MGT 6203 Group #73 Final Paper

Overview of Project:

Team #: 73

Team Members:

1. **Archana Premkumar; GTID: 903031628:** Archana Premkumar is a Sales Engineer at Ribbon Health. She graduated from GT with a BS in Biomedical Engineering.
2. **Bryant Phan, GTID: 903662148:** Bryant Phan is a Sr. Manager in Global Health Economics & Outcomes Research (HEOR) at Pfizer, Inc. He graduated from Stanford University with a BA in Human Biology and Columbia University with an MPH in Epidemiology.
3. **Pranay Shah; GTID: 903755704:** Pranay Shah is a Data Scientist at JP Morgan Chase. He graduated with a BS in Finance and Business Analytics from Rutgers University.
4. **Rishi Kanoongo; GTID: 903655876:** Rishi Kanoongo is a Sr. Data Modeler in Data Analytics consulting at KPMG LLP. He graduated from Rutgers with a BS in Business Analytics and Information Technology.
5. **Sarah Goldrup; GTID: 903738172:** Sarah Goldrup is a lead SDET at General Motors. She graduated with BS in mathematics from St.Edward's University

OBJECTIVE/PROBLEM

Project Title: Predicting Flight Delays and Cancellations in the United States Using Weather Data

Background:

Extreme weather conditions, such as severe rain, fog or snow pose a threat to flight operations and often result in flight cancellations. However, according to the Bureau of Transportation Services, non-extreme weather conditions account for 45.8% of delays in 2020[1]. There are over 45,000 flights carrying approximately 2.9 million passengers per day[2]. A significant delay or cancellation in flights in one area can cause a downstream effect that results in over 30 billion dollars of costs[3] from disrupting flight schedules and triggering additional subsequent delays in other areas.

Predicting the likelihood of flight delays or cancellations can help inform airline operations (e.g., scheduling additional employees to assist customers during anticipated delays) to minimize further revenue loss. Furthermore, if airline operators can predict when delays and cancellations are most likely to happen, they can schedule a buffer time between flights to minimize the disruption. If we can determine a rudimentary solution then that may be scaled to provide improved planning for customers, airports and airlines alike.

Problem Statement:

The primary objective of this research is to predict flight cancellations or delays based on weather conditions using historical weather and airline cancellation data for three major airports - Chicago O'Hare International Airport (ORD), Los Angeles International Airport (LAX), New York City John F. Kennedy Airport (JFK). The scope of the research will be narrowed down to address four primary questions:

1. What weather conditions cause flights to be delayed or canceled?
2. Do weather conditions significantly impact airline cancellations?
3. Do airlines delay/cancel flights based on the severity of the weather? If so, are they more likely to delay/cancel when severity is lower or higher?
4. Which metropolitan area is more prone to delays and cancellations?

Hypothesis:

Our initial hypothesis is that there will be a lagging correlation between weather events and flight delays/cancellations. In other words, severe weather events will precede a flight delay/cancellation. We speculate that there may be a causal relationship between weather events and flight delays/cancellations. We also hypothesize that higher severity, greater precipitation amounts, and precipitation weather events (snow, rain, etc.) will be correlated with longer delays and an increased chance of cancellations.

Overview of Data

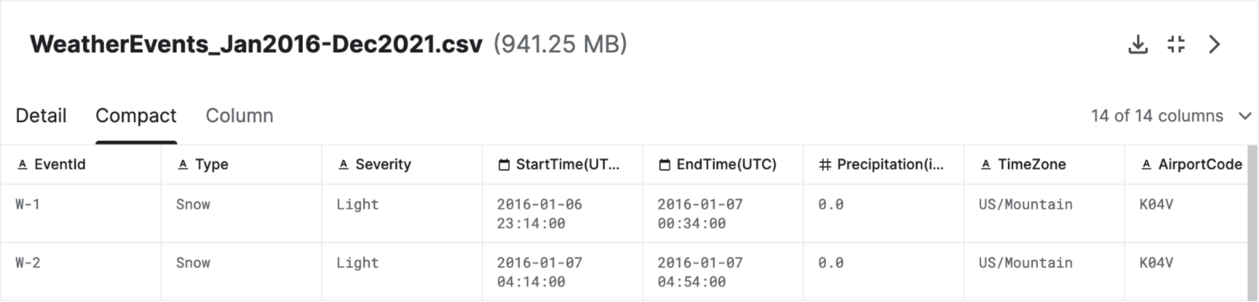
Data Background:

The analysis will utilize two different data feeds, both of which are sourced from Kaggle. The first is US Weather Events (2016 - 2021), a csv file which contains 7.5 million weather events collected from 2,071 airport-based weather stations in the United States (**Table 1, Figure 1**). The second source is Airline Delay & Cancellation Data (2009-2018), a csv file which contains data on approximately 15 million flights per year (**Table 2, Figure 2**).

Data Overview:

Variable Name	**Description**
Type	Defines the type of weather event. This categorical variable can take one of the following string values: ● Cold: The case of having extremely low temperature, with temperature below -23.7 degrees of Celsius. ● Fog: The case where there is low visibility condition as a result of fog or haze. ● Hail: The case of having solid precipitation including ice pellets and hail. ● Rain: The case of having rain, ranging from light to heavy. ● Snow: The case of having snow, ranging from light to heavy. ● Storm: The extremely windy condition, where the wind speed is at least 60 km/h. ● Other Precipitation: Any other type of precipitation which cannot be assigned to previously described event types.
Severity	Severity of the weather event
Start Time	Start time of the weather event in UTC time zone
End Time	End time of the weather event in UTC time zone
Precipitation	Total precipitation in inches, if applicable
Airport Code	The airport station the weather event was reported at

Table 1 *- Relevant variables from US Weather Events (2016 - 2021)*



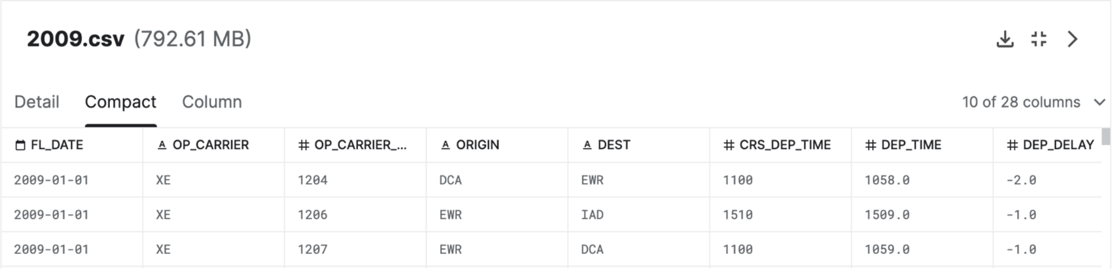
WeatherEvents_Jan2016-Dec2021.csv (941.25 MB)

EventId	Type	Severity	StartTime(UT...	EndTime(UTC)	Precipitation(i...	TimeZone	AirportCode
W-1	Snow	Light	2016-01-06 23:14:00	2016-01-07 00:34:00	0.0	US/Mountain	K04V
W-2	Snow	Light	2016-01-07 04:14:00	2016-01-07 04:54:00	0.0	US/Mountain	K04V

