TAXI OUT TIME ANALYSIS AT SAN DIEGO INTERNATIONAL AIRPORT

Chandhnee Karthikeyan Iyer

AGENDA

Understanding the Data

Exploratory Data Analysis

Forecasting

Takeaways

APPROACH

Understanding data:

Python (Google Colab), MS Excel

Exploratory Data Analysis:

Tableau

Forecasting:

Python (Google Colab)

UNDERSTANDING THE DATA

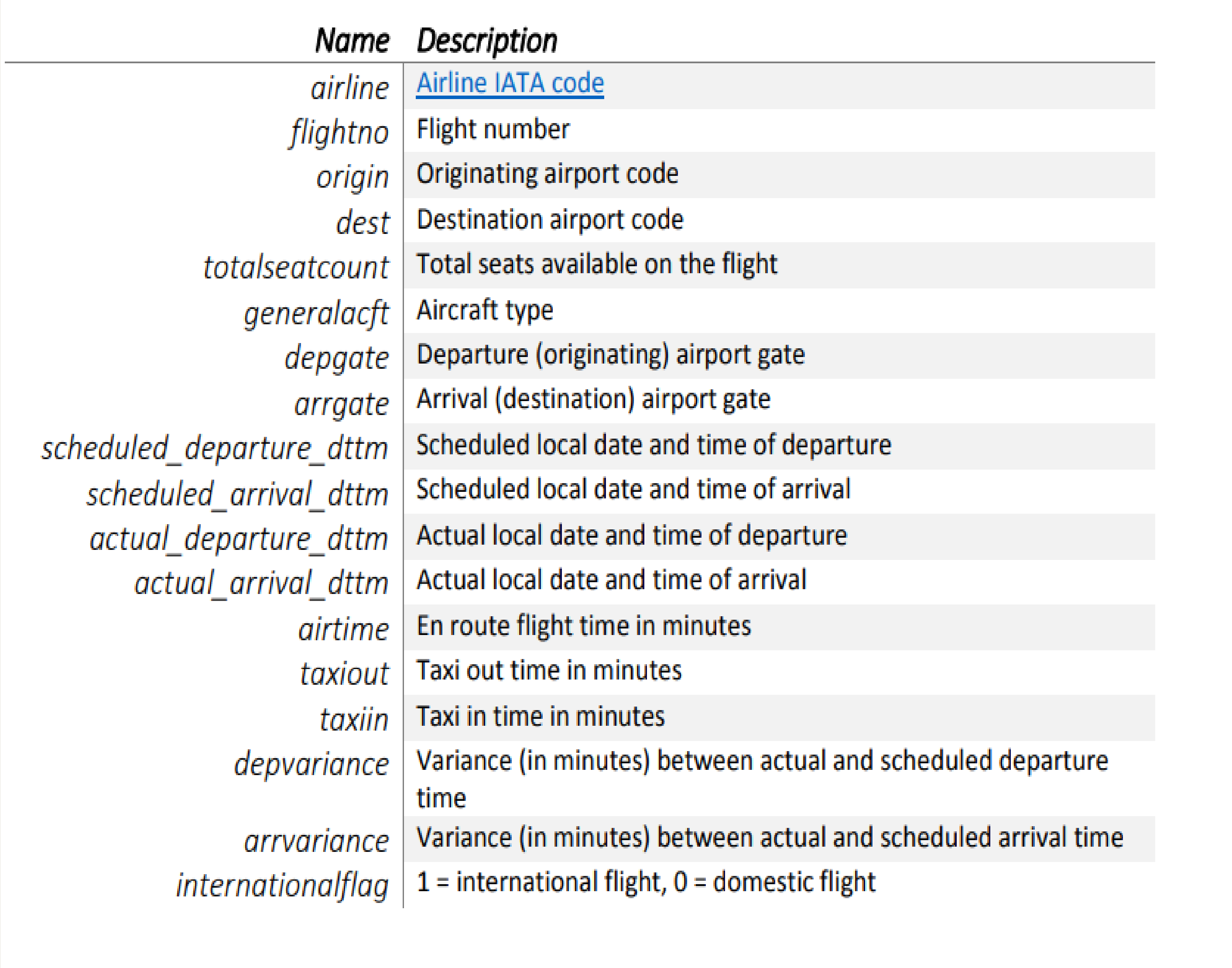
DATA — A record of all Flights departing from San Diego International Airport in 2017 and 2018

GRAIN — A Record represents a departure from San Diego International Airport

VOLUME — 188525 records

VARIETY — Across 14 different Airlines

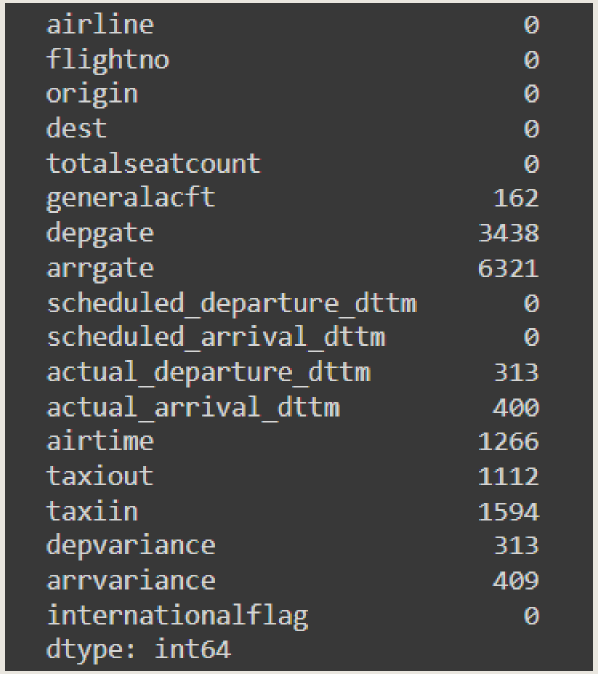
THE DATASET



Name	Description
airline	[Airline IATA code]()
flightno	Flight number
origin	Originating airport code
dest	Destination airport code
totalseatcount	Total seats available on the flight
generalacft	Aircraft type
depgate	Departure (originating) airport gate
arrgate	Arrival (destination) airport gate
scheduled_departure_dttm	Scheduled local date and time of departure
scheduled_arrival_dttm	Scheduled local date and time of arrival
actual_departure_dttm	Actual local date and time of departure
actual_arrival_dttm	Actual local date and time of arrival
airtime	En route flight time in minutes
taxiout	Taxi out time in minutes
taxiin	Taxi in time in minutes
depvariance	Variance (in minutes) between actual and scheduled departure time
arrvariance	Variance (in minutes) between actual and scheduled arrival time
internationalflag	1 = international flight, 0 = domestic flight

ANOMALIES

NULL VALUES IN DATA



```
airline                       0
flightno                      0
origin                        0
dest                          0
totalseatcount                0
generalacft                 162
depgate                    3438
arrgate                    6321
scheduled_departure_dttm      0
scheduled_arrival_dttm        0
actual_departure_dttm       313
actual_arrival_dttm         400
airtime                    1266
taxiout                    1112
taxiin                     1594
depvariance                 313
arrvariance                 409
internationalflag             0
dtype: int64
```

ANALYSING NULLS IN VARIABLES of INTEREST

- TAXIOUT: 0.005% of records
- Departure Gates: 0.018% of records
- Actual Departure Time and Actual Arrival Times for < 0.001% of records

POTENTIAL INCORRECT VALUES IN DATA

OUTLIERS

Values greater than 1.5* IQR

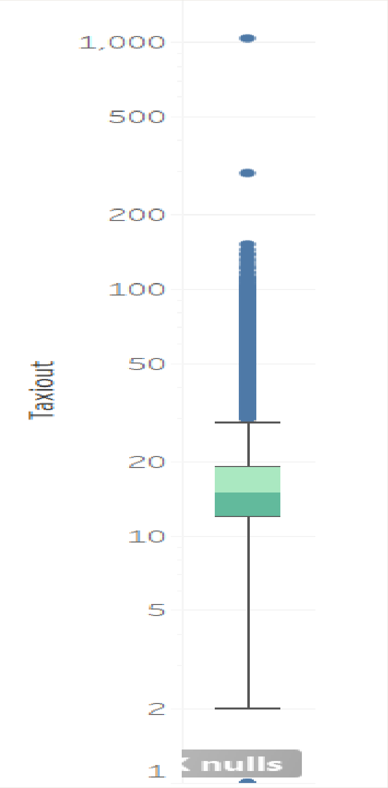
- TAXI IN
- TAXI OUT

NON-CONFIRMING DATA

- AIRTIME has negative values
- Actual Arrival Time has values in 1970
- Departure Gate has gates that do not belong to San Diego Airport

