# Market Demand Forecasting

Anupama Lakshmanan
Sabre Labs

10 May 2021

Copyright © by Sabre. Published by AGIFORS with permission.

# Speaker Biography

## Anupama Lakshmanan

Anupama Lakshmanan received a Ph.D. in Decision Sciences in March 2020 from the Indian Institute of Management Bangalore. Her dissertation is titled “Time Series with Multiple Levels of Seasonality”.

She is interested in business forecasting and time series analysis. She works on demand, schedule and market share forecasting problems in Sabre.

# INTRODUCTION

Air travel demand estimation

# Background

- Market Size Estimation → global air travel demand
- Air Traffic dataset is generated at many different levels (city pair, country pair) and in many different formats (Origin-Destination, Leg, Segment)
- Different varieties of information available for estimating future demand
- Different use cases for future demand data, short term route planning to long term aircraft ordering
- To cater to all use cases and to use different sources of data we support three varieties of future demand data
  - Short Term – up to 12 months – at origin-destination level
  - Medium Term – up to 36 months – at city pair and cabin level
  - Long Term – up to 120 months – at quarterly country pair level

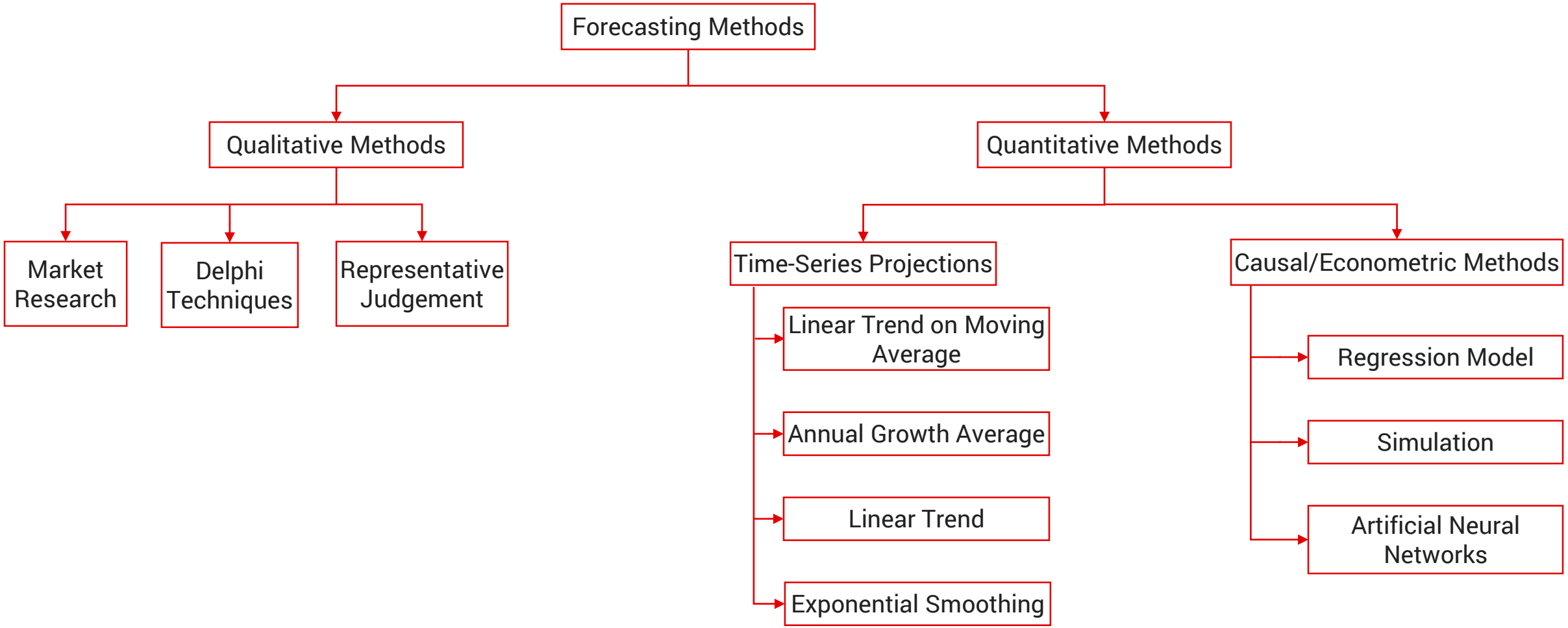
# Problem description

- If we let the historical passenger count be denoted by $y_1, \dots, y_{T-1}$ for a market $M$, then we can write the forecasts as $y_{T+h|H}$.

