CONESTOGA

Connect Life and Learning

Data Analytics Final Project

Analyze factors contributing to flight delays and propose strategies for improvement.

Bhamani, Darshit

Business Analytics

INFO8066 - Winter 2024 - Section 3

Data Analytics

Chris Jung

April 17, 2024

INTRODUCTION

In an endeavor to elevate operational efficiency and flight punctuality within the aviation sector, our comprehensive analysis leverages a dataset detailing flight delays, carrier specifics, and airport operations. This exploration is driven by the goal of identifying and mitigating the underlying causes of flight delays, thereby enhancing customer satisfaction, and reducing financial burdens on airlines. Key objectives include improving on-time arrivals, diminishing average delay durations, and fostering positive customer experiences. Success will be gauged through measurable improvements in these areas, aiming to strengthen the airlines' operational performance and customer perception. Through a data-driven approach, this analysis seeks to propose actionable strategies that address the complexities of flight delays, setting a path toward a more efficient and reliable air travel experience.

Dataset: Airline delay dataset containing flight delay information, carrier details, and airport data of USA.

LINK: https://www.kaggle.com/datasets/eugeniyosetrov/airline-delays

CLIENT

The ideal end user of our analysis is the Director of Operations for a leading national airline, tasked with the critical responsibility of streamlining flight operations to ensure punctuality and reliability. This director is specifically concerned with the systemic issue of flight delays that not only tarnish the airline's reputation but also lead to substantial financial losses and diminished customer satisfaction. Their role encompasses the oversight of all operational aspects that directly impact flight schedules, including but not limited to crew management, fleet allocation, and coordination with airports.

OPERATIONAL ANALYTICS PROBLEM STATEMENT

In recent years, airline has grappled with an increasing trend of flight delays, culminating in a tangible decline in customer satisfaction scores and a noticeable erosion of brand loyalty. These delays have not only imposed additional operational costs — ranging from crew overtime to compensatory vouchers for affected passengers — but have also deterred potential customers, affecting both market share and revenue growth adversely. The escalation of such incidents, especially during peak travel seasons, has highlighted inefficiencies within the current operational framework and the urgent need for a data-driven overhaul.

This challenge is multifaceted, stemming from both controllable factors, such as aircraft turnaround times and crew scheduling, and uncontrollable elements like weather conditions and security issues. The complexity of the airline's network, combined with the variability in airport facilities and operations, further complicates efforts to minimize delays and improve reliability.

Recognizing the critical need to address these operational hurdles, the Director of Operations has initiated a strategic project aimed at harnessing advanced analytics to diagnose the root causes of delays. The objective is to employ predictive modeling and operational data analysis to not only identify the primary drivers of delays within their control but also to forecast potential disruptions. This initiative is envisioned to equip the airline with actionable insights, enabling the implementation of targeted interventions to enhance operational efficiency, reduce costs associated with delays, and, most importantly, elevate the customer experience. The

goal is to reclaim the airline's position as a leader in punctuality and reliability, thereby driving competitive advantage and fostering long-term customer loyalty.

CRISP DM

Based on the given context of the dataset, which involves analyzing airline delay data, here's how we can adapt the CRISP-DM framework:

Business Understanding:
Objective: Understand the factors contributing to flight delays in the airline industry and identify strategies to minimize them.
Goals: Reduce overall delay times, improve flight punctuality, enhance customer satisfaction, and optimize operational efficiency.
Success Criteria: Decrease in average delay times, increase in on-time arrivals, and positive feedback from customers.

Data Understanding:
Data Source: Obtain the airline delay dataset containing information on flight delays, reasons for delays, carrier information, and airport details.
Initial Assessment: Review the dataset to understand its structure, size, and variables.
Exploratory Data Analysis (EDA): Analyze data distributions, summary statistics, and correlations between variables to gain insights into the dataset.

Data Preparation:
Data Cleaning: Handle missing values, outliers, and inconsistencies in the dataset.
Feature Engineering: Extract relevant features such as delay times, carrier information, airport details, and delay reasons.
Data Transformation: Normalize or scale features as needed for modeling.

Modeling:
Model Selection: Choose appropriate modeling techniques such as linear regression, decision trees, or ensemble methods to predict flight delays.
Model Training: Train the selected models using the prepared dataset, considering factors like carrier performance, weather conditions, and airport congestion.
Model Evaluation: Evaluate model performance using metrics like R-squared, mean squared error, and accuracy.

Evaluation:
Model Performance: Assess the effectiveness of the trained models in predicting flight delays accurately.
Business Impact: Determine the potential impact of deploying the models on operational efficiency, customer satisfaction, and cost savings.
Feedback Loop: Gather feedback from stakeholders and domain experts to refine the models and improve their performance.

Deployment:

Implementation Strategy: Develop a deployment plan for integrating the models into airline operations.

Monitoring Mechanism: Establish mechanisms to monitor model performance in real-time and detect anomalies or deviations from expected outcomes.

Integration with Operations: Integrate the models with existing airline systems to facilitate decision-making and optimize resource allocation.

Iterative Process:

Continuous Improvement: Iterate on the data preparation, modeling, and evaluation phases based on feedback and new insights gained from deployment.