ASSUMPTIONS

TAXIOUT



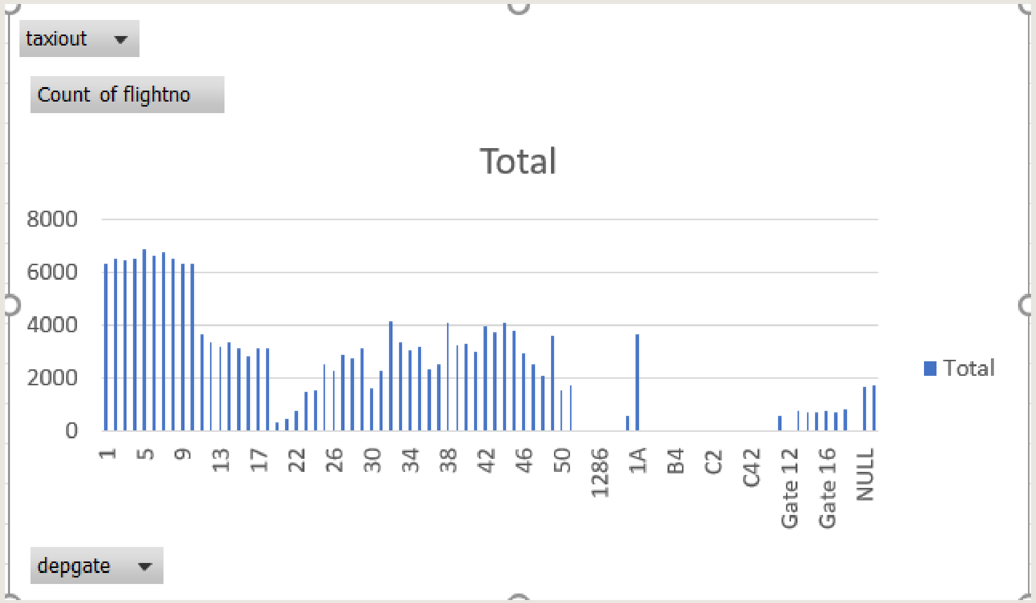
Most values are distributed between 12 and 19 minutes

There are several valid outliers. One invalid outlier above 1000 has been ignored

1112 null values which I attribute to:

- Flight Cancellation (missing actual arrival and departure time)
- Data recording issue
- Airline G4 does not record Taxiout time (58% of null taxi out come from gate 30 and G4 Airline)

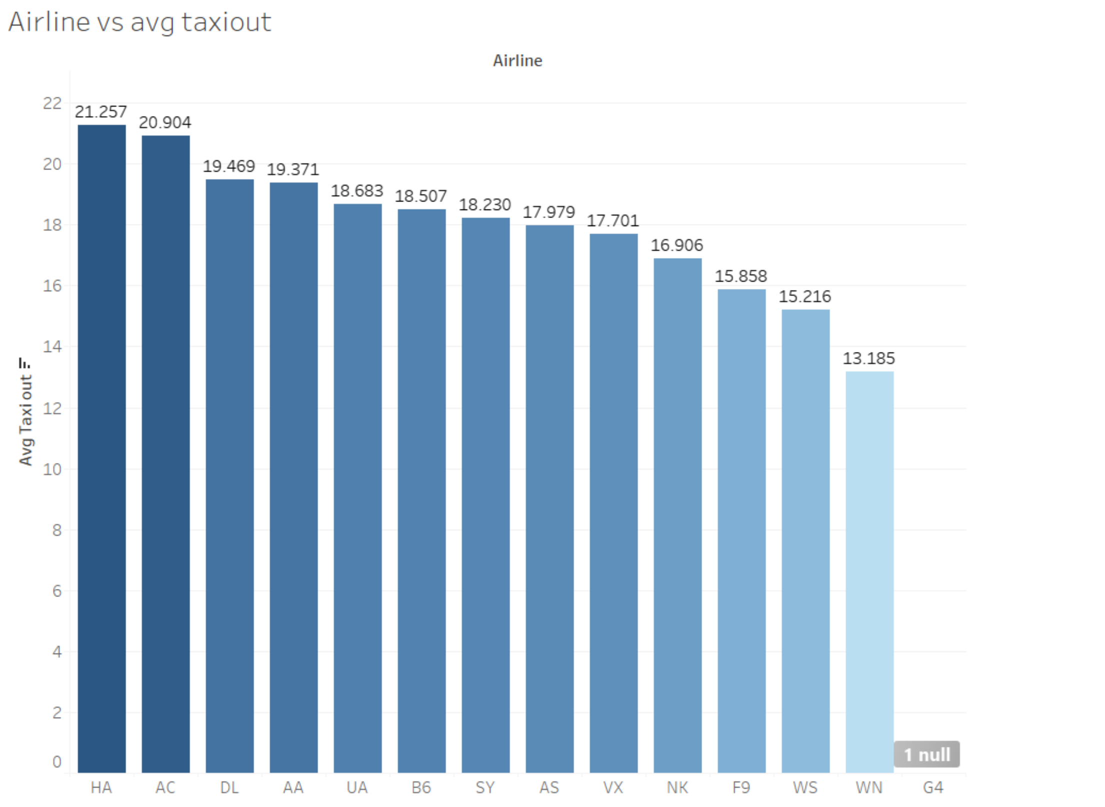
DEPARTURE GATE



- Remove Suffix ‘GATE’
- As per publicly available information SAN DIEGO airport only has gates 1-51 and gate 1A. Therefore all other gates listed in the graph are marked incorrect and ignored.
- Terminal 1: Gates 1-18 and 1A
- Terminal 2 : Gates 19-51

