Predicting Flight Departure Delays in an Air Traffic Network

Capstone Project 1

The problem

Potential Clients

- Companies distributing use of the Aircraft Situation Display to Industry (ASDI)
- Air Route Traffic Control Centers (ARTCC) looking to make better predictions of AAR's and plan efficient holding patterns.
- Implementers of Ground Delay Programs at Domestic Airports
- Traffic Management Personnel seeking to get the earliest Expect Departure Clearance Time (EDCT)

Context - Cost of Delay

- $40 billion in annual cost due to delay s in 2015
- $62.55 per minute average cost of aircraft block (taxi plus airborne) in 2016
- In 2016 nearly 40% of delays were due to the delayed arrival of the incoming aircraft, reflecting the high levels of interdependence in the delay dynamics

Problem statement

Predict whether or not the departure delay on an Origin-Destination pair, with a 6-hour prediction horizon, will exceed a 15 min threshold; such that the future delay will fall into one of two classes, ‘above threshold’ (1) and ‘below threshold’ (0).

Methodology

Clean Data

Wrangling

1. Improve readability and address missing values
2. Aggregate to create mock network
3. Create target variable of departure delay state 6 hours in the future

Visualize

Storytelling

1. Flight Frequency
2. Delay Distribution
3. On-Time Performance
4. Network Visualization

Analyze

Inferential Statistics

1. Test for Normality and CLT in variables of interest
2. Regression Analysis
3. Hypothesis Tests

Construct Model

Binary Classification

1. Baseline Logistic Regression Classifier
2. Resample data to address target class imbalance, both under- and over-sampling
3. Train Logistic Regression and Random Forest Classifiers under both resampling conditions

Wrangling

Surface Cleaning and Pre-Processing

1. Import data and address missing values
2. Compartmentalize columns and evaluate by category, then merge results into final Flights dataframe
3. Create NetworkX Digraph from components
4. Create mock air traffic network, a dataframe named Links_d, containing target class for model construction

Flights DataFrame

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 1824403 entries, (ABE-ATL, 2016-01-01 07:00:00) to (YUM-PHX, 2016-12-31 19:15:00)
Data columns (total 39 columns):
quarter                  int64
month                    int64
day_of_month             int64
day_of_week              int64
fl_date                  object
Day_of_Week              object
Month                    object
dt_index                 datetime64[ns]
hour_of_day              object
unique_carrier           object
fl_num                   int64
origin_airport_id        int64
origin_city_market_id    int64
origin                   object
origin_city_name         object
origin_state_abr         object
origin_state_nm          object
dest_airport_id          int64
dest_city_market_id      int64
dest                     object
dest_city_name           object
dest_state_abr           object
dest_state_nm            object
link                     object
unique_carrier_nm        object
crs_dep_time             int64
dep_time                 float64
dep_deviation            float64
dep_delay                float64
wheels_on                float64
taxi_in                  float64
crs_arr_time             int64
arr_time                 float64
arr_deviation            float64
arr_delay                float64
crs_elapsed_time         float64
actual_elapsed_time      float64
air_time                 float64
distance                 float64
dtypes: datetime64[ns](1), float64(12), int64(11), object(15)
memory usage: 554.5+ MB
```

Links_d DataFrame

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 882815 entries, (ANC-SEA, 2016-01-01 00:00:00) to (TPA-ATL, 2016-12-31 19:00:00)
Columns: 561 entries, crs_dep_time to dd_binary
dtypes: float64(15), int32(1), uint8(545)
memory usage: 566.6+ MB
None
Index(['crs_dep_time', 'dep_time', 'dep_deviation', 'dep_delay', 'wheels_on',
       'taxi_in', 'crs_arr_time', 'arr_time', 'arr_deviation', 'arr_delay',
       ...
       'dest_city_name_San Diego', 'dest_city_name_San Jose',
       'dest_city_name_San Juan', 'dest_city_name_Seattle',
       'dest_city_name_St. Louis', 'dest_city_name_Tampa',
       'dest_city_name_Washington',
       'dest_city_name_West Palm Beach/Palm Beach', 'dd_in_6hrs', 'dd_binary'],
      dtype='object', length=561)
```

Storytelling

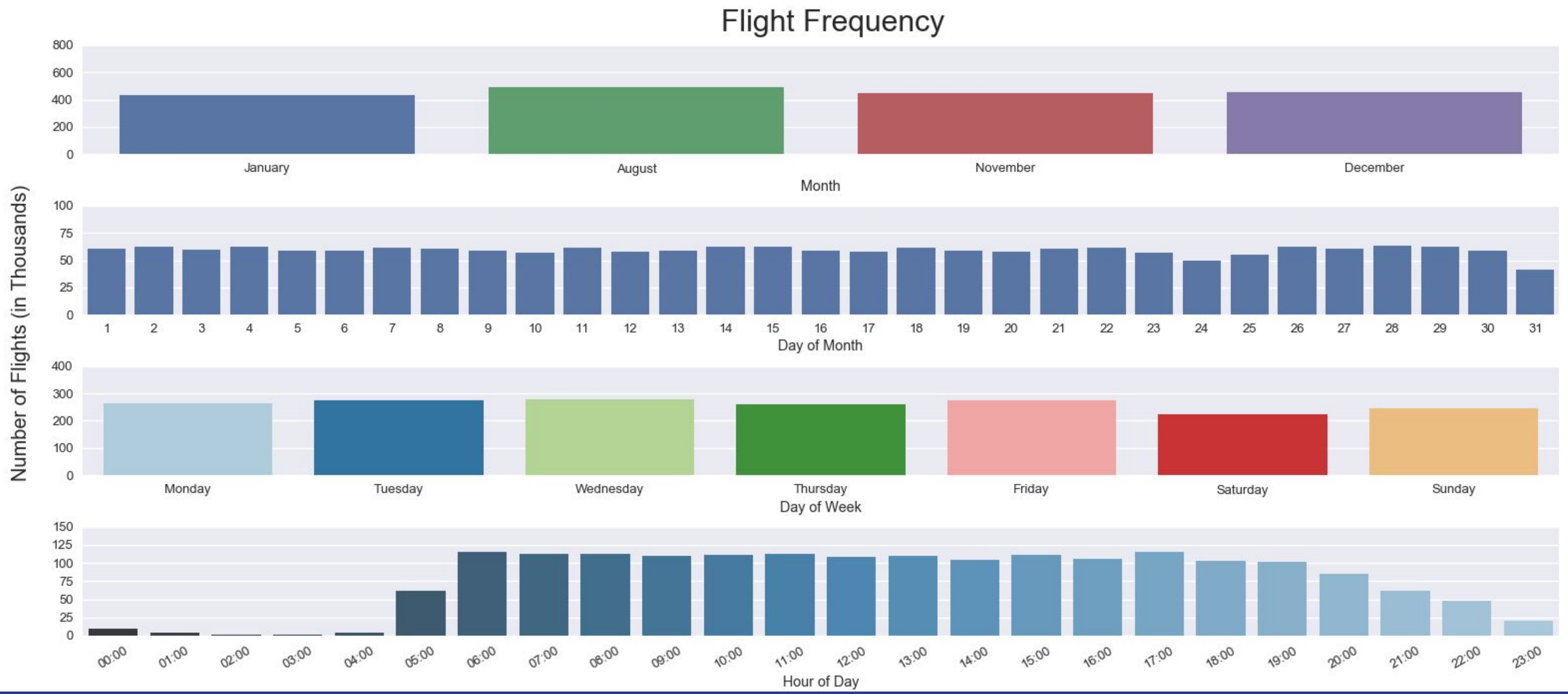
Exploratory Data Analysis

1. Flight Frequency
   - By Temporal Categories
   - By Commercial Airline Carrier
   - By Location Attributes
2. Delay Distribution
   - Overall
   - By Commercial Airline Carrier
   - For Highest Traffic Airports
3. On-Time Performance
4. Network Visualization

