XB_0018.Introduction to Data Science

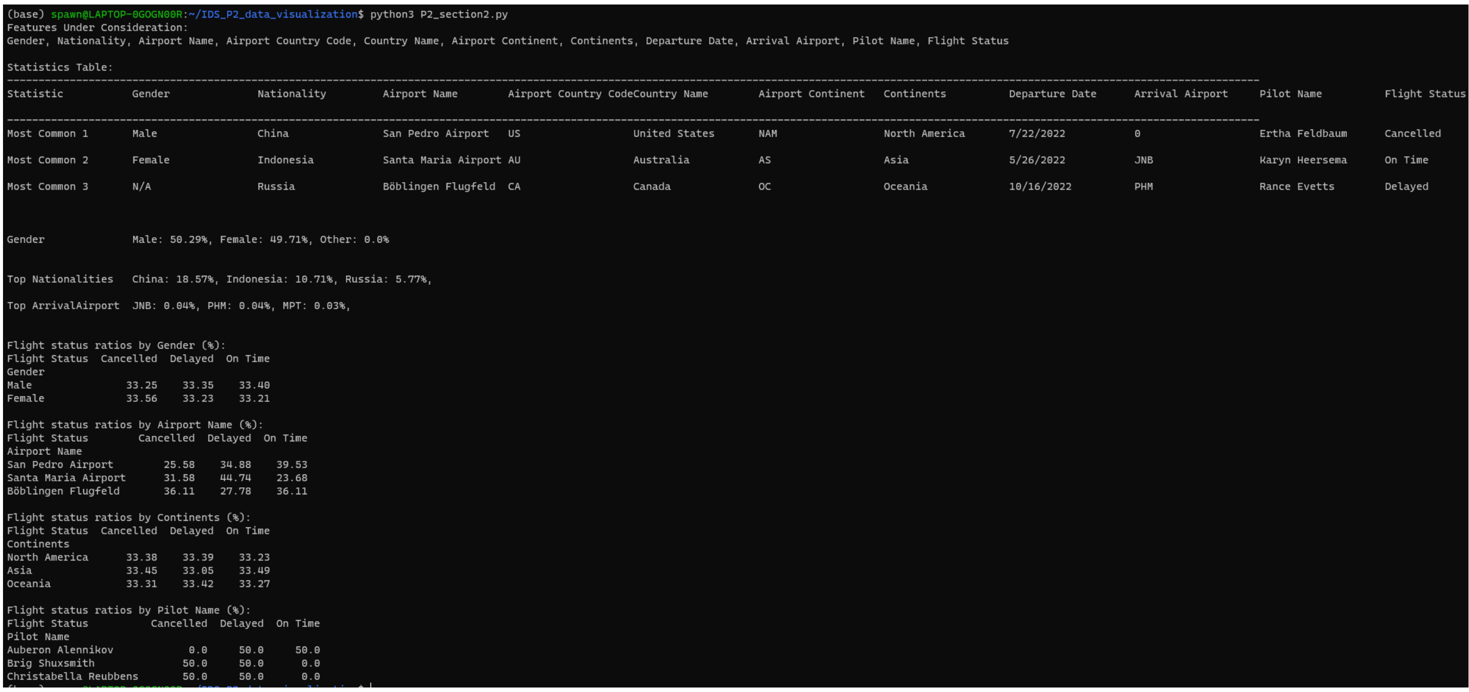
P2.Data visualization: “Exploring Insights Airline Data”

1 Introduction.

The Data set has been found on Kaggle, labeled as license"Public Domain". The original set is from www.mockaroo.com and has been posted 30/08/2023 on Kaggle according to the page description. A recent update of the .csv file has also been made on 13/09/2023. The airline dataset contains a rich set of 15 features, spanning passenger demographics, travel details, flight information, and airport specifics. With 98000 non null entries across all columns, the dataset is quite large, enhancing its utility for robust statistical and machine learning analyses. Features such as Passenger ID, Gender, Age, Nationality, and Flight Status allow for a detailed understanding of customer demographics and flight operational status. The inclusion of Airport and Pilot details adds layers of granularity that could be beneficial for optimizing operations and "regulatory compliance": airline industry is dealing with multifaceted issues that involve adhering to a complex set of laws and guidelines designed to ensure safety, fairness, and efficiency. Non-compliance can result in hefty fines, legal action, and damage to the airline's reputation. Furthermore, the dataset includes temporal information through the 'Departure Date' field, with dates as recent as December 2022, adding relevance for time-sensitive analytics; the dataset is recent, feature-rich, large, and complete. It is well-suited for a range of applications, from enhancing the smoothness of operations and elevating passenger satisfaction to assisting governing agencies and conducting industry analysis. Given its comprehensiveness and size, it offers a strong foundation for insightful analytics in the aviation sector.