Figure 1 - Screenshot of Weather Events Data from Kaggle

Variable Name	Description
FL_DATE	Date of flight in yyyy/mm/dd format
ORIGIN	Departure airport code
CRS_DEP_TIME	Planned departure time in HHMM format
DEP_DELAY	Total delay on departure in minutes
WEATHER_DELAY	Delay in minutes caused by weather
CANCELED	Binary variable for whether the flight was canceled
CANCELLATION_CODE	Code that classifies the reason for cancellation. “B” means the flight was canceled due to weather

Table 2 *- Relevant variables from US Weather Events (2016 - 2021)*



2009.csv (792.61 MB)

FL_DATE	OP_CARRIER	OP_CARRIER_...	ORIGIN	DEST	CRS_DEP_TIME	DEP_TIME	DEP_DELAY
2009-01-01	XE	1204	DCA	EWR	1100	1058.0	-2.0
2009-01-01	XE	1206	EWR	IAD	1510	1509.0	-1.0
2009-01-01	XE	1207	EWR	DCA	1100	1059.0	-1.0

Figure 2 *- Screenshot of Flights and Relevant Variables from 2009 Dataset*

Data Cleaning:

Data cleaning is an integral part of the entire modeling process as it identifies errors and duplicates, but also only facilitates decision making by producing quality, standardized data and models. All data cleaning and analyses were conducted using R.

Following the initial steps of reading in the data and setting the working directory, the raw weather data was filtered to the period between January 2016 and December 2018. This dataset was then further filtered to exclusively include airports of interest, including Los Angeles, New York, and Chicago, which were identified using their respective airport codes. It was necessary to mutate the airport codes to ensure compatibility with the airline delay data, which contained variations such as KJFK and JFK. This was achieved by identifying and resolving the differences between IATA and ICAO codes. To ensure optimal standardization, the date was transformed into a POSIX object, facilitating the accurate matching of records in subsequent analyses. Only relevant columns, such as city, county, state, start time, end time, zip code, and airport code, were retained for analysis.

Throughout the data cleaning process, several checkpoints such as ensuring the sample size for flights per year were sufficient in our final dataset for a thorough analysis and all of the variables maintained integrity by matching the appropriate column names across both the raw weather dataset and the airport/flights dataset. The consistent checkpoints provided confidence that we could bind the datasets by appending or unioning, depending on the programming language used.

A CRS_DEP_TIME variable was created in the raw airline data using regular expressions (REGEX) by adding a in the timestamp. Since the original timestamp was in “hhmm” format, it needed to be converted to a “hh:mm” format for analysis. Furthermore, not all of the timestamps are consistently standardized. For example, sometimes hours less than 10 did not have a leading 0, which was rectified in this step of the cleaning process as well. The data and time fields were concatenated and the resulting date-time-stamp was converted into a POSIX object with the time zone of the respective airport. The date-time-stamps were subsequently transformed into UTC time to achieve standardization across airports and datasets. The resulting dataset, named "raw_airline_combined", contained 1.8 million rows at this stage of the data cleaning process. In order to filter out irrelevant data, only records with weather delays of at least one minute or canceled flights were retained. To achieve this, NA values in the "total" and "weather delay" columns were converted to numeric 0s, and the dataset was filtered using an inclusive OR condition. A subset of columns from the filtered dataset was then selected to enable the subsequent joining of the data with the weather data.

The row count was reduced to 51.9k in the previous step, providing a manageable yet sufficient sample size for robust analysis. An inner join of the cleaned weather and airline data resulted in a combined data frame of 182k rows, despite the weather data consisting of only 4k rows. This discrepancy can be explained by fanning, where one date record can correspond to multiple airline records. Approximately 50% of the combined records had a cancellation flag. The data was then joined to the airport code and airline flight date, with the additional condition of including only those airline records occurring after the weather event since it did not make logical sense for a flight to be delayed or canceled prior to a weather event. The resulting dataset was saved to a csv file for further exploratory data analysis. The training data consisted of rows from 2016 and 2017, while the testing data was from 2018, with sample sizes of 89.5k and 93.2k respectively.

Prior to proceeding to the next section, it is important to acknowledge potential data biases that could affect our analysis. It is important to note that using training and testing models based on time can be risky as factors and circumstances may change over time. This could result in a model that is not as relevant as it could be, as the effects present in a past year may be resolved in a future year. This approach was chosen due to the nature of airline and meteorology modeling techniques, which is constantly evolving over time. The objective was to test whether a model created in 2017 could improve airline delays or cancellations in 2018. In addition to biases and limitations in predictive models using time components, joining disparate data sources poses a risk, especially when relying on a single data source for prediction. Although there are precedents and use cases for this in the real world, it is not without risk. Lastly, human error is always a possibility, from faulty labeling during data ingestion to slipshod logic in the data analytics phase. However, we used the metadata and schema provided, and with additional resources, we could have refined the logic further. The next section will delve into the exploratory data analysis resulting from the data cleaning process.

Key Variables:

Dataset	**Independent Variables (Weather Dataset)**	**Dependent Variable (Airline Dataset)**	**Notes**
Weather	Date, airport code, type of weather event, severity of event, geographic location, precipitation amount	Arrival delay (continuous), flight cancellation (binary)	Interaction variables may be considered (e.g. type of event * severity). Categorical values/flags may be created from continuous values as well.

Exploratory Data Analysis:

The merged data is composed of 182,682 total observations and sixteen attribute columns (Event ID, Type, Severity, Precipitation, Timezone, LocationLat, LocationLng, Airport Code New, New Date, Use Date, Canceled, Weather Delay, New Time, Final Time, PosixTime and UTCTime. In this dataset, two of our sixteen attribute columns are response variables - Canceled and Weather Delay. Canceled is a binary categorical response variable; it contains a 1 if the flight was canceled and a 0 if the flight was not canceled. Weather Delay is the number of minutes a flight was delayed.

As the dataset size was adequate, the dataset was split into training and testing sets with approximately 50% of the data being assigned to each part. The reason for this was the temporal aspect of the data, where the training group was developed by splitting the full dataset into only departures in 2016 and 2017, while departures in 2018 constituted the testing group. Although we originally planned for a 67/33 split, the actual split surprisingly turned out to be 50/50 due to a larger number of observations in 2018. We understand that this might be considered risky, but for real-world predictive models, it is necessary to apply past data to future scenarios. The training group contained 93,150 observations, while the testing group contained 89,532 observations. To ensure the inclusion of the entire datasets, ceiling functions were used.

What weather conditions cause flights to be delayed or canceled?

To better visualize the weather conditions that caused flights to be delayed or canceled, two treemaps were created - one for weather conditions that caused delays and the other for weather conditions that caused cancellations. **Figure 3** below shows the first treemap, which highlights the weather conditions that caused the highest average delays. Interestingly, precipitation was the top condition that caused delays, followed by rain, fog, hail, cold, storm, and snow.