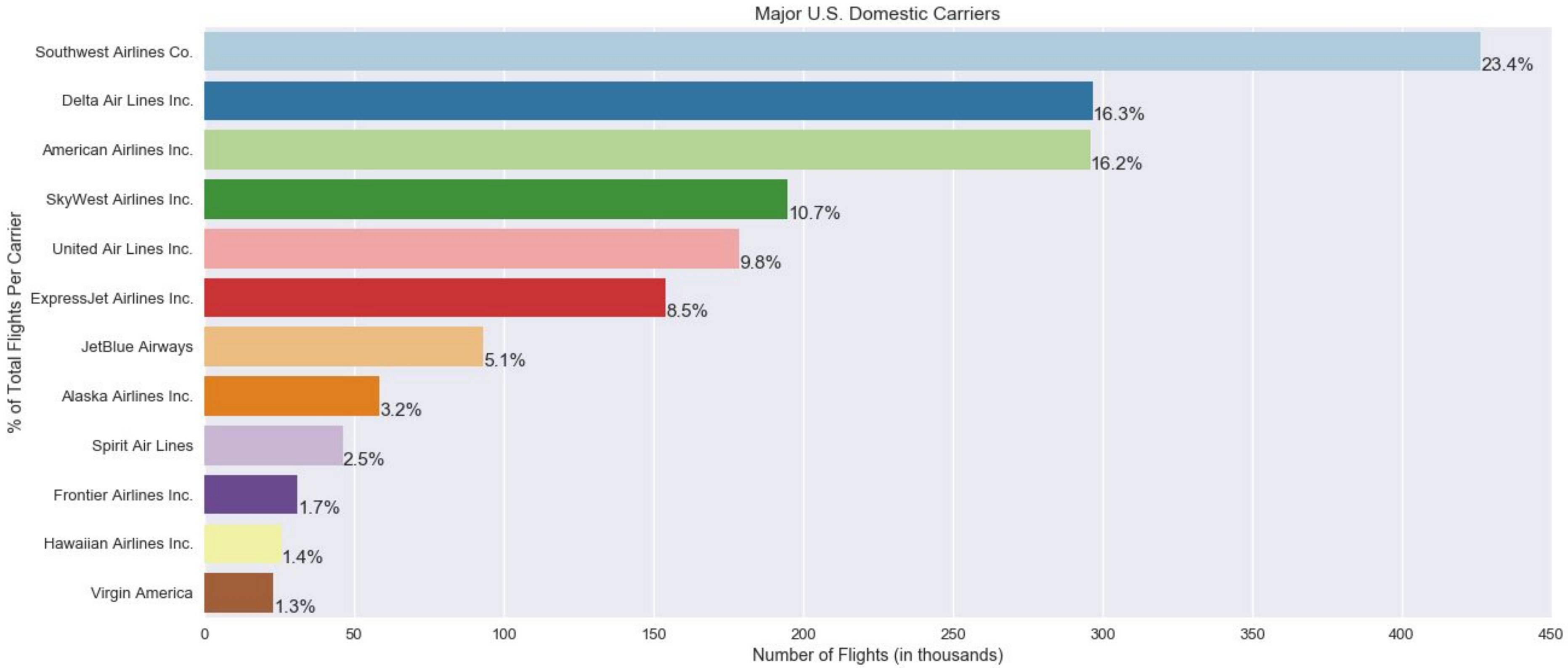
Air Traffic Statistics Explained

- Sample contains all commercial flights in January, August, November, and December of 2016, from the 12 Major Domestic Passenger Airlines.
- The U.S. Department of Transportation defines a major carrier or major airline carrier as a U.S.-based airline that posts more than $1 billion in revenue during fiscal year, grouped accordingly as “Group III” (“Air Carrier Groupings 2016”, U.S. Bureau of Transportation)
- A flight is counted as “on time” if it operated less than 15 minutes after the scheduled time show in the carriers’ Computerized Reservation Systems (CRS), with departure and arrival times being calculated from gate to gate, not including taxi or airtime.