Adaptation to Changes: Adapt the models to changes in flight patterns, regulations, and external factors that may affect delay predictions.

By following the CRISP-DM framework tailored to the airline delay dataset, we can systematically analyze the data, develop predictive models, and deploy solutions that address the challenges of flight delays in the airline industry.

PROJECT DELIVERABLES:

1. Finding Percentage of delayed flight

Python Query Output:
Percentage of Delayed Flights in 2019: 20.29%
Percentage of Delayed Flights in 2020: 11.72%

Insights:
- In 2019, every 5th flight arrived with a delay (20.29%)
- In 2020, every 10th flight arrived with a delay (11.72%)

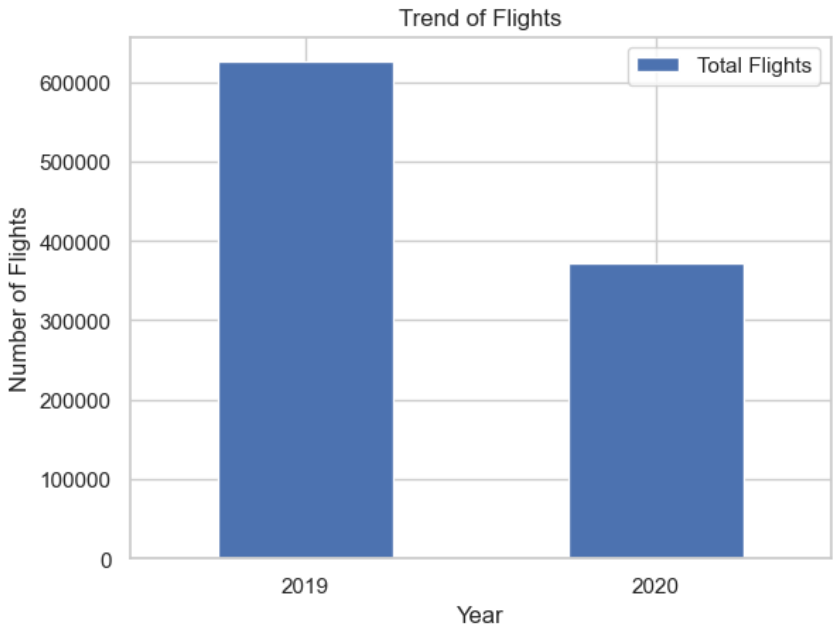
2. Finding Total flight delay time

Python Query Output:
Total delay time in 2019: 143073 hours
Total delay time in 2020: 42678 hours

Insights:
- In 2019, the total duration of the delay was 143,073 hours.
- In 2020, the total duration of the delay was 42,678 hours.
- The difference between 2019 and 2020 is extremely significant and amounts to 100,395 hours of total delays. The difference looks staggering.

3. Total Number of Flight

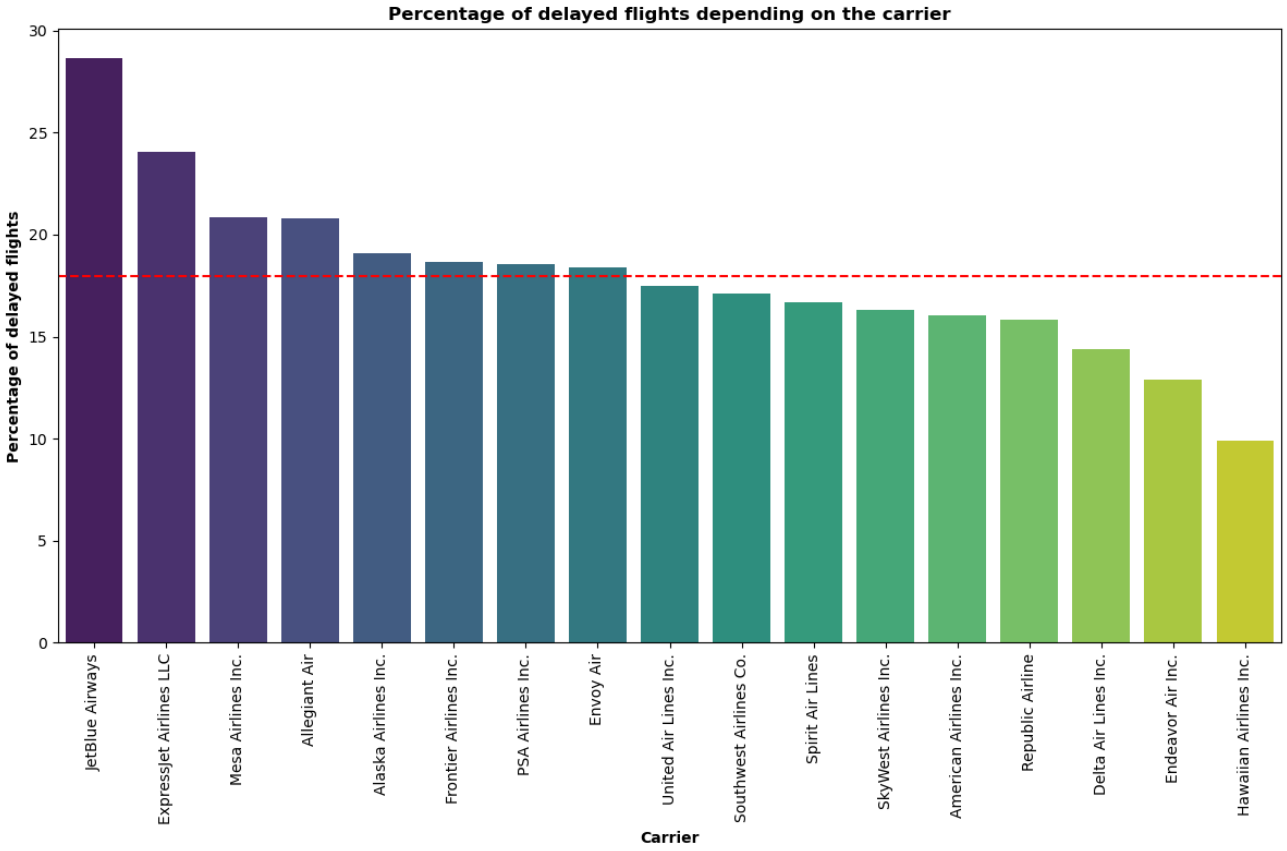
Python Query Output:



Insights:
- The number of flights is decreased in year 2020 compared to 2019 due to natural pandemic of COVID.

4. Flight carriers' performance and comparison between them based on Percentage of delayed flights.

Python Output Query:



Insights:
Mean Percentage of delayed flights: 17.98

- Top 5 carriers by percentage of delayed flights (BEST):

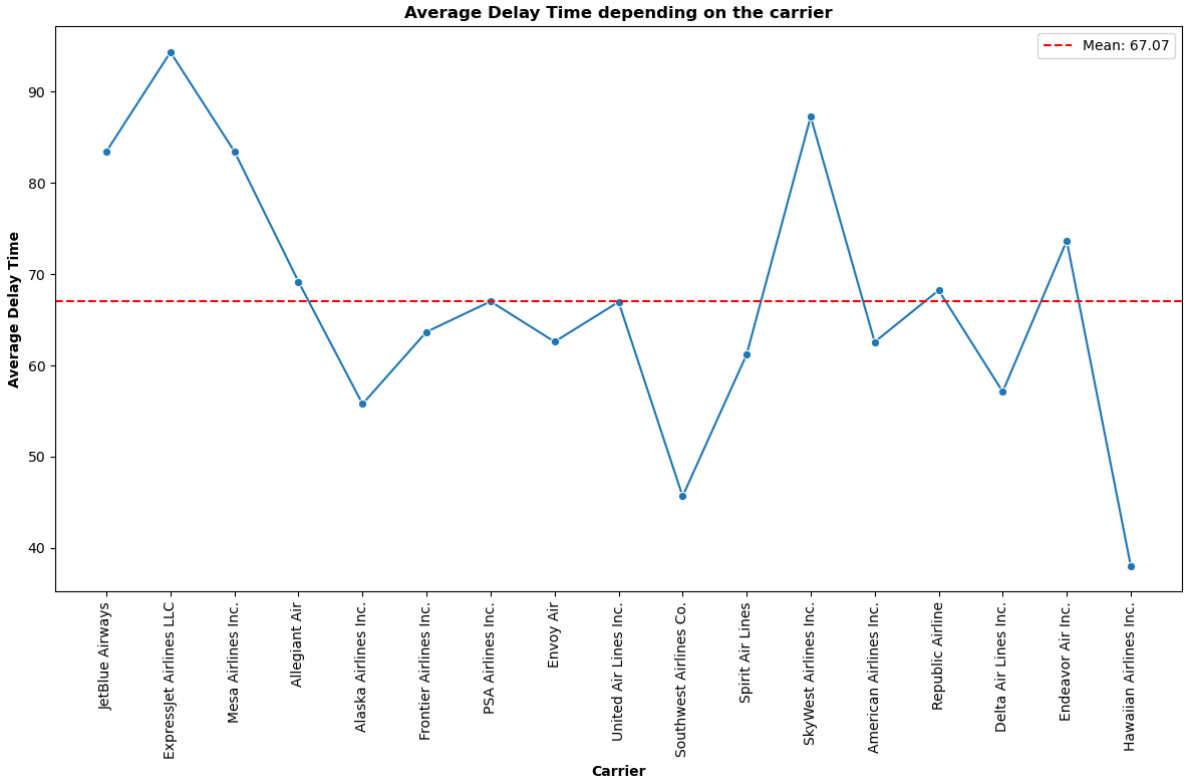
Hawaiian Airlines Inc. - 10.0%
Endeavor Air Inc. - 13.0%
Delta Air Lines Inc. - 14.0%
Republic Airline - 16.0%
American Airlines Inc. - 16.0% Given the choice, I would recommend using these carriers.

- Top 5 carriers by percentage of delayed flights (WORST):

JetBlue Airways - 29.0%
ExpressJet Airlines LLC - 24.0%
Mesa Airlines Inc. - 21.0%
Allegiant Air - 21.0%
Alaska Airlines Inc. - 19.0%

5. Flight carriers' performance and comparison between them based on average delay time.

Python Query Output:



Insights:
Overall Average Delay Time: 67.07

- Top 5 carriers by average flight delay duration (BEST):

Hawaiian Airlines Inc. - 38.0 min
Southwest Airlines Co. - 46.0 min
Alaska Airlines Inc. - 56.0 min
Delta Air Lines Inc. - 57.0 min
Spirit Air Lines - 61.0 min Given the choice, I would recommend using these carriers.