From the raw data, it can be seen that the only category that LA leads both NYC and Chicago in is precipitation which doesn't mean that it gets more precipitation but that they may be ill equipped to handle it which causes weather delays. Chicago leads NYC in Rain, Cold, and Storm while NYC leads Chicago in Fog, Hail, and Snow.

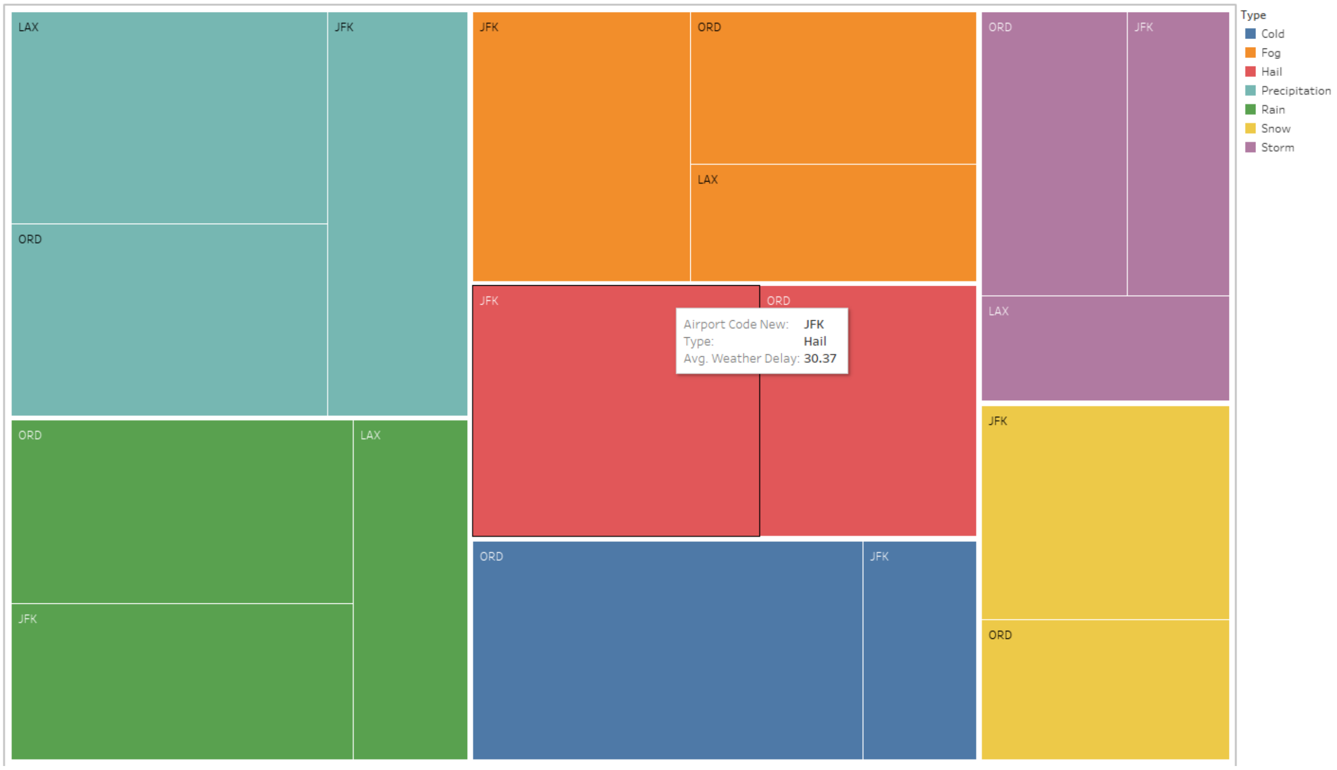


***Figure 3** - Treemap of Weather Conditions Causing Flight Delays in JFK, ORD, and LAX*

The second treemap (**Figure 4**), seen below, outlines weather conditions that caused cancellations:

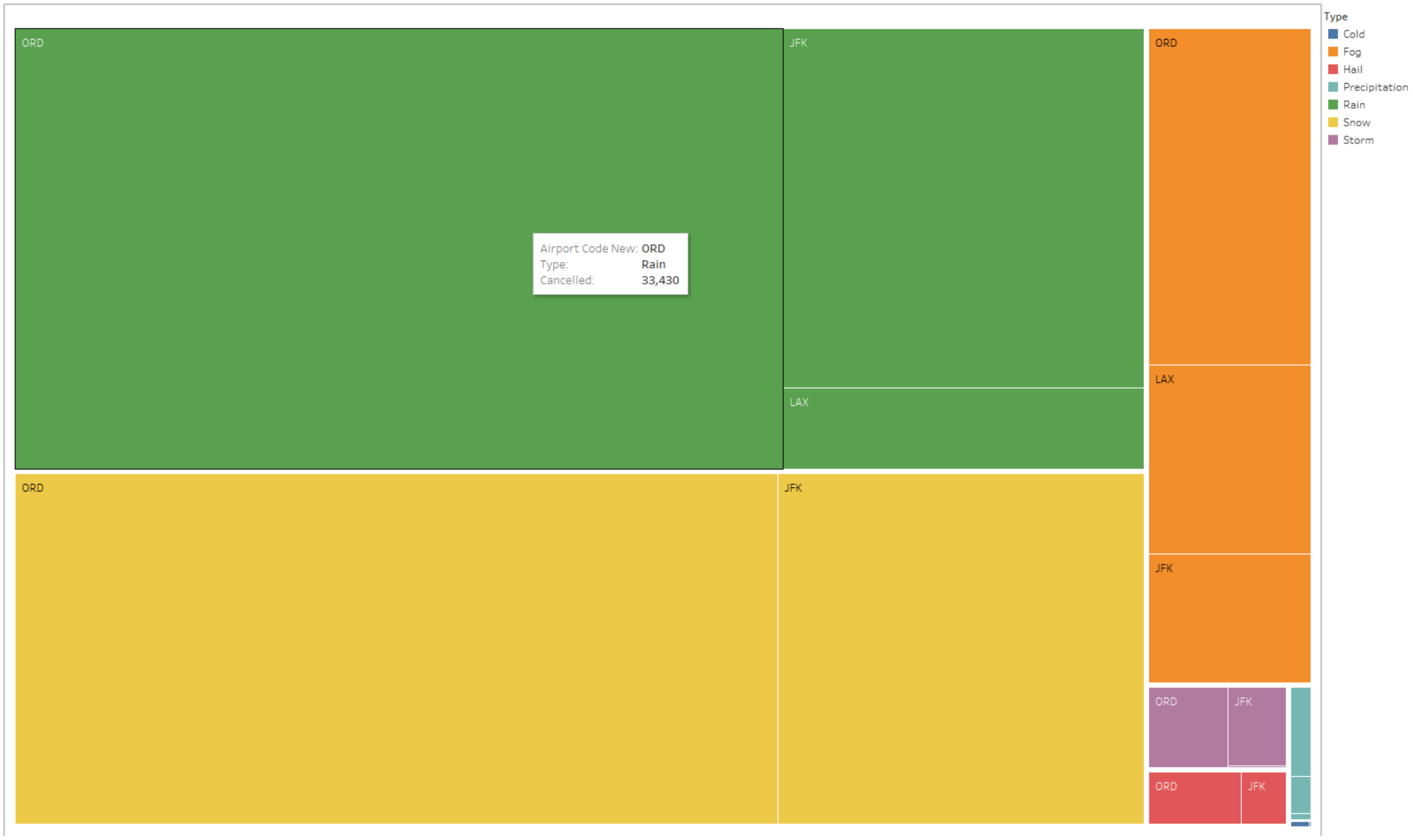


Figure 4 *- Treemap of Weather Conditions Causing Flight Cancellations in JFK, ORD, and LAX*

