Predicting Flight Cancellations Departing LaGuardia Airport

Messy Data and Machine Learning
May 11th, 2023
Kelly Xiao and Rongze Yan

Contents

*** Codes available at https://github.com/RZYan20/MDML_FlightCancellation**

I. Research Question and Literature Review

What can be worse than having your flight canceled at the last minute when you are heading to something important? According to the Bureau of Transportation Statistics (BTS), the number of flight cancellations has been steadily rising over the last four years, not including the COVID-ravaged year of 2020. Reuters reported an overall cancellation rate of 2.57% in 2022, which works out to 144,515 canceled flights. The number of canceled flights in just the first half of 2022 alone — about 128,000 — exceeded the number of cancellations seen in all of 2019. Flight cancellations can happen for a variety of reasons and can be unpredictable, especially during peak summer travel time and winter storms. BTS categorizes the reason for flight cancellations as four types: carrier, weather, national air system, and security. For example, Southwest Airline is always top ranked for cancellations (Smith, 2023; Fox, 2022). During 2022, the year with exceptionally high flight disruptions, New York's newly renovated LaGuardia (LGA) and Newark Liberty International in New Jersey experienced 7.7% and 7.6% cancellations, while the average American airport saw 2.6% of its flights canceled during the summer time (Song, 2022).

Our motivated goal for this project is to use historical flight information to predict whether a flight would be canceled or not for a specific airport, LGA, for its future flights. The model should be one that is able to be generalized into different flight routes and/or airports. Furthermore, we wish to understand whether certain airlines or airports are more prone to flight cancellations and why. What are the main factors that contribute to flight cancellations? Can machine learning models accurately predict flight cancellations? We will be assessing the effectiveness of different machine learning models in predicting flight cancellations and determine which model is best suited for this task. This report could contribute to developing recommendations for airlines or airports to identify what needs to change in regards to flight scheduling and operations, weather forecasting technologies, and security checks. Travelers could also utilize these models to decide whether they want to choose a certain flight based on their traveling information.

II. Dataset and Descriptive Statistics

Overview of Data

The original dataset was downloaded from the Bureau of Transportation Statistics, under the form of “On-Time: Reporting Carrier On-Time Performance (1987-present)”, along with a variable created by scraping the Skytrax website for airline reviews. The data was downloaded by months, and later appended into a full dataset containing information in 2017-2019 and 2021-2022, with a total number of 33,027,375 observations and 38 variables. The 2017, 2018 data will be used as train data to predict 2019 and flights, and additional predictions with 2021 (year of covid) and 2022 were also made to see if accuracy varies for later years. The full dataset was then reviewed, cleaned, and new variables were generated for further analysis and predictions, with 692900 observations and 23 variables. The final dataset only contains observations that depart from the LGA airport, and more justification for the choice will be explained in detail in the descriptive statistics section.

The process of cleaning includes feature engineering variables number 17 to 22, with more details explained in the descriptive statistics section, and variable number 23 is a variable composed of data scraped from the airline review website.

The final dataset is composed by the following variables:

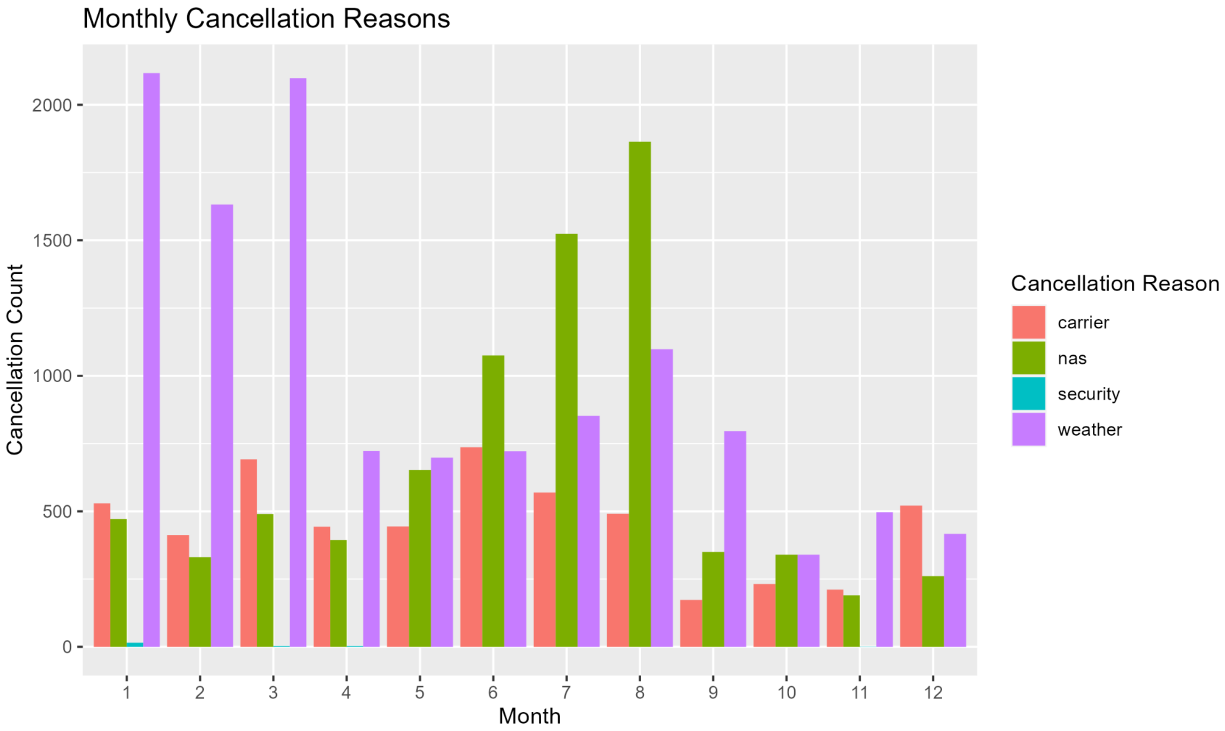
1. **year** 2017, 2018, 2019, 2021, 2022
2. **quarter** 1-4
3. **month** 1-12
4. **day_of_month** 1-31
5. **day_of_week** 1-7
6. **fl_date** flight date (yyyy/mm/dd)
7. **airline** airline carrier id ("AA" "B6" "DL" "EV" "F9" "NK" "OO" "UA" "VX" "WN" "9E" "MQ" "OH" "YV" "YX")
8. **fl_number** flight number
9. **origin** origin airport
10. **dest** destination airport
11. **dep_del15** (1 if delay>15 min; 0 otherwise)
12. **dep_time** (hhmm-hhmm)
13. **canceled** (outcome variable) whether a flight was canceled (0 or 1)
14. **cancellation_code** cancellation reasons (carrier, weather, national aviation system, security)
15. **air_time** flight time, in minutes
16. **distance** distance between origin/destination airports (miles)
17. **weekend** whether the day of the flight was a weekend (0 for Mon-Fri; 1 for Sat-Sun)

18. **holidays** whether the flight date was a holiday (0 or 1)
19. **delay_rate** percentage of delays (=delay if delay>15 min) each day and by airline
20. **cancellation_rate** percentage of cancellations each day and by airline
21. **dep_time_blk** scheduled departure time block (0 - 6)
22. **num_flight** number of flights each day by airline
23. **mean_score** review score of customer satisfaction by airline and month

Descriptive Statistics

According to the bar charts showing top cancellation numbers by different origin airports (Chart2.1) and ranking of top cancellation rates by airport, among airports with top 5 high cancellation numbers, LGA has the highest cancellation rate, supporting the existing literature's argument in the severity of cancellation issue at New York City's LGA airport. Therefore, this project will solely focus on the predictions for cancellation in flights departing from LGA airport, and the following analysis and summarizing statistics will also be based on observations that departed from the LGA airport.

Cancellation Reasons



Four levels were coded for the cancellation_code variable, indicating four different reasons to the flight being canceled, respectively being carrier, national aviation system, weather and security. From the monthly cancellation reasons chart above, we can see that weather and NAS (non-extreme weather conditions and airport traffic and operations) is unsurprisingly the most common reason for flight cancellations, followed by issues with the carrier. As we are unable to include precise historical weather data by flight dates, the model in this project will try to account for the pattern with time indicating variables.