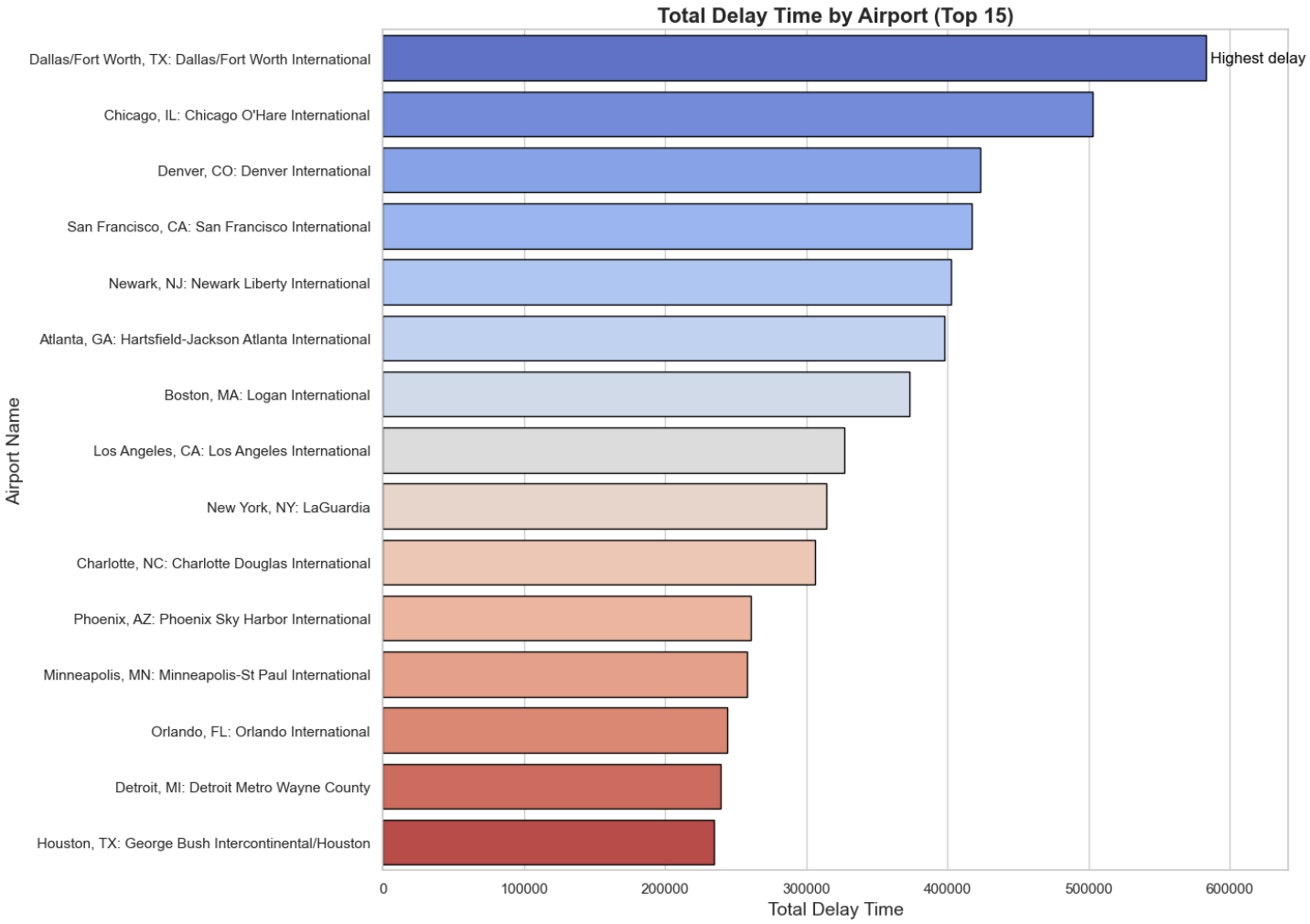
- Top 5 carriers by average flight delay duration (WORST):

ExpressJet Airlines LLC - 94.0 min
SkyWest Airlines Inc. - 87.0 min
JetBlue Airways - 83.0 min
Mesa Airlines Inc. - 83.0 min
Endeavor Air Inc. - 74.0 min

- If there is a need to minimize flight delays, Hawaiian Airlines Inc. is the best carrier in terms of both average delay and percentage of delayed flights. However, Southwest Airlines, Alaska Airlines, Delta Air Lines, Endeavor Air Inc are also good choices, as they have good performance in both categories.