2 Dataset overview:

We are especially interested in the delays" of operations.
We will not consider the following columns/features: ID, Age, First Name, Last Name .
A python script has been written to retrieve and display the information.



```
(base) spawn@LAPTOP-0GOGN00R:~/IDS_P2_data_visualization$ python3 P2_section2.py
Features Under Consideration:
Gender, Nationality, Airport Name, Airport Country Code, Country Name, Airport Continent, Continents, Departure Date, Arrival Airport, Pilot Name, Flight Status

Statistics Table:
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
Statistic           Gender              Nationality         Airport Name        Airport Country CodeCountry Name        Airport Continent   Continents          Departure Date      Arrival Airport     Pilot Name          Flight Status

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
Most Common 1       Male                China               San Pedro Airport   US                  United States       NAM                 North America       7/22/2022           0                   Ertha Feldbaum      Cancelled

Most Common 2       Female              Indonesia           Santa Maria Airport AU                  Australia           AS                  Asia                5/26/2022           JNB                 Karyn Heersema      On Time

Most Common 3       N/A                 Russia              Böblingen Flugfeld  CA                  Canada              OC                  Oceania             10/16/2022          PHM                 Rance Evetts        Delayed



Gender              Male: 50.29%, Female: 49.71%, Other: 0.0%


Top Nationalities   China: 18.57%, Indonesia: 10.71%, Russia: 5.77%,

Top ArrivalAirport  JNB: 0.04%, PHM: 0.04%, MPT: 0.03%,


Flight status ratios by Gender (%):
Flight Status  Cancelled  Delayed  On Time
Gender
Male               33.25    33.35    33.40
Female             33.56    33.23    33.21

Flight status ratios by Airport Name (%):
Flight Status        Cancelled  Delayed  On Time
Airport Name
San Pedro Airport        25.58    34.88    39.53
Santa Maria Airport      31.58    44.74    23.68
Böblingen Flugfeld       36.11    27.78    36.11

Flight status ratios by Continents (%):
Flight Status  Cancelled  Delayed  On Time
Continents
North America      33.38    33.39    33.23
Asia               33.45    33.05    33.49
Oceania            33.31    33.42    33.27

Flight status ratios by Pilot Name (%):
Flight Status          Cancelled  Delayed  On Time
Pilot Name
Auberon Alennikov            0.0     50.0     50.0
Brig Shuxsmith              50.0     50.0      0.0
Christabella Reubbens       50.0     50.0      0.0
```

3 Charts:

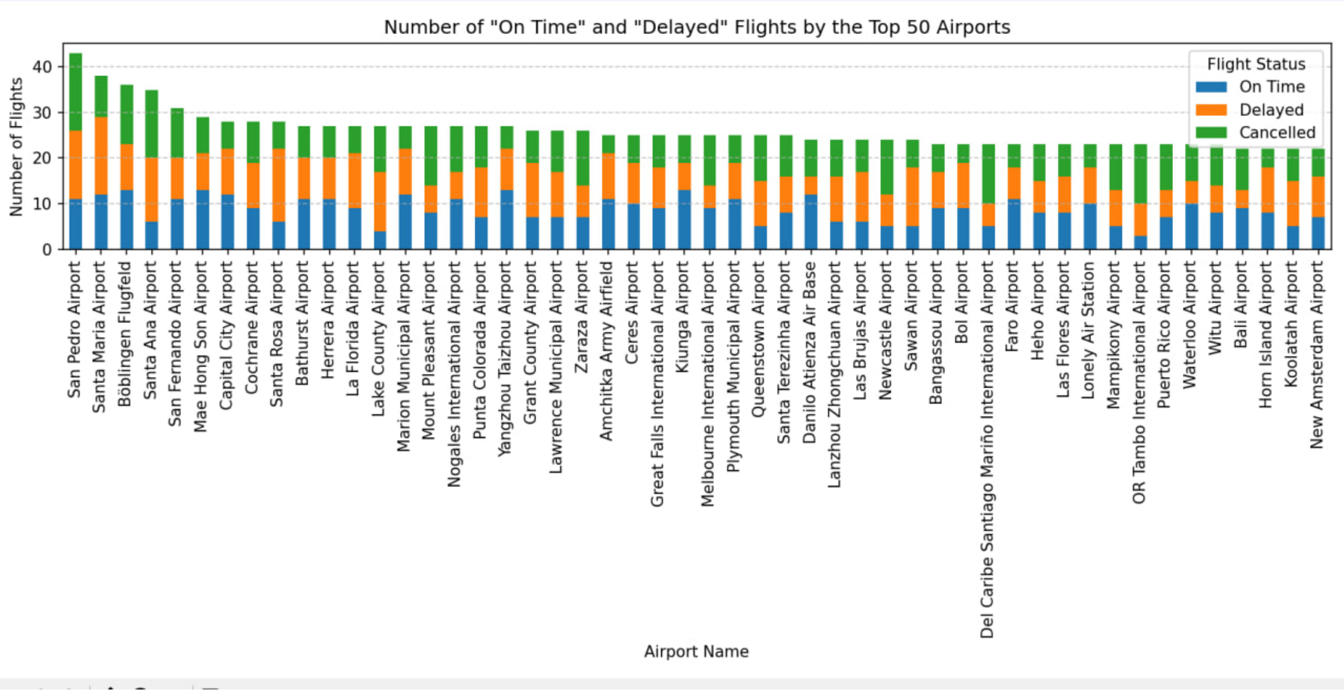


Figure 1:

This chart is showing the relation between number of flights and the name of the airport that the plane started from[Figure 1]. The blue bar represents the flight amount that is on time, the orange bar represents the flight amount that is delayed, the green bar represents the flight amount that is canceled. Looking at this chart it is seen that some airports such as Lake County Airport, OR Tambo International Airport, De Caribe Santiago Marino International Airport should focus more on the processes of collecting passengers and releasing the plane on time. It also can be noticed that there are some airports that have more on time planes then all the others. It means that to make processes better, the administration of airports with bad rates of “On Time” flights can look into the processes implemented in the airports with good rates, for example Mae Hong Son Airport.

