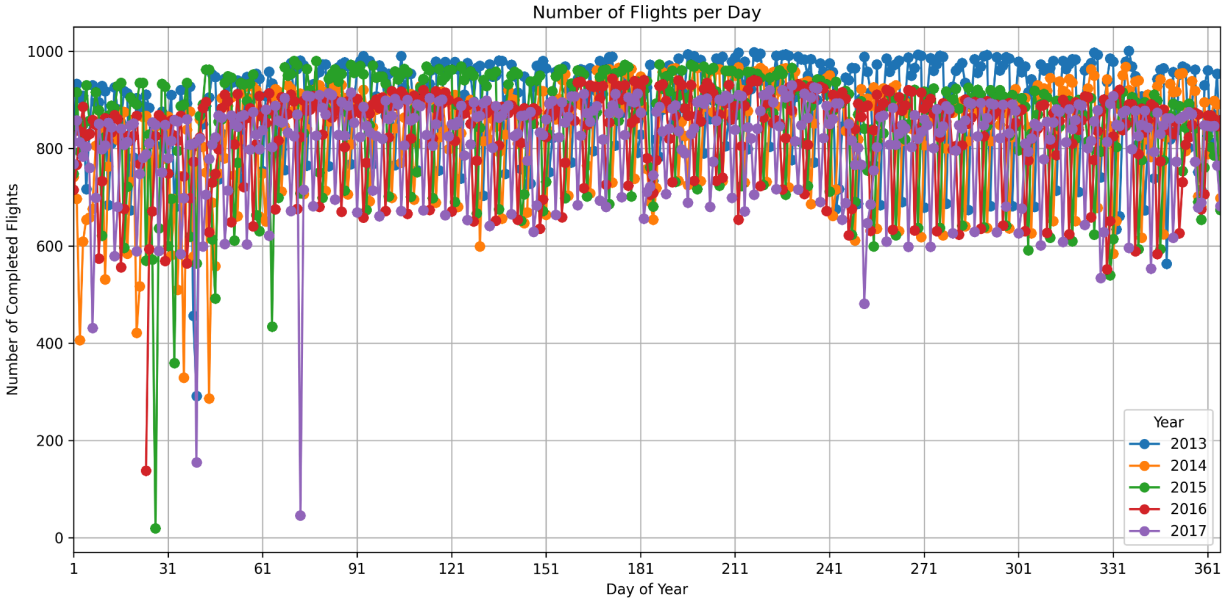
Chris Troop
2024-12-09
CSC 5610

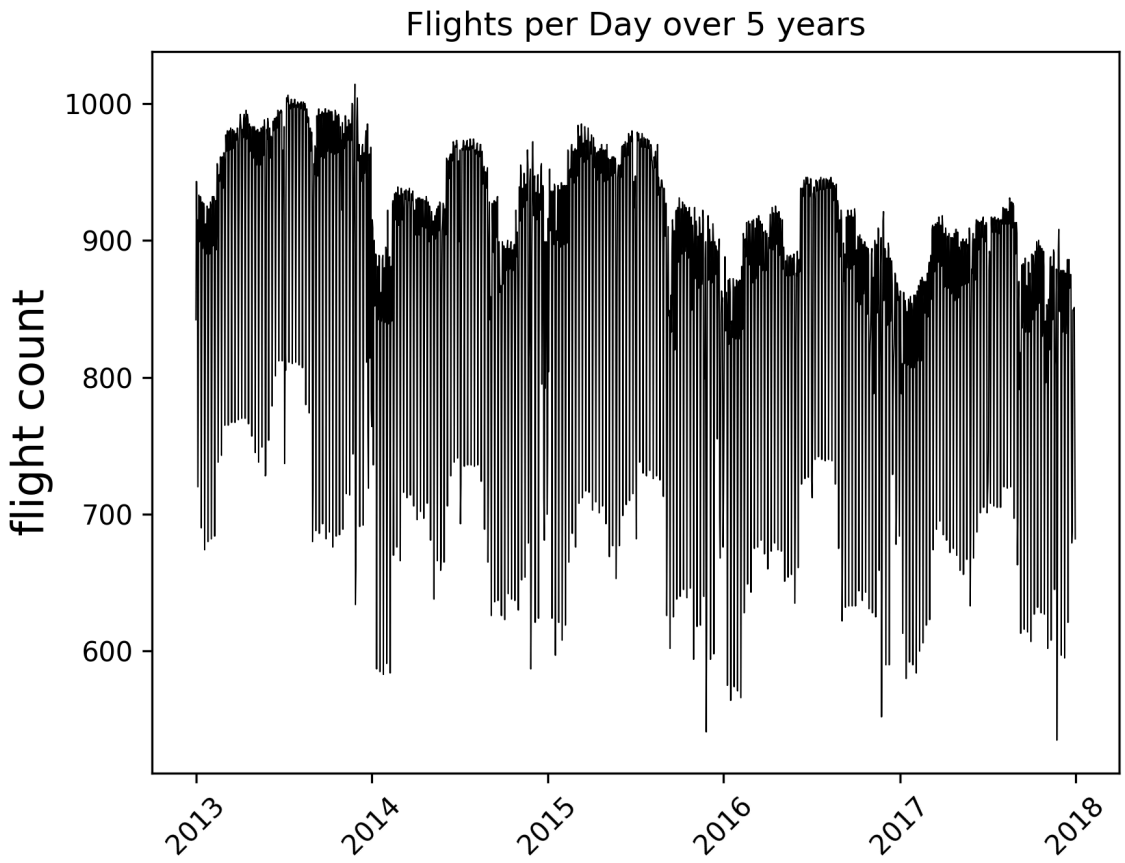
Project M5

The data provided was five years worth for flights incoming and outgoing from the three big airports in New York City(JKF, Newark Liberty, and LaGuardia). The flights were provided in 1 csv file per year from 2013 to 2017. The data was combined into one data frame and then a new variable was created to summarize the number of flights per day. Each year's daily flight numbers are plotted below.



There was not much to takeaway besides that there are outliers in the earlier part of the year, which may line up with significant winter weather storms. The only other notable observation is what looks to be a decrease in flights per day as the years progress. This takeaway is more obvious in the graph below where all the years are plotted as one

dataset.



This decreasing trend is worth further investigation and the rest of the report will detail that. To start, the variables that were used are below in the table.

Variable	Descriptions
Date	Date
flight_count	Number of flights per day
Holiday	Engineered feature that marks date true of false for major holidays
month	Month (Jan, Feb,...)
day_of_week	Day of week (Monday, Tuesday,...)

