PREDICTING AIRLINE DELAY AND TIME OF DELAY IN MAJOR AIRLINES, UB

SARANG A KATKAR

INTRODUCTION TO THE PROBLEM:

- Airline delays cost the airline company a significant amount of money and is one of the major reasons of financial losses. We want to dig some of the popular airlines and their cause of delay and then predict what factors contribute to a flight arriving late.

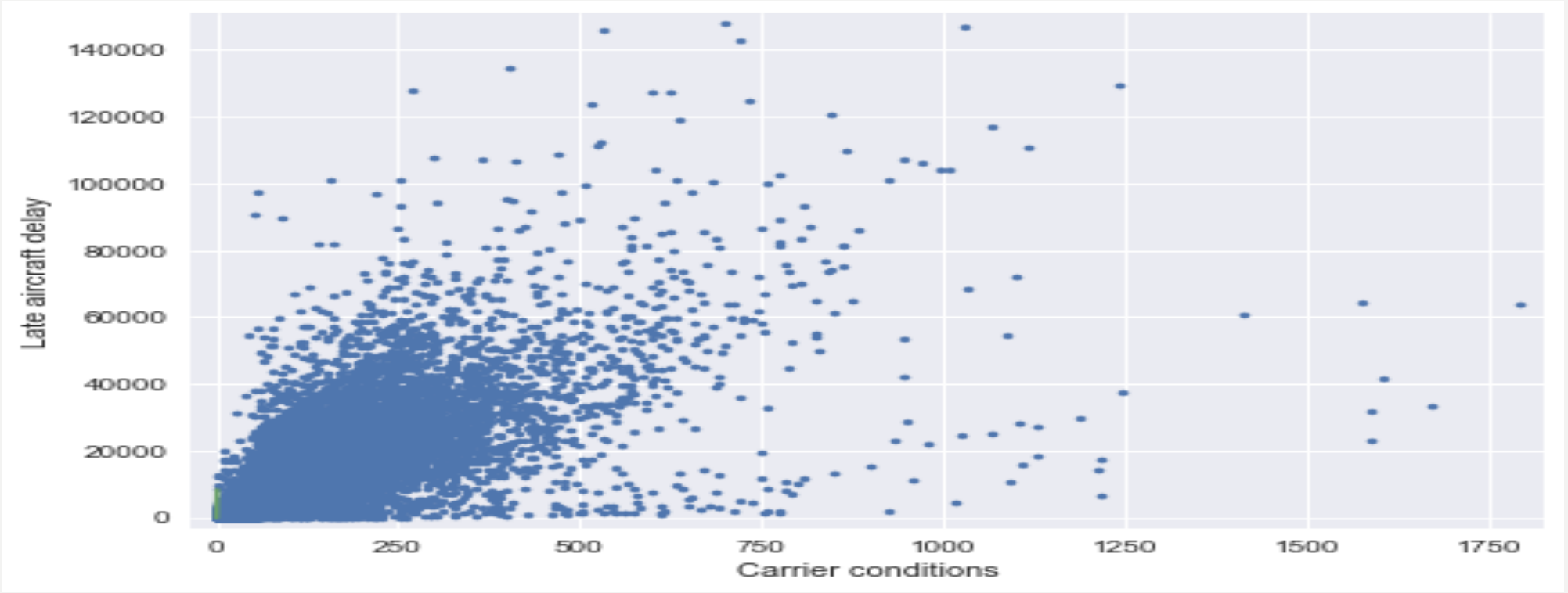
DATA LOADING

- *#Data Loading*
- airdf=pd.read_csv('/Users/sarang/Desktop/318750629_72017_4040_airline_delay_causes.csv')