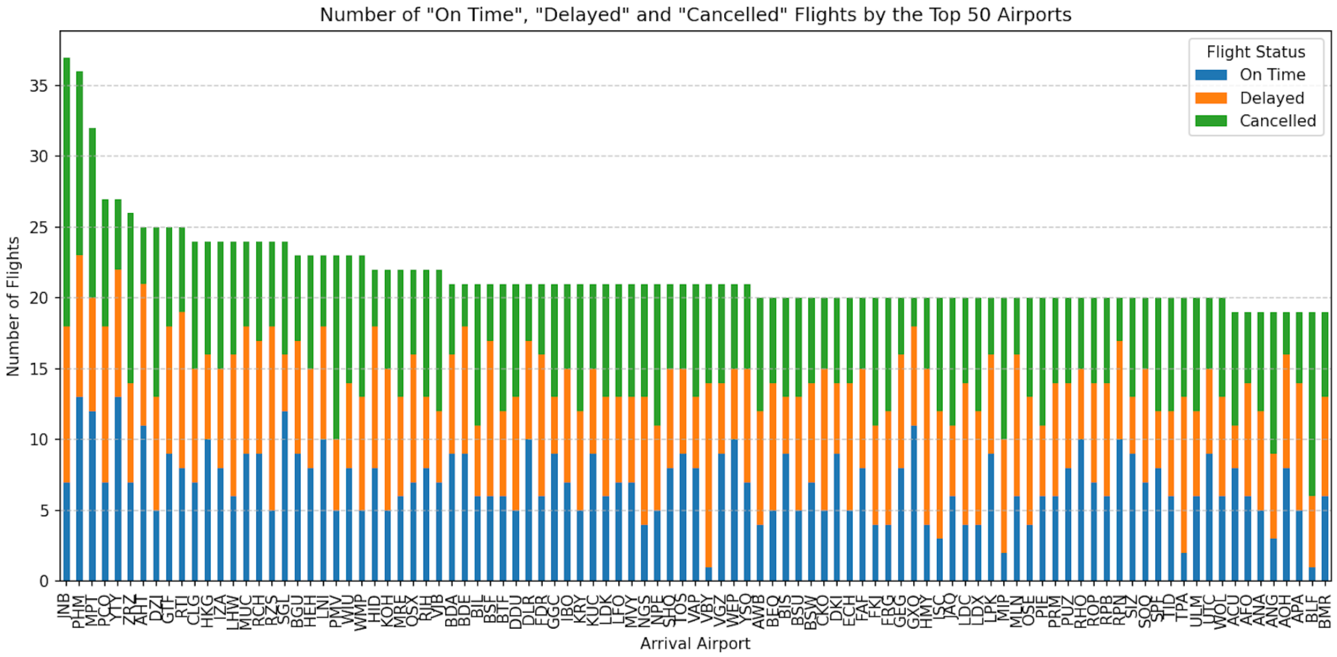


Figure 2:

The second chart displays the relation between number of flights and its statuses and the arrival airport code[Figure 2]. From this chart it is clear that some airports have awful processes of receiving planes, for example VBY airport has only 1 “On Time” flight from 21 overall recorded flights. Furthermore, this chart shows the importance of comparing with the overall load of the airport. INB airport has 7 “On Time” flights, however, by comparing with other airports, and by looking at the number of the overall flights, it is clear that this number should be bigger.
In the next charts we decided to analyze the age of travelers in the top and worst traveling countries.