Day and Time of Flight

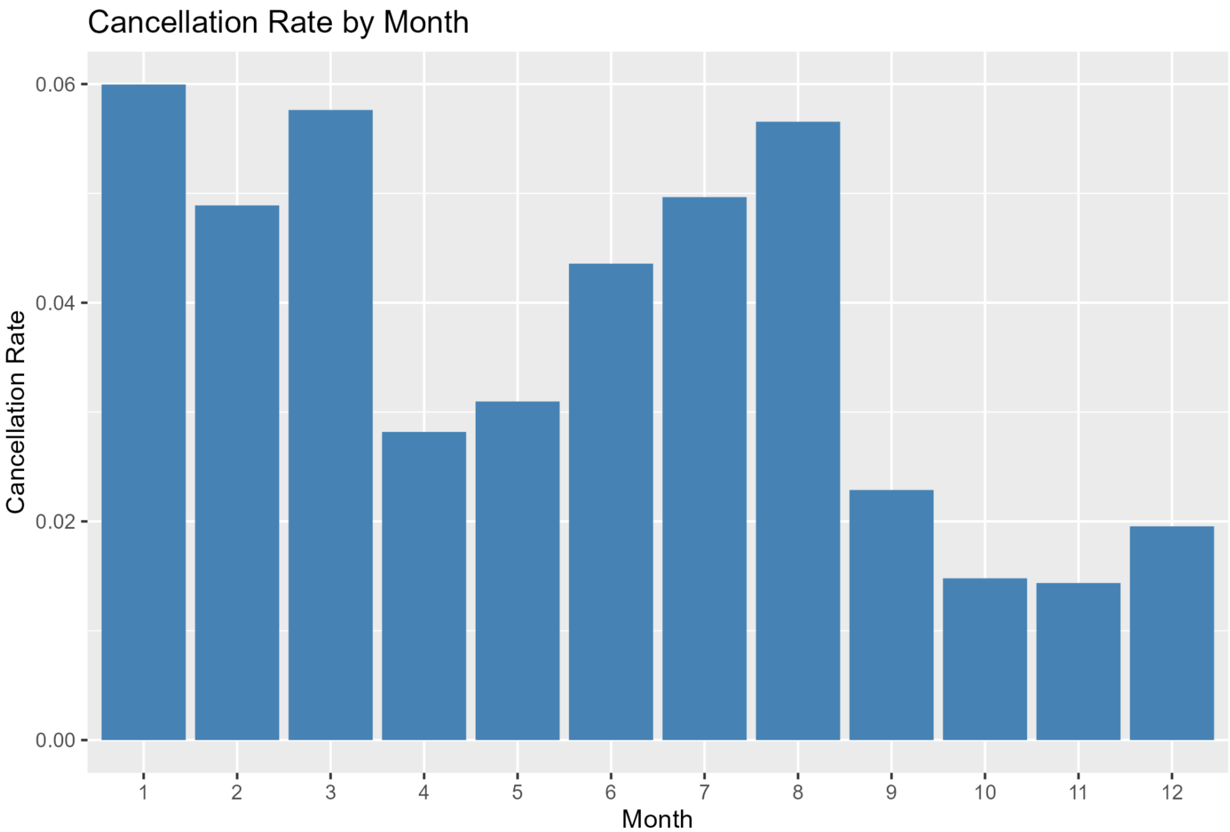
Day of Week

Looking at the cancellation rate by day_of_week (Chart 2.2), Monday through Friday show a higher cancellation rate in general, while cancellation rates on Saturdays and Sundays are lower. Therefore, a weekend indicator is included to indicate whether the flight date is on weekends(1) or weekdays(0), replacing the original day_of_week variable.

Day of Month

There are no clear patterns of cancellation rate by day of month (Chart 2.3), and we therefore did not include the variable in the model. A holiday variable was then generated to indicate whether a flight date is a holiday or not.

Month in Year



From this bar chart of cancellation rate by month we can see that there is roughly a seasonal pattern with cancellation rates by months, and therefore, we decided to include the variable quarter in the model.

Departure Time Block

We wanted to explore whether there is a correlation between the time period of the scheduled departure time of each flight, so we generated a dep_time_blk variable with values from 1-6, evenly dividing 24 hours in a day into 6 blocks.

Airlines

There is also a great variance of cancellation number and rate by airlines (Chart 2.4 & 2.5), Envoy Air (MQ), ExpressJet Airlines (EV), Endeavor Air (9E), and American Airlines (AA) are on the top of both lists. A satisfaction score of all airlines by month was scraped from the skytrax airline review website where individual customers post their review and scoring. Scores posted by individual customers were scraped from the website and a mean was calculated by month for each airline and then appended to the final dataset. This variable had a considerable amount of missing values in the originally scraped data, and we replaced it either with monthly mean or quarterly mean.

Delay Rate & Cancellation Rate Correlation

To investigate whether the compiling of delayed flights will result in flight cancellations, we wanted to see if delay rates by airlines each day are correlated with cancellations rates by airlines each day. The dep_rate variable was generated under this purpose, and it has a correlation coefficient of 101.0863 with cancellation rate each day by airlines, with a p-value of 0. This indicates that there is a strong positive relationship between delay rate and cancellation rate each day of individual airlines, where higher delay rates are associated with higher cancellation rates.