DATA CLEANING

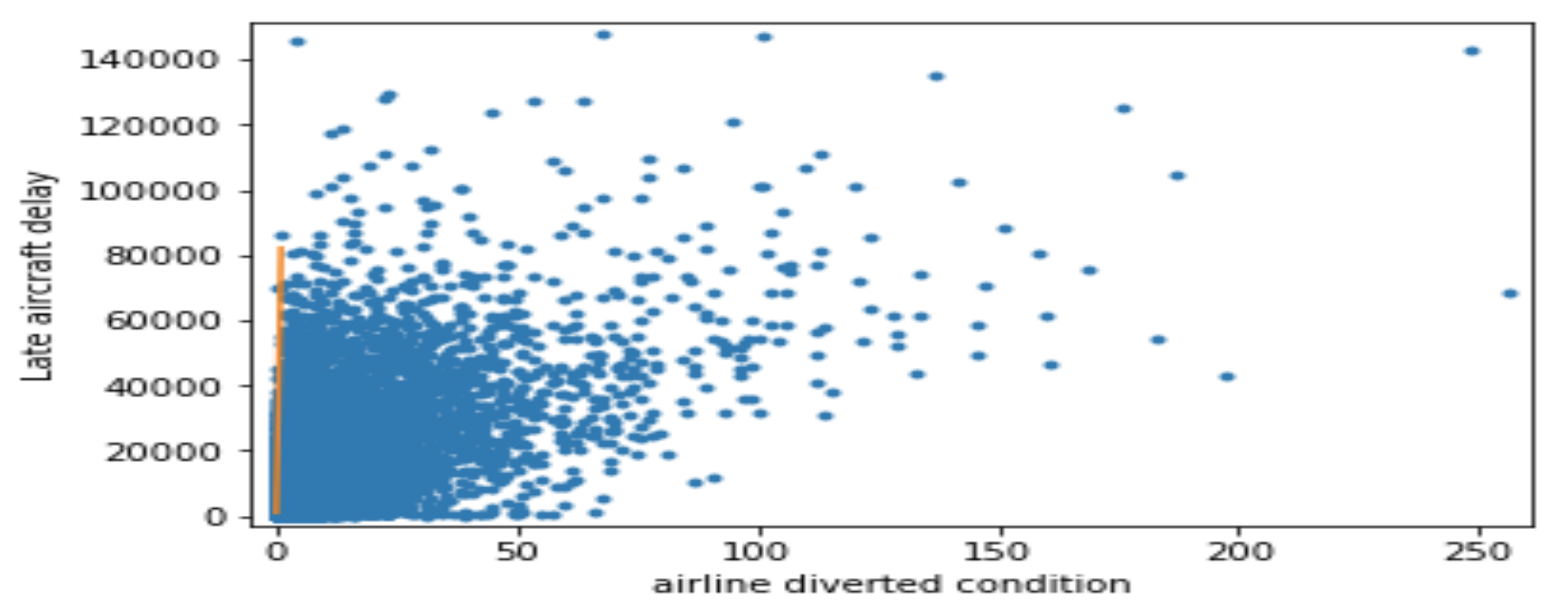
- *#Data Cleaning*
- 1.x=airdf.dropna(subset=['year',' month','carrier','carrier_name','airport', 'airport_name','arr_flights','arr_del15','carrier_ct',' weather_ct','nas_ct','security_ct', 'late_aircraft_ct','arr_cancelled','arr_diverted',' arr_delay',' carrier_delay','weather_delay', 'nas_delay','security_delay','late_aircraft_delay'],how='any')
-
- 2. x.airport.astype('category')
- 3. x.carrier_name.astype('category')
-

EXPLORATORY DATA ANALYSIS



INFERENCE: Shows that carrier conditions doesn't actually follow a linear trend hence it wont be a significant factor in late aircraft delay.(As many of data points fall below even if carrier conditions gets worse)

EDA 1



Inference: Given the slope=811 and given how slant it is, we can say that a diverted flight is certainly going to cause delay.

EDA 2

- **#Correlation between security conditions and late aircraft delay**
- 1.correlation=np.corrcoef(x['security_ct'],x['late_aircraft_delay'])
- correlation
- Out[98]:
- array([[1.   , 0.42960708],
- [0.42960708, 1.   ]])
- **Inference: shows some correlation between security conditions and delay. But it is weak 0.42.**