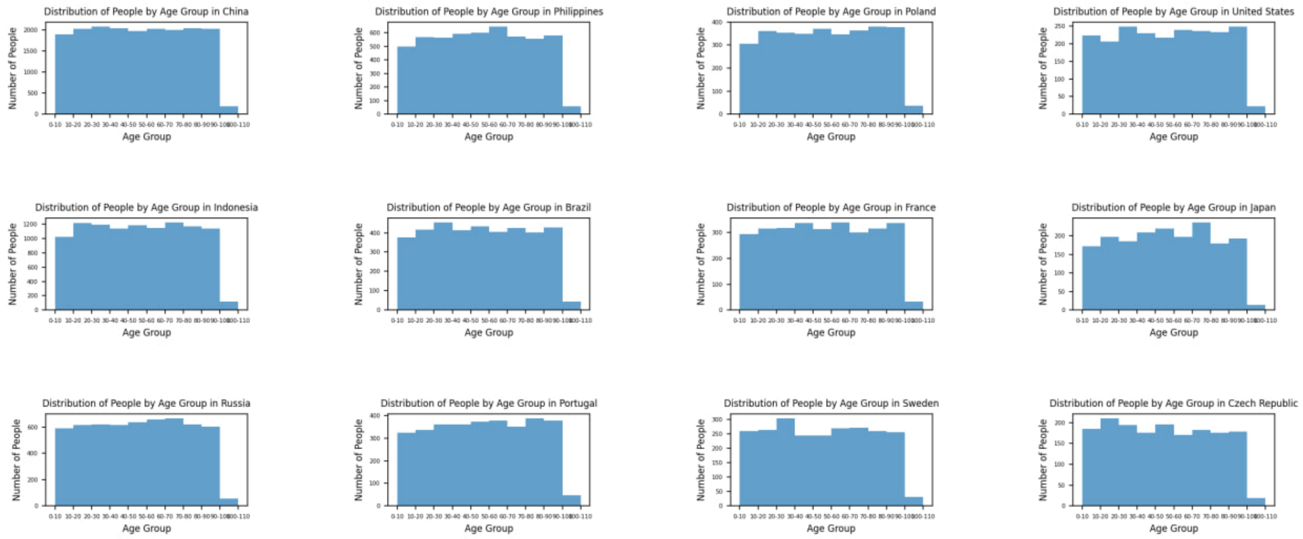


Figure 3:

This is the chart of age distribution over several countries that have the most flights based on this dataset[Figure 3]. Overall, the histograms look almost similar, but in some countries like for example Philippines, or Japan, there are more elderly people traveling. And in some countries like Brazil, there are more young people traveling. The reason why it is the case is that countries like Japan have a longer life expectancy and better life quality, so elderly people are able to travel. And in countries like Brazil, it is reversed.

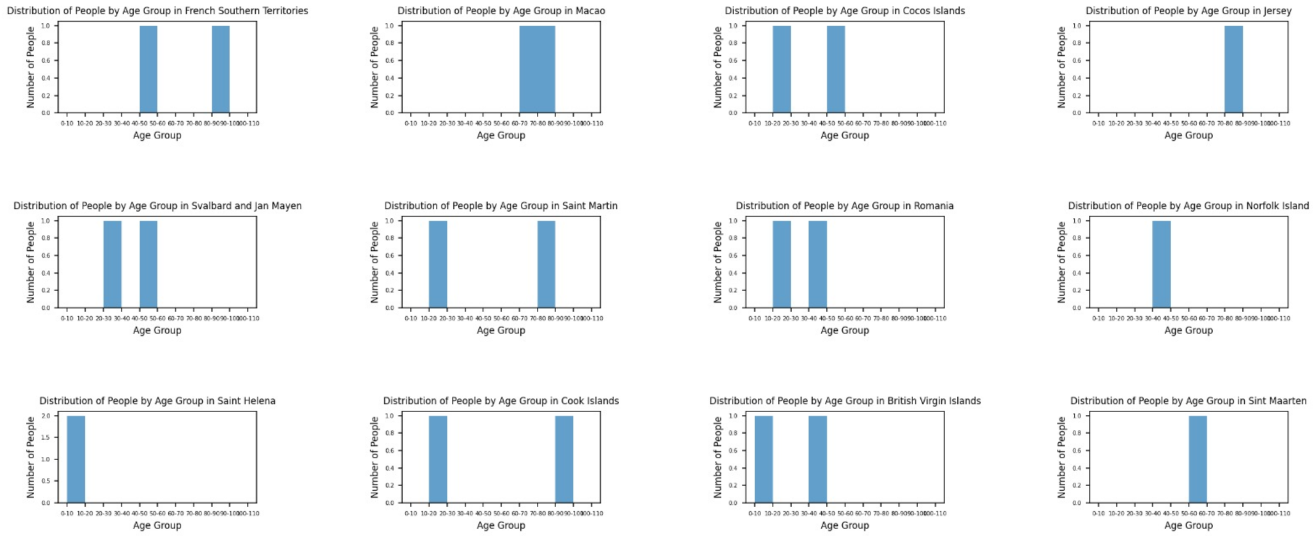


Figure 4:

This is the chart of age distribution over several countries that have the least flights based on this dataset[Figure 4]. The reason why this chart looks so different is because those airports are not as popular as top airports. Using this data it is hard to analyze anything. But we can see that there are more young people traveling to those countries than the elderly. It can mean that these people like to experiment and find new ways to travel.