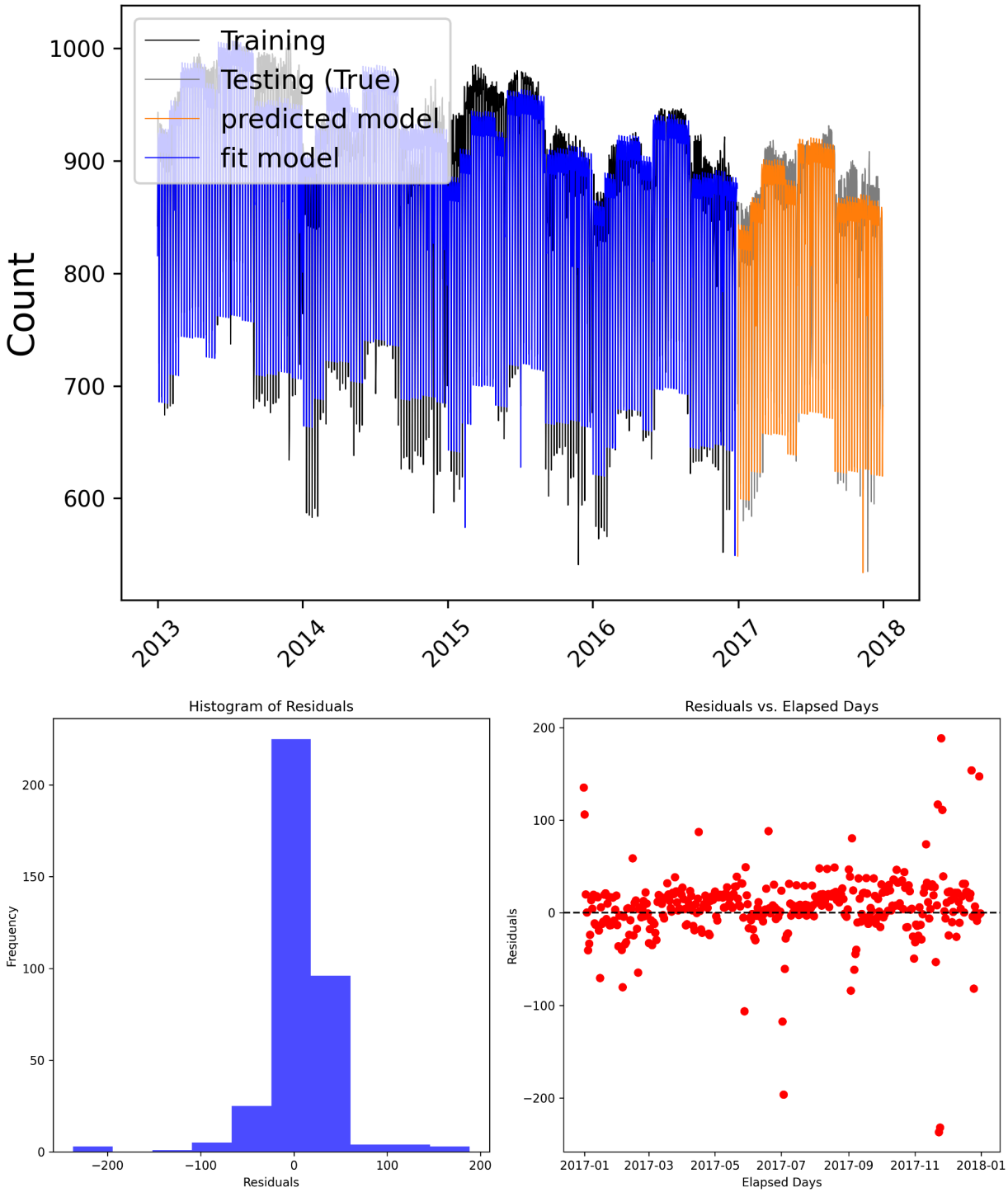
To verify the negative trend observed, linear regression was leveraged to verify the overall trend of the five years of flight date. The slope of the fit line was -0.0563774875584799, the negative slope indicates a negative trend. The p-value was

2.5812714220214682e-20, which means the trend is significant and can be further investigated.

Four models were created and evaluated for RMSE and MAPE. The models and metrics are detailed in the table below.

Model	RMSE	MAPE
Ridge	89	11%
Ridge (+month)	85	10%
Ridge (+month +day of week+holiday)	37	11%
Random Forest (+month +day of week +holiday)	41	5%

The model that best captured the seasonality of the data was the “Ridge (+month +day of week)” and it is shown below with its residual plots as well.