Based on the treemap visualization, it is evident that the major causes of flight cancellations were rain and snow while fog was the third cause. Moreover, analysis of the raw data revealed that Chicago (ORD) and New York (JFK) were most affected by rain and fog. On the other hand, snow mostly occurred in New York (JFK) followed by Chicago (ORD). Our initial hypothesis suggested that precipitation events such as rain, snow, etc. would cause longer delays and an increased number of cancellations. However, from the treemaps created, it appears that fog was the primary cause of delays, and snow did not rank among the top three conditions. Nevertheless, rain and snow were the top two reasons for cancellations. Based on this initial data analysis we can conclude that for airports located in New York (JFK), Chicago (ORD), and Los Angeles (LAX), delays are usually caused by fog, storms, and rain while cancellations are typically caused by rain and snow (excluding LAX for snow).

Do weather conditions significantly impact airline cancellations?

Approximately 50% of our records or 100,000 rows have cancellations in them. This is not a small sample and we can reasonably tell already that weather certainly has some correlation with cancellations even if it doesn't have a causal effect. To identify specific attributes of causality and see which type of weather event or the severity of these events' impacts on cancellations and delays, we will have to use handy statistical models for it. We will discuss in our results section below.

<u>*Do airlines delay/cancel flights based on the severity of the weather? If so, are they more likely to delay/cancel when severity is lower or higher?*</u>

As stated in the hypothesis, common sense would lead to the assumption that the higher the severity of weather, the greater the likelihood of a delay or cancellation. However, does the data reflect this belief? To investigate further, two bar charts were created to visualize the relationship between weather severity and delays or cancellations. **Figure** 5 visualizes weather severity that caused total delays.

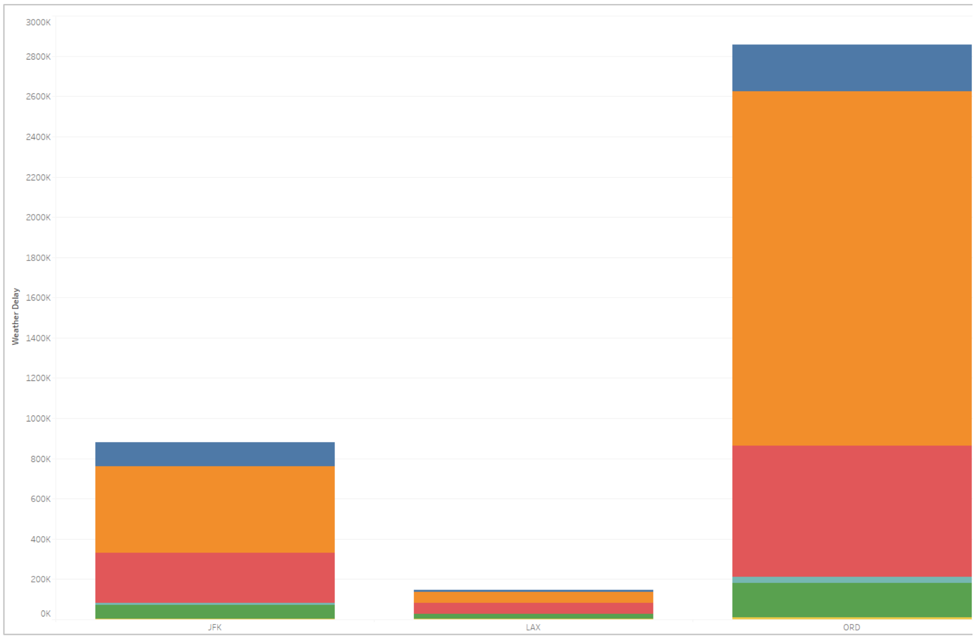


Figure 5 *- Barchart of Total Delays Caused By Weather Severity at JFK, LAX, and ORD*

The bar chart in **Figure 5** provides a few insights. Firstly, the majority of the delays occurred at the ORD airport in Chicago. Secondly, the delays were mostly caused by light weather severity, which is unexpected as we anticipated that more severe weather would lead to more delays. After light weather conditions, moderate weather was the next most common cause of delays. This confirms that the severity of weather is correlated with delays and lighter/moderate weather conditions are more likely to cause delays than heavy or severe weather.

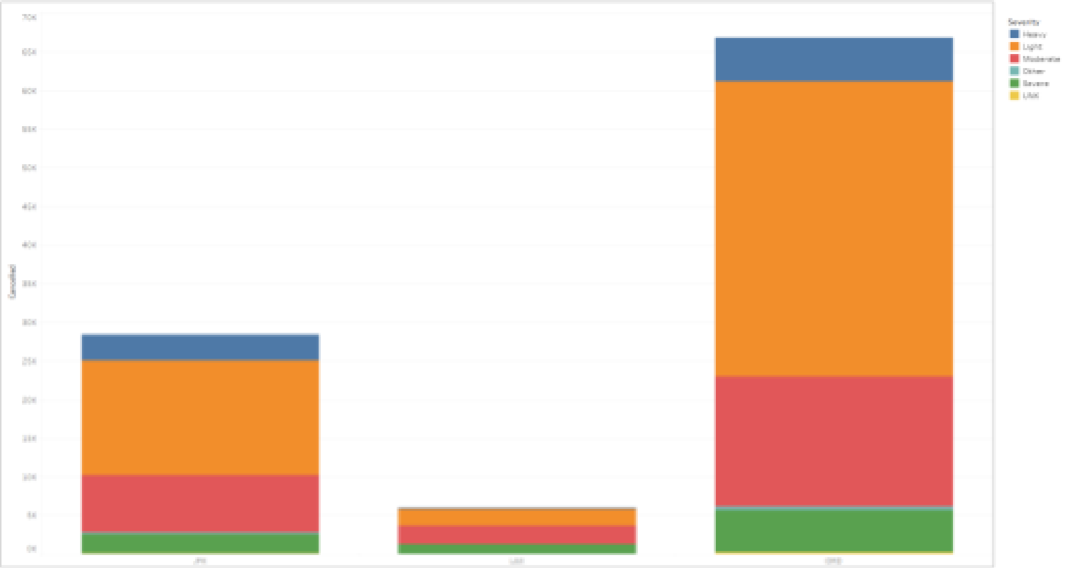


Figure 6 *- Barchart of Total Cancellations Caused By Weather Severity at JFK, LAX, and ORD*

The next bar chart we will take a look at is weather severity that caused cancellations (**Figure 6**). We are also considering the fact that light and moderate severity weather events occur more frequently than stormy severe ones. This may lead to sampling bias in our analysis and we're being cognizant of that as we approach this problem. If that's the case then light and moderate may seem to account for more delays due to their sheer ubiquity.