1.1 Frequency by Time



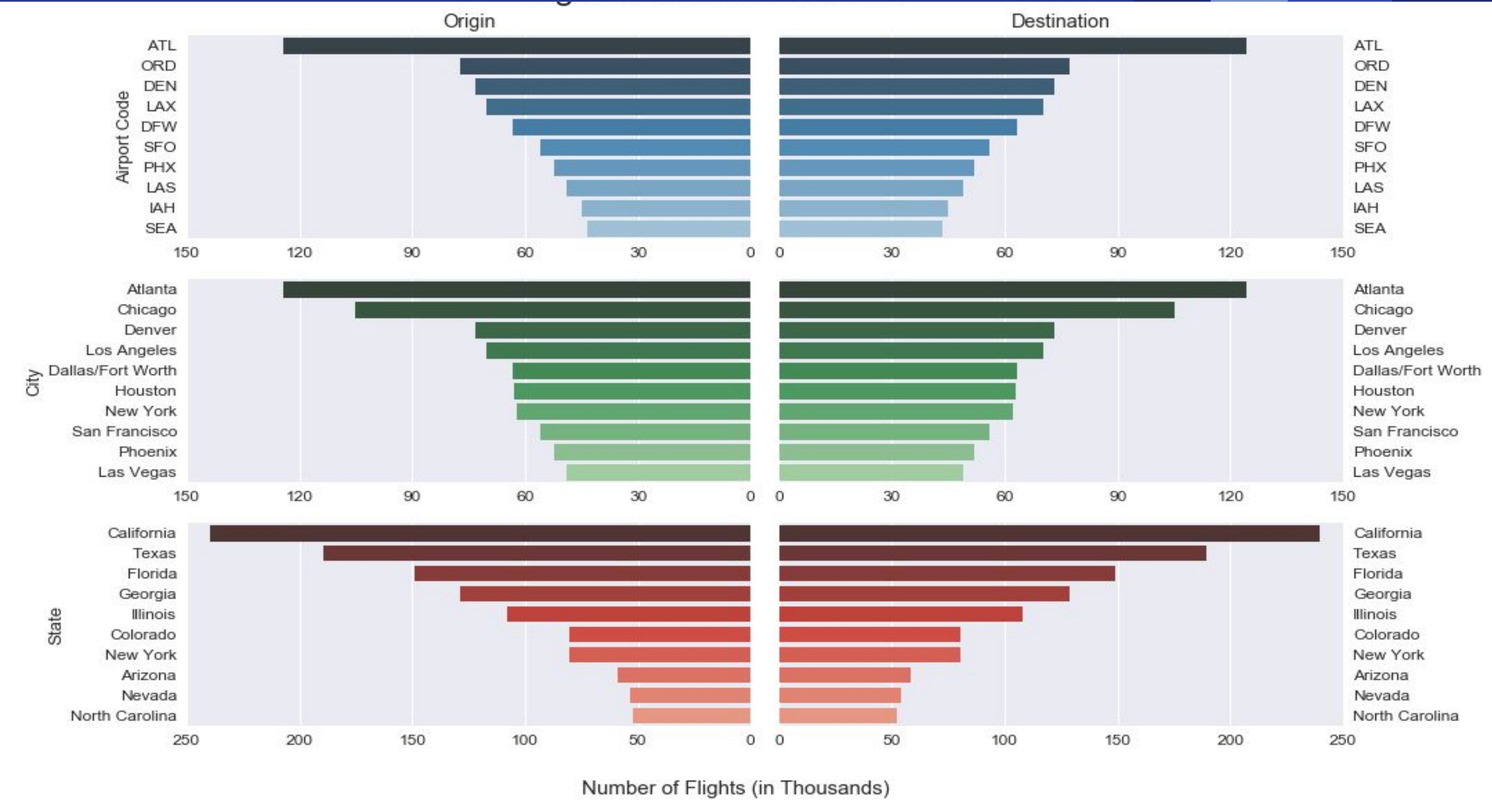
1.2 Frequency by Carrier



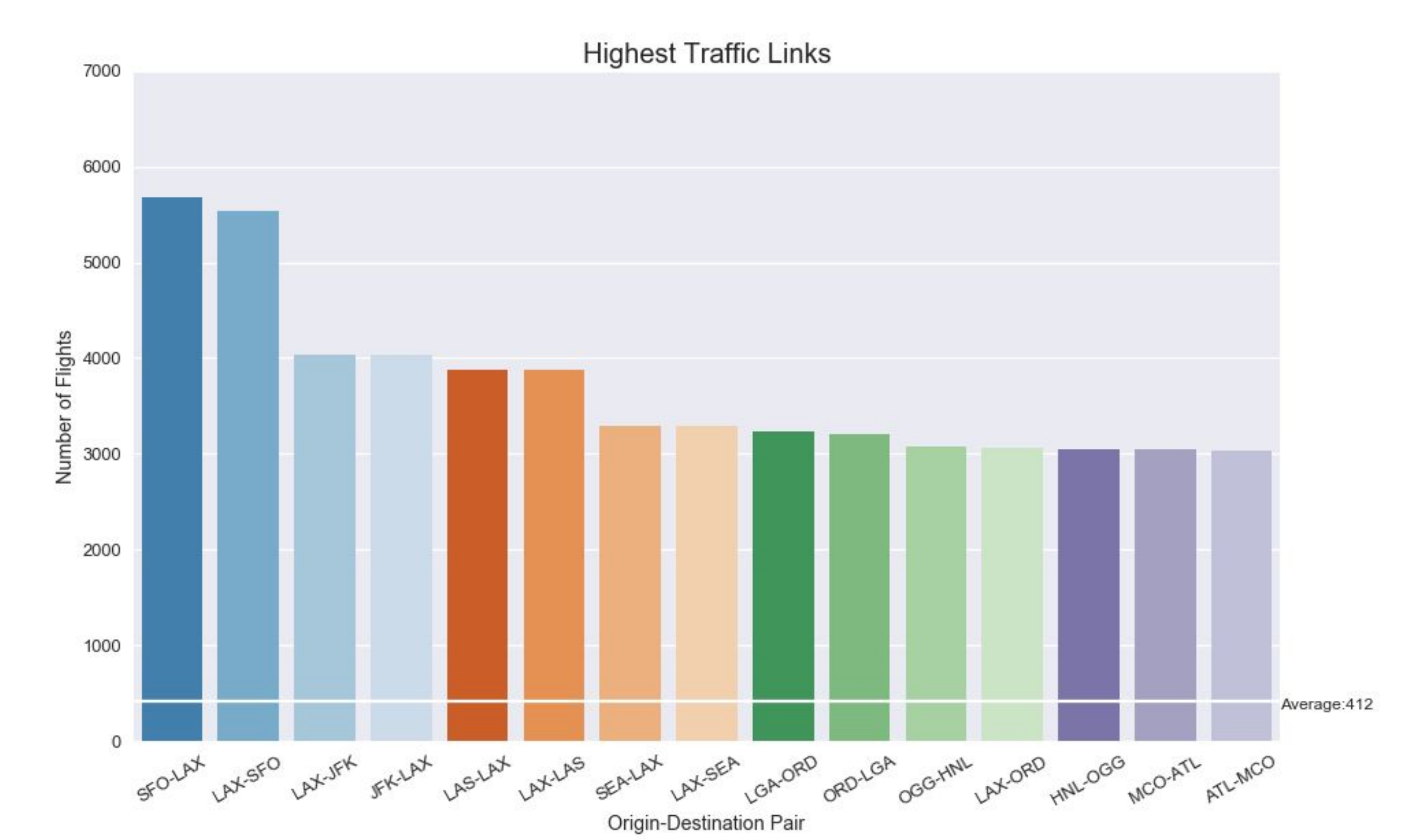
1.3 Frequency by Location

	Number of Locations Reported
Origin Airport	309
Origin City	297
Origin State	52
Destination Airport	308
Destination City	297
Destination State	52
Origin-Destination Pair (Link)	4431