6. Airport performance and comparison between them based on total delay time.

Python Query Output:

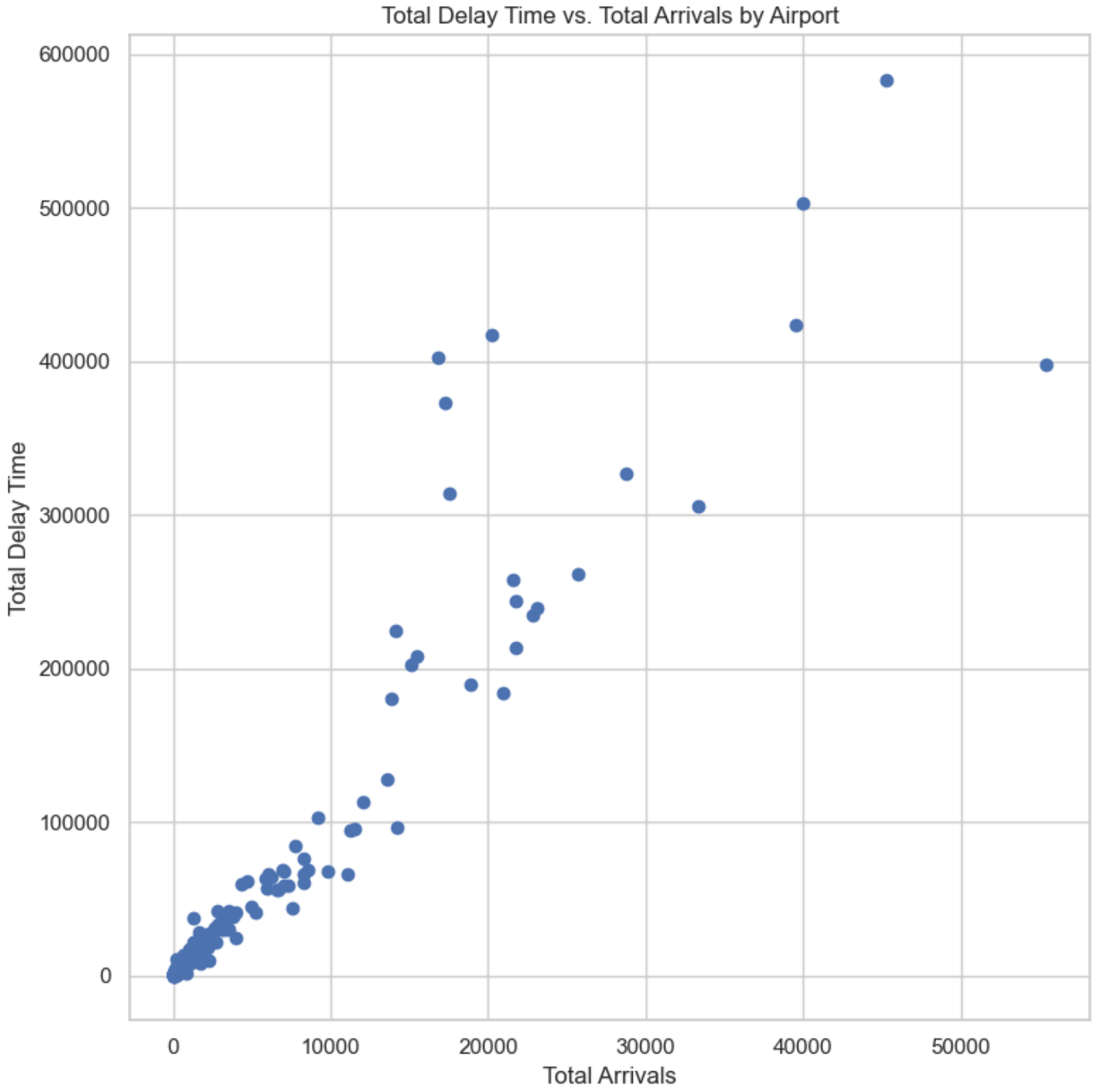


Insights:

- From the graph we can say that Dallas/Fort Worth is one of the busiest airports which has the highest delay time and Huston has the lowest and both are based on Texas.

7. Comparing total delay time and total arrival at the airports

Python query output:

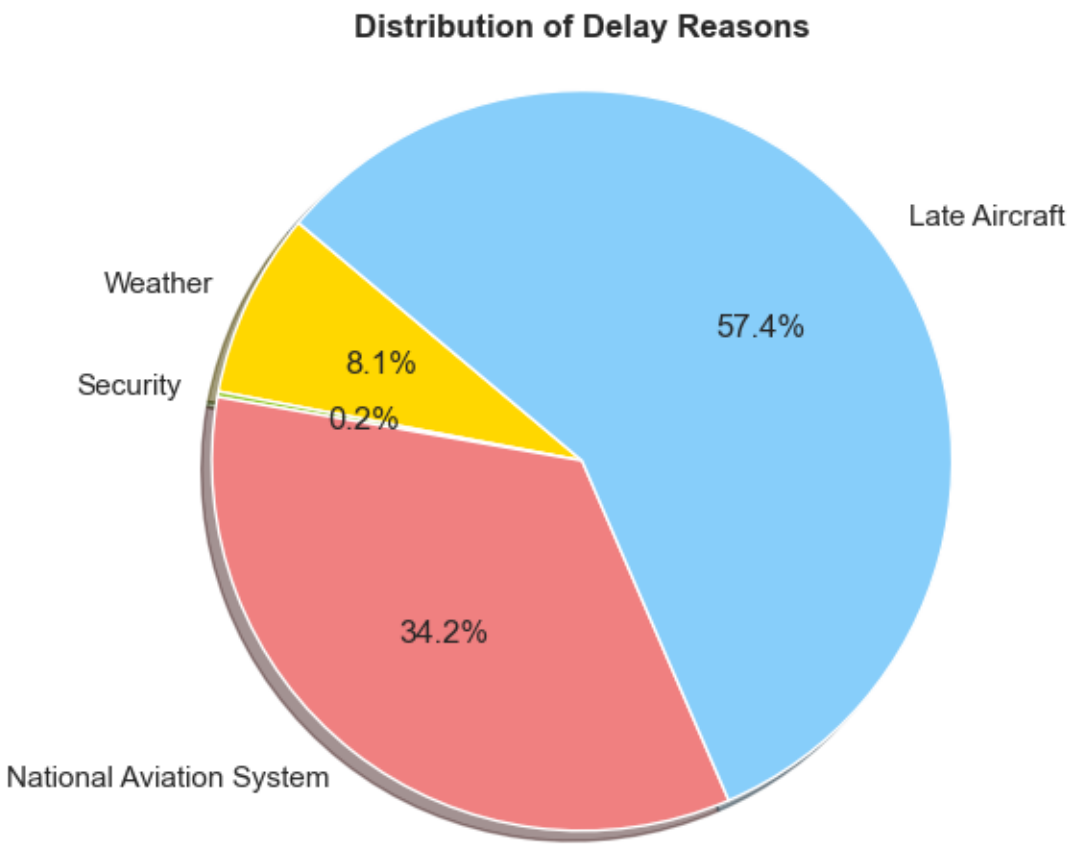


Insights:

- Base on the graph we can say that Total Delay Time and Total Arrivals are in positive correlation as Total Arrivals increases which impact increase in Total Delay Time as well.

8. Analysis based on Delay Reasons

Python Query Output:



Insights:

- Late Aircraft 4200591
  National Aviation System 2505844
  Weather 593688
  Security 18055

- More than half of all delays were caused by late arrival of flights (57.4% or 4,200,591 minutes)
- The National Aviation System is the second largest cause of delays at airports (34.2% or 2,505,844 minutes)
- Weather conditions cause less delay than I expected (8.1% or 593,688 minutes)
- Security issues are an extremely rare cause of flight delays (0.2% or 18,055 minutes)

9. Linear Regression Model

Python Query Output:

```
R-squared: 1.00
Mean Squared Error: 0.00
                                                 Coefficient
Number of flights arriving at airport           2.236145e-16
Number of flights more than 15 minutes late    -1.656891e-14
Number of flights delayed due to air carrier    1.000000e+00
Number of flights due to weather                1.000000e+00
Total time (minutes) of delay due to National A...  1.000000e+00
Total time (minutes) of delay as a result of a ...  1.000000e+00
Total time (minutes) of delay flights as a resu...  1.000000e+00
```

Insights:

- R-squared: 1.00

The R-squared value, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
An R-squared value of 1.00 indicates a perfect fit, meaning the model explains 100% of the variance in the target variable from the features. This is highly unusual in real-world scenarios and suggests that the model might be overfitting or that there's a direct calculation or linear relationship between the features and the target.

- Mean Squared Error (MSE): 0.00

MSE measures the average squared difference between the estimated values and the actual value. An MSE of 0.00 indicates that the model's predictions perfectly match the actual data, further supporting the notion of a potentially too good to be true model performance.

- Coefficients:

The coefficients describe the relationship between each feature and the target variable. They represent the expected change in the target variable for a one-unit change in the feature, holding all other features constant.
arr_flights coefficient is extremely close to 0, suggesting that the total number of arriving flights has no significant impact on the total delay time in the model's current formulation.

arr_del15 also has a coefficient close to 0, indicating it doesn't significantly affect the total delay time in the model's prediction.
carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay each have coefficients of 1, suggesting a one-to-one direct contribution to the total delay time. This implies that the total delay time (arr_delay) is a direct sum of these delay types, which explains the perfect model fit.

- The model's perfect fit and the coefficients indicate that arr_delay might be directly calculated from the sum of the individual delay types (carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay). This is further suggested by their coefficients being exactly 1.
- In practice, a perfect R-squared value and an MSE of 0 should prompt a review of the data and model. It could be that the model is simply "learning" a direct calculation present in the data rather than capturing underlying patterns that would generalize well to unseen data.
- If the goal is predictive modeling or understanding the impact of various factors on flight delays, you might want to reassess the included features. For instance, other variables not directly involved in calculating arr_delay could offer more insights into what influences delays.
- Consider the possibility of data leakage, where information about the target variable is inadvertently included in the features, leading to overly optimistic performance metrics. This is often addressed by ensuring the features used for prediction would be available at the time of prediction in a real-world scenario and are not inherently part of the target variable's calculation.

- Recommendations:
  Further Investigation: Since the total delay is directly calculated from the sum of the individual delays, future models might benefit from exploring other aspects of flight delays. Consider factors not included in the current dataset, such as time of day, day of the week, seasonality, specific airline efficiency metrics, and airport congestion levels.

- Model Complexity: Investigate models that can capture nonlinear relationships and interactions between variables. Given the complex nature of airline operations and delays, linear regression might be too simplistic to uncover deeper insights. Techniques like polynomial regression, decision trees, or even ensemble methods and neural networks could reveal more nuanced relationships.