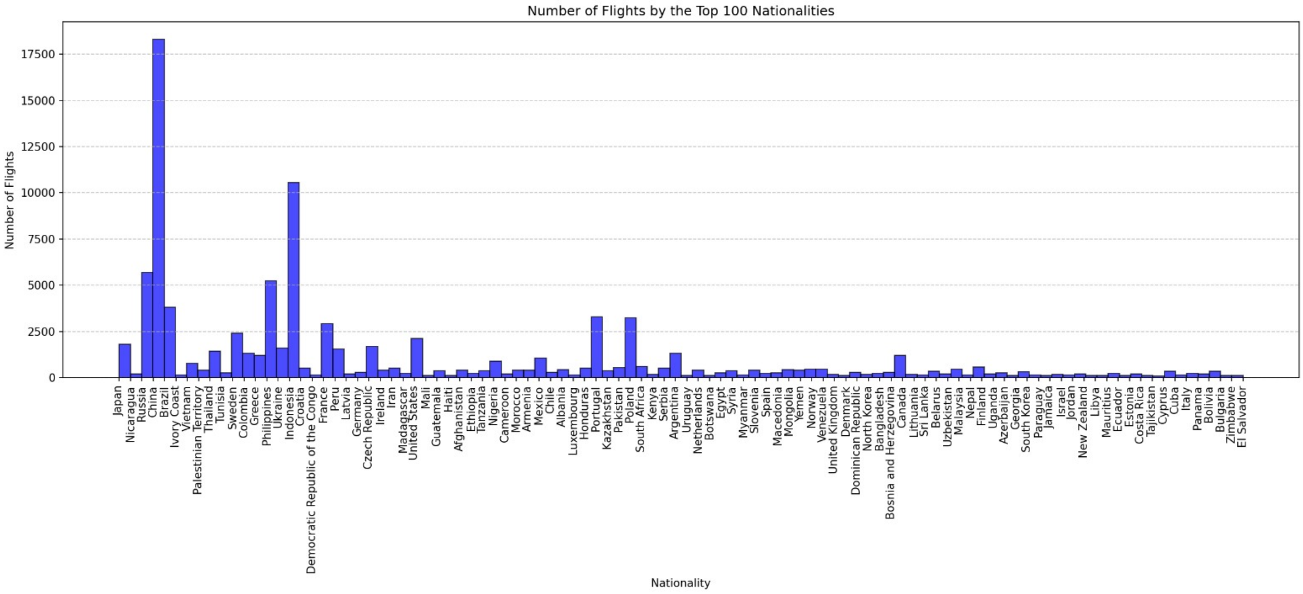


Figure 5:

This bar graph represents the relation between the country and the amount of the flights made in them[Figure 5]. The reason why some of them have more flights then the others, is in the amount of people living there.

4 Problem description:

There are several problems that could be solved with this data set. One of which is the problem of airports with bad flight management. From the first two charts it is possible to identify the most underperforming airports, the ones with the biggest number of delayed and canceled flights. This indicates bad flight management and influences customer experience in the airport. Using this data set it is possible for a data scientist to make a list of good and bad airports and give this list to people in the form of a website or an app, and a user will be able to choose an airport based on this information. Also the management of the airports can look at the better performing airports and try to replicate their processes or knowledge to achieve the less number of delayed and canceled flights. The other problem that can be solved using this data set is bad customer orientation. For an airport business, as for any business, it is important to target the correct customers. The charts 3 and 4 can help the advertising management of airports of certain countries to correctly target the correct age group. This can increase the profits of a given airport because there will be no need to spend money on advertisements that will attract non-flyers.

5 References:

Predicting Flight Delays.
”Selection of Best Machine Learning Model to Predict Delay in Passenger Airlines.”
Authors: R. Kothari, R.Kakkar, G. Sharma
Year: 2023
Type: research article
DOI: r 10.1109/ACCESS.2023.3298979
Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10194905
The paper aims to address a critical issue in the aviation industry flight delays. These delays not only inconvenience passengers but also result in significant financial losses for airlines. The study focuses on leveraging machine learning algorithms to predict flight delays more accurately, thereby enabling airlines to take proactive measures. The primary objective is to identify the most effective machine learning model for predicting flight delays. The study compares various machine learning algorithms, including: decision trees, random forests, support Vector Machines (SVM), and neural networks, to determine which model offers the highest accuracy and efficiency. The ”researchers” collected data from multiple airlines, including variables like weather conditions, air traffic, mechanical issues, and crew availability. This data was then pre-processed to remove any inconsistencies or outliers. The study employed a 70-30 split for training and testing the models. Feature selection techniques like Recursive feature elimination (RFE) were used to identify the most relevant variables. Then the paper presents a comparative analysis of the performance of different machine learning models. Metrics such as accuracy, precision, recall, and F1-score were used for evaluation. Random forest algorithm emerged as the most effective, with an accuracy rate of 92%, followed by Neural Networks at 89%. SVM and Decision trees lagged behind with accuracy rates of 85% and 82%, respectively. The study emphasizes the real-world applicability of the selected model. The random forest algorithm’s high accuracy and low computational cost make it ideal for real time prediction systems. Airlines can integrate this model into their existing infrastructure to predict delays and inform both passengers,crew,online users in advance, thereby minimizing inconveniences and operational costs.While the study provides valuable insights, it acknowledges certain limitations. The data used was limited to specific airlines and routes, which may not be universally applicable. Future research could focus on incorporating more diverse data sets and exploring ensemble methods for even more accurate predictions. They conclude that machine learning algorithms, particularly the random forest model, can significantly improve the prediction of flight delays. By adopting such technology, airlines can enhance customer satisfaction, reduce financial losses, and optimize operational efficiency.