EXPLORATORY DATA ANALYSIS

HOW DOES TAXI OUT TIME VARY BY AIRLINE

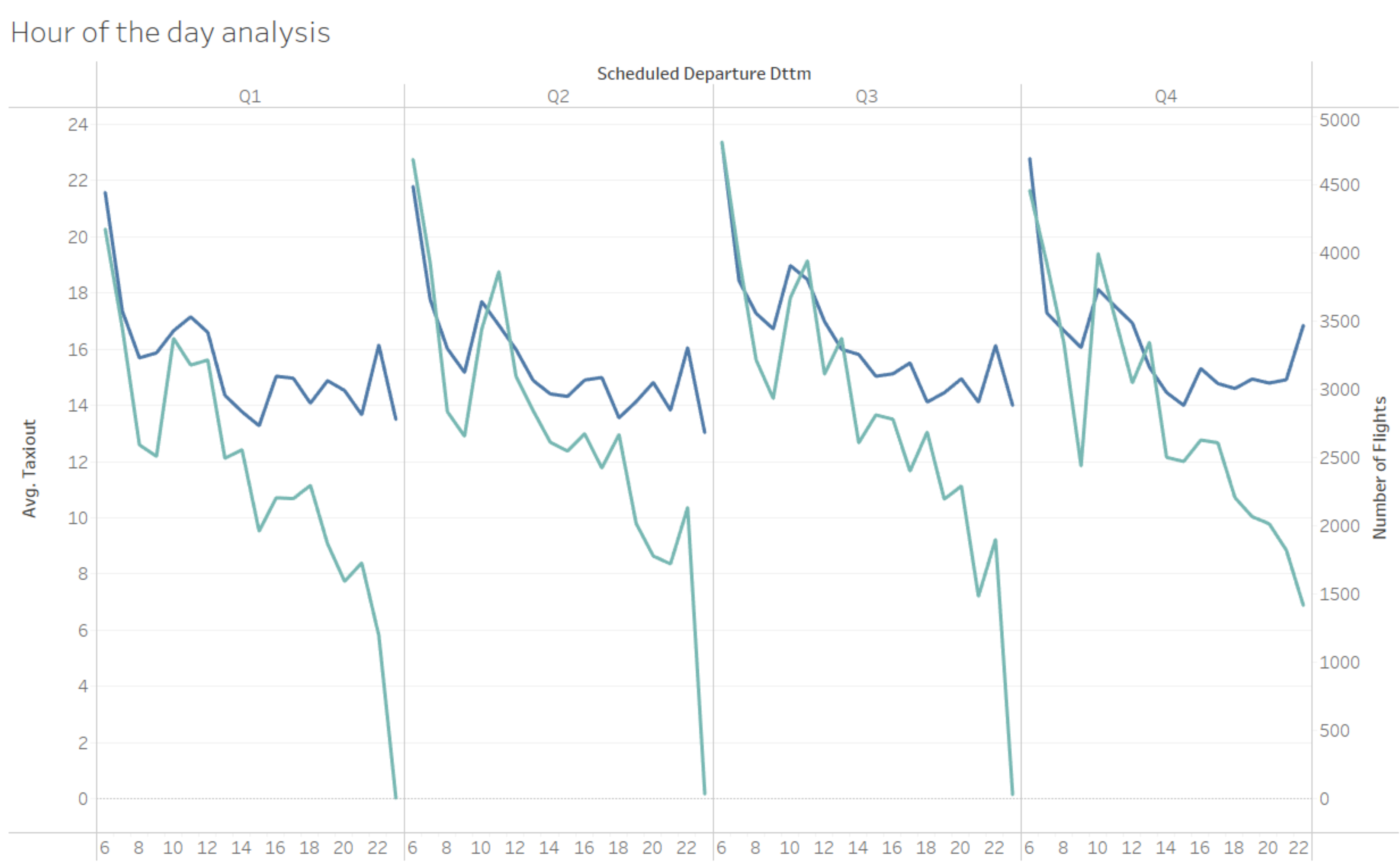


Southwest Airline (WN) has the lowest Taxiout time

Hawaiian Airlines has the highest

Alaska Airlines (AS) has a comparatively low Taxi out time

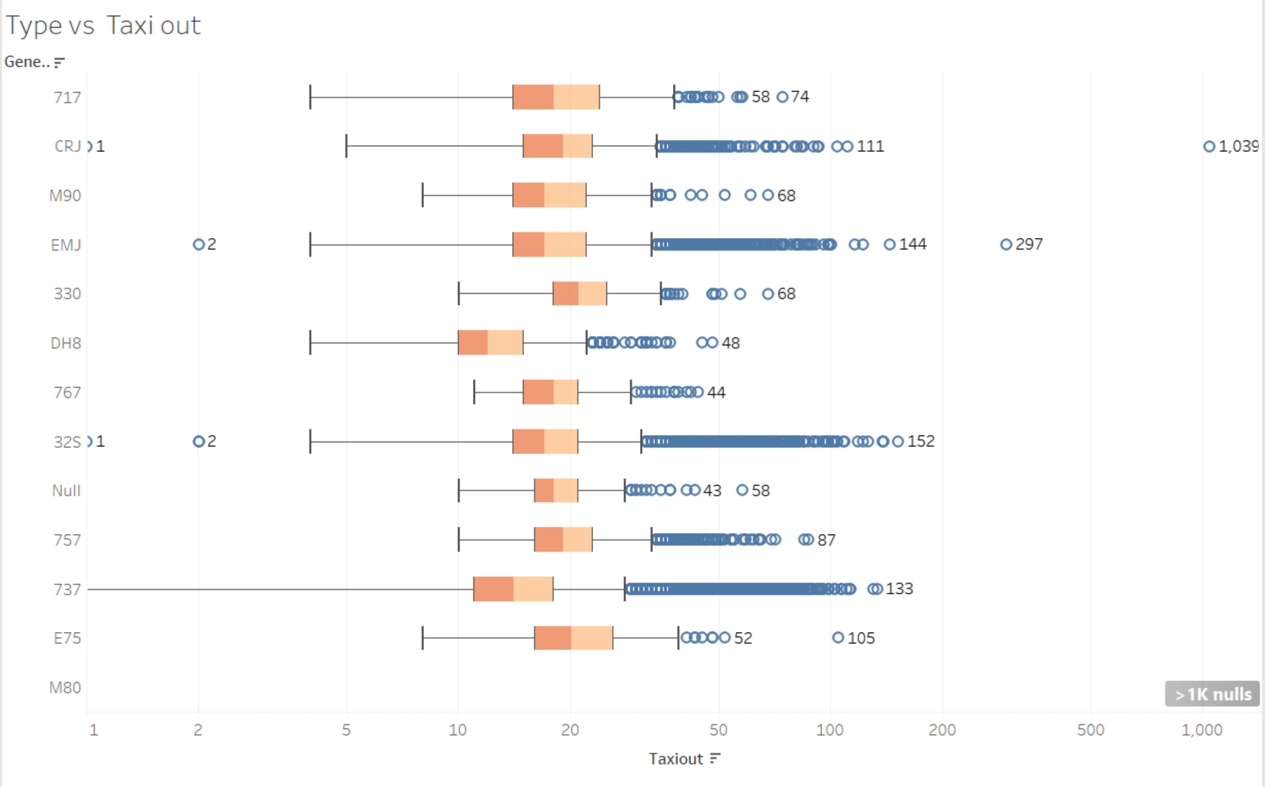
HOW DOES TAXI OUT TIME VARY THROUGH THE DAY?



Influenced by hour of the day.

Also shows some correlation with number of flights departing at that time

IS AIRCRAFT TYPE SIGNIFICANT?



Certain Aricrafts have lower

Taxiout time than others:

737

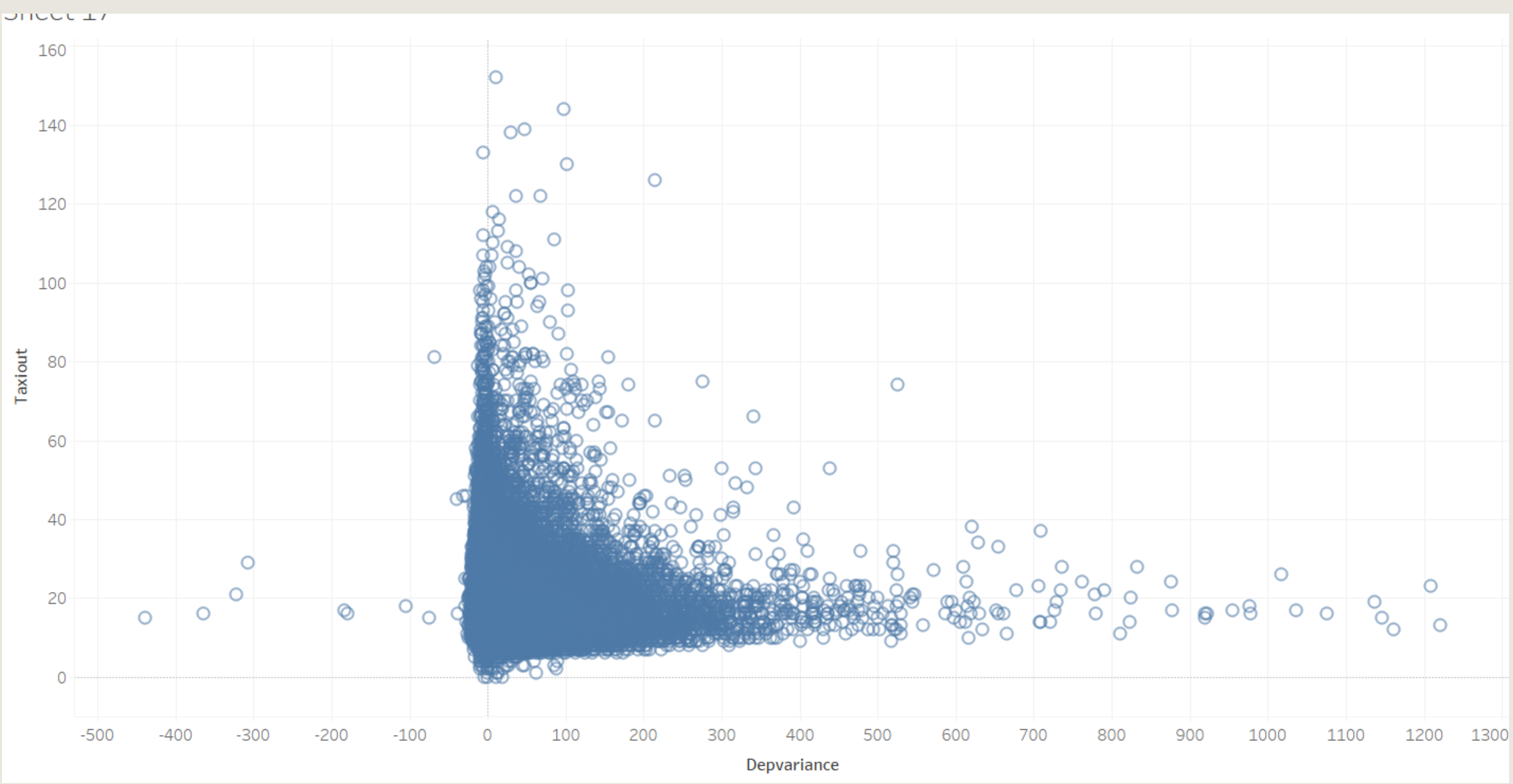
32s

DH8

More than 99% of Southwest

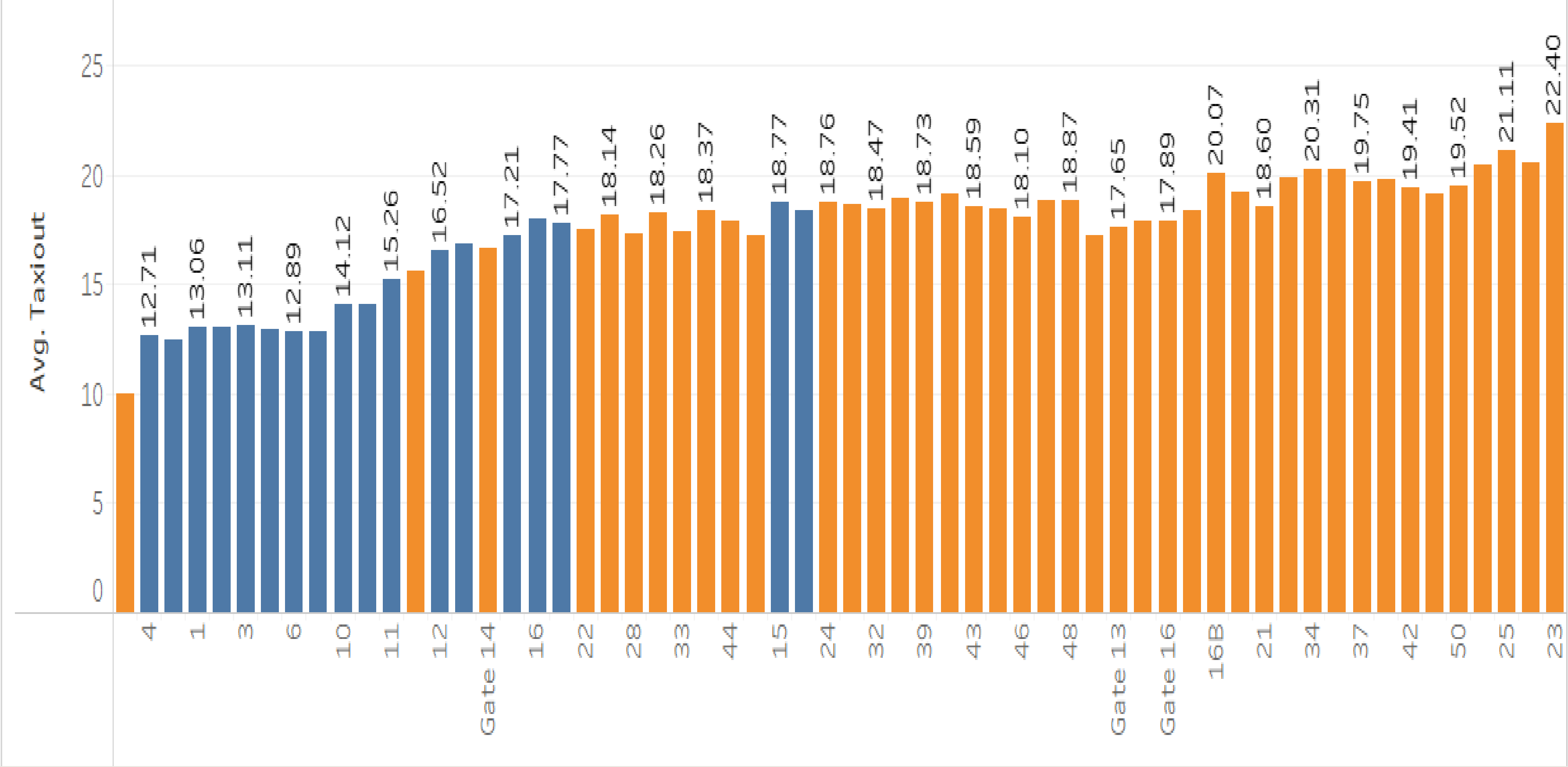
Flights are 737.

DEPARTURE VARIANCE AND TAXIOUT



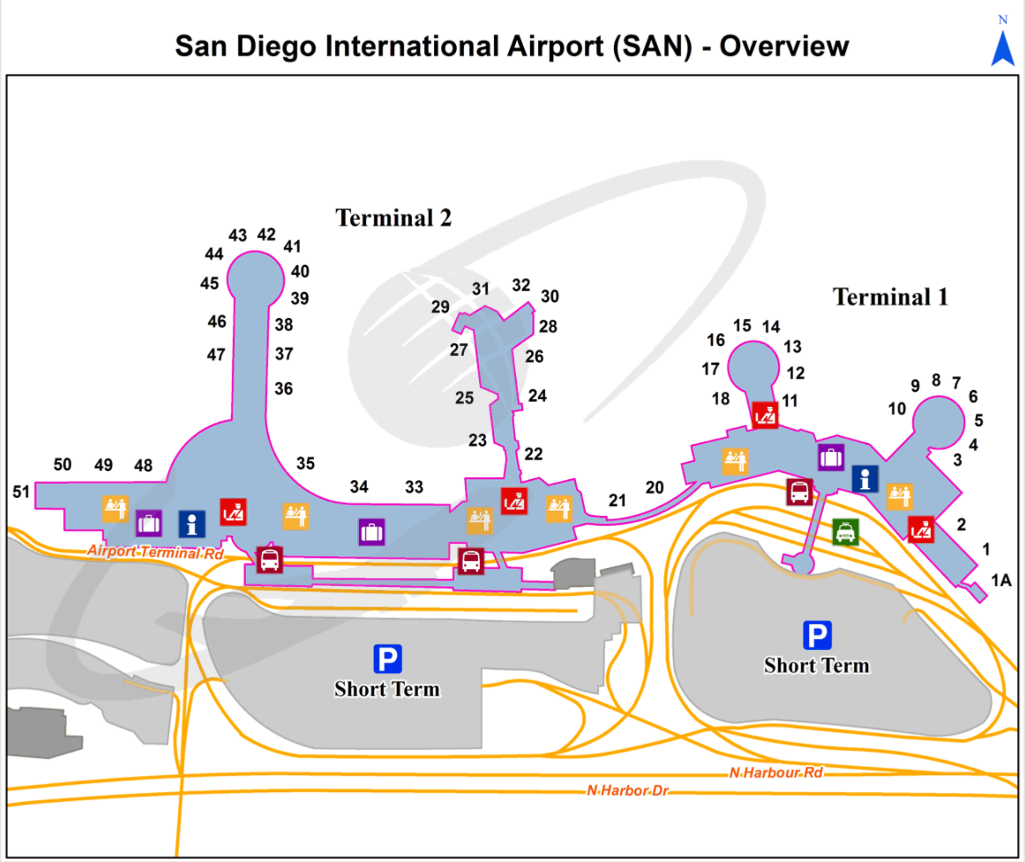
Flights which leave very early and flights which leave very late have low Taxi out times