where

- $H$ is for training set of passenger count from $y_1$ $to$ $y_{T-1}$ for a market $M$
- $T$ is for time with monthly frequency
- $h$ is for forecasting horizon taking account of all observations up to time $T-1$ for a market $M$

# Forecasting Methods




Source:

*Grosche T. (2009) Integrated Airline Scheduling. In: Computational Intelligence in Integrated Airline Scheduling. Studies in Computational Intelligence, vol 173. Springer, Berlin, Heidelberg*

# Data

- Passenger Demand Data

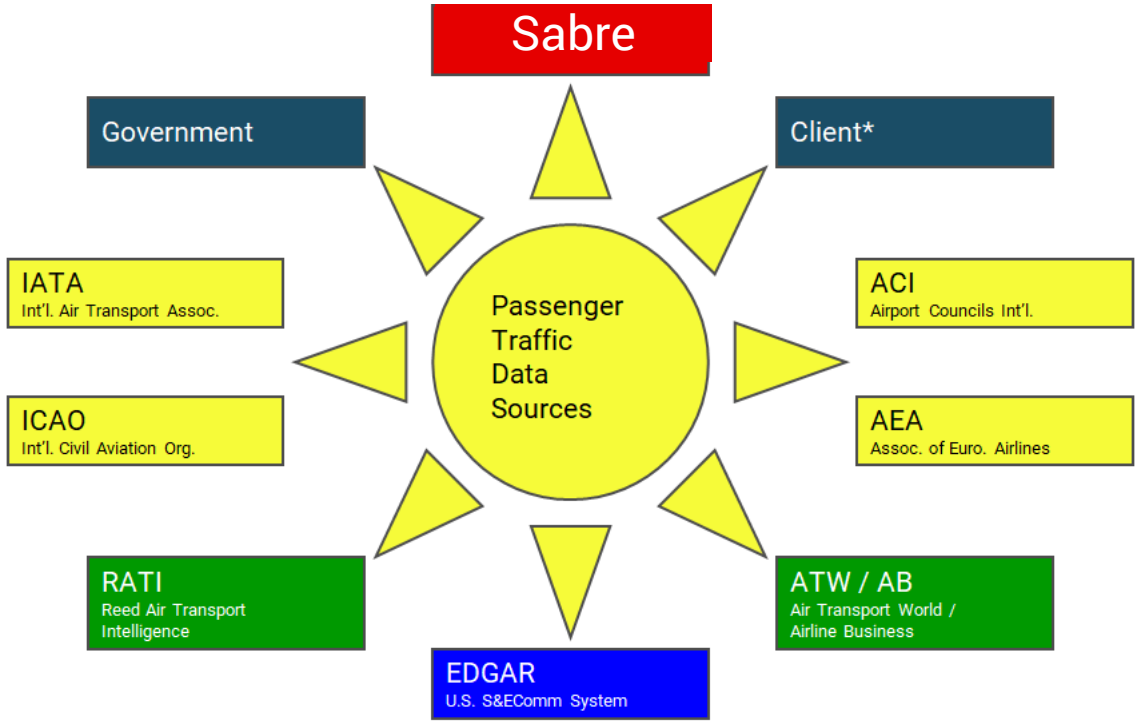
- The passenger demand dataset offers valuable market information for many types of Aviation analyses, including market sizes, route performance, traffic flows, and fares.