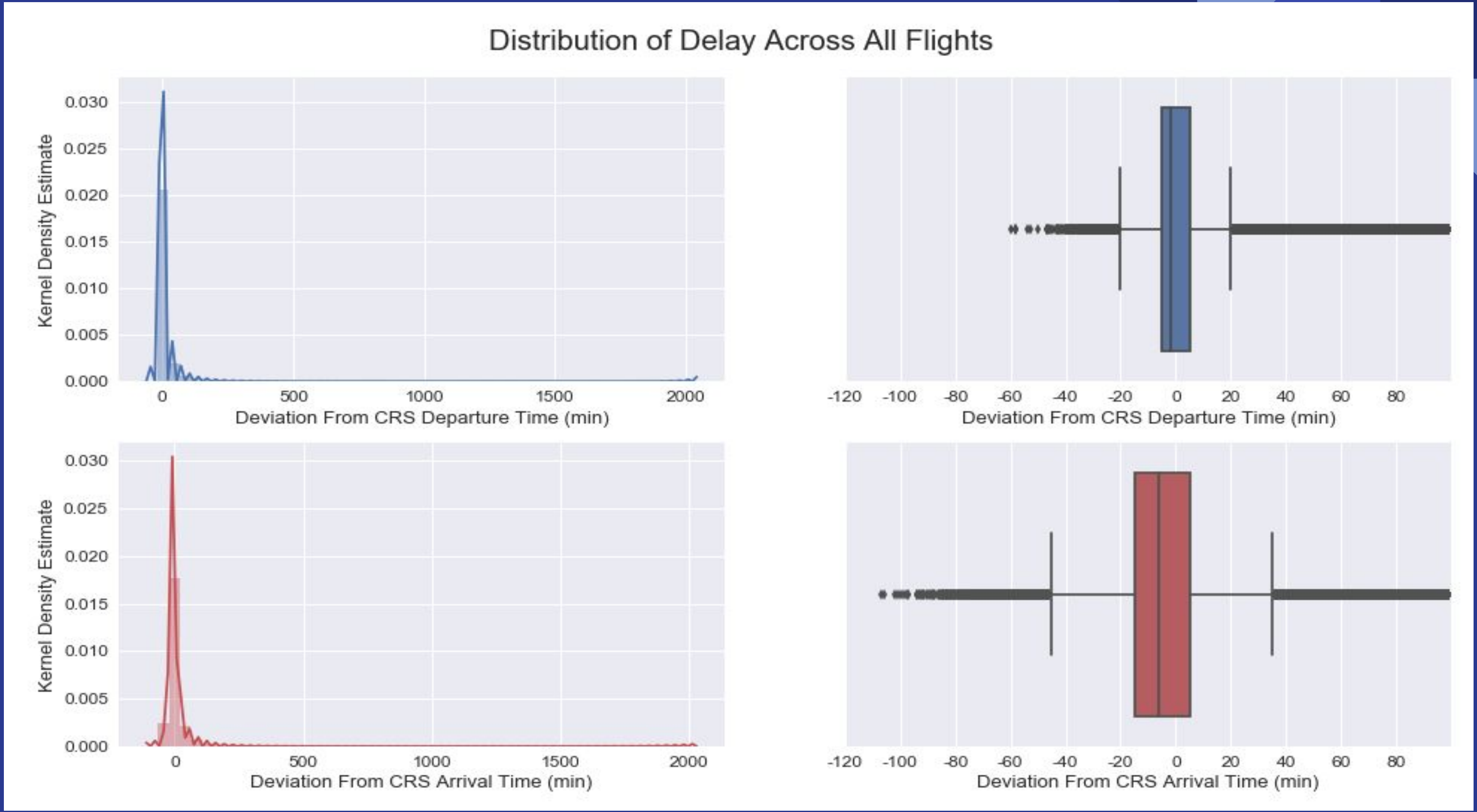
1.3 Highest Traffic Locations



1.3 Highest Traffic Origin-Destination Pairs

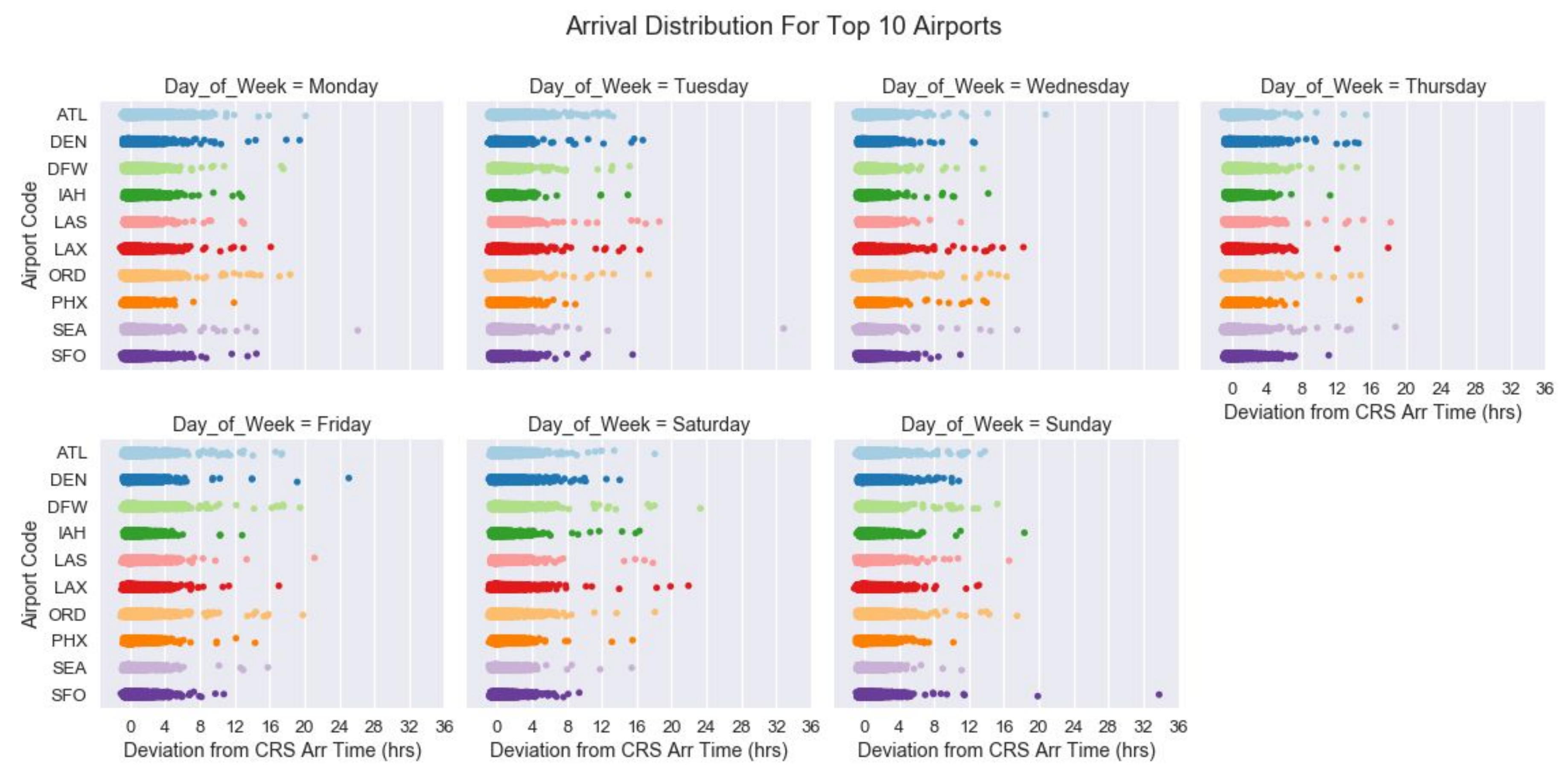


2.1 Overall Delay Distribution

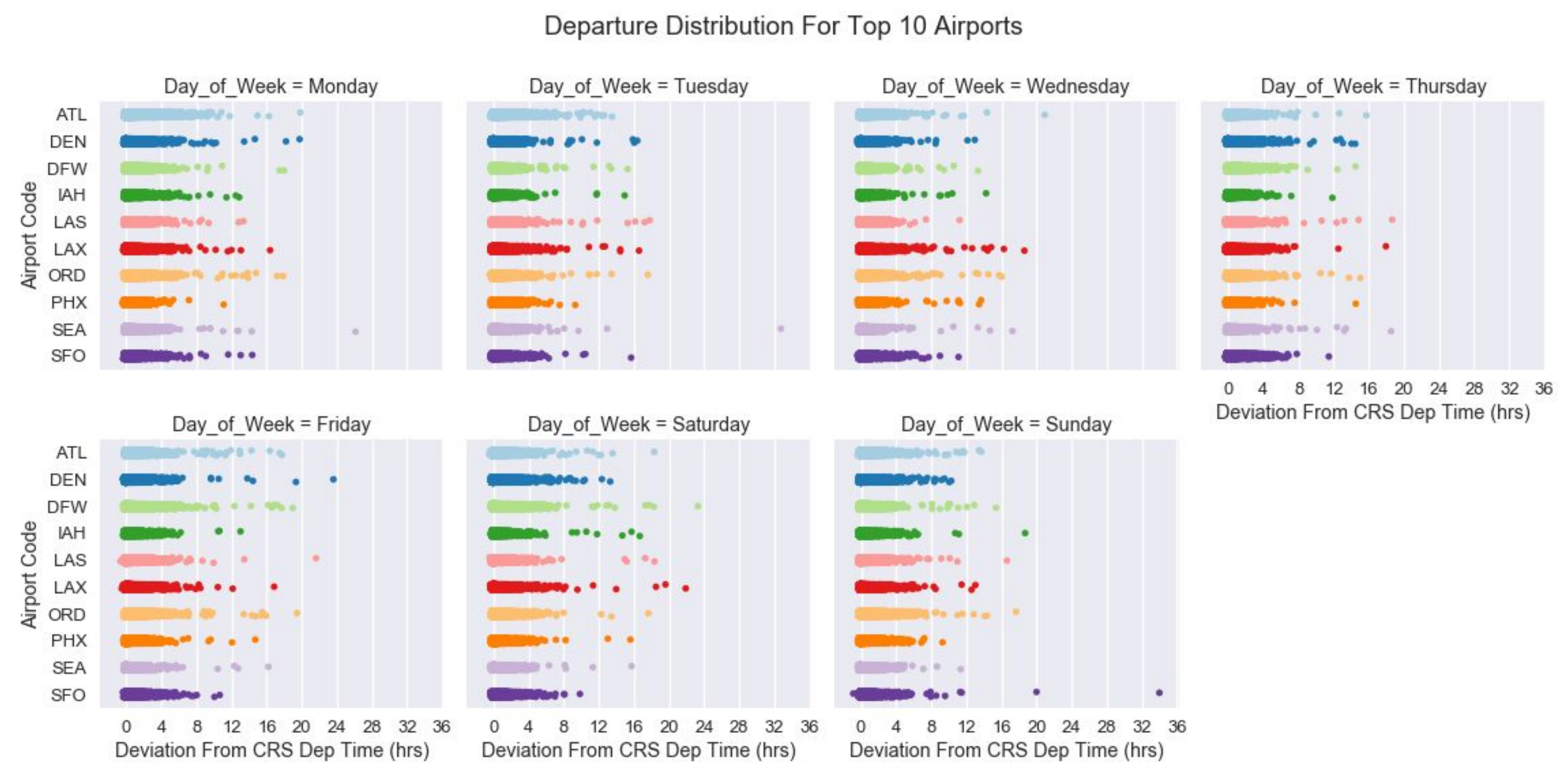


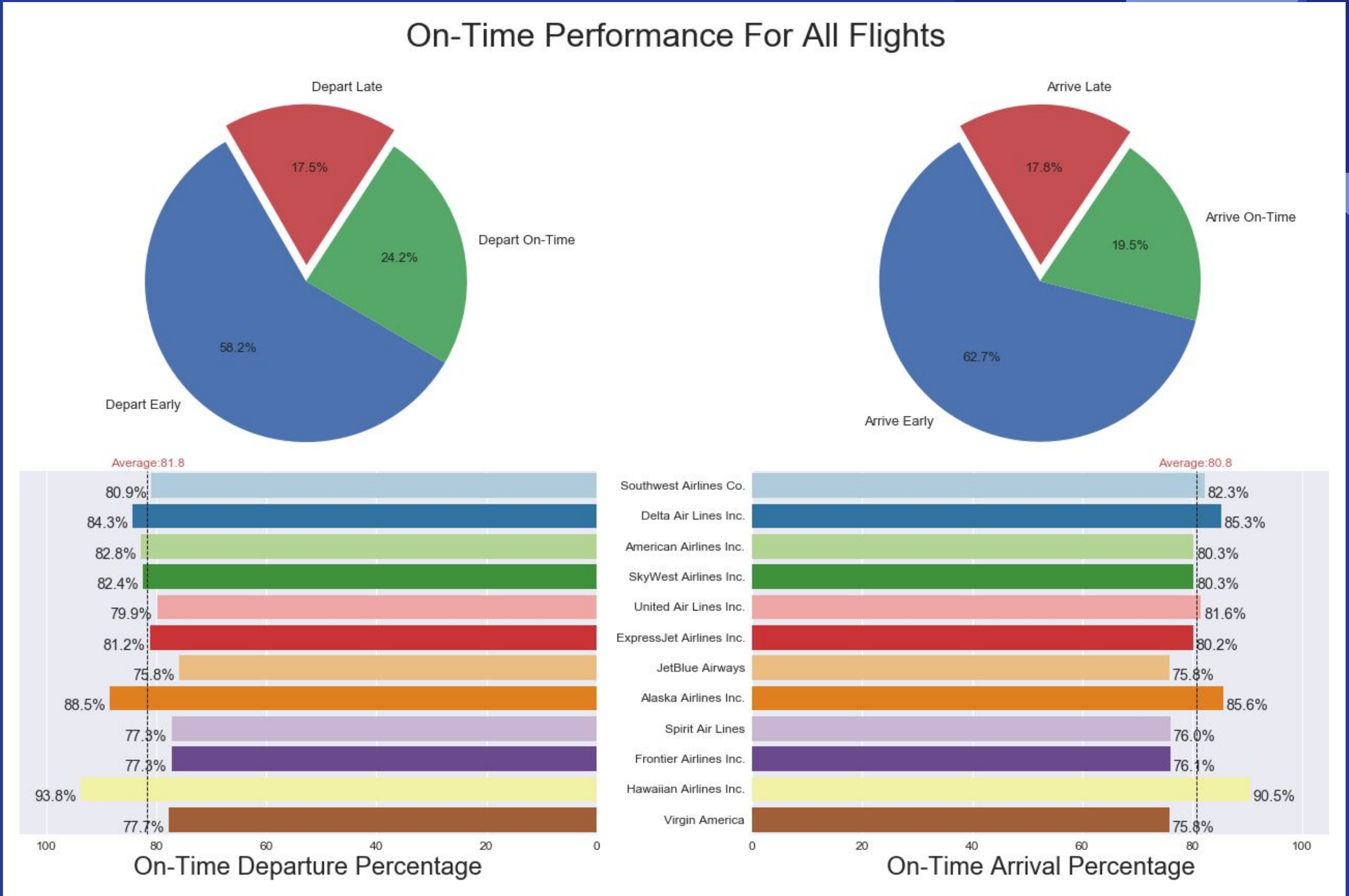
2.2 Distribution of Delay by Carrier

2.3 Distribution of Delay for Highest Traffic Airports

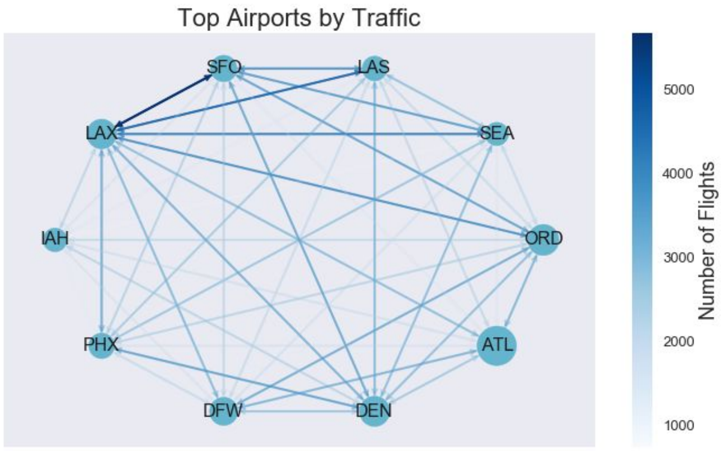


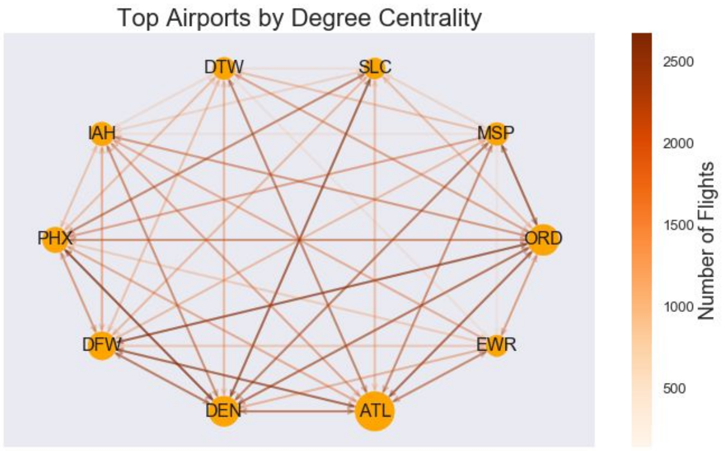
2.3 Distribution of Delay for Highest Traffic Airports

