arrival Delay an. Da es einige Flughäfen mit nur einem Flug und andere mit 78.453 Flügen gibt, wurde das Scaling angepasst (s. Chartauswahl und Design). Zudem wurden zwei Traces mit den Flughäfen ohne Ankunftsflüge hinzugefügt, bei denen die Punkte ausgegraut sind.

Chartauswahl und Design

Die Karte kann als Ergänzung zu der anderen Karte gesehen werden. Sie schlüsselt nochmal spezifisch nach Flughafen die Verspätungszeit auf und zeigt zudem die Flughafengröße im Verhältnis zur Verspätungszeit an. Damit kann der Betrachter selbst Rückschlüsse auf 1. den Zusammenhang zwischen Flughafen-Größe und Ankunftsverspätung und 2. indirekt auch auf die Abflugsverspätung ziehen, da es für Flughäfen unmöglich ist, größere Ankunftsverspätungen beim Gepäck umladen u.ä. auszugleichen (wenn ein Flugzeug verspätet ankommt und danach gleich weiterfliegen soll, wird es auch verspätet starten müssen). Die sinnvolle Gruppierung der Flughäfen in verschiedene Traces erwies sich als schwierig, da sehr viele Flughäfen eine Verspätung zwischen -5 und 15 Minuten haben, aber nur wenige Outlier eine längere. Nach einigen Überlegungen wurden drei Optionen ausprobiert: 1. eine lineare Einteilung, die sich an einer geometrischen Reihe ausgehend von -20 bzw. 80 orientiert, bei der jeweils durch zwei dividiert wird, wobei jedoch die Schritte Nahe 0 zusammengefasst wurden – damit wird eine gleiche Einteilung in beide Richtungen geschaffen, bei der die Schritte größer werden, um Outlier besser zusammenzufassen, 2. eine Einteilung nach der Fibonacci-Reihe, ausgehend von 0 (jeweils durch Addition bzw. Subtraktion der nächsten Zahl in der Fibonacci-Reihe), wodurch die Schritte mit Zunehmen des absoluten Wertes der Verspätung auch in Größe zunehmen (Werte um 0 genauer eingeteilt, äußere Werte mit weniger Flughäfen zusammengefasst) und 3. eine Einteilung durch visuelle

Abschätzung der Sigma-Werte der Standardabweichung mit Hilfe eines Histogramms. Dabei erwies sich letztere als die beste Option (nicht zu viele Schritte, aber trotzdem eine relativ gute Verteilung über die Schritte hinweg). Zudem musste das Scaling der Bubbles angepasst werden, da es einige Flughäfen mit nur einem Flug und andere mit 78.453 Flügen gibt, damit die kleinen Flughäfen nicht verschwinden oder die großen alles überlagern. Nach dem Versuch logarithmische Skalierungen der Flugzahl vorzunehmen, wurde zur Normalisierung der Werte letztendlich eine einfache Min-Max-Skalierung gewählt. Da die kleinen Flughäfen trotz Skalierung immer noch zu klein waren, wurde eine Mindestgröße festgelegt. Die Farbwahl fiel auf eine Reihe an Farben, die einen Farbverlauf von grün zu rot in vier Schritten abbilden. Die Farben wurden mit dem Colordesigner Gradient Generator erzeugt. Die überpünktlichsten Flughäfen sind knallgrün, die Flughäfen, bei denen Flüge nur etwas zu früh, pünktlich oder nur leicht verspätet waren, sind leicht-gelblichen grün. Flughäfen mit längeren Verspätungen sind orange und bei enorm hohen Verspätungen rot. Diese Farbwahl wurde getroffen, da sie sehr intuitiv für den Betrachter ist. Ein Hovertext zeigt die Flughafennamen, die Anzahl gelandeter Flüge und die durchschnittliche Ankunftsverspätung an.