Airfare Prediction

Wooseok Kim

Progress

What I did

- Environment Setup
- Studying Hadoop
- Algorithm analysis

Things to do until end of semester

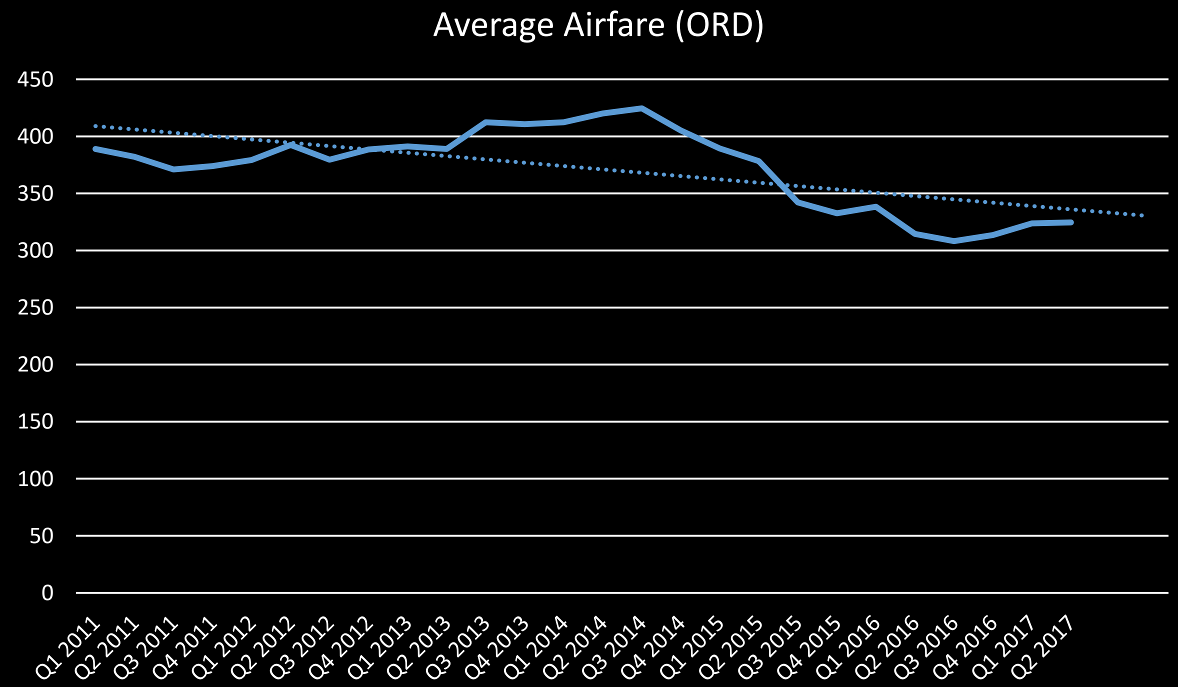
- How to run the Hadoop on AWS
- Do the programming for the algorithm
- More familiar with map -reduce

Algorithm Analysis

- Data-driven Forecasting methods
  - There is no difference between a predictor and a target
- Model-driven forecasting methods
  - Similar to conventional predictive models, which have a predictor and a target
  - Based on the data from adjacent time periods
- Decomposition
  - Trend
  - Seasonality
  - Noise