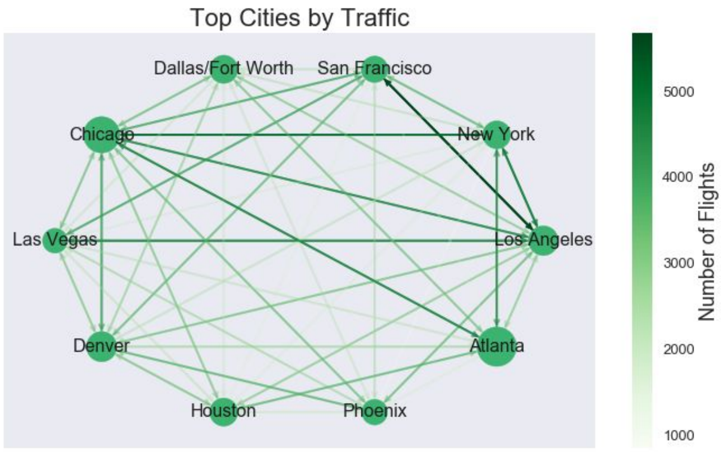


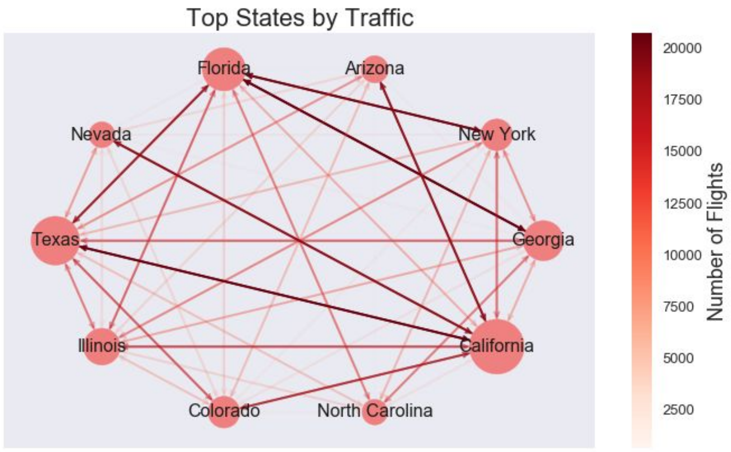

On-Time Performance For All Flights
Depart Late
17.5%
Depart On-Time
24.2%
Depart Early
58.2%
Arrive Late
17.8%
Arrive On-Time
19.5%
Arrive Early
62.7%
Average:81.8
Average:80.8
Southwest Airlines Co. 80.9% 82.3%
Delta Air Lines Inc. 84.3% 85.3%
American Airlines Inc. 82.8% 80.3%
SkyWest Airlines Inc. 82.4% 80.3%
United Air Lines Inc. 79.9% 81.6%
ExpressJet Airlines Inc. 81.2% 80.2%
JetBlue Airways 75.8% 75.8%
Alaska Airlines Inc. 88.5% 85.6%
Spirit Air Lines 77.3% 76.0%
Frontier Airlines Inc. 77.3% 76.1%
Hawaiian Airlines Inc. 93.8% 90.5%
Virgin America 77.7% 75.8%
100 80 60 40 20 0
0 20 40 60 80 100
On-Time Departure Percentage
On-Time Arrival Percentage

4. Network Visualization









Inferential Statistics

Initial Data Analysis

1. Test for Normality and CLT in variables of interest
2. Regression Analysis
3. Hypothesis Tests
   - T-Test of Arrival and Departure Deviation Means
   - Pearson’s r Permutation of Median Arrival and Departure Delays