Based on the bar chart on the top we can come to the following conclusions: The majority of canceled flights due to weather severity were in Chicago (ORD), followed by New York (JFK) and Los Angeles (LAX). Similarly to delays, light weather conditions were responsible for the most cancellations, followed by moderate conditions. This finding is surprising as it is logical to expect heavy/severe weather conditions to cause more cancellations rather than lighter/moderate conditions. Delaying flights due to lighter/moderate conditions is more understandable and perhaps could be explained as a precautionary measure, but canceling flights under such conditions can be frustrating for consumers. There were a few possible explanations for light and moderate weather conditions causing more cancellations than severe conditions, with one possible explanation being that airlines may not have allowed sufficient time to assess the weather event's full impact before canceling the flight. Another potential explanation can be due to the rarity of severe weather events and the analysis being somewhat biased due to the sheer number of light and moderate conditions. A third potential explanation is that light and moderate conditions may have caused long enough delays that flights towards the end of the day end up being canceled. These are further analyses that can be conducted with a more robust dataset.

<u>Which metropolitan area is more prone to delays and cancellations?</u>

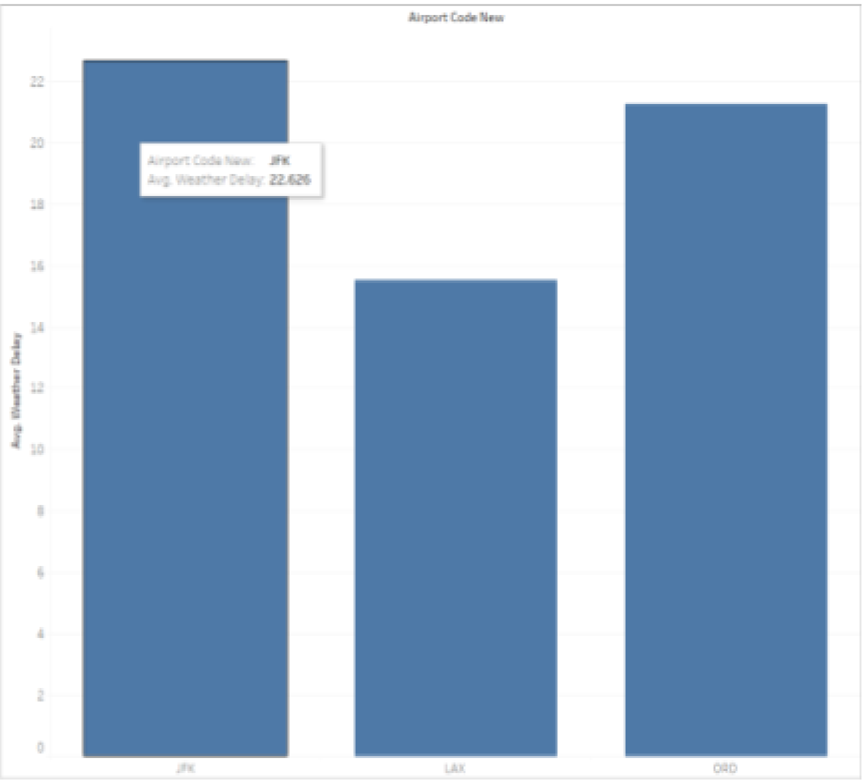


Figure 7: *Average Weather Delays by Airport*

Two bar charts were created to determine which of the three metropolitan areas (JFK, ORD, or LAX) is more prone to cancellations, one of which represents airports against average delay times and the other which compares these same airports to the total number of reported cancellations in that city. We are assuming that the delay and cancellation trends we see at these airport codes are similar to other airports based in the same metropolitan area. From these metrics, we gained insight into which cities tend to pose the greatest risk in terms of delays/cancellations and if weather patterns have some sort of direct relationship in that risk. The first barchart to the bottom (**Figure 7**) shows airports compared to average delays. From this chart we can clearly see that JFK in New York had the greatest length of average delays. We can attribute this to a few factors - perhaps the weather conditions in New York are more volatile than LAX in Los Angeles and ORD in Chicago? From **Figure 2**, we can see that JFK is more susceptible to fog, storms, rain, snow and hail. However, from **Figure 5** we can see that JFK does not really suffer from heavy/severe weather. Could JFK airport in fact have the highest average delays due to weather or could there be additional factors involved like TSA wait times, supply chain issues, etc.?

It’s particularly interesting to note that JFK and ORD are relatively close - could this mean that although ORD has significantly more delays by volume as we saw above, those delays are relatively shorter than the ones JFK sees. In other words, JFK would have 50 delays at 30 minutes a piece while ORD would have 500 delays at 25 minutes a piece. We believe this is the case and it may have additional factors as well which we hope to identify based on the results of our modeling.

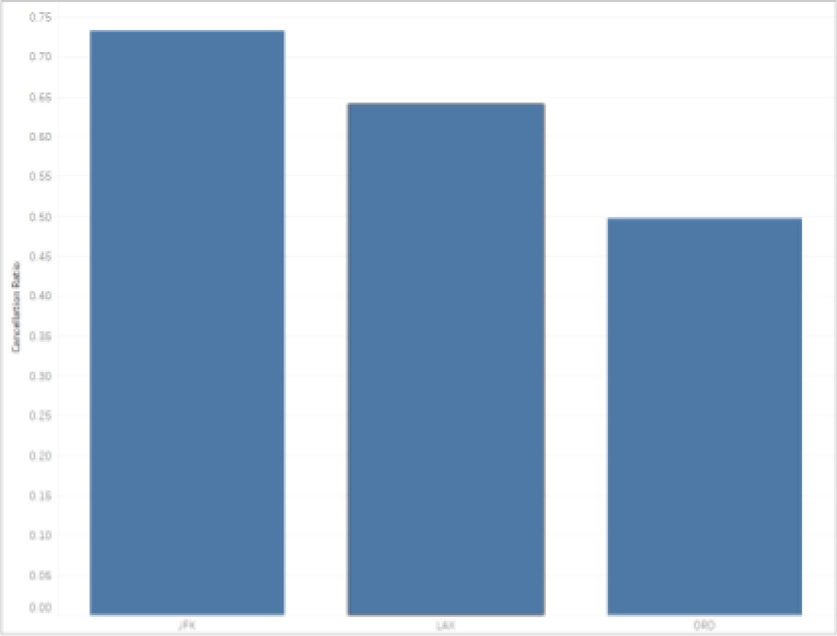


***Figure 8** - Percentage of Cancellations based on Airports*

From the chart in **Figure 8,** the highest percentage of flight cancellations occurred at JFK in New York, followed by LAX in Los Angeles and ORD in Chicago. Possible reasons for these cancellations could be the severe weather conditions at JFK. **Figure 4** shows that JFK receives the highest amount of snow out of the three airports, which could indicate that snow is a difficult condition to operate flights in. Additionally, ORD has the highest amount of rainfall followed by JFK, but it has the lowest percentage of total flights canceled when compared to the other airports. This suggests that JFK received more rainfall than ORD, leading to more cancellations. However, further analysis is required to confirm this hypothesis. The results section attempted to answer some leading questions that arose from the analysis. These questions are as follows:

- Can an argument be made that LAX in Los Angeles is less susceptible to inclement weather (due to their location when compared to ORD and JFK) they are less likely to cancel flights?
- Is weather severity a significant explanatory variable that affects delays and cancellations?
- Is weather type a significant explanatory variable that affects delays and cancellations?
- Can we confirm that JFK has the most delays and cancellations due to weather?