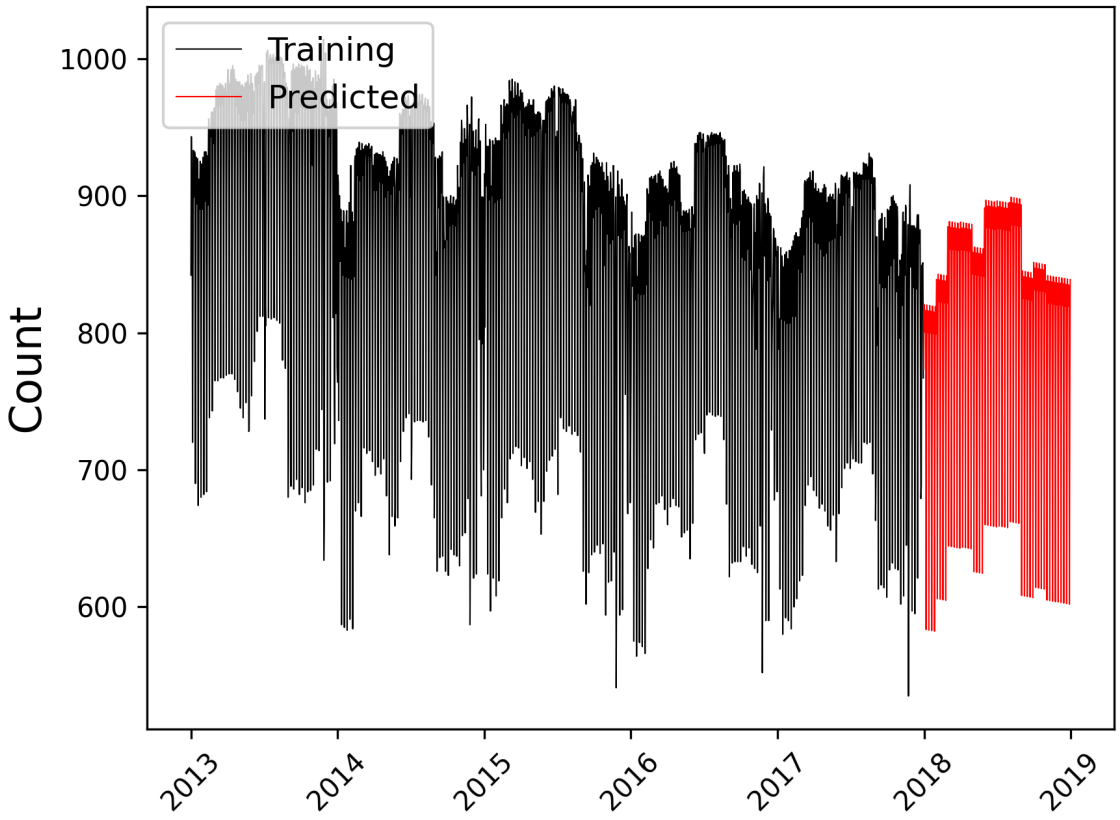
The residuals are normally distributed around zero and there are very few outliers. The outliers may be explained by the aforementioned winter weather that New York battles every year.

The top features of the best performing ridge and random forest are in the table below.

Ridge	Random Forest
day_of_week_0 54.315425 day_of_week_3 49.815818 day_of_week_4 45.334952 day_of_week_2 39.926423 month_7 36.182538 month_6 33.259112 day_of_week_1 33.155940 month_8 33.092291 month_4 11.374009 month_3 9.777188	day_of_week_5 0.590450 elapsed_days 0.215102 day_of_week_6 0.071919 Holiday 0.039897 day_of_week_3 0.013509 month_1 0.009882 month_11 0.008531 month_12 0.008189 month_7 0.008138 month_6 0.004080

The ridge model favorest features that contribute strongly to predicting the outcome where the random forest places importance on features that reduce the complexity of the model. Based on that, it seems like the day of the week was a feature that was highly important in the ridge model for predicting the number of flights that day. The random forest model found day of the week important but it also leveraged the elapsed days since the start of the model to predict the number of flights. The random forest did not seem to predict the seasonal data as well as the ridge so over time the model may lose its predictive value for long horizon applications.

The ridge model was selected for the future data set prediction. The model is shown below.



While the forecasted number of flights per day for 2018 looks a little rigid and not as variable as the training data set; the overall shape of the year looks realistic and maintains the downward trend in number of flights. I feel like the model came out decent and has good predictive value.