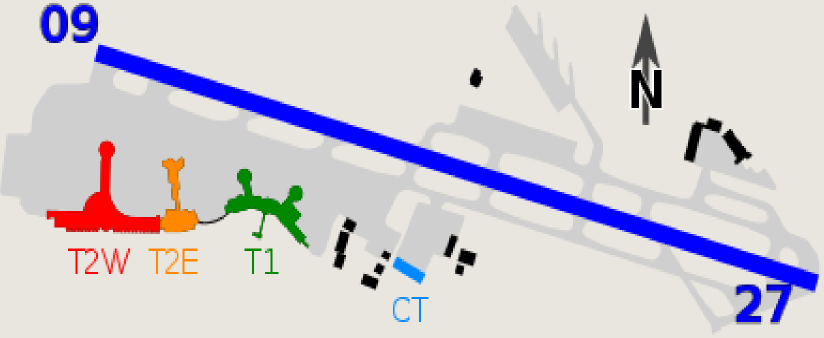
DEPARTURE GATE AND TAXIOUT



In General Terminal 1(Blue) Gates have lower Taxi out than Terminal 2 (Orange) Gates.

WHY IS TERMINAL NUMBER SIGNIFICANT?





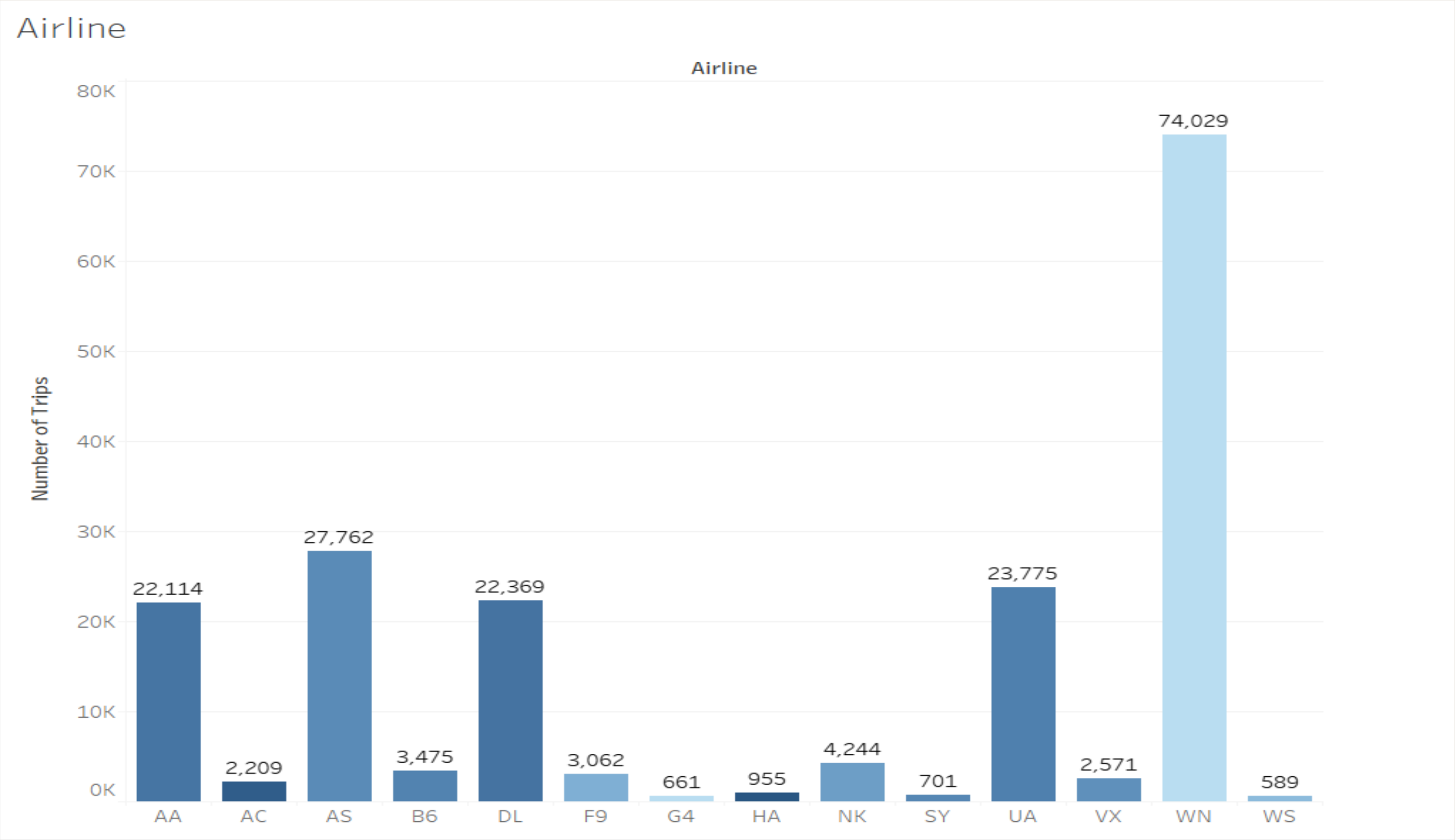
Terminal 1 is closer to Runway 27

which is the commonly used

Runway. The shorter distance helps

reduce Taxi out

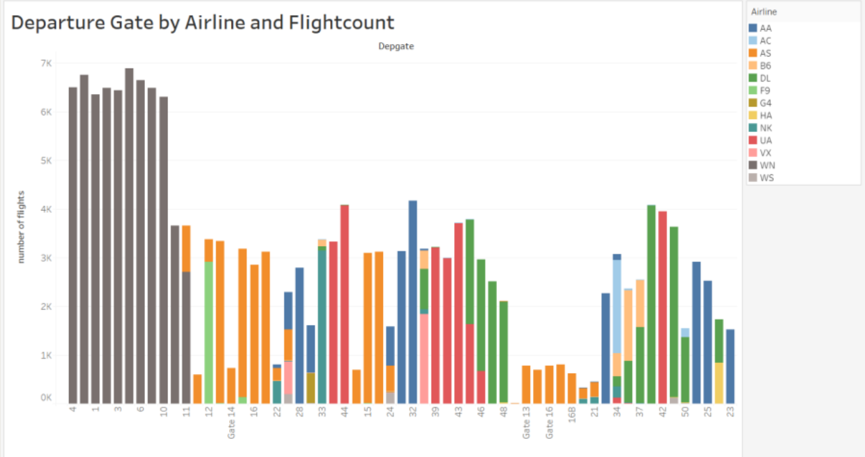
WHY DOES SOUTHWEST AIRLINE HAVE LOW TAXIOUT?