III. Machine Learning Methods

Model 1: Logistic Regression

Logistic regression is a classification algorithm that uses a logistic function to predict the probability of a binary outcome variable. In our case, we will fit the model to predict the probability of a flight being canceled based on 8 predictors. The continuous variables include number of flights per day per airline, customer average rating per airline, flight distance, and delay rate. The categorical variables are airlines, weekend, holidays, and departure time block.

Predictions

To predict the precision of our models, we used three different years of data as individual test dataset, respectively been year of 2019, 2021, 2022, to see how well the model predicts cancellation in a year right next to the training dataset (2017-2018), with effects of Covid-19, and during a period of travel recovery.

Formula

We have two formulas in our logistic regression model, the first one containing the scrapped variable mean score of airlines per month, the second without. We experimented with the model by taking out different variables one at the time, but only the change in whether including mean_score or not influenced the accuracy of the model prediction in a consistent manner throughout different years of test data prediction.

Formula 1

formula1 = cancelled ~ airline + weekend + holidays + distance + delay_rate + dep_time_blk + num_flight + mean_score

After fitting the model to the test dataset, we had the list of predicted probabilities of each flight being canceled, and the AUC scores we computed for the logistic model's precision are respectively 0.6541 in the 2019 test dataset, 0.5581 in 2021, and 0.5843 in 2022, suggesting that the model will have good discriminatory ability about 65% of the time in the year of 2019 and so on.

Formula 2

formula2 = cancelled ~ airline + weekend + holidays + distance + delay_rate + dep_time_blk + num_flight

Formula 2 removed the mean_score predictor, and after fitting the new model to the test dataset, we had the list of predicted probabilities of each flight being canceled, and the AUC scores we computed for the logistic model's precision are respectively 0.6786 in the 2019 test dataset, 0.5641 in 2021, and 0.5978 in 2022, suggesting that the model will have good discriminatory ability about 68% of the time in the year of 2019 and so on. The second formula exemplates a slightly better performance of the model on all years of the test datasets.

Model 2: Random Forest

We also used a Random Forest model to predict cancellation outcomes in this part. Random forest models are developed by decision trees, while at the same time it corrects for decision trees' overfitting to the training dataset. During the prediction phase, each tree in the forest independently makes a prediction, and the final prediction is determined by aggregating the individual tree predictions. This aggregation can be done by taking the majority vote in classification problems or averaging the predictions in regression problems.

In this project, we used the package "ranger" to fit the model on the training dataset (2017-2018), using 1,000 trees, and ensuring that both respect for unordered factors and probability are TRUE. For this model, we also applied both formulas and valiates them with the three test datasets.

Formula 1

After fitting the model with formula 1 and validating it through test datasets, we get the following AUC scores: test year 2019 being 0.6103, test year 2021 with a score of 0.6819, and test year 2022 with a score of 0.6106. This suggests that the model will have good discriminatory ability about 61% of the time in the year of 2019, 69% in 2021, 61% in 2022.

Formula 2

After removing mean_score, fitting the model with formula 2 and validating it through test datasets, we get the following AUC scores: test year 2019 being 0.6401, test year 2021 with a score of 0.6859, and test year 2022 with a score of 0.6091. This suggests that the model will have good discriminatory ability about 64% of the time in the year of 2019, 69% in 2021, 61% in 2022.

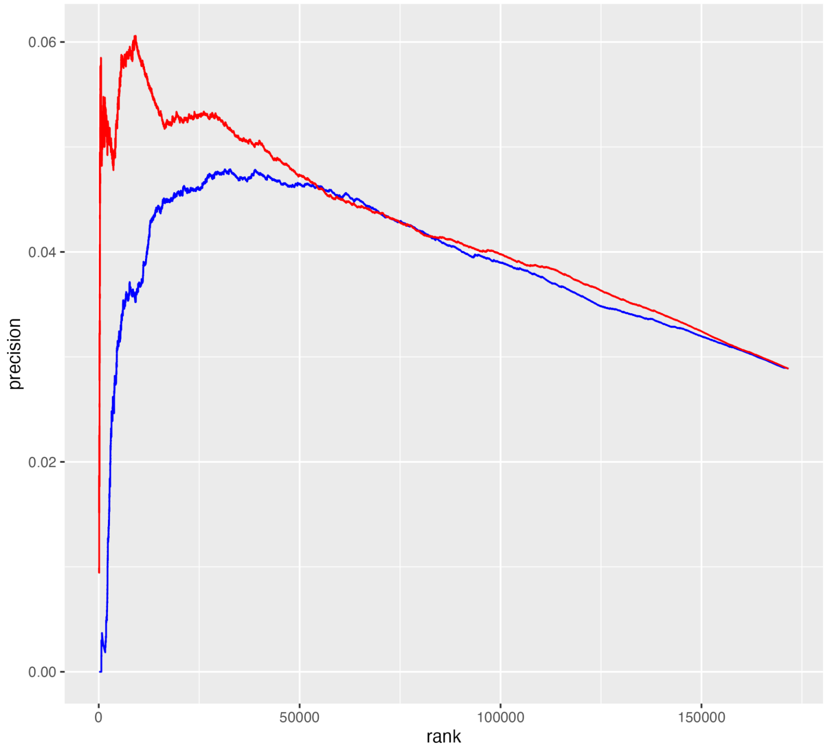
IV. Results, Limitations, and Implications