- Traffic
  - Fully-routed itineraries with cabin level details
  - By airline -- Ticketing and operating
  - 3 connection points
  - Directional data

- Fares
  - Sabre in house
  - By booking code

- Updated Monthly

- Available back to 2001



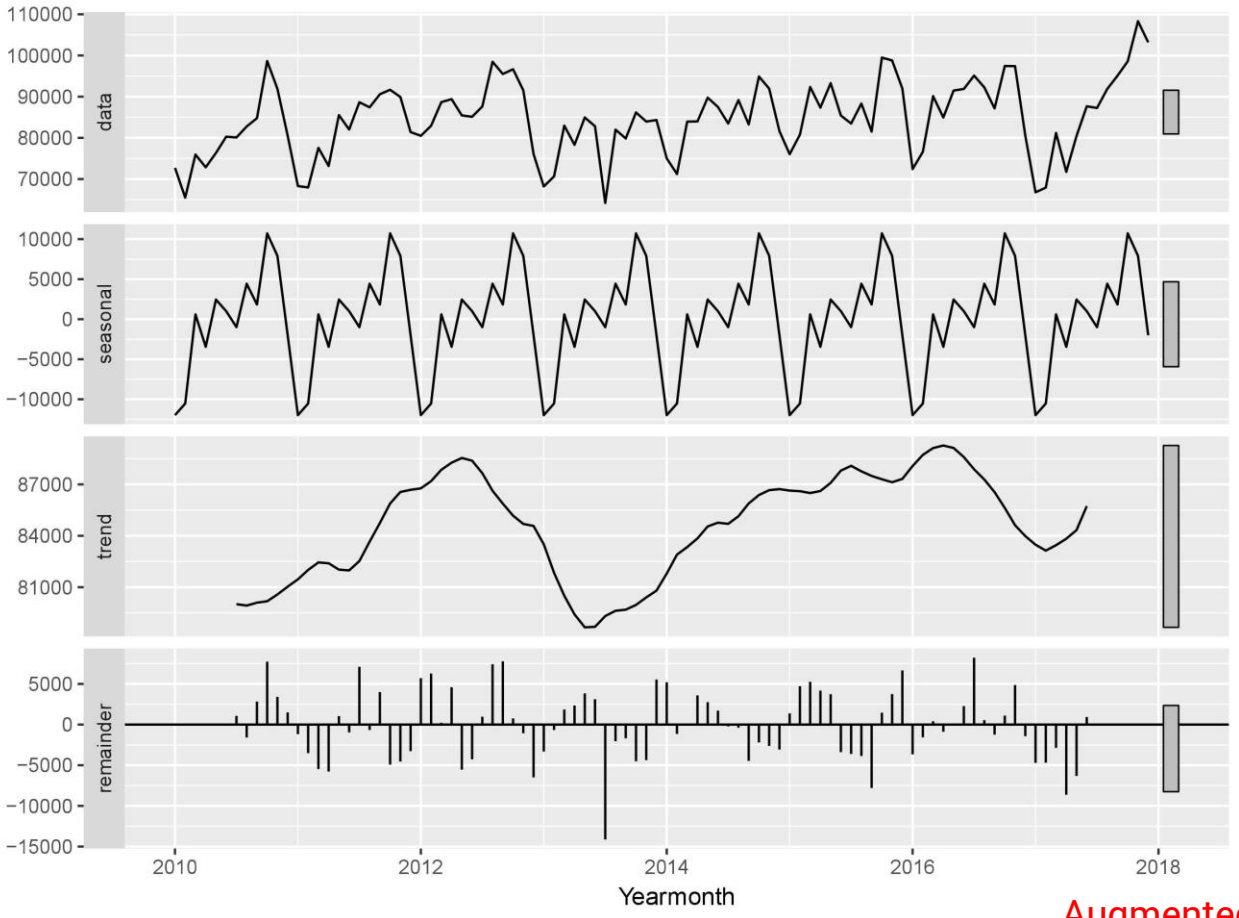
# Data Description

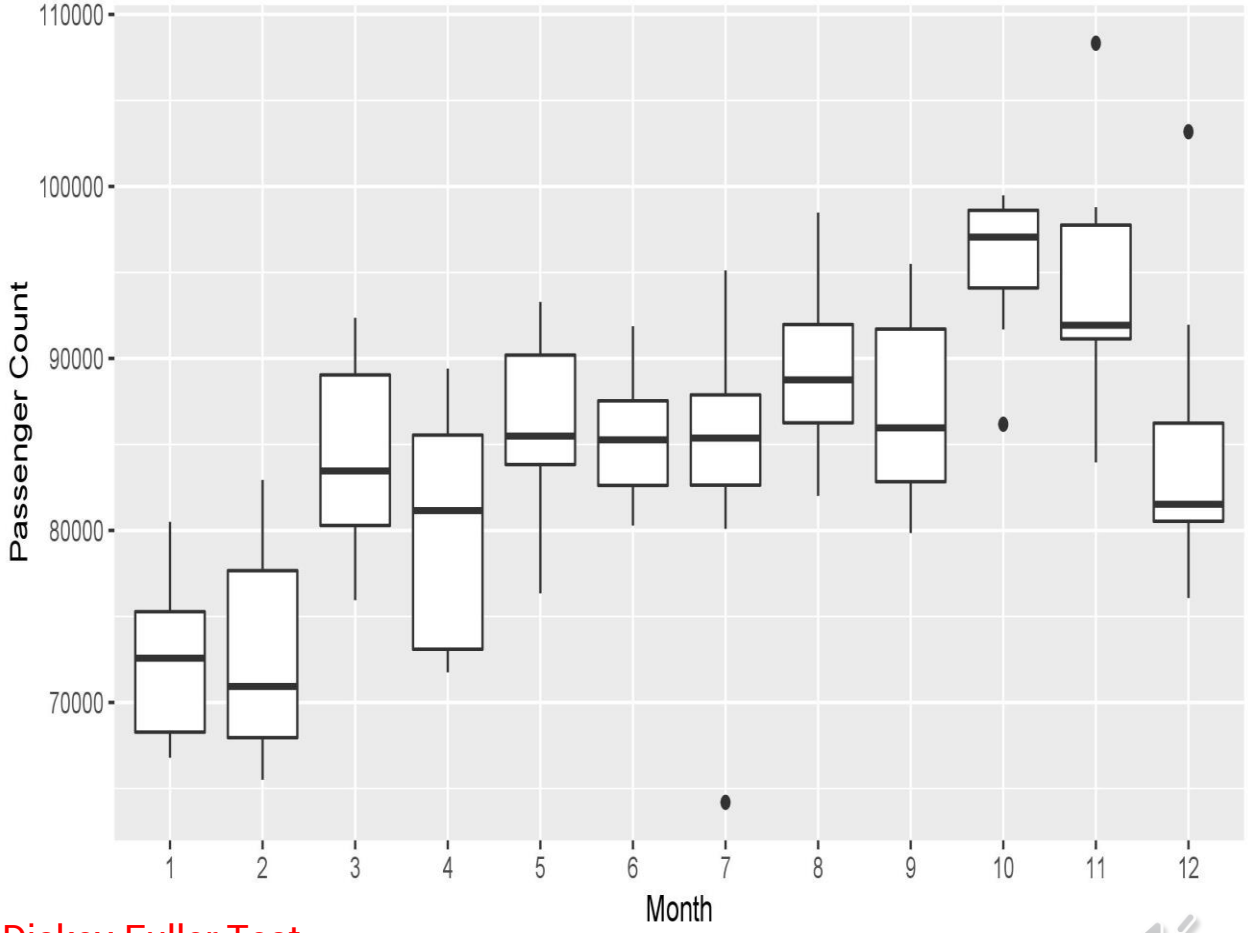
- Directional Origin-Destination level data is aggregated at City Pair level.
- In this work, we follow Euro-control's definition of flight length. Euro-control defines
  - Short-haul routes as shorter than 1,500 km (930 mi; 810 nmi)
  - Medium-haul between 1,500 and 4,000 km (930 and 2,490 mi; 810 and 2,160 nmi)
  - Long-haul routes as longer than 4,000 km (2,500 mi; 2,200 nmi)
- The demand data has 10 regions.
- Top 100 markets (by passenger count – in December 2018) in short-haul, medium- haul, and long-haul originating from all 10 regions in the passenger demand data. Therefore, we are considering 3000 markets.
- Training Data:
  - January 2010 to December 2015.
- Test Data:
  - January 2016 to December 2018.