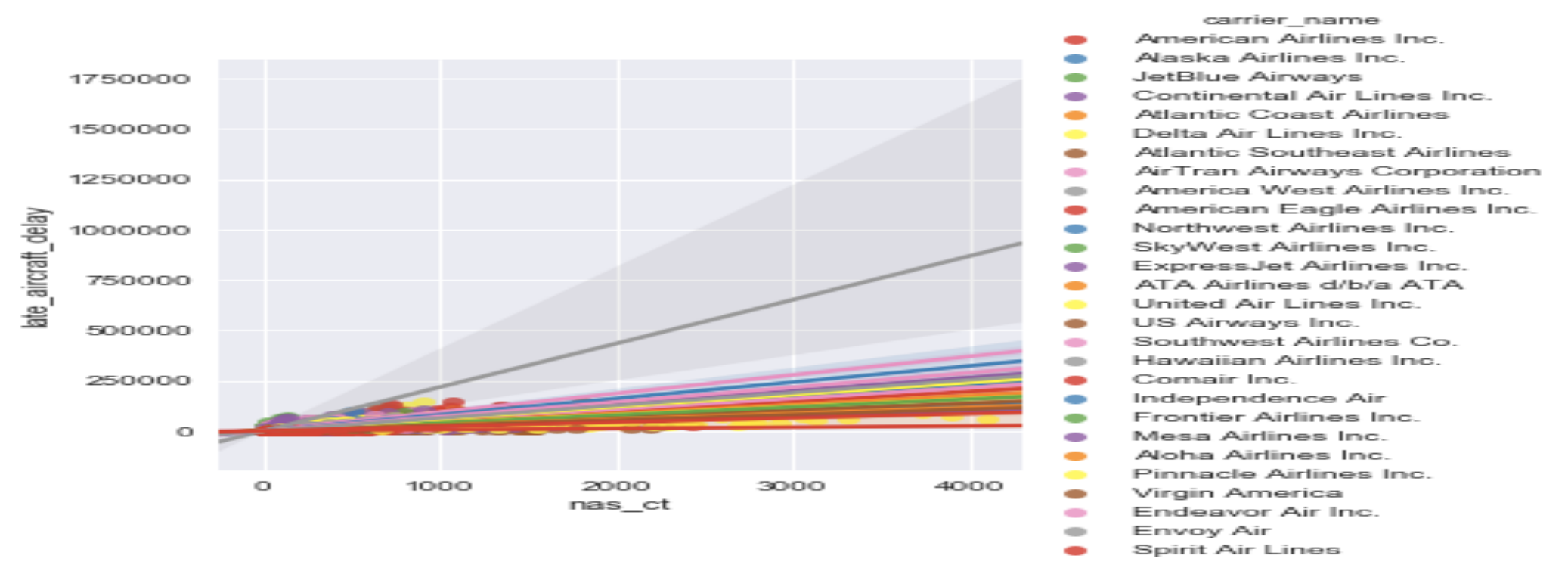
EDA 3

- **#covariance between carrier conditions and late aircraft delay**
- 2.covariance_matrix=np.cov(x['carrier_ct'], x['late_aircraft_delay'])
- covariance_matrix
- Out[100]:
- array([[2.35531691e+03, 2.02648252e+05],
- [2.02648252e+05, 2.62728747e+07]])
-
- **Inference: shows carrier conditions and late aircraft don't vary the same way. Hence not a factor to be relied upon for prediction. We can use lasso regressor to find it out.**

EDA 4

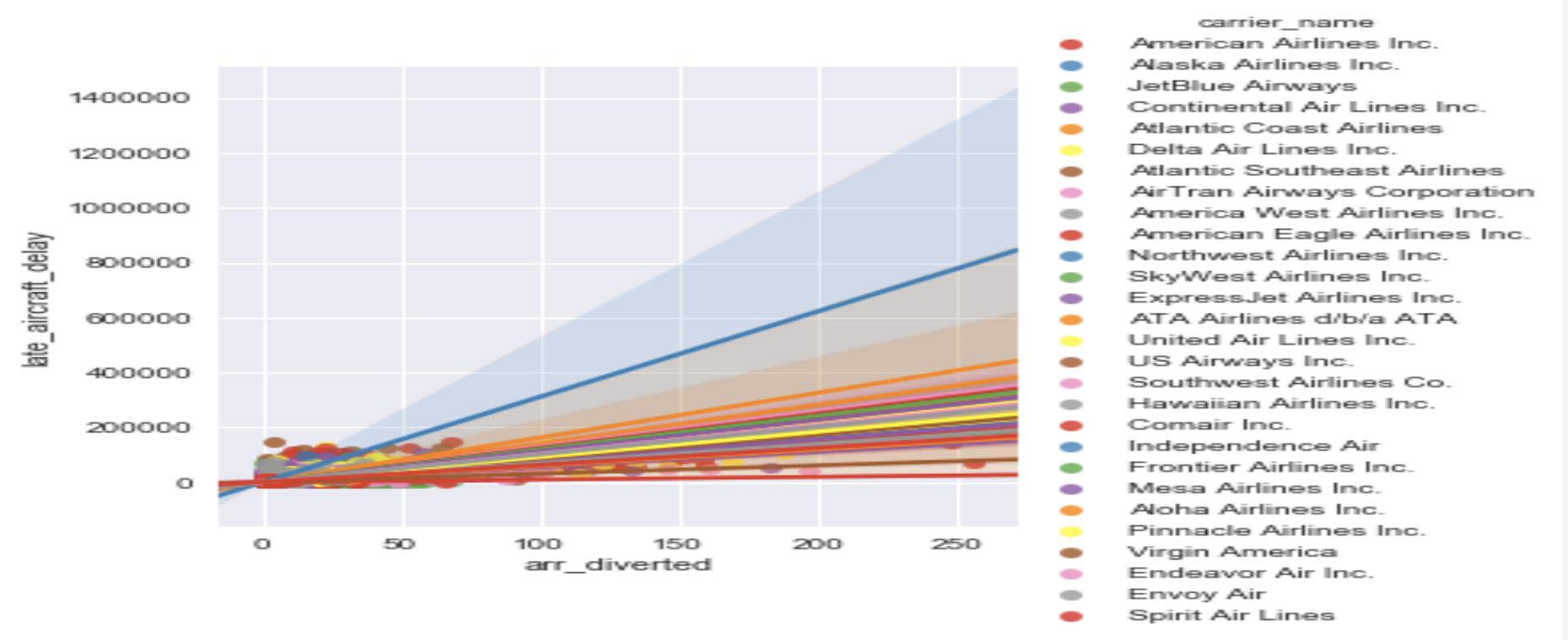
- **#covariance between security conditions and late aircraft**
- 3. covariance_matrix_security=np.cov(x['security_ct'], x['late_aircraft_delay'])
-
- covariance_matrix_security
- Out[105]:
- array([[7.91576198e-01, 1.95916758e+03],
- [1.95916758e+03, 2.62728747e+07]])
-
- **Inference: It confirms that they don't vary similarly as increase in security delay wont necessarily cause aircraft to get late.**
-