Variables of Interest

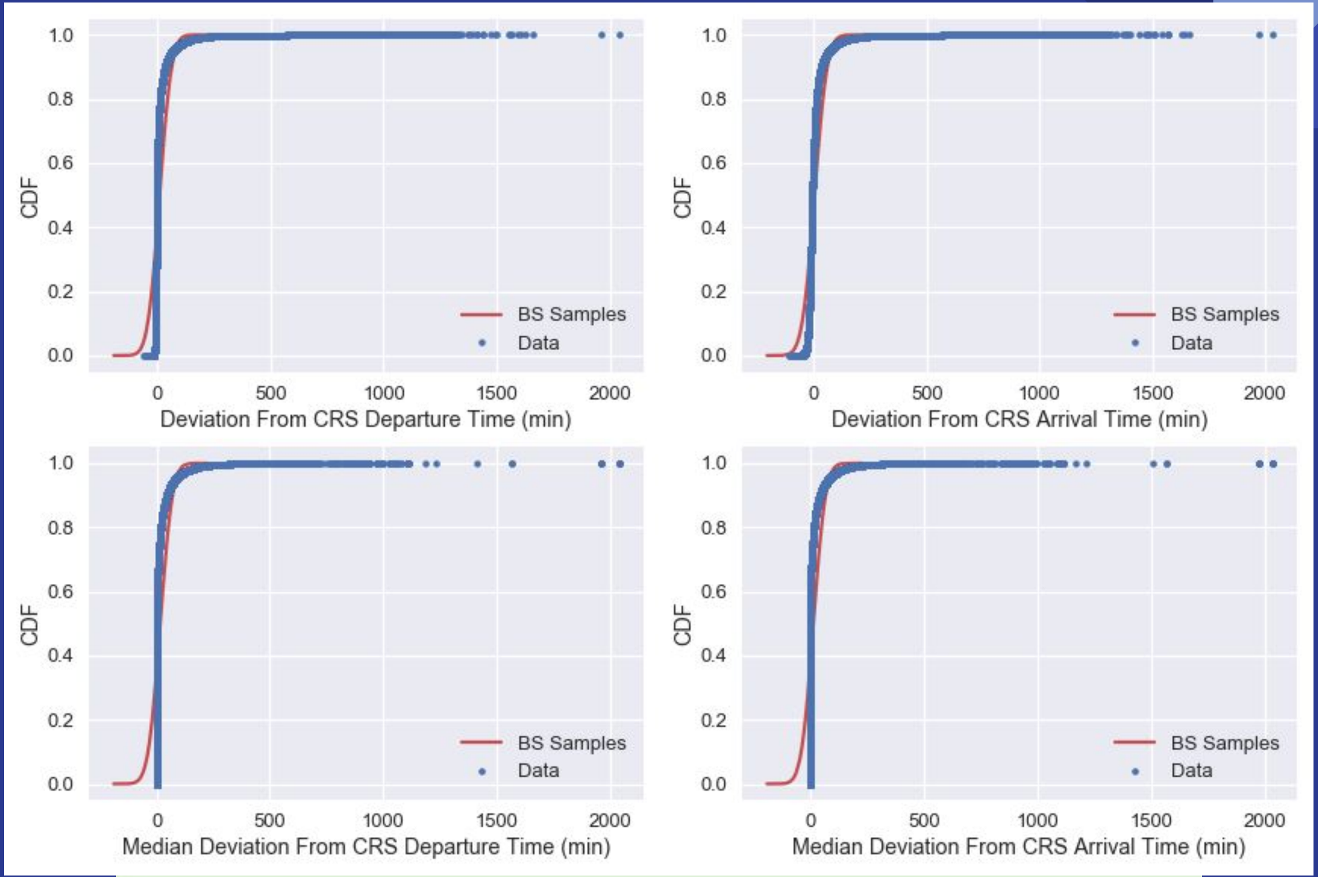
Flights DataFrame:

- dep_deviation: deviation from the actual departure time to the scheduled (CRS) departure time
- arr_deviation: deviation from the actual arrival time to the scheduled (CRS) arrival time

Links_d DataFrame:

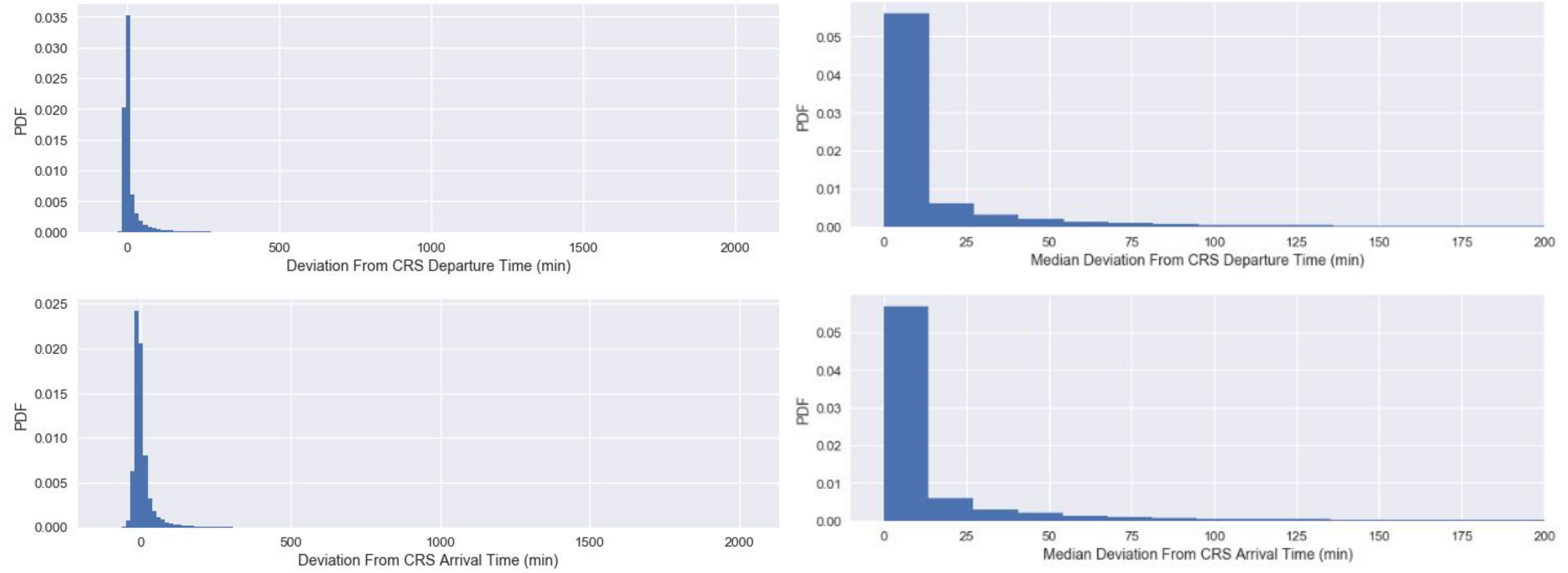
- dep_delay: median deviation of the actual departure time from the scheduled (CRS) departure time, for an origin-destination pair, in an hour of day that had non-zero traffic
- arr_delay: median deviation of the actual arrival time from the scheduled (CRS) arrival time, for an origin-destination pair, in an hour of day that had non-zero traffic

Test for Normality in Variables of Interest



The Cumulative Density Function of each variable shows they all possess normal distributions