# Data Analysis

## Short Haul

Ex: *San Francisco to Los Angeles (Air line: 543.36 km)*



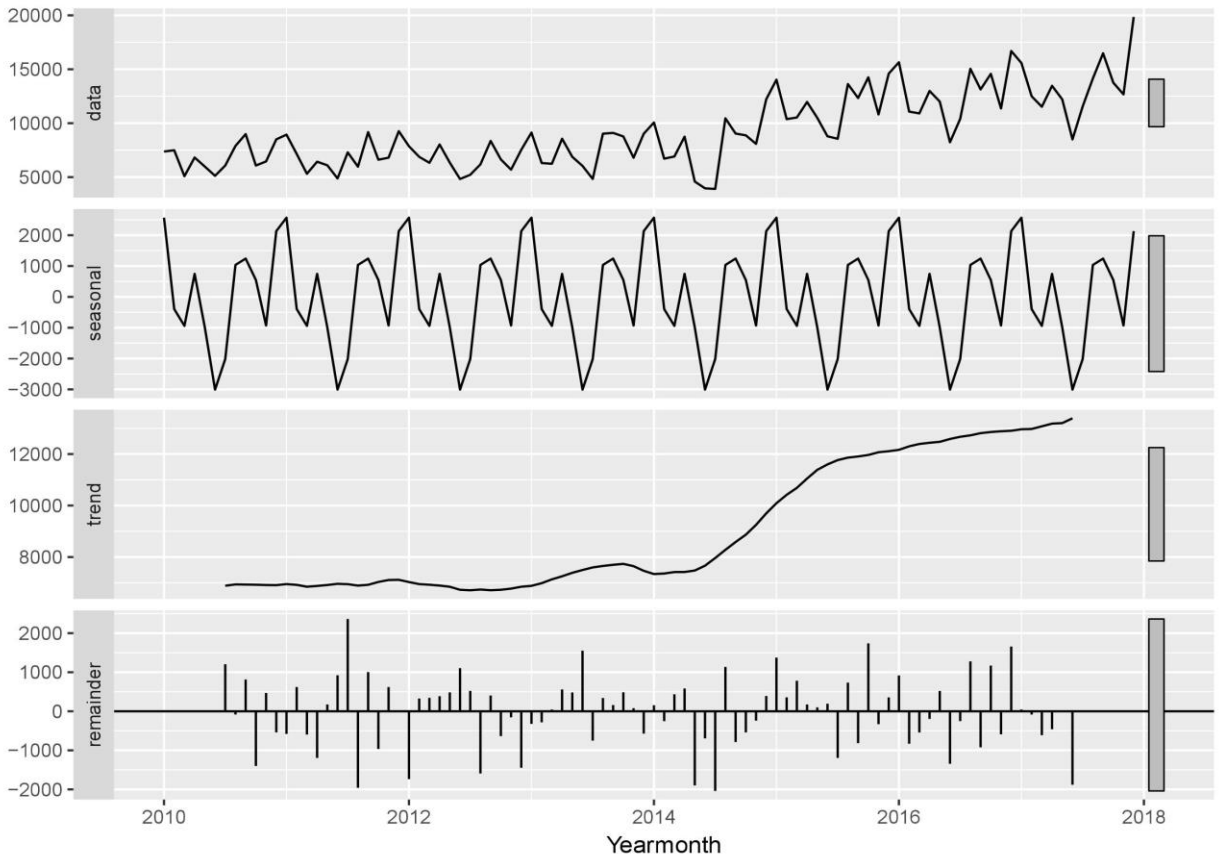

### Augmented Dickey-Fuller Test