Passenger Behavior Analysis.
”An aircraft boarding model accounting for group behavior.”
Authors: Tie-Qiao Tang, Shao-Peng Yang, Hui Ou, Liang Chen, Hai-Jun Huang.
Year: 2018
Journal: Journal of Air Transport Management
DOI: 10.1016/j.jairtraman.2018.03.004
https://www.sciencedirect.com/science/article/abs/pii/S0969699717304933
This paper delves into the challenges posed by the rapid increase in civil aviation, particularly focusing on aircraft boarding efficiency. The authors propose a new model that accounts for group behaviour during the boarding process. The study aims to explore the impacts of group behaviour on various aspects such as passenger motion, seat conflict,check-in time, time of handling luggage and overall boarding time; starting by discussing the existing models and strategies for aircraft boarding, highlighting that none consider the impact of group behaviour, which is often present during the boarding process. The authors lay down several assumptions for their model, such as the boarding scenario, the number of passengers, and the conditions under which seat conflicts occur. The research defines time parameters like T1 (time for ticket checking) and Tlugg (time for luggage handling). These parameters are adjusted based on whether passengers are boarding as a group or individually. The authors conduct numerical tests to study the influences of group behaviour on passenger trajectory, seat conflicts, and boarding time. The results indicate that group behaviour positively impacts boarding efficiency. Researchers use Ordinary Differential Equations (ODE) to describe passenger movements, incorporating variables like passenger speed, optimal speed, and distance between passengers and concludes that encouraging group behaviour can enhance efficiency during the boarding process. The impacts become more prominent with an increasing number of groups.

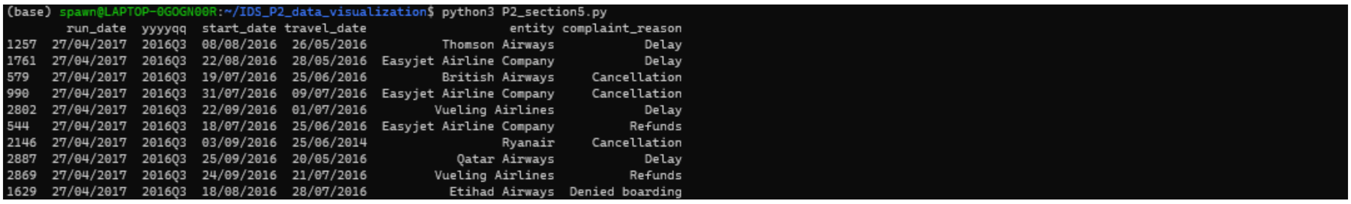
Airline Fuel Efficiency.
”Improving airline fuel efficiency via fuel burn prediction and uncertainty estimation”
Authors: Kang, Lei; Hansen, Mark
Year: 2018
Journal: Transportation research.
DOI: 10.1016/j.trc.2018.10.002
https://www.sciencedirect.com/science/article/abs/pii/S0968090X18314153
The article addresses the critical issue of fuel consumption in the aviation industry. The study tends to enhance fuel efficiency by providing more ”accurate discretionary fuel loading recommendations to dispatchers”. The paper employs ensemble learning techniques to improve fuel burn prediction and constructs prediction intervals to capture the uncertainty of these predictions. The paper highlights that reducing fuel consumption is a unifying goal across the aviation industry. It mentions that dispatchers generally overload discretionary fuel to account for various uncertainties like weather and traffic congestion. The authors propose a novel discretionary fuel estimation approach using ensemble learning techniques. They also introduce prediction intervals to capture the uncertainty of model predictions. Data were collected from a major U.S.-based airline, the FAA’s ASPM database, and the NOAA database. The main target variable is the actual fuel burn for a flight. Statistical Contingency Fuel (SCF): The paper discusses the limitations of the current SCF estimation procedure, which is widely used but often leads to overloading of fuel. The authors propose a more dynamic and reliable SCF estimation approach based on machine learning techniques. The potential benefit of this approach is estimated to be 61.5 million dollars in fuel savings and 428 million kilograms of Co2 reduction per year for the study airline. This is a significant contribution to the field of aviation fuel efficiency, providing a data-driven approach to mitigate the challenges of fuel consumption and its environmental impact.

Air Traffic Control Optimization.
”TRACON; Aircraft Arrival Planning and Optimization through Spatial Constraint Satisfaction”
Authors: Bergh, Christopher P.; Krzeczowski, Kenneth J.; Davis, Thomas J.
Year: 1995 !
Journal: American Institute of Aeronautics and Astronautics
DOI: 10.2514/atcq.3.2.117
https://ntrs.nasa.gov/citations/20020017252
This research introduces a new algorithm for aircraft arrival planning and optimization. The algorithm is integrated into the ”Final Approach Spacing Tool” (FAST) in the Center Terminal Radar Approach Control (TRACON) automation system developed at NASA-Ames Research Center. The paper aims to address the inefficiencies and complexities in the current air traffic control arrival planning procedures.The primary objective for the authors is to create an operationally acceptable aircraft arrival plan that considers the entire arrival airspace, not just the runway. The authors argue that generating efficient and conflict-free aircraft arrival plans at the runway does not guarantee an operationally acceptable plan upstream from the runway. The new design aims to reduce air traffic controller workload and improve efficiency in TRACON procedures. The authors conducted FAST simulations involving full-proficiency, level five air traffic controllers from the Dallas-Fort Worth (DFW) TRACON. Based on these simulations, they designed and coded an algorithm called ”spatial constraint satisfaction,” which underwent testing and will soon begin field evaluation at DFW and Denver International Airport facilities. The new algorithm uses ”spatial constraint satisfaction to create a comprehensive plan for the entire terminal airspace. It includes ”representations of controller preferences, operationally required amounts of extra separation, and integrates procedures for aircraft conflict resolution”. The same algorithm aims to reduce the workload of both ”feeder” and ”final” controllers by providing a more balanced and efficient plan for aircraft arrival. The paper discusses the complexities of runway allocation, especially during high-traffic ”rushes.” The algorithm aims to balance aircraft across multiple runways efficiently. The design was the result of an iterative process involving real time simulation and controller input, making it more aligned with practical needs. The research emphasizes that planning decisions should be based on information that is both local to particular aircraft and global to the entire TRACON to create an efficient and operationally acceptable plan; this field will soon feed the airline industry and improve its global organization.