Feature Engineering:

Despite attempting feature engineering with the available variables, a suitable clean variable was not easily obtainable, and we have decided to abandon this approach. Instead, we plan to conduct a stepwise regression analysis using the available variables. This approach is expected to yield a more comprehensive analysis and enable us to identify specific features that have explanatory or predictive abilities, rather than combining unrelated variables to achieve small improvements in modeling accuracy.

Overview of Modeling

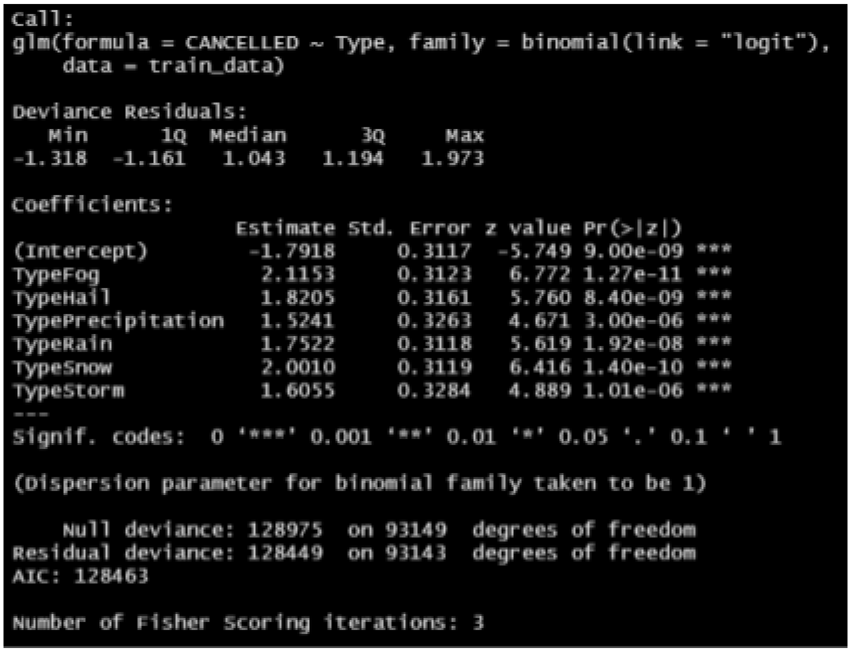
Methodology:

The first model used in this analysis is a logistic regression model, also known as a generalized linear model (glm) with a binomial distribution. The purpose is to determine whether the type of weather has any predictive value for cancellations. The model summary is presented below as **Figure 9**.

We can see that all of the variables are statistically significant at a 99.9% confidence interval. As we can also see, the model establishes TypeCold as the intercept or the baseline value (when all of the other categorical variables are set to zero). This can be interpreted as: the log odds of a cancellation happening when it is cold is -1.7918. Using **Formula 1**, we can convert the log odds to probability, which is easier to interpret. The probability of a flight cancellation when it is cold is 14.2%. When there is foggy weather, the log of the odds increases compared to the intercept since it’s more likely to be canceled as we saw in our EDA above. All of the weather events have the same directional impact as fog, with snow having nearly the same impact, while precipitation and storm have relatively low impacts. The AIC of this model is incredibly high however that’s par for the course with real world models on disparate data sources and we’ll use this to benchmark against our other models.

$$E(Y) = P = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

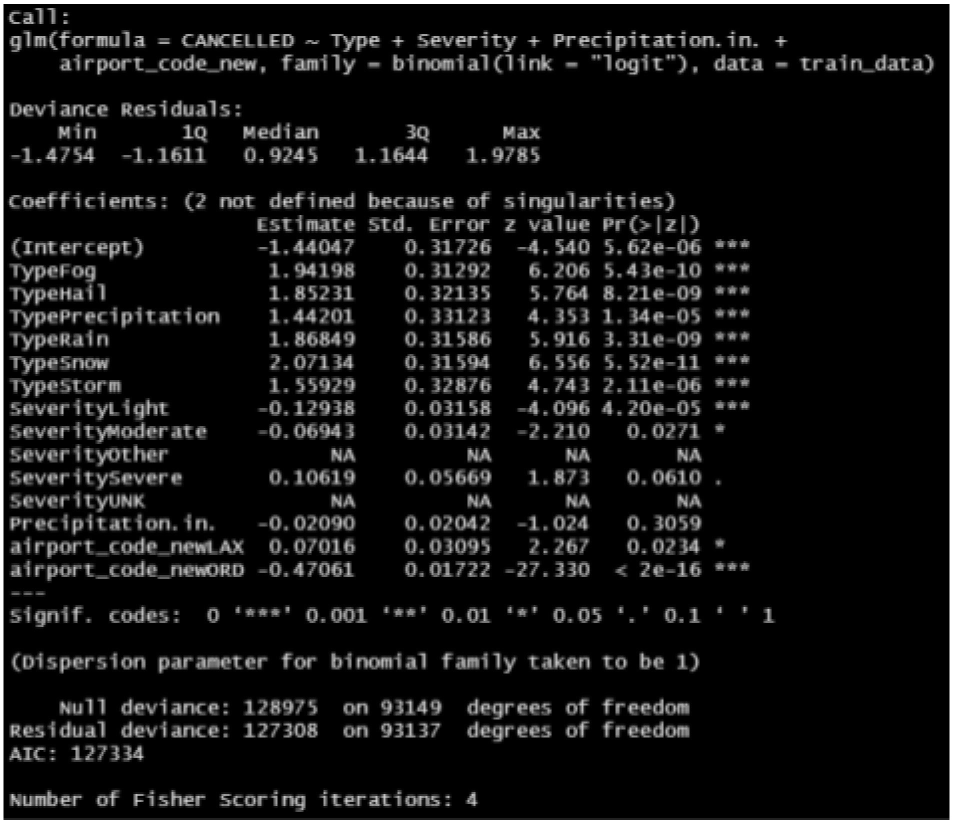
Formula 1 - Converting Log Odds to Probability



```
Call:
glm(formula = CANCELLED ~ Type, family = binomial(link = "logit"),
    data = train_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.318  -1.161   1.043   1.194   1.973

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.7918     0.3117  -5.749 9.00e-09 ***
TypeFog             2.1153     0.3123   6.772 1.27e-11 ***
TypeHail            1.8205     0.3161   5.760 8.40e-09 ***
TypePrecipitation   1.5241     0.3263   4.671 3.00e-06 ***
TypeRain            1.7522     0.3118   5.619 1.92e-08 ***
TypeSnow            2.0010     0.3119   6.416 1.40e-10 ***
TypeStorm           1.6055     0.3284   4.889 1.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128975  on 93149  degrees of freedom
Residual deviance: 128449  on 93143  degrees of freedom
AIC: 128463

Number of Fisher Scoring iterations: 3
```