Results and Limitations

Model	**Logistic Regression**			**Random Forest**		
Year	2019	2021	2022	2019	2021	2022
AUC	0.6786	0.5641	0.5978	0.6401	0.6819	0.6091

Based on the reported AUC of the two sets of predictors of our choice, the model without adding the customer review had a better result. Our prediction for why this happened is because of the data quality and sample size. For example, while American Airlines has more than 4000 reviews, SkyWest has only two reviews for the past 5 years. To handle these missing values for certain airlines, we replaced it with the average score across quarters or across years. Our initial intention to include the customer review score is to differentiate between different airlines , and we were also interested in exploring the association between customer's experience at the flight and how well an airline generally performs on its on time performance. However, the result shows that it might not be the most relevant predictor for flight cancellations. Therefore, we decide to leave out the predictor for our final model presentation.

Based on the AUC scores of the logistic regression and random forest models for predicting flight cancellations, we can draw a few conclusions. First, we can see that Random Forest provides a slightly better result for year 2022 but much better for year 2021. On the other hand, Logistic Regression outperformed in 2019. Both models performed better in 2019 compared to 2022. This may be because the models were trained on data from previous years, which may not have fully captured the unique challenges and trends of air travel developed during the pandemic. Additionally, factors such as changing regulations and customer behaviors during the pandemic may have made it more difficult to accurately predict flight cancellations.



There could be other reasons why the random forest model performed better than the logistic regression model in 2021. It is possible that the travel patterns during the pandemic year were different and the random forest model was better able to capture the non-linear relationships between the predictor variables and the response variable. Random forest models are known to be effective in handling non-linear relationships and interactions between variables.

Regarding the performance of both models being better in 2019 compared to 2022, it could be due to the fact that the models were trained on data up until 2019 and may not have been able to capture any changes or trends that occurred after that year. Additionally, travel patterns and cancellation rates may have changed over time, which could also contribute to the decrease in performance in 2022. Overall, both models showed some predictive power but with very low AUC scores, and this is largely due to lack of data on important factors that can impact flight disruptions. One of the main limitations of this project is the absence of certain crucial data such as weather conditions, airline operations data, and staffing information. The inclusion of these factors would likely improve the performance of the model and provide a more accurate prediction of flight cancellations. However, these data are more difficult to extract from publicly available sources and therefore we were not able to include them for the purpose of this project. The random forest model may be the better model to build upon with more comprehensive data for predicting flight cancellations during times of uncertainty and rapidly changing trends. However, further analysis and model refinement may be necessary to fully capture the complex and evolving nature of air traveling.

Implications of Our Model

By understanding and leveraging the predictions of flight cancellations, our project can contribute to improving operational efficiency, customer experience, cost management, risk mitigation, and overall performance in the aviation industry.

Operational Efficiency

Accurate predictions of flight cancellations can help airlines and airports optimize their operations. By anticipating cancellations, airlines can proactively manage resources, such as crew scheduling, aircraft allocation, and gate assignments, to minimize disruptions and ensure efficient operations.

Customer Experience

Predicting flight cancellations can assist in providing timely information to passengers and mitigating inconvenience. Airlines can proactively notify affected passengers, provide alternative travel arrangements, and offer personalized customer support, thereby improving the overall customer experience.

Cost Reduction and Resource Allocation

Effective prediction of flight cancellations can lead to cost savings for airlines. By anticipating cancellations, airlines can optimize their resource allocation, such as fuel consumption, crew utilization, and maintenance scheduling, which can result in reduced operational costs. Reliable cancellation predictions can also support long-term planning and resource allocation for airlines and airports. This includes capacity planning, fleet management, infrastructure investments, and optimizing staffing levels to align with expected cancellation rates.