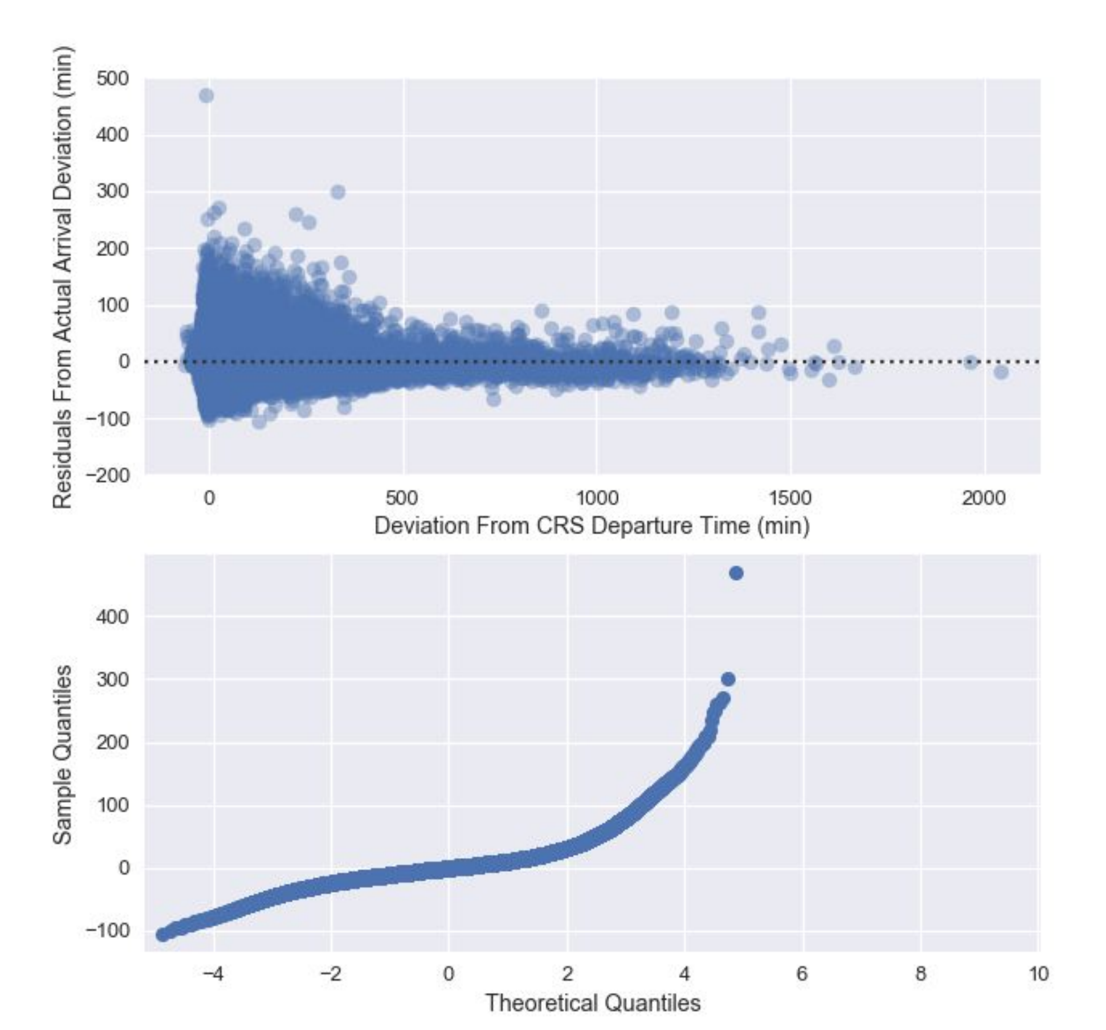
Test for Central Limit Theorem



Central Limit Theorem applies as sample size is very large and the probability density function of each variable shows observations are independent

Regression Analysis

- For the purpose of validating a directed network approach utilizing only departure delay as a predictor of overall (departure and arrival) delay
- The residuals between departure and arrival deviation are in a random pattern, supporting the use of a linear model. The quantile plot shows that the datasets are heavily skewed, and robust methods should be used in model construction to lessen the influence of extreme values.



T-Test of Arrival and Departure Deviation Means

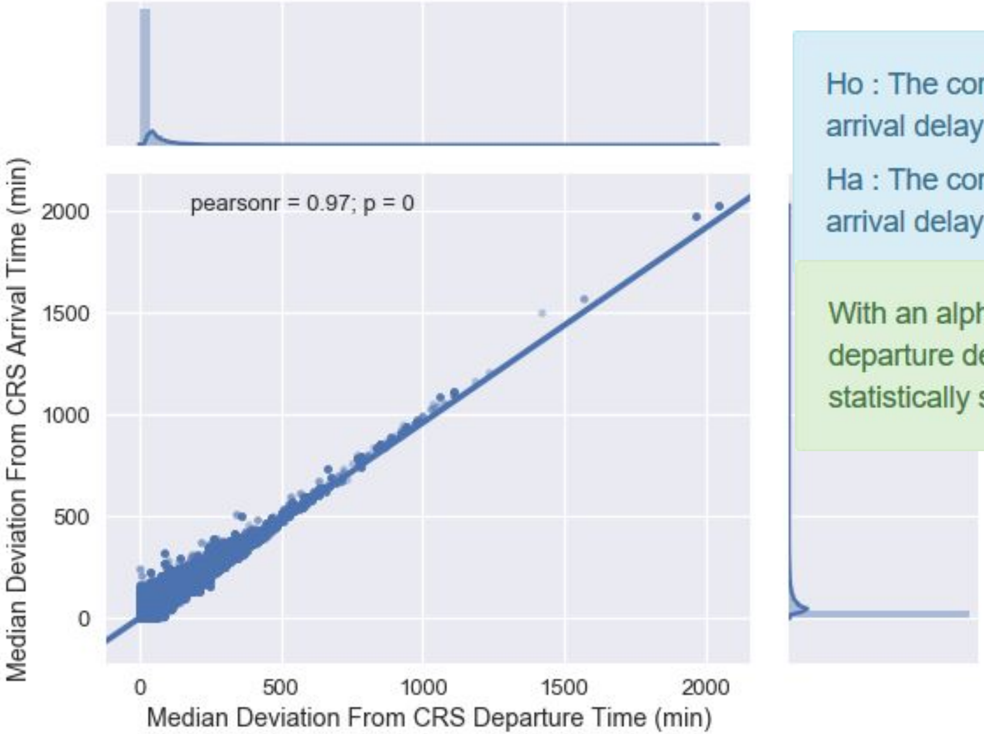
Ho : $\mu_d = \mu_a$

Ha : $\mu_d \neq \mu_a$

```
Margin of Error:  0.045
Difference of Means:  5.549
Departure Deviation  95% Confidence Interval:  [-11. 112.] min
Arrival Deviation  95% Confidence Interval:  [-31. 112.] min
T-Test: tstat = -124.015, P-value = 0.0000000000
```

With an alpha level of .01 (α = .01), the difference between the means of departure and arrival deviation from scheduled (CRS) time was statistically significant, $p < .01$

Pearson's r Permutation Test of Median Arrival and Departure Delays



Ho : The correlation between the current median departure delay and median arrival delay for an Origin-Destination pair is not significant

Ha : The correlation between the current median departure delay and median arrival delay for an Origin-Destination pair is significant

With an alpha level of .01 (α = .01), the correlation between the median departure delay and the median arrival delay for Origin-Destination pairs is statistically significant, p < .01

Model Construction

Binary Departure Delay Classification

1. Baseline Logistic Regression Classifier
2. Resample data to address target class imbalance, both under- and over-sampling
3. Train Logistic Regression and Random Forest Classifiers under both resampling conditions