- Operational Efficiency: The direct summing up of delays into total delay time underscores the importance of minimizing each type of delay to improve overall operational efficiency. Airlines and airports could focus on specific delay mitigation strategies, such as improving carrier scheduling, bolstering infrastructure to deal with adverse weather, and investing in technologies that reduce NAS delays.

- Preventive Measures: Given that each delay type contributes directly to total delay, implementing preventive measures and quick responses to initial delays can significantly reduce the cumulative effect. For example, better weather forecasting and more flexible scheduling can reduce weather and NAS delays, respectively.

- Data Exploration: Explore the data further for patterns that were not captured by the linear model. For example, examining the relationship between delays and specific airports or carriers might identify efficiency bottlenecks or best practices worth emulating.

- Avoiding Data Leakage: Ensure that future models avoid data leakage by excluding features that directly contribute to or are part of the target variable calculation. This will help develop models that are more predictive and generalize better to unseen data.

What If Analysis

Problem - How do different factors (, weather, national aviation system, late aircraft) contribute to airline delays, and what can be done to minimize delays and improve operational efficiency?

Scenario 1 – Improved Weather Forecasting

Insights-

1. Overall Reduction in Delay Times: The increase in accuracy from safe landing weather forecasting to 30% would save 17,107 minutes of delay time. Although it may seem like a paltry reduction on a very big scale, those minutes will multiply when more people are involved or the period getting longer; thus, these minutes represent real savings of time and possible costs.

Weather Delays	Baseline	10%	20%	30%
Total Delay Time	11145121	11085752	11026383	10967014
Total Delay Time Reduction (%)	0%	0.5%	1.06%	1.5%

2. Incremental Efficiency Gains: The data demonstrates the consistent movement to more efficiency stems from increasing accuracy of forecasting weather. 10% better reduces delay times by 0.5%, 20% more improved reduces by 1.06%, and 30% more improved brings about 1.5%. It seems like, even little improvements in accuracy of forecasting gives enormous overall reduction in the delays.

3. The Importance of Every Percentage Point: Specifically, since the times are in minutes, every percentage point in the delays should be vital, especially where time is the basis of efficiency, customer delight, and operational costs; transportation, logistics, and utilities, among many others, are the industries.

Scenario 2 - Late Aircraft Recovery Plan

Late Aircraft Delays	Baseline	10%	20%	30%
Total Delay Time	11145121	10725061	1305002	9884943
Total Delay Time Reduction (%)	0	3.7%	7.5%	11.3%

1. Substantial Reduction in Delays: The information expresses a large drop-off of the overall delay time when the effectivity of the plan of recovery of late aircraft expands. Starting from a baseline of 11,145,121 minutes, we have a reduction to 9,884,943 minutes with an improvement of 30%. It should be noted that this reduction signifies the effectiveness of the prevention measures that have been taken.

2. Progressive Improvement: The aggregated delays time compression percentages of 3.7%, 7.5% and 11.3% for 10%, 20%, and 30% upgrades convey a directly proportional impact of improving a late aircraft recovery plan. This implies that trying to smooth ice aging almost always pays off.

3. Greater Impact than Weather Forecasting: Whereas the weather forecasting scenario contribute to shorter delay times, the delays reduce considerably more with the aircraft recovery improvements toward the end of operation. It can be implied that late aircraft are more flexible and consequential on the overall delay from weather that is naturally unpredictable.

Scenario 3 - NAS Optimization

NAS Delays	Baseline	10%	20%	30%
Total Delay Time	11145121	10894536	10643952	10393367
Total Delay Time Reduction (%)	0	2.2%	4.49%	6.74%

Insights:

1. Consistent Delay Reduction: Moving towards an improved management in the NAS performance results in a gradual decrease in the total delay times from the baseline value of 11,145,121 minutes to 10,393,367 minutes with 30% optimization. This therefore shows that the function of NAS efficiency towards reduction of operational delays is direct.

2. Progressive Impact: The 2.2%, 4.49%, and 6.74% delay time reductions for 10%, 20%, and 30% improvements, respectively, show a definite, accordingly growing impact. The larger the share of operations handled by the NAS, the lower the total waiting time, which implies that the relationship between the efficiency of NACS and delay reduction is somewhat linear.

3. Critical Role of NAS in Delay Mitigation: However, if NAS inefficiencies caused most of those delays, moderate progress could save the passengers a considerable amount of time. This certainly points to the extent to which NAS operations have an impact on the productivity of air traffic management. Therefore, emphasis on such interventions will undoubtedly bring about noticeable improvements.

Overall,

1. Weather Forecasting Improvements: Improving the forecasting accuracy gives airports and airlines the advantage of making decisions proactively and therefore better manage any harmful weather conditions; passenger delays are minimized, and stress is reduced.
2. Late Aircraft Recovery Plan: Through shuffling an aircraft crew, going ahead with the plane maintenance, and perfect in-flight crew service, airlines have better possibility to bring back from the late planes, which, consequently, encouraging smoother and on-time operations and thus preventing extended delays in other flights.