Risk Management

Accurate cancellation predictions enable airlines to manage risks associated with disrupted flights. This includes proactive measures such as adjusting flight schedules, rerouting passengers, and mitigating potential impacts on connecting flights. Effective risk management can help maintain the overall reliability and reputation of the airline. By anticipating cancellations, airlines can make necessary adjustments to ensure compliance with passenger rights, compensation policies, and regulatory reporting requirements.

Appendix

Chart 2.1

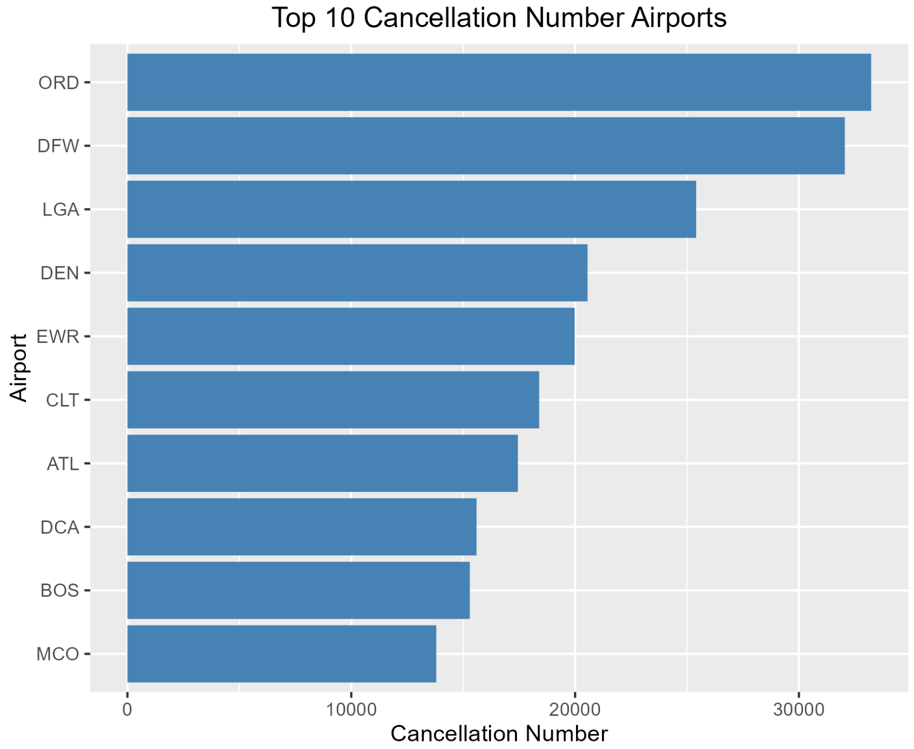


Chart 2.2