EDA 5



Inference:
Grouped by different airlines we see the relation that National Aviation security/ systems contributes in getting a airline to get late (linearly).

EDA 6



Inference: it supplements to the hypothesis that a diverted aircraft will cause an aircraft to be late.

ANALYSIS

- ***#Group by different airlines and their count on cancelled flights***
- Groupby
- by_class = x.groupby('carrier_name')
-
- count_by_class = by_class['arr_cancelled'].count()

ANALYSIS 2

- ***#Group by different airlines and their average traffic in the airports***
- by_carrier_sub=by_carrier['arr_flights']
-
- aggregated=by_carrier_sub.agg(['mean'])
-
- Aggregated

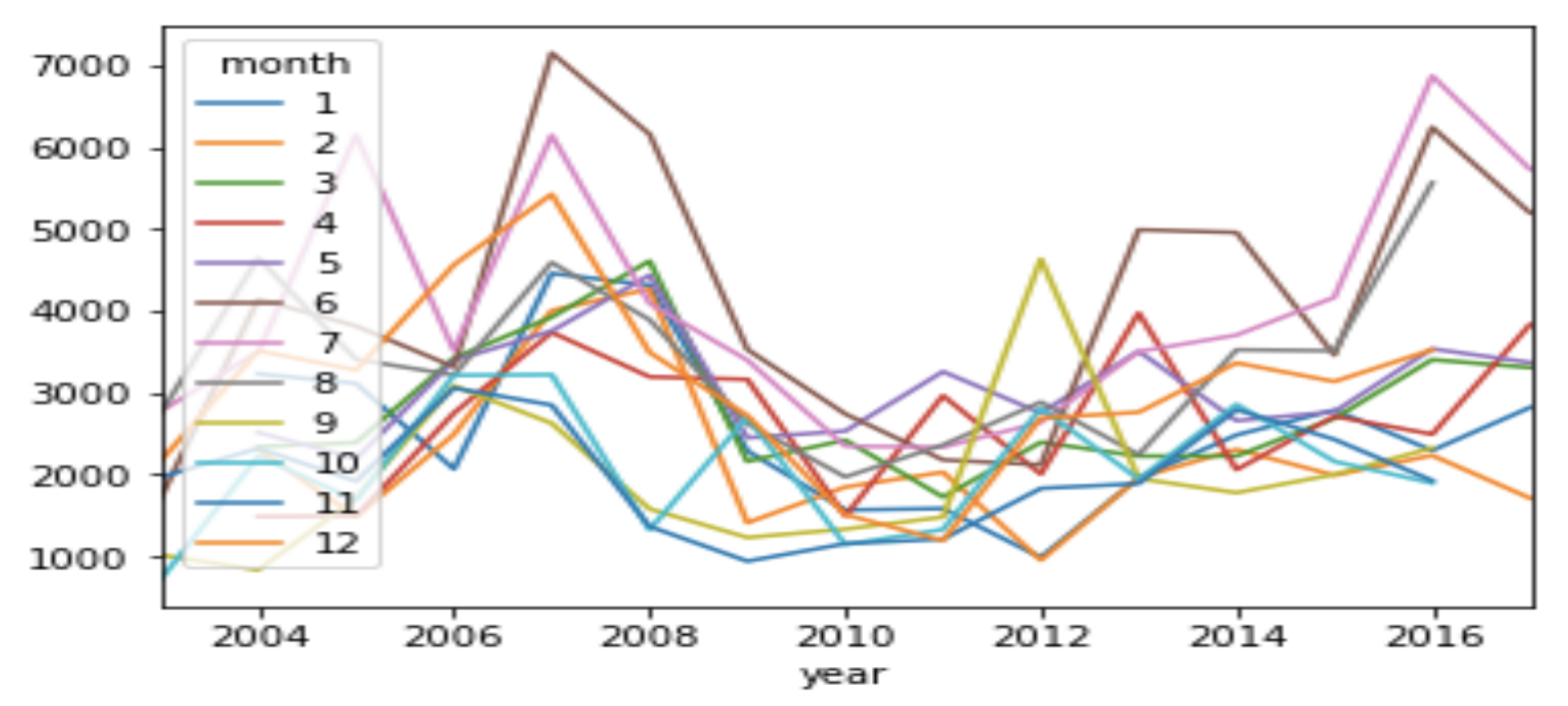
ANALYSIS 3