3. NAS Optimization: The upgrade of air traffic control systems, adoption of advance airspace management systems and airport capacity optimization should be employed for delays reduction of the national airspace. It also exploits technology to increase efficiency and interfering between and among all the airspace management stakeholders.

Pivot Table

We have taken all the flights detail along with the sum of their total delay time in minutes and total number of delays.

Row Labels	Sum of arr_delay2	Count of arr_delay
Alaska Airlines Inc.	351035	144
Allegiant Air	240714	240
American Airlines Inc.	1185345	199
Delta Air Lines Inc.	1077429	255
Endeavor Air Inc.	431683	226
Envoy Air	505165	285
ExpressJet Airlines LLC	254148	106
Frontier Airlines Inc.	232696	184
Hawaiian Airlines Inc.	39243	37
JetBlue Airways	864790	120
Mesa Airlines Inc.	530595	219
PSA Airlines Inc.	479952	187
Republic Airline	541222	196
SkyWest Airlines Inc.	1804000	482
Southwest Airlines Co.	1396083	181
Spirit Air Lines	307232	97
United Air Lines Inc.	903789	193
Grand Total	**11145121**	**3351**

Insights-
There were the highest total delay time (up to 1,804,000 minutes) and also the highest number of delayed flights (it was 482) for SkyWest Airlines, Inc. Such a thing means not only that they have a great impact by itself as for their delay, but also that such delays of many flights cause significant turbulence to passengers.

On the contrary, Hawaiian Airlines Inc. owns an airline fleet with the least delay time and the lowest stream of delayed flights, which could be the evidence of the more streamlined operations or less flights overall. Airlines as American Airlines Inc. and Delta Air Lines Inc. have hold longer average delays but such delays per flights are less frequent than this case of SkyWest, indicating that more time is used for a single flight. This could possibly reveal an intense delay issue caused delays are handled or from the effect of more international flights of this kind that they operate.

HOW SOLUTION PROVIDE VALUE

The solution, leveraging advanced data analysis and predictive modeling, delivers substantial value by enhancing operational efficiency, improving passenger experience, and supporting strategic decision-making for airlines and airports:

Enhancing Operational Efficiency

- Predictive Insights: Advanced modeling techniques, such as Random Forests or Gradient Boosting, go beyond linear relationships, offering more accurate predictions of delays. This allows airlines and airports to anticipate and mitigate potential disruptions before they escalate, optimizing resource allocation (e.g., crew scheduling, gate assignments) and minimizing the domino effect of delays across the network.
- Focused Interventions: By pinpointing the most significant predictors of delays, such as carrier issues or airport-specific challenges, operators can target interventions more effectively. This targeted approach ensures that efforts and resources are directed where they can have the most impact, improving overall efficiency.

Improving Passenger Experience

- Reduced Delays: With more accurate predictions and targeted interventions, the frequency and duration of delays can be significantly reduced. This leads to a more reliable flight schedule, enhancing passenger satisfaction as travelers experience fewer disruptions and uncertainties.
- Proactive Communication: The ability to predict delays more accurately also enables airlines to improve communication with passengers, providing timely updates about potential delays and allowing passengers to adjust their plans accordingly. This transparency can greatly enhance passenger trust and satisfaction.

Strategic Decision-Making

- Data-Driven Strategies: The insights gained from the analysis empower airlines and airports to make informed strategic decisions. Whether it's adjusting flight schedules to avoid known delay triggers or investing in infrastructure improvements at certain airports, data-driven strategies can lead to more sustainable operational improvements.
- Real-Time Operational Adjustments: Implementing a real-time dashboard that leverages predictive modeling for delay forecasts allows operators to make dynamic, on-the-ground decisions. This agility in operations can be the difference between a minor delay that is contained and a major delay that impacts numerous flights.

Long-Term Benefits

- Competitive Advantage: Airlines and airports that can minimize delays through effective predictive modeling and operational strategies can distinguish themselves in the market. This competitive advantage can translate into increased market share as passengers prefer airlines with a reputation for reliability.

- Operational and Financial Improvements: Reducing delays not only improves the passenger experience but also leads to significant operational cost savings. Fewer delays mean less need for compensations, reduced fuel costs from idling aircraft, and better utilization of crew and airport facilities.

In essence, this comprehensive approach not only addresses immediate challenges related to flight delays but also lays the groundwork for a more efficient and passenger-friendly future in air travel.

CONCLUSION

The strategic application of advanced data analysis and predictive modeling to the airline's flight delay challenges has illuminated actionable pathways for enhancing operational efficiency and customer satisfaction. By pinpointing and addressing the root causes of delays, the airline is positioned to significantly reduce operational disruptions and costs. This approach not only promises to elevate the customer travel experience but also to secure a competitive advantage in the aviation market, leading to sustained growth and brand loyalty.

CONTRIBUTION

1. Selection of Data set, Preparing Project Introduction, forming operational analytics problems, Excel part analysis with making scenarios and performing pivot table analysis, making final recommendations and conclusion.
2. Creating CRISP-DM framework analysis of Airline Industry, Performing Data cleaning and processing with Python.
3. Performing Data analysis EDA, Visualizations, Linear Regression model and insight by performing Python codes.
4. Analysing the What If analysis and performing the scenario analysis by creating various scenarios and provides the insights and analysis.