Customer Experience and Loyalty.
”An analysis of the concept and its performance in airline brands.”
Authors: L. Calum, M. Keith
Year: 2014
Journal: Research in Transportation Business & Management
DOI: 10.1016/j.rtbm.2014.05.004
https://www.researchgate.net/publication/262923057
This one delves into the intricate relationship between customer experience and brand performance in the airline industry. It aims to clarify the often-misunderstood concept of customer experience and assess its impact on brand loyalty and advocacy. The article investigates how the airline industry applies the concept of customer experience and measures its performance. It uses data from the International Air Transport Association (IATA), collected over a year from 18,567 passengers across fifteen major full-service airlines in Europe, the Middle East, and Asia. Furthermore it notes that the concept of customer experience is not well-understood and lacks a clear and consistent definition. It is more than just a one-time transactional experience and involves multiple ”touchpoints” between the company and the customer. The ultimate goal of providing a good customer experience is to achieve brand loyalty and advocacy. Satisfied customers are more likely to remain loyal to the brand even under less-than-ideal circumstances and are more likely to recommend the brand to others. The article argues that there is no generally accepted method for measuring customer experience. The key indicator is customer satisfaction, which needs to be understood at various stages of the customer journey to refine the experience over time. Despite its importance, many airlines either do not apply the concept of customer experience or do not understand passenger satisfaction levels across their entire customer journey. Some airlines like ”Etihad Airways” have started focusing on customer experience as a key strategy for differentiation and growth. The authors highlights that achieving excellent customer experience is a challenge for airlines, especially when many of the staff involved in the customer journey (e.g., airports, security, customs) are outside an airline’s direct control. To conclude, understanding and applying the concept of customer experience is crucial for airlines to differentiate themselves in a competitive market and to foster brand loyalty and advocacy.

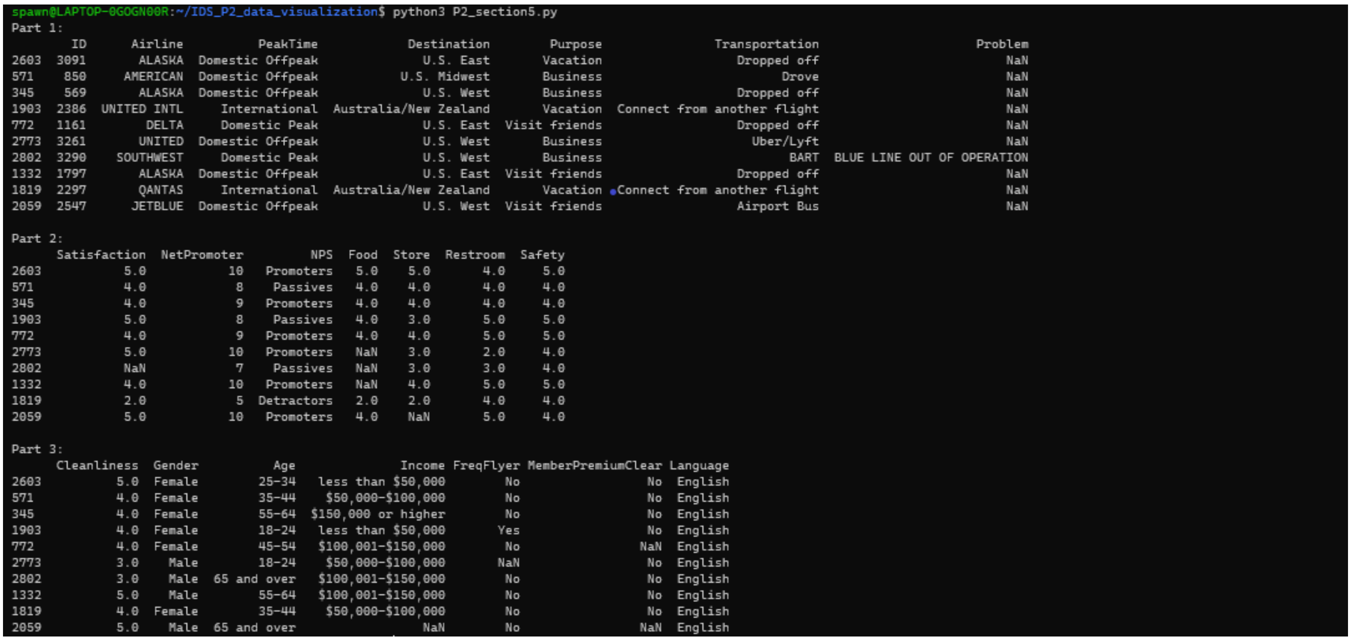
Alternatives datasets found through GHDB(used to find publicly available information):