- ***#Group by different airlines and their average weather conditions, carrier delay, late_aircarft_delay***
-
- by_factors=by_carrier[[' weather_ct',' carrier_delay','late_aircraft_delay']]
-
- aggregated=by_factors.agg(['mean','max'])
-

INFERENCE FROM ANALYSIS

- *INFERENECE:*
- *we came to know that southwest airlines had the highest*
- *average delay in aircrafts of 6726mins. Followed by American airlines with*
- *an average of 2859.But since the airline traffic is higher in American*
- *airlines i.e-651, so we want to dig deeper into American airlines.*

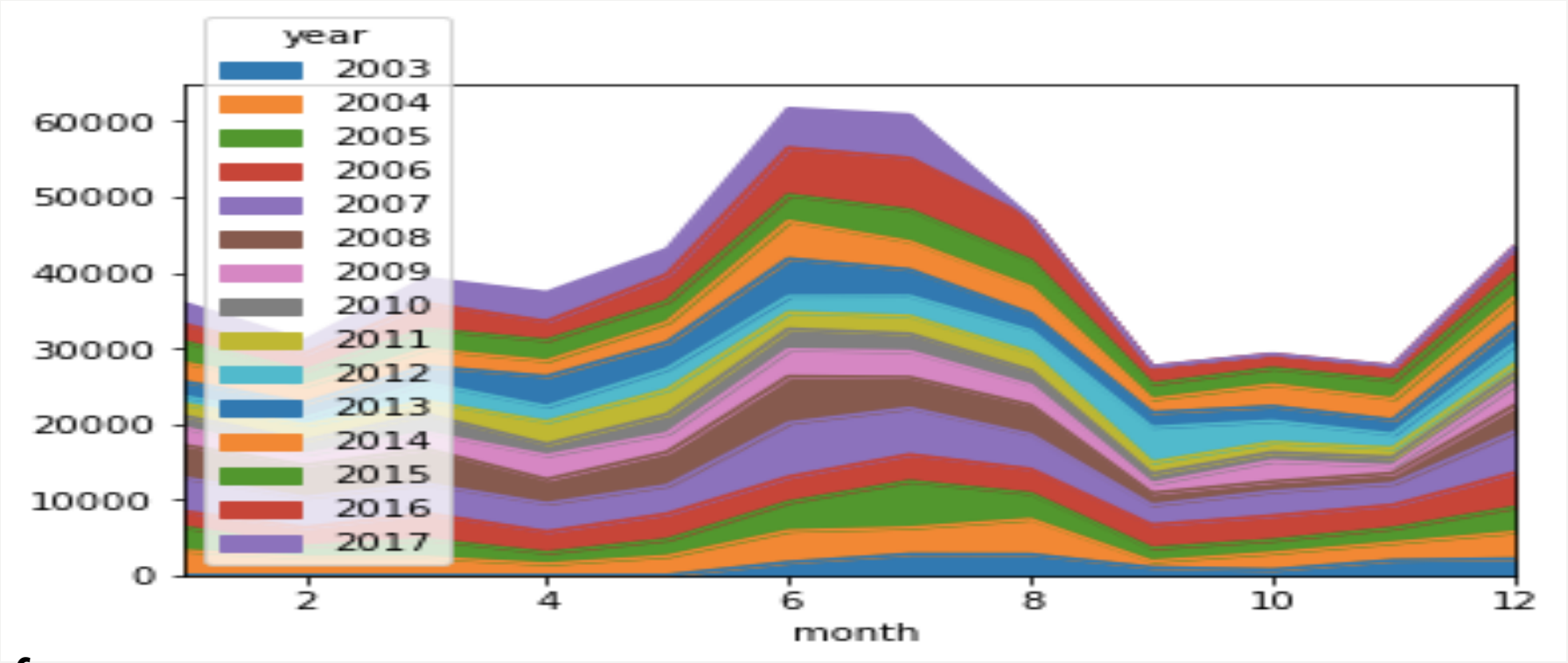
#TIME SERIES PLOT OF AMERICAN AIRLINES BY YEAR AND MONTH



Inference:

It can be observed that in the year 2007 and especially in the month of june to august for almost all the following years the delay was maximum.