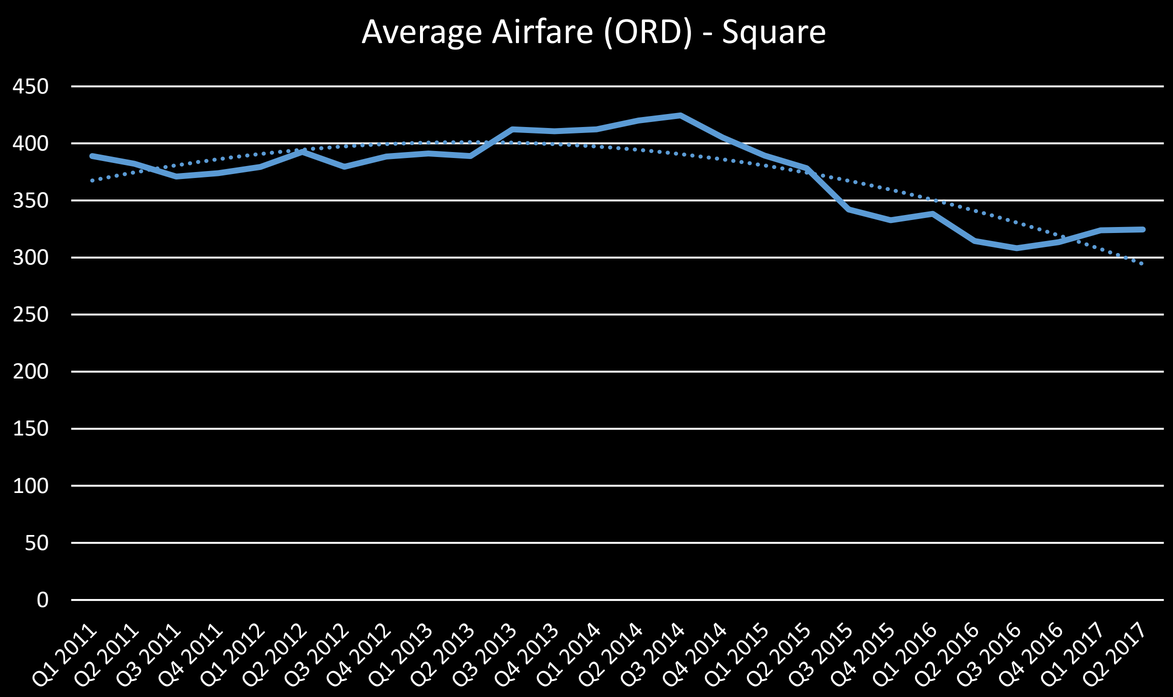
Model-Driven Approaches

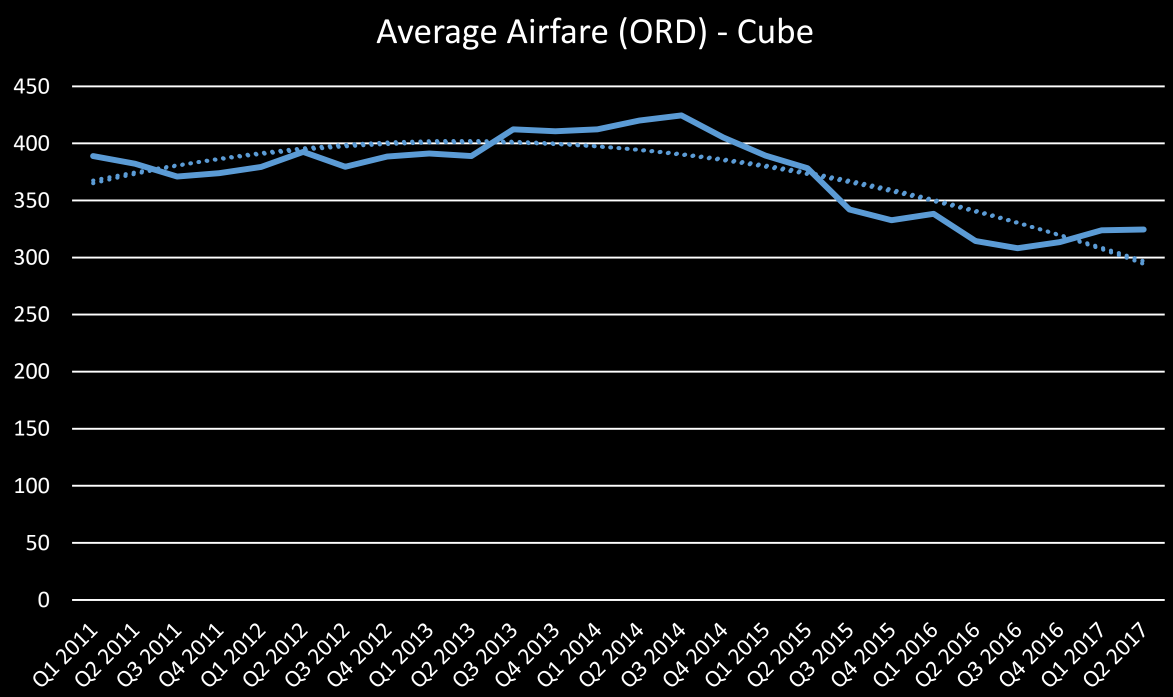
- Linear Regression
  - The simplest approaches
  - Can capture the long-term tendency, but it does a very poor job of fitting data