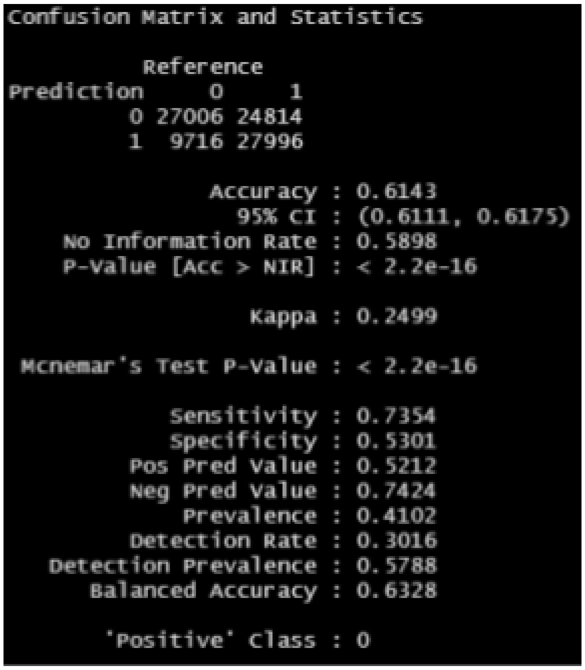
Figure 9 *- Logistic Regression Model Summary for Flight Cancellations based on Weather Event Type*

There are several intermediary models with AIC metrics that do not perform as well as the model discussed earlier. The model summary in **Figure 10** is the best performing one, based on its variables and its AIC, which is much lower than the model presented in **Figure 9**. A lower AIC score indicates a better model. This model includes all the available variables, such as Type, Severity, Precipitation inches, and airport. This shows that the Type of weather is statistically significant, while Severity and airport code are less important when considered together. We also examined a model that excluded the Type variable, which resulted in a higher AIC score. However, we did not investigate this further due to time constraints. Using a threshold probability of 52%, our confusion matrix (**Figure 11**) yielded an accuracy score of 61.43%, which is slightly better than a coin flip.



```
Call:
glm(formula = CANCELLED ~ Type + Severity + Precipitation.in. +
    airport_code_new, family = binomial(link = "logit"), data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4754  -1.1611   0.9245   1.1644   1.9785

Coefficients: (2 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.44047    0.31726  -4.540 5.62e-06 ***
TypeFog               1.94198    0.31292   6.206 5.43e-10 ***
TypeHail              1.85231    0.32135   5.764 8.21e-09 ***
TypePrecipitation     1.44201    0.33123   4.353 1.34e-05 ***
TypeRain              1.86849    0.31586   5.916 3.31e-09 ***
TypeSnow              2.07134    0.31594   6.556 5.52e-11 ***
TypeStorm             1.55929    0.32876   4.743 2.11e-06 ***
SeverityLight        -0.12938    0.03158  -4.096 4.20e-05 ***
SeverityModerate     -0.06943    0.03142  -2.210   0.0271 *
SeverityOther              NA         NA      NA       NA
SeveritySevere        0.10619    0.05669   1.873   0.0610 .
SeverityUNK                NA         NA      NA       NA
Precipitation.in.    -0.02090    0.02042  -1.024   0.3059
airport_code_newLAX   0.07016    0.03095   2.267   0.0234 *
airport_code_newORD  -0.47061    0.01722 -27.330  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128975  on 93149  degrees of freedom
Residual deviance: 127308  on 93137  degrees of freedom
AIC: 127334

Number of Fisher Scoring iterations: 4
```

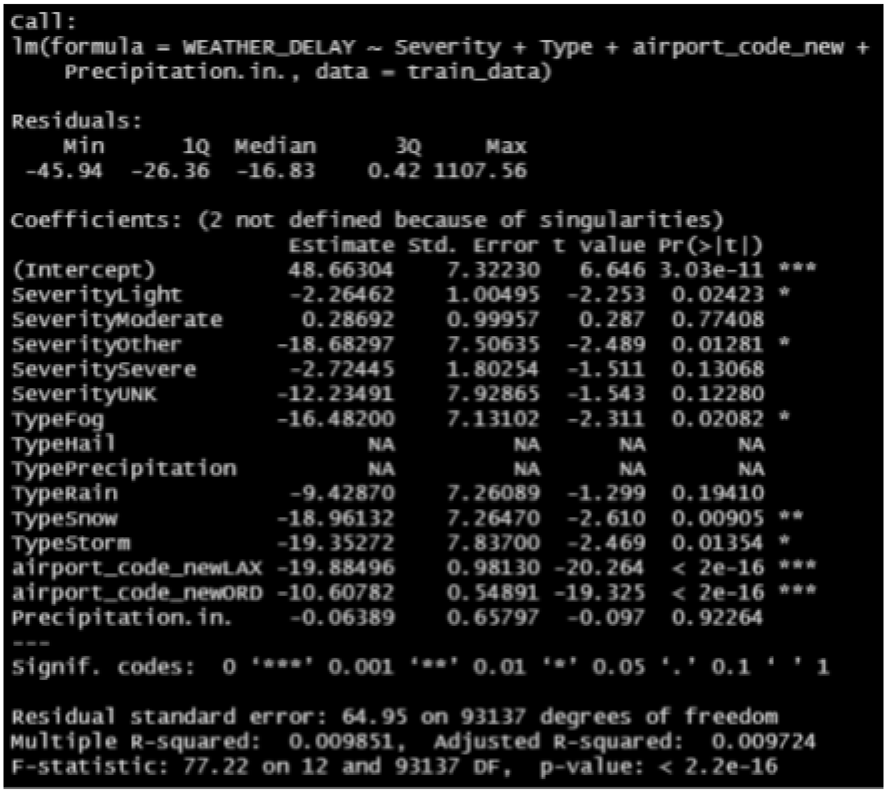
Figure 10*- Logistics Regression Summary for Flight Cancellations based on Type and Severity of Weather*



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27006 24814
         1  9716 27996

               Accuracy : 0.6143
                 95% CI : (0.6111, 0.6175)
    No Information Rate : 0.5898
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2499

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7354
            Specificity : 0.5301
         Pos Pred Value : 0.5212
         Neg Pred Value : 0.7424
             Prevalence : 0.4102
         Detection Rate : 0.3016
   Detection Prevalence : 0.5788
      Balanced Accuracy : 0.6328

       'Positive' Class : 0
```

Figure 11- Confusion Matrix for Figure 10 (Model 2)

Additionally, a linear regression model was constructed using severity, type, airport code, and precipitation in inches as predictors, with weather delay as the response variable. However, the resulting model showed a low R-squared value which meant that it had low predictive power based on the variables in the model. Thus, we concluded that the logistic regression model was more useful for predicting cancellations and more valuable than determining the exact duration of a delay. As previously mentioned, there may be numerous other variables that influence both cancellations and delays, which would require additional data and time to identify in order to improve our model. The summary of the linear regression model can be found in **Figure 12**.