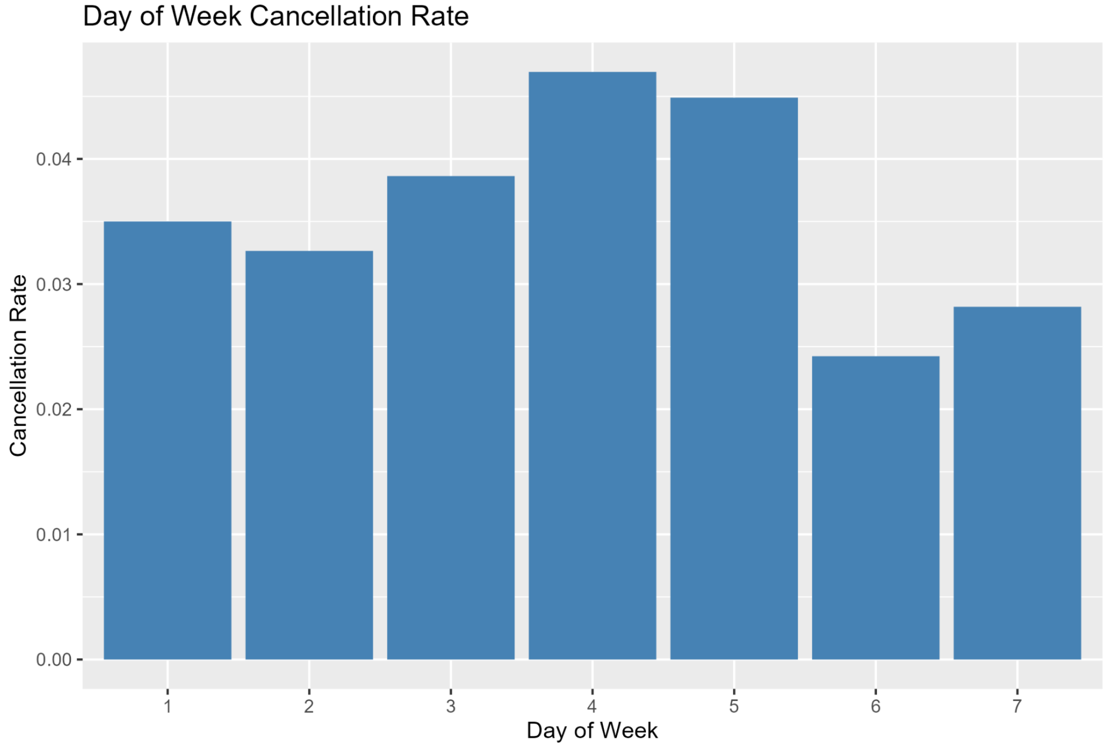


Chart 2.3

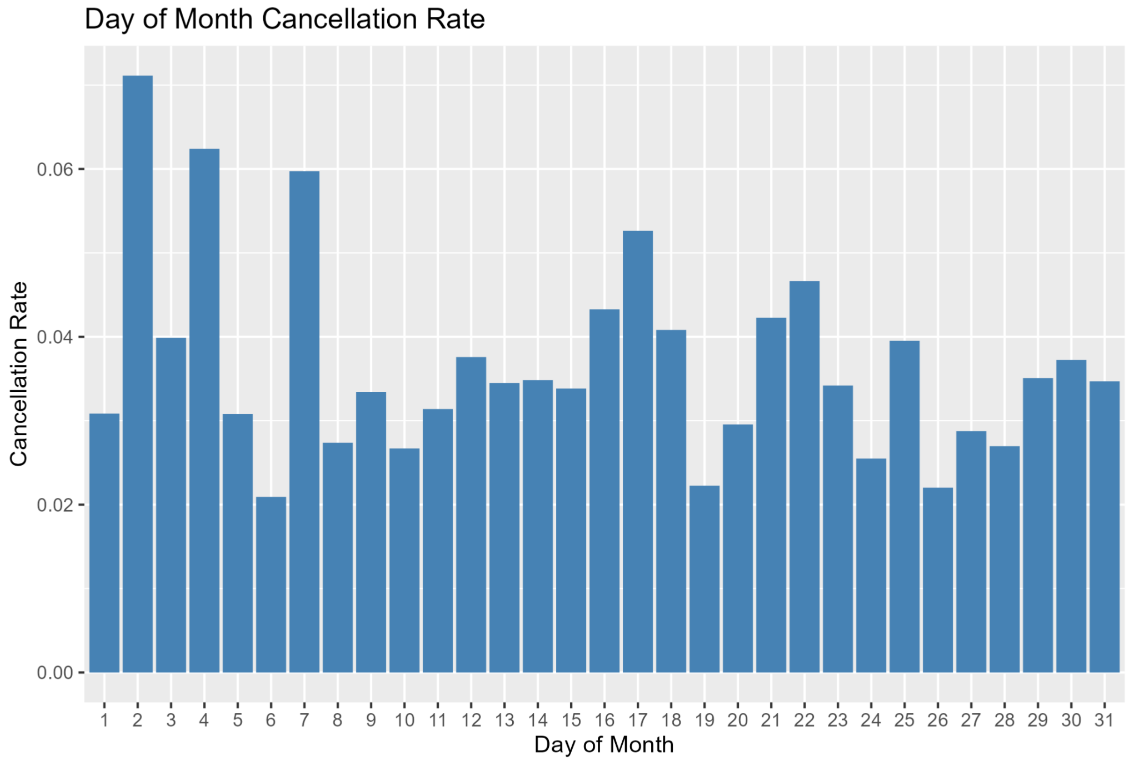


Chart 2.4

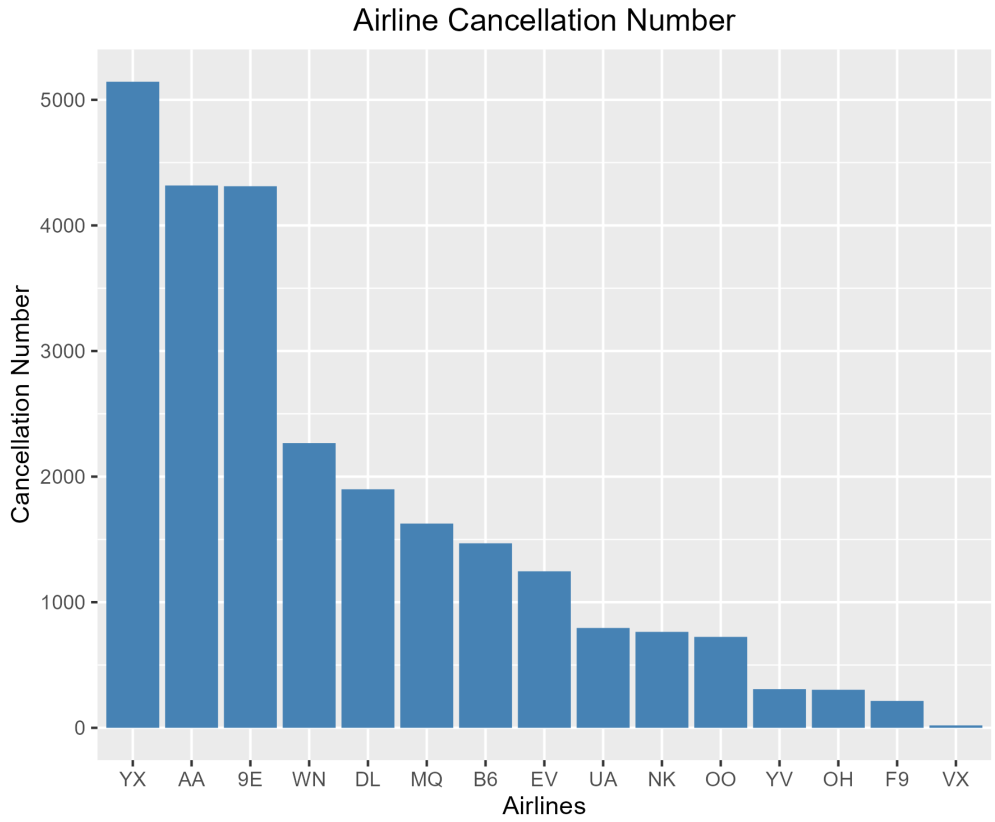
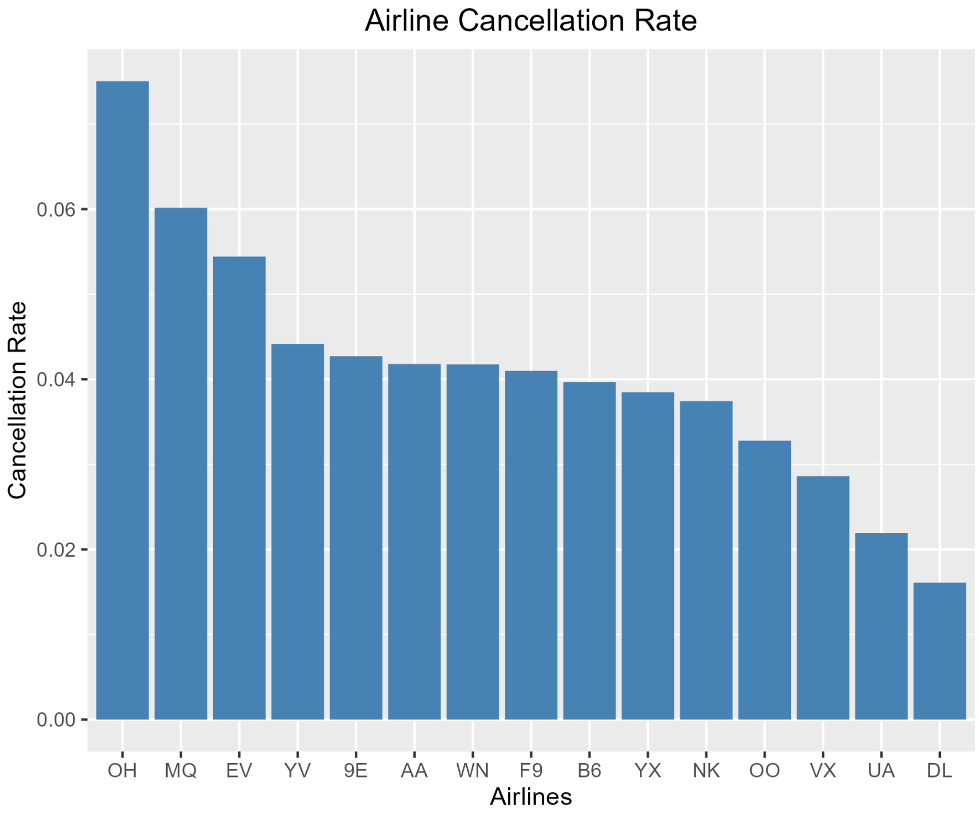


Chart 2.5


Airline Cancellation Rate
Cancellation Rate
0.06
0.04
0.02
0.00
OH
MQ
EV
YV
9E
AA
WN
F9
B6
YX
NK
OO
VX
UA
DL
Airlines