Model-Driven Approaches

- Polynomial Regression
  - Similar to linear regression except that higher-degree functions of the independent variable are used squares and cubes





Model-Driven Approaches

- Linear Regression with seasonality
  - The time-independent variable captures the trend and the dummy variables capture seasonality
  - Can be used for predicting any future value beyond n+1
- Autoregression Models
  - Regression models applied on lag series where each lag series is a new predictor used to fit the dependent variable, which is still the original series value
  - Create a lag series involving forecast errors and use this as another predictor.

Data-Driven Approaches

- Naïve Forecast
  - The simplest forecasting model
  - F_{n+1}, the forecast for the next period, is given by the last data point
- Simple Average
  - Compute the next data as an average of all the data points
  - F_{n+1}=AVG(y_n, y_{n-1}, … , y_1)
- Moving Average
  - Select a window of the last k periods for the average (n, … , n-k+1)
  - Window keeps moving forward and thus returns a moving average

Data-Driven Approaches

- Weighted Moving Average
  - $F_{n+1} = a * y_n + b * y_{n-1} + c * y_{n-2}$, where typically a > b > c
  - Assume that a = 0.6, b = 0.3, c = 0.1

	Airfare Avg (ORD)	Simulated Forecast
Q2 2016	314.45	
Q3 2016	308.14	
Q4 2016	313.45	
Q1 2017	323.80	**311.957** = 0.6*313.45 + 0.3*308.14 + 0.1*314.45
Q2 2017	324.54	**319.129** = 0.6*323.80 + 0.3*313.45 + 0.1*308.14

Data-Driven Approaches

- Exponential Smoothing
  - $F_{n+1} = \alpha * y_n + (1 - \alpha) * F_n$, α is generally 0~1
  - If α is close to 1, then the previously forecasted value of the last period has less weight than the actual value of the last period. Ex) α=1, Naïve Forecast
  - Can't make forecast more than one-step ahead because of data requirement for the previous forecasted value, F_n

Data-Driven Approaches

- Need more sophisticated techniques than the ones described in order for trend and seasonality
- Once capturing trend and seasonality, can forecast the value at any time in the future, not just one step ahead value

Data-Driven Approaches

- Two-parameter Exponential Smoothing
  - One-parameter exponential smoothing equation simply calculates the average value
  - If the series has a trend, an average slope can be estimated as well
  - L_n: avg value or length of Seasonality, T_n: Trend
  - $F_{n+1} = L_n + T_n$
  - $L_n = \alpha * y_n + (1-\alpha) * (L_{n-1} + T_{n-1})$
  - $T_n = \beta * (L_n - L_{n-1}) + (1-\beta) * T_{n-1}$

Data-Driven Approaches

- Two-parameter Exponential Smoothing
  - $F_{n+1} = L_n + T_n$, $L_n = \alpha * y_n + (1-\alpha) * (L_{n-1} + T_{n-1})$
  - $T_n = \beta * (L_n - L_{n-1}) + (1-\beta) * T_{n-1}$
  - Assume that α=0.3, β=0.6, Forecast for Q2 2016 = 320

	Airfare Avg (ORD)	L_n	T_n	F_{n+1}
Q2 2016	314.45	320	0	
Q3 2016	308.14	318.335 = 0.3*314.45 + 0.7*320	-0.999 = 0.6(318.335-320)+0.4*0	317.336 = 318.335 + (-0.999)
Q4 2016	313.45	314.276 = 0.3*308.14 + 0.7*(318.335-0.999)	-2.835 = 0.6(314.276-318.335)+0.4(-0.999)	311.441 = 314.276+(-2.835)

Next-Steps

- Implement the code for Two-parameter exponential smoothing
- How to run the Hadoop on AWS
- More familiar with the Hadoop especially Map-Reduce