#TIME SERIES PLOT OF AMERICAN AIRLINES BY YEAR AND MONTH

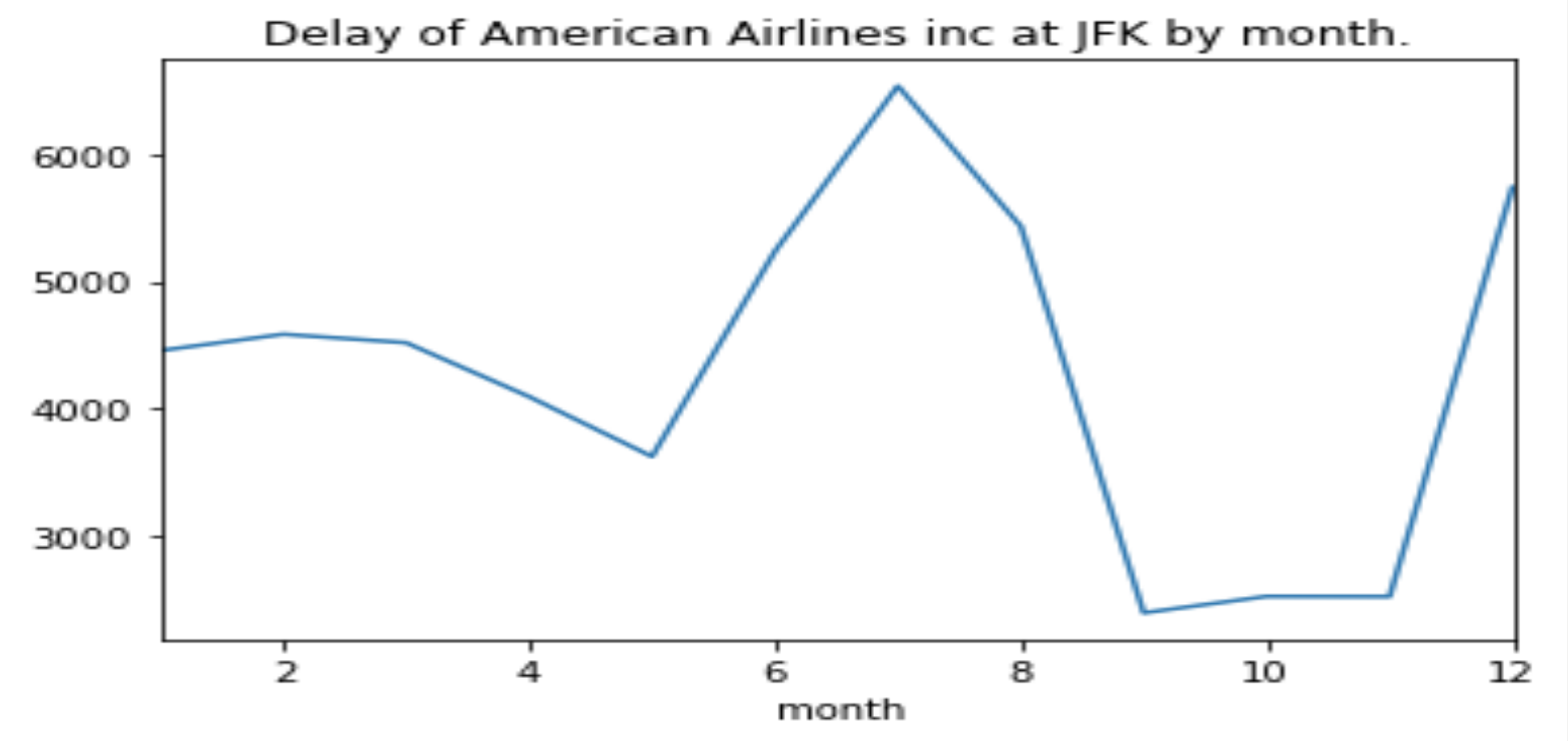


Inference:

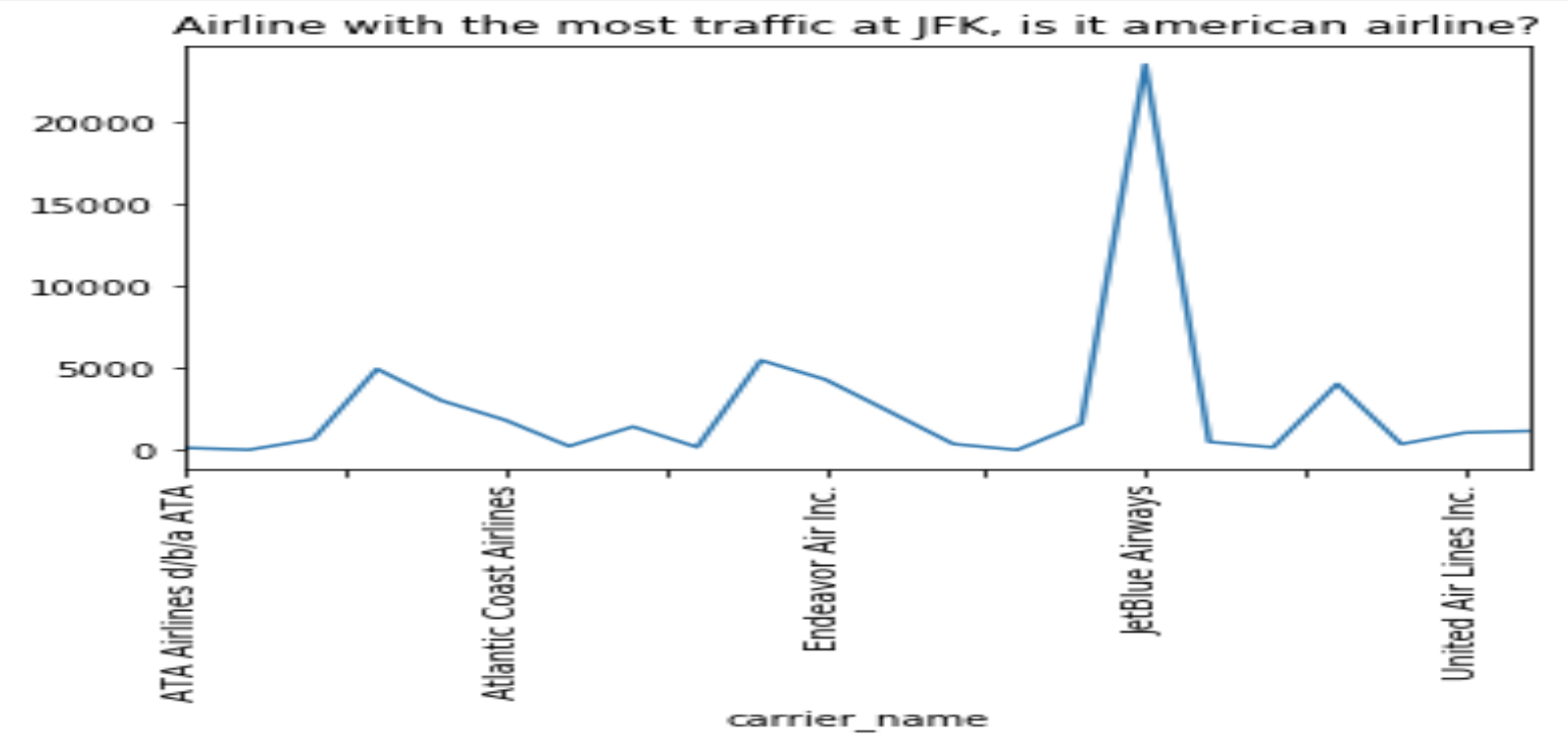
It can be inferred that chances that of the American airlines flight getting late are high in the month of June, July and August.