Passenger_Complaints_all_data_2016.csv
Random 10 rows sample:



```
(base) spawn@LAPTOP-0GOGN00R:~/IDS_P2_data_visualization$ python3 P2_section5.py
        run_date  yyyyqq  start_date travel_date                   entity complaint_reason
1257  27/04/2017  2016Q3  08/08/2016  26/05/2016          Thomson Airways            Delay
1761  27/04/2017  2016Q3  22/08/2016  28/05/2016  Easyjet Airline Company            Delay
579   27/04/2017  2016Q3  19/07/2016  25/06/2016          British Airways     Cancellation
990   27/04/2017  2016Q3  31/07/2016  09/07/2016  Easyjet Airline Company     Cancellation
2802  27/04/2017  2016Q3  22/09/2016  01/07/2016         Vueling Airlines            Delay
544   27/04/2017  2016Q3  18/07/2016  25/06/2016  Easyjet Airline Company          Refunds
2146  27/04/2017  2016Q3  03/09/2016  25/06/2014                  Ryanair     Cancellation
2887  27/04/2017  2016Q3  25/09/2016  20/05/2016            Qatar Airways            Delay
2869  27/04/2017  2016Q3  24/09/2016  21/07/2016         Vueling Airlines          Refunds
1629  27/04/2017  2016Q3  18/08/2016  28/07/2016           Etihad Airways  Denied boarding
```

Airport_SFO_satisfaction.csv
Random 10 rows sample(split in three parts):



```
spawn@LAPTOP-0GOGN00R:~/IDS_P2_data_visualization$ python3 P2_section5.py
Part 1:
        ID      Airline          PeakTime            Destination        Purpose              Transportation                     Problem
2603  3091       ALASKA  Domestic Offpeak              U.S. East       Vacation                 Dropped off                         NaN
571    850     AMERICAN  Domestic Offpeak           U.S. Midwest       Business                       Drove                         NaN
345    569       ALASKA  Domestic Offpeak              U.S. West       Business                 Dropped off                         NaN
1903  2386  UNITED INTL     International  Australia/New Zealand       Vacation  Connect from another flight                         NaN
772   1161        DELTA     Domestic Peak              U.S. East  Visit friends                 Dropped off                         NaN
2773  3261       UNITED  Domestic Offpeak              U.S. West       Business                   Uber/Lyft                         NaN
2802  3290    SOUTHWEST     Domestic Peak              U.S. West       Business                        BART  BLUE LINE OUT OF OPERATION
1332  1797       ALASKA  Domestic Offpeak              U.S. East  Visit friends                 Dropped off                         NaN
1819  2297       QANTAS     International  Australia/New Zealand       Vacation  Connect from another flight                         NaN
2059  2547      JETBLUE  Domestic Offpeak              U.S. West  Visit friends                 Airport Bus                         NaN

Part 2:
      Satisfaction  NetPromoter         NPS  Food  Store  Restroom  Safety
2603           5.0           10   Promoters   5.0    5.0       4.0     5.0
571            4.0            8    Passives   4.0    4.0       4.0     4.0
345            4.0            9   Promoters   4.0    4.0       4.0     4.0
1903           5.0            8    Passives   4.0    3.0       5.0     5.0
772            4.0            9   Promoters   4.0    4.0       5.0     5.0
2773           5.0           10   Promoters   NaN    3.0       2.0     4.0
2802           NaN            7    Passives   NaN    3.0       3.0     4.0
1332           4.0           10   Promoters   NaN    4.0       5.0     5.0
1819           2.0            5  Detractors   2.0    2.0       4.0     4.0
2059           5.0           10   Promoters   4.0    NaN       5.0     4.0

Part 3:
      Cleanliness  Gender          Age               Income FreqFlyer MemberPremiumClear Language
2603          5.0  Female        25-34   less than $50,000        No                 No  English
571           4.0  Female        35-44    $50,000-$100,000        No                 No  English
345           4.0  Female        55-64  $150,000 or higher        No                 No  English
1903          4.0  Female        18-24   less than $50,000       Yes                 No  English
772           4.0  Female        45-54   $100,001-$150,000        No                NaN  English
2773          3.0    Male        18-24    $50,000-$100,000       NaN                 No  English
2802          3.0    Male  65 and over   $100,001-$150,000        No                 No  English
1332          5.0    Male        55-64   $100,001-$150,000        No                 No  English
1819          4.0  Female        35-44    $50,000-$100,000        No                 No  English
2059          5.0    Male  65 and over                 NaN        No                NaN  English
```

All data sets are related to the airline or aviation industry. They likely contain data that is customer focused, such as passenger complaints and satisfaction levels, which can be correlated with the original data set's passenger demographics. About data-granularity, the original dataset seems to have a broader scope, covering various aspects like routes, pilots, and departure dates. The new datasets might focus specifically on complaints or satisfaction metrics. Time periods differ: the complaints dataset specifies data "Passenger_Complaints_all_data_2016.csv". The original dataset's time period is from 2023, this could lead to misinterpretation. The geographical focus differs: the satisfaction dataset is specifically about SFO Airport, whereas the original dataset seems to be more globally oriented. The complaints and satisfaction datasets can be correlated with the original dataset based on "Passenger ID" or "Nationality" to get insights into what kinds of passengers are more likely to complain or be satisfied. By correlating 'Flight Status' from the original dataset with complaints or satisfaction metrics, we could gain operational insights. These kinds of correlations could provide more precise insights into customer behavior, operational efficiencies, and areas needing improvement.