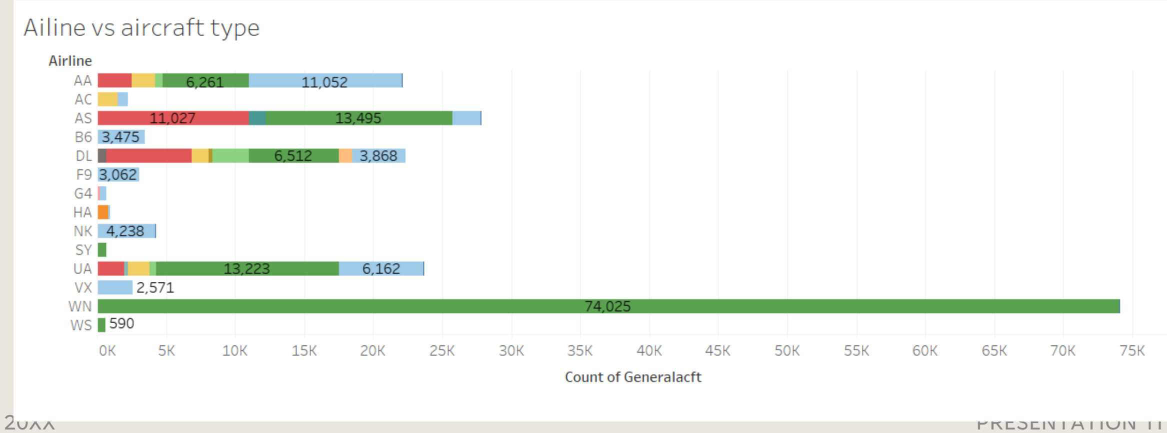
Highest Number of Flights

but lowest Taxi out

POSSIBLE REASONS



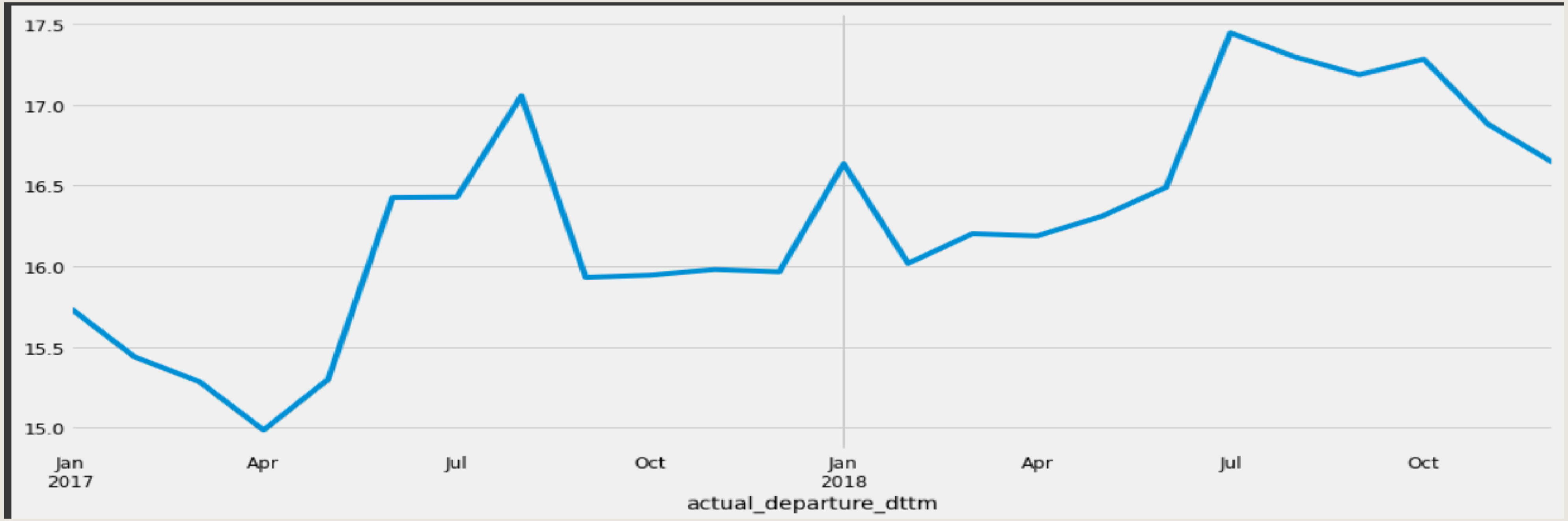
Soutwest Airlines departure gates lie within gates 1- 11 of terminal 1 which is closer to the Runway.



99% of flights are 737s which typically have low taxi outs

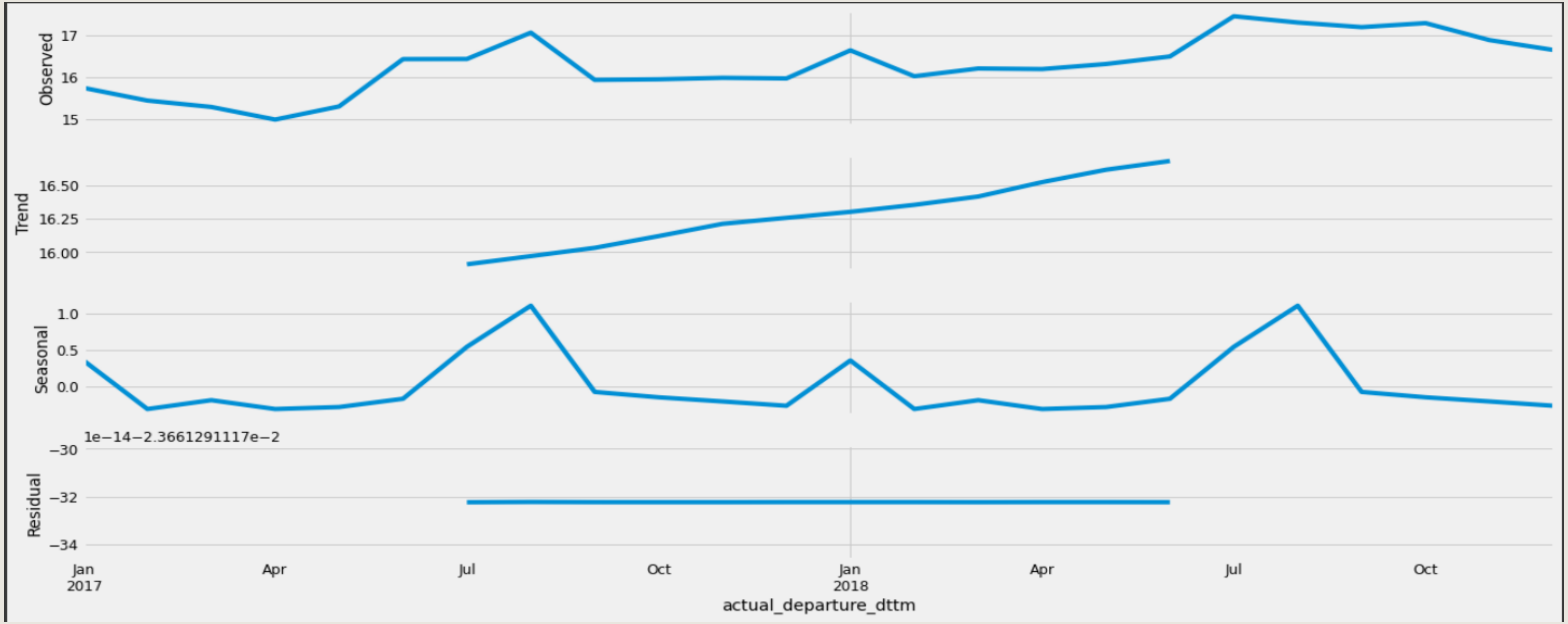
FORECASTING TAXIOUT

- Taxiout is time series Data since we have the value for Distinct Points in time
- Aggregate Taxiout to find monthly average between 2017 and 2018