AMERICAN AIRLINES AT JFK

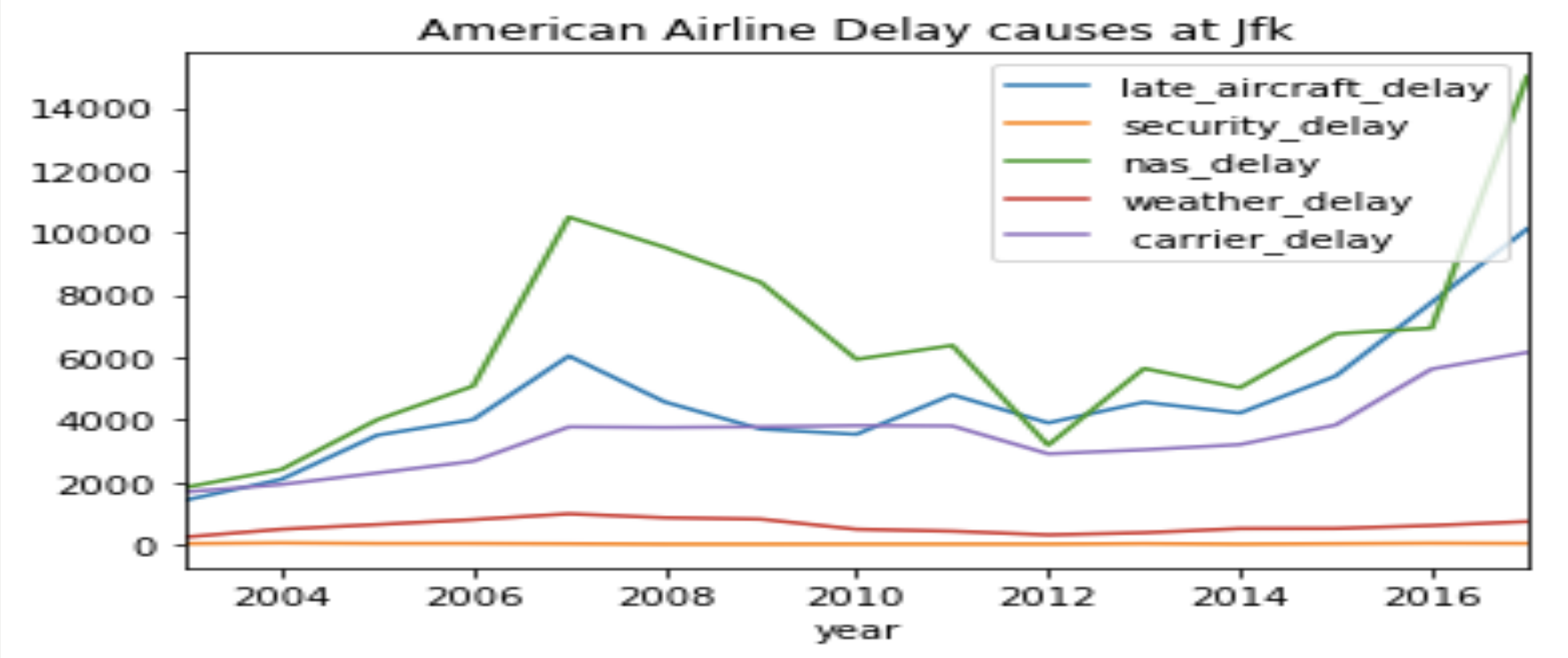
DELAY AT JFK BY MONTH



AIRLINE WITH HIGH TRAFFIC AT JFK IS IT JFK?



AMERICAN AIRLINE DELAY CAUSES AT JFK



INFERENCE

The cause of delay of American airlines at JFK airport being majorly due to national aviation system/security process followed by carrier delay. Which they might have to improve upon. However it is observed there was no significant delay through weather at all (almost negligible).

DIMENSIONAITY REDUCTION

- ***#Feature Selection/Feature Engineering and Dimensionality reduction***
- from sklearn.linear_model import Lasso
- a=x[['arr_cancelled','arr_diverted','carrier_ct',' weather_ct','nas_ct','security_ct', 'late_aircraft_ct']]
- # Instantiate a lasso regressor: lasso
- lasso = Lasso(alpha=0.4,normalize=True)
-
- # Fit the regressor to the data
- lasso.fit(a,y)

LINEAR REGRESSION

- ***# Linear Regression***
- # Create training and test sets
- X_train, X_test, y_train, y_test = train_test_split(a, y, test_size = 0.3, random_state=42)
- reg_all = LinearRegression()
- # Fit the regressor to the training data
- reg_all=reg_all.fit(X_train,y_train)
- # Predict on the test data: y_pred
- y_pred = reg_all.predict(X_test)
- # Compute and print R^2 and RMSE
- print("R^2: {}".format(reg_all.score(X_test, y_test)))
- rmse = np.sqrt(mean_squared_error(y_test,y_pred))
- print("Root Mean Squared Error: {}".format(rmse))
- R^2: 0.964577207896

CROSS VALIDATION

- ***5 Fold cross validation***
- from sklearn.linear_model import LinearRegression
- from sklearn.model_selection import cross_val_score
- # Create a linear regression object: reg
- reg = LinearRegression()
- cv_scores = cross_val_score(reg,a,y, cv=5)
- # Print the 5-fold cross-validation scores
- print(cv_scores)
- print("Average 5-Fold CV Score: {}".format((np.mean(cv_scores))))
- [0.94524707 0.96828198 0.97113696 0.97147914 0.95313221]
- Average 5-Fold CV Score: 0.96185547261

CONCLUSION

- **Conclusion:**
- We conclude that factors such as national aviation system security, diverted flight, cancelled
- Flight and not variables such as weather conditions, security check predicts the delay time of
- the aircraft accurately.