```
Call:
lm(formula = WEATHER_DELAY ~ Severity + Type + airport_code_new +
    Precipitation.in., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
 -45.94  -26.36  -16.83    0.42 1107.56

Coefficients: (2 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           48.66304    7.32230   6.646 3.03e-11 ***
SeverityLight         -2.26462    1.00495  -2.253  0.02423 *
SeverityModerate       0.28692    0.99957   0.287  0.77408
SeverityOther        -18.68297    7.50635  -2.489  0.01281 *
SeveritySevere        -2.72445    1.80254  -1.511  0.13068
SeverityUNK          -12.23491    7.92865  -1.543  0.12280
TypeFog              -16.48200    7.13102  -2.311  0.02082 *
TypeHail                    NA         NA      NA       NA
TypePrecipitation           NA         NA      NA       NA
TypeRain              -9.42870    7.26089  -1.299  0.19410
TypeSnow             -18.96132    7.26470  -2.610  0.00905 **
TypeStorm            -19.35272    7.83700  -2.469  0.01354 *
airport_code_newLAX -19.88496    0.98130 -20.264  < 2e-16 ***
airport_code_newORD -10.60782    0.54891 -19.325  < 2e-16 ***
Precipitation.in.     -0.06389    0.65797  -0.097  0.92264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.95 on 93137 degrees of freedom
Multiple R-squared:  0.009851,  Adjusted R-squared:  0.009724
F-statistic: 77.22 on 12 and 93137 DF,  p-value: < 2.2e-16
```

Figure 12 - Linear Regression Model Summary

Challenges:

Some of the early challenges we faced as a team were related to joining the two datasets based on a primary key. The two datasets had different types of airport codes - International Air Transport Association (IATA) codes and International Civil Aviation Organization codes. The difference between the two codes is that ICAO is a four letter code built with the first letter representing a region, the second letter representing a country, and the third and fourth letters representing the airport. IATA is built for the passenger of the flight and makes referencing the airports much easier. In order to mitigate the effects of this initial issue, we joined our datasets on cities instead, picking a few of the most busy areas - New York, Chicago and Los Angeles.

Another issue we foresaw in the early stages of merging our datasets was the component of time. Since we were trying to determine if weather events were directly correlated with delays or cancellations we needed to ensure that we factored in the time component in our join along with the airport component.

Weather is definitely a reason as to why flights are delayed or canceled, but there are many moving parts when it comes to an airline being on time - flight crew, TSA lines, connecting flights, baggage claims, air traffic, etc. In the future, we would try to find datasets that incorporate a more holistic view of airport data to create a more accurate model in order to understand flight delays/cancellations and what impact weather truly has on it.

Discussions:

As we evaluated our modeling results, we saw that our models were far from perfect in terms of their accuracy. They were based on disparate data sources for a very small timeframe with a longitudinal (time-wise) training and testing split. We also focused on a very narrow spectrum of airport codes to tighten up our analysis and reduce any noise due to geography. If we had more time we could've tuned our model to improve its accuracy from 60% which is relatively decent given the real-world applications we hoped to achieve. Additionally, if given more time and resources we would have added more data elements/attributes to improve our predictive performance.

Our original hypothesis was that weather events would precede airline delays and cancellations, our predictive model suggests that this may be possible so we can't reject the hypothesis just yet. We would need more analysis and modeling to identify whether this is truly predictive but it certainly outpaces any randomness in airline delays so we feel like the model is relatively competent and useful. We did however hypothesize that severe weather events would cause higher rates of airline delays and cancellations which we haven't seen. For example, the intercept in the logistic regression model is based on: Cold weather type, heavy severity, and JFK airport; if we hold all other variables constant then severe severity doesn't increase the log odds of a cancellation by much and it's not statistically significant. We would've wanted to see a significant impact which we didn't see so we can safely say that severe weather has minimal impact however weather as a whole has some predictive impact on cancellations.

The linear regression piece of trying to predict approximate weather delay was also fruitless as we were unable to get a meaningful model that helped predict the magnitude of delays. This is tough to predict given that different conditions, airports, and many other factors may influence this and there may not be a cookie-cutter approach requiring nuanced approaches. This analysis may not be suitable for a linear regression model or it may require additional data elements that we were unable to source.

We innovated more on the data cleaning and exploratory data analysis piece. Data cleaning innovation included advanced regular expressions, use of complicated POSIX and UTC time elements, and involved data manipulation using dplyr. We also used Tableau to do our data visualizations and identify any trends and gaps in our data which we corrected to improve our modeling efforts. We could've used fancier models or hyper-tuning but chose not to as logistic regression gave us some lead and we realized the gaps were primarily in our data aggregation as opposed to optimizing on the model and just overfitting which was a true risk.

Conclusion/Takeaways:

Despite the best performing model having only ~60% predictive power, the business impact of this analysis is still significant for airlines, airports, and travelers. By understanding the impact of weather on flight delays and cancellations, airlines can improve their scheduling and resource allocation, resulting in increased customer satisfaction and reduced costs. For example, airlines may choose to proactively cancel flights in anticipation of severe weather events to minimize disruptions to their schedule, or adjust flight routes to avoid airports that are more prone to delays and cancellations. Similarly, airports can use this information to better manage their resources during periods of inclement weather and reduce the negative impact on travelers.

Additionally, our analyses highlight the importance of data quality and availability in predictive modeling. While the results of this study are promising, the limitations of the data used for the analysis indicate the need for more comprehensive and accurate data sources. Investing in data collection and management processes can improve the accuracy and effectiveness of predictive models, leading to better decision-making and ultimately, a competitive advantage in the industry.

In conclusion, this analysis suggests that the type of weather experienced by airports varies and has an impact on the frequency and magnitude of flight delays and cancellations. Some of the results were intuitive, such as snow or fog causing heavy delays and cancellations. However, there were some results that were surprising such as light to moderate weather severity contributing to the majority of delays and cancellations.

Although there is some predictive ability when considering all variables together, individually they do not provide meaningful predictive models. A future direction for this project would be to build models with interaction terms, and perhaps a more robust time series analysis.

References

1. Bureau of Transportation Statistics. (n.d.). Understanding the reporting of causes of flight delays and cancellations. Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics. Retrieved March 30, 2023, from https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
2. Federal Aviation Administration. (n.d.). Air traffic by the numbers. Air Traffic By The Numbers | Federal Aviation Administration. Retrieved March 30, 2023, from https://www.faa.gov/air_traffic/by_the_numbers
3. Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A., &amp; Zou, B. (2010, October). Total Delay Impact Study. Retrieved March 30, 2023, from https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_ 10_18_10_V3.pdf