SEASONALITY AND TREND

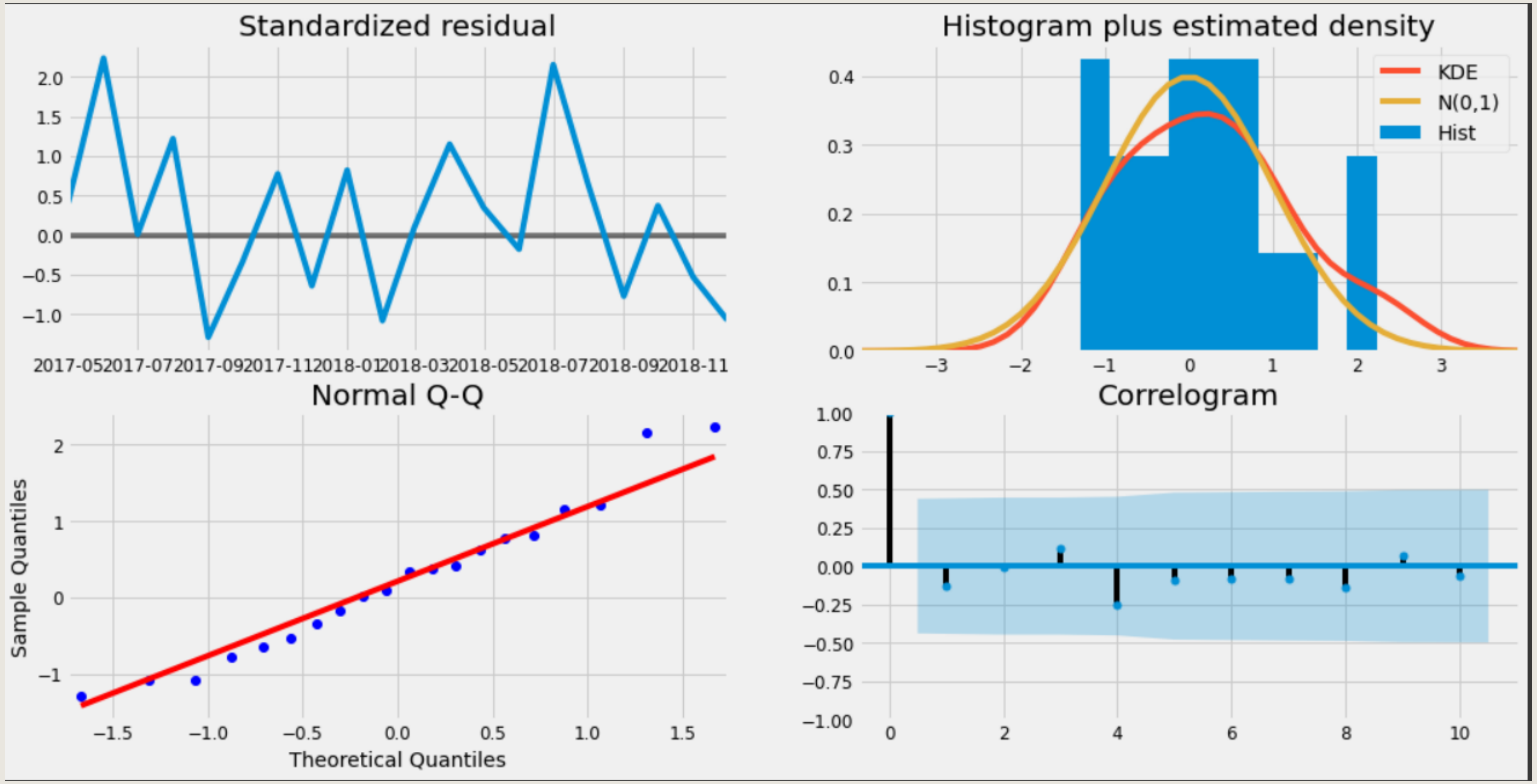
- Taxiout has an upward trend time over time
- It shows minimal seasonality over a period of 12 months
- The data is stationary



LIMITATIONS AND CONSIDERATIONS

- Method: Auto Regressive Integrated Moving Average
- In particular the SARIMAX model in the STATSMODELS Library is used
- The model requires the data to be stationary.
- Trend and seasonality are accounted for
- The autoregressive parameter p and the moving average parameter q and the lag d are chosen by looking at the Akaike Information Criterion (AIC). A lower AIC is preferred and grid search is used to find possible values of p,d and q
- The optimal p,d and q values are chosen finally based on the model evaluation metric ‘Mean Squared Error’
- Final Values of (p,d,q) chosen are (3,1,2) with a seasonality of 12 to represent yearly seaonality.

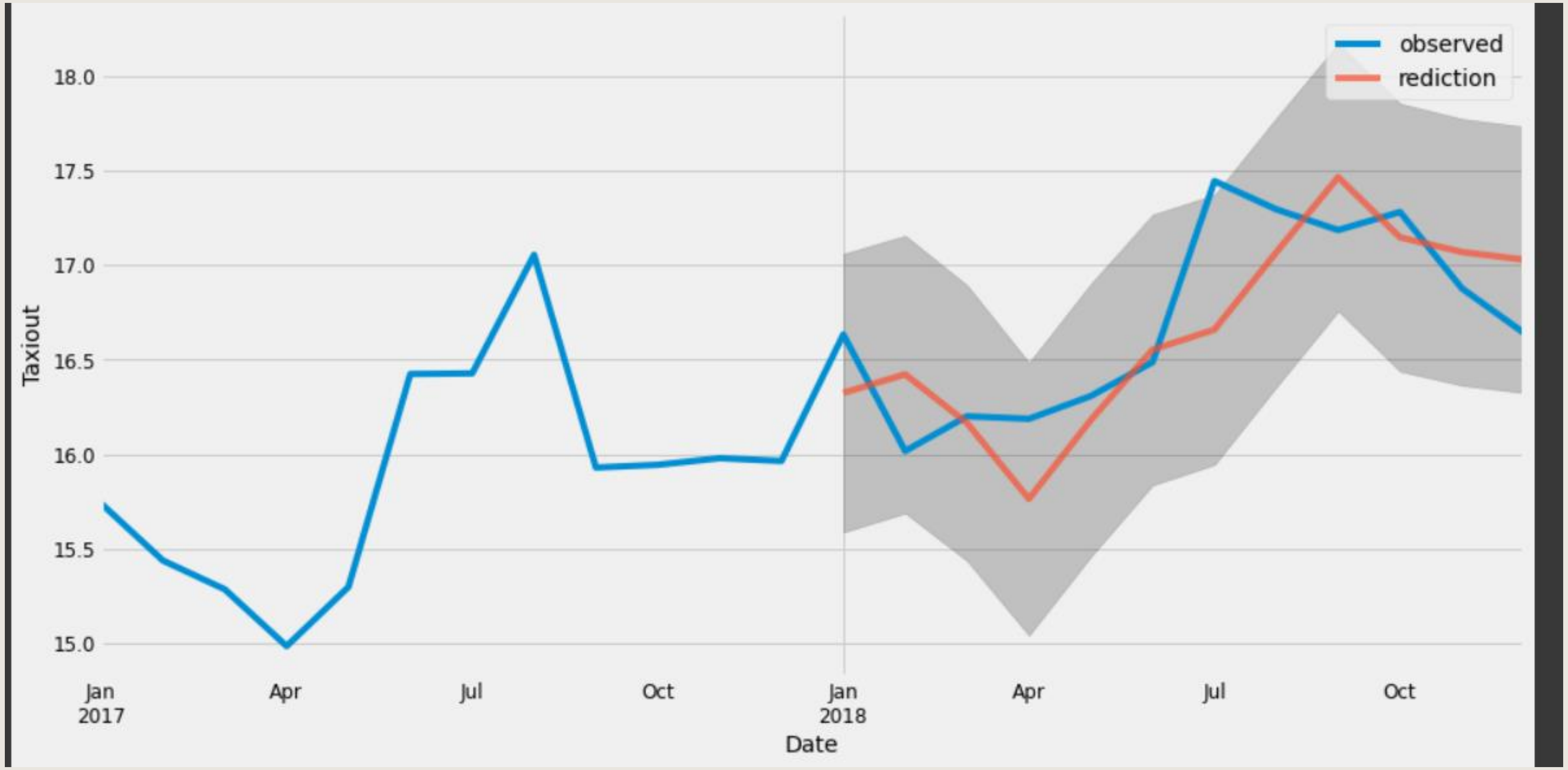
MODEL DIAGNOSTICS



INDICATORS OF A ROBUST MODEL

- Standardized residual resembles white noise centered around 0
- Smoothed Histogram of Residual closely resembles a normal curve
- Points in the Normal Q-Q plot lie along the trend line
- More than 95% of the correlations for lag greater than zero are insignificant (98% of values lie in the blue shaded region)

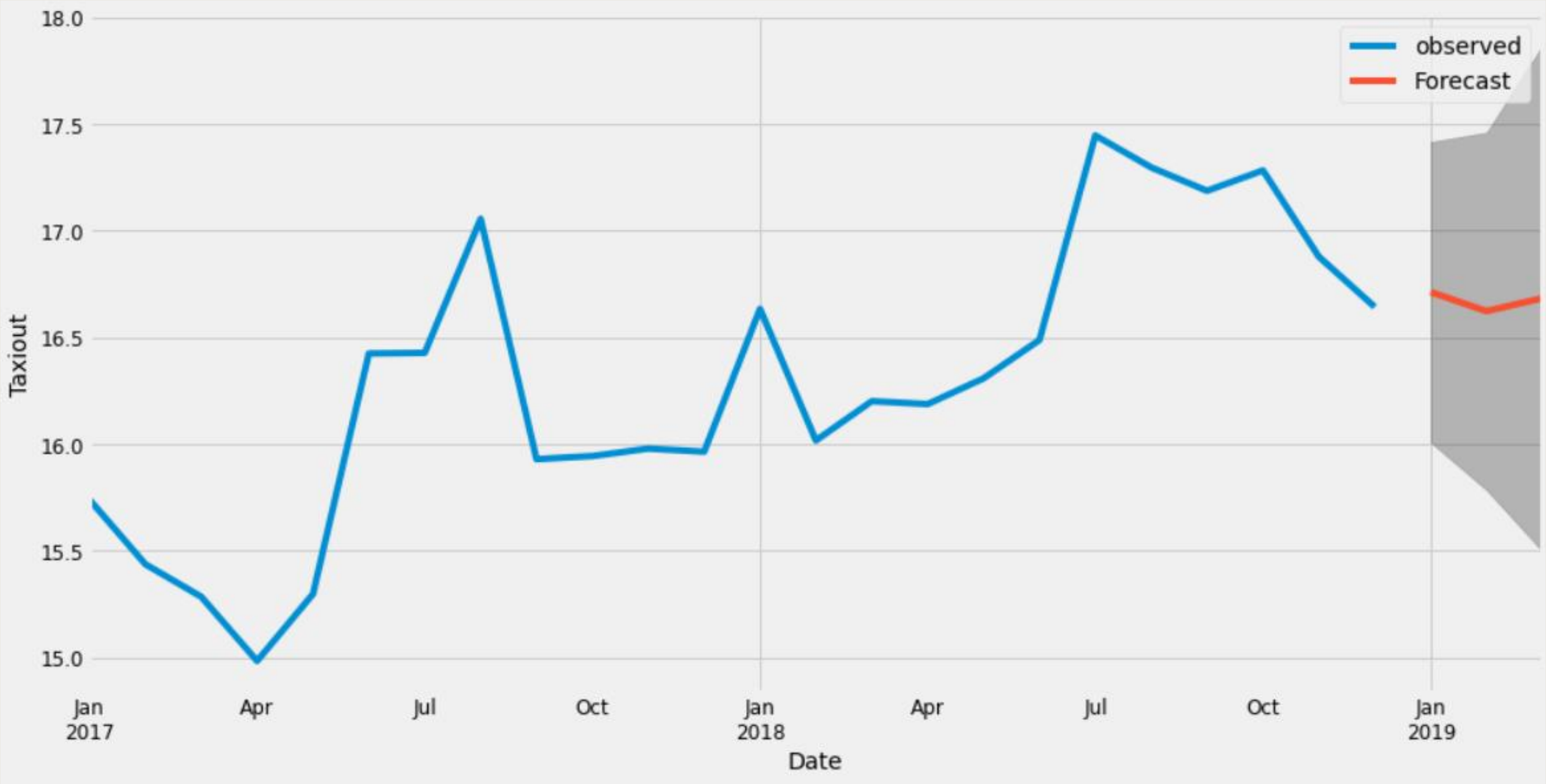
TESTING THE MODEL: PREDICTING TAXIOUT FOR 2018



Taxiout values for 2018 are predicted using the model and compared with observed known values:

- The Mean squared Error (MSE) is 0.12
- Root Mean Squared Error (RMSE) is 0.34

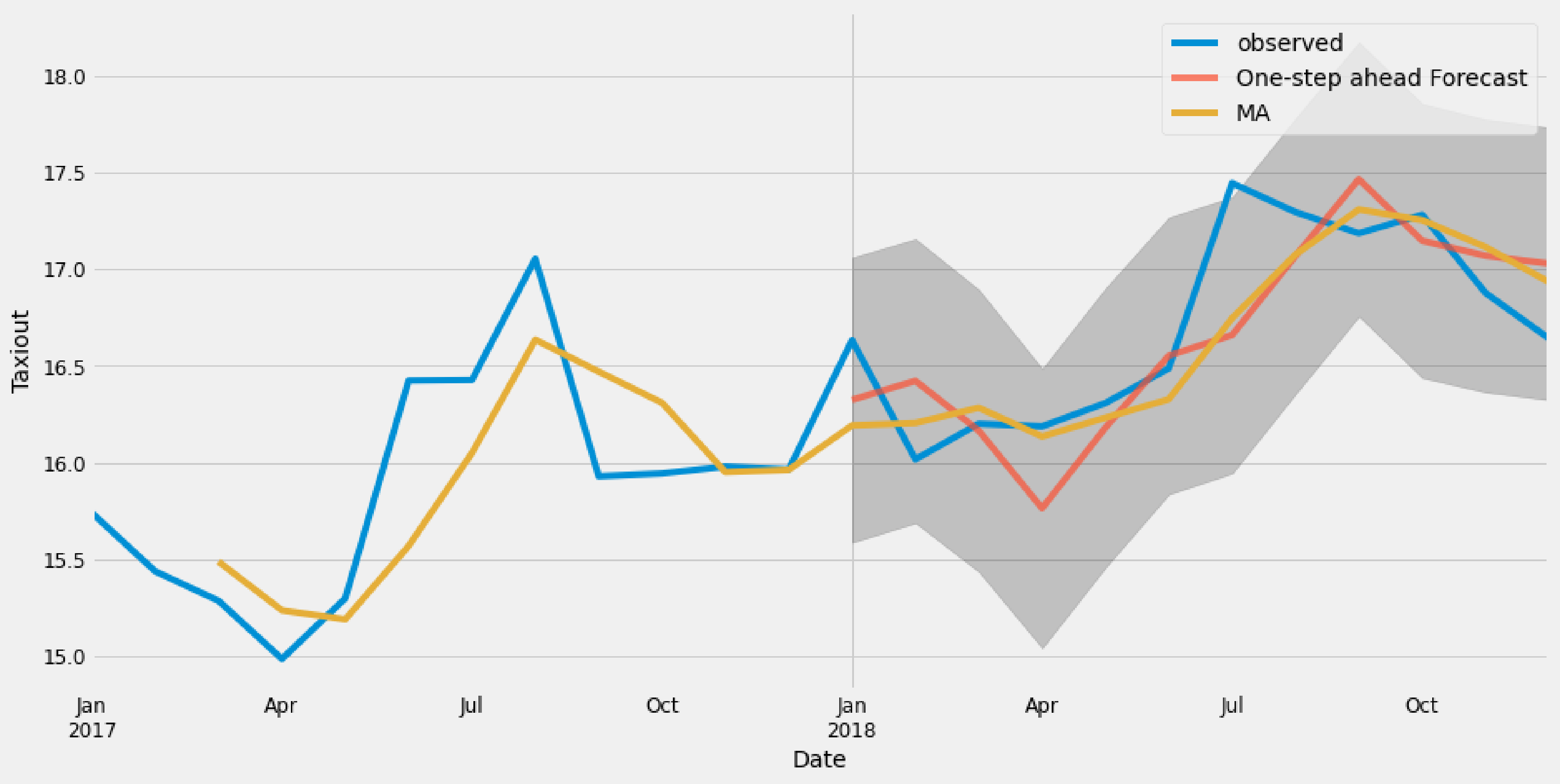
FORECASTING FOR 2019 JAN-MAR



The model forecasts the following values of Taxiout for 2019 (JAN-MAR)

January : 16.712616 minutes
February : 16.623003 minutes
March : 16.683487 minutes

ALTERNATIVE APPROCHES



Moving average calculated over recent values(past 3 months) to calculate the Taxiout value.

Comparing with ARIMA, both methods have the same MSE of 0.12 however Moving Average does not consider seasonality.

TAKEAWAYS

MOST SIGNIFICANT PARAMETERS FOR TAXIOUT

- Time of the Day
- Number of Flights
- Departure Gate

HOW TO IMPROVE FORECAST

- More Data (5 to 10 years worth of Data)
- Including other significant variables such as Number of Flights in SARIMAX model as an exogenous variable to increase accuracy

THANK YOU

Chandhnee Karthikeyan Iyer

ckiyer@uw.edu