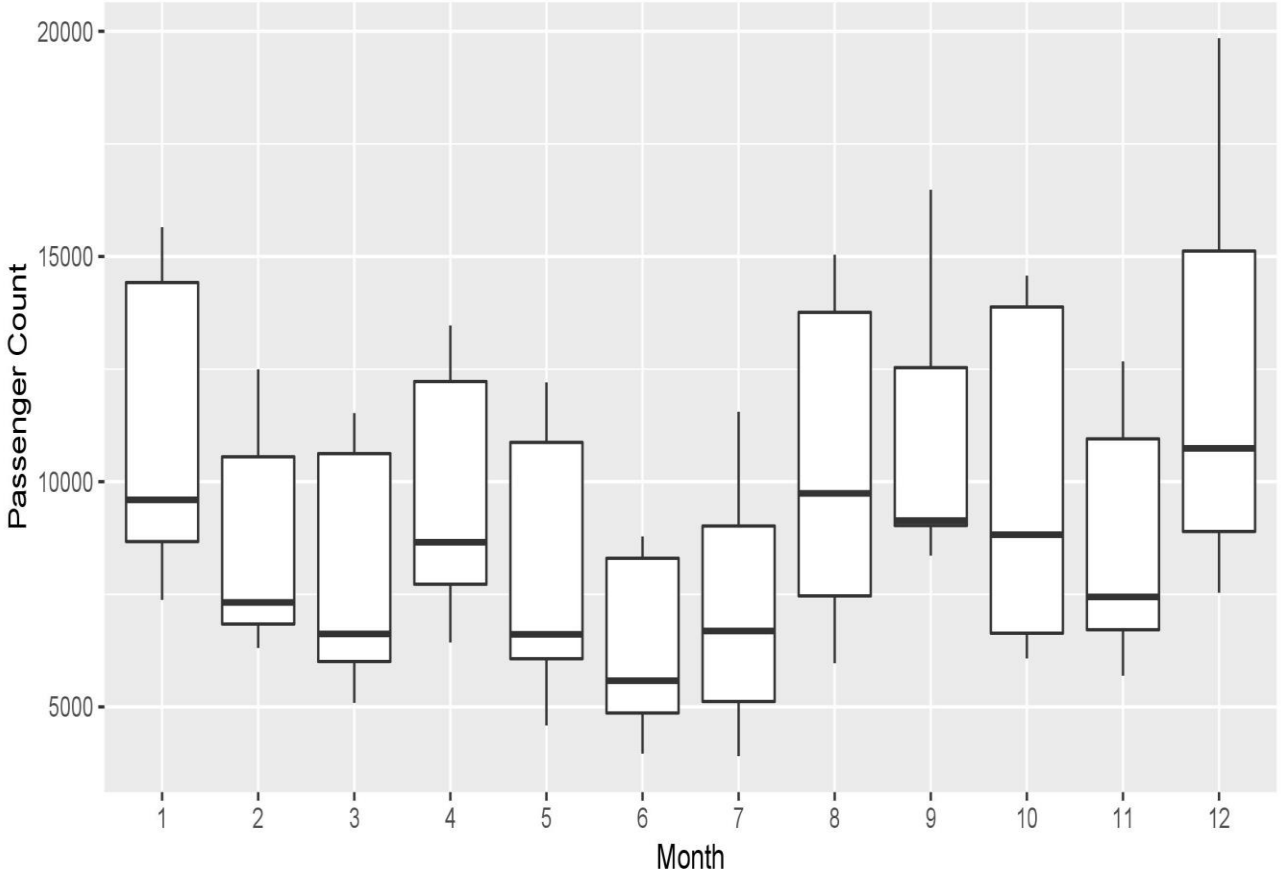
- Dickey-Fuller = -3.7587
- Lag order = 4
- p-value = 0.02396

# Data Analysis

## Medium Haul

Ex: *Bangalore to Dubai (Air line: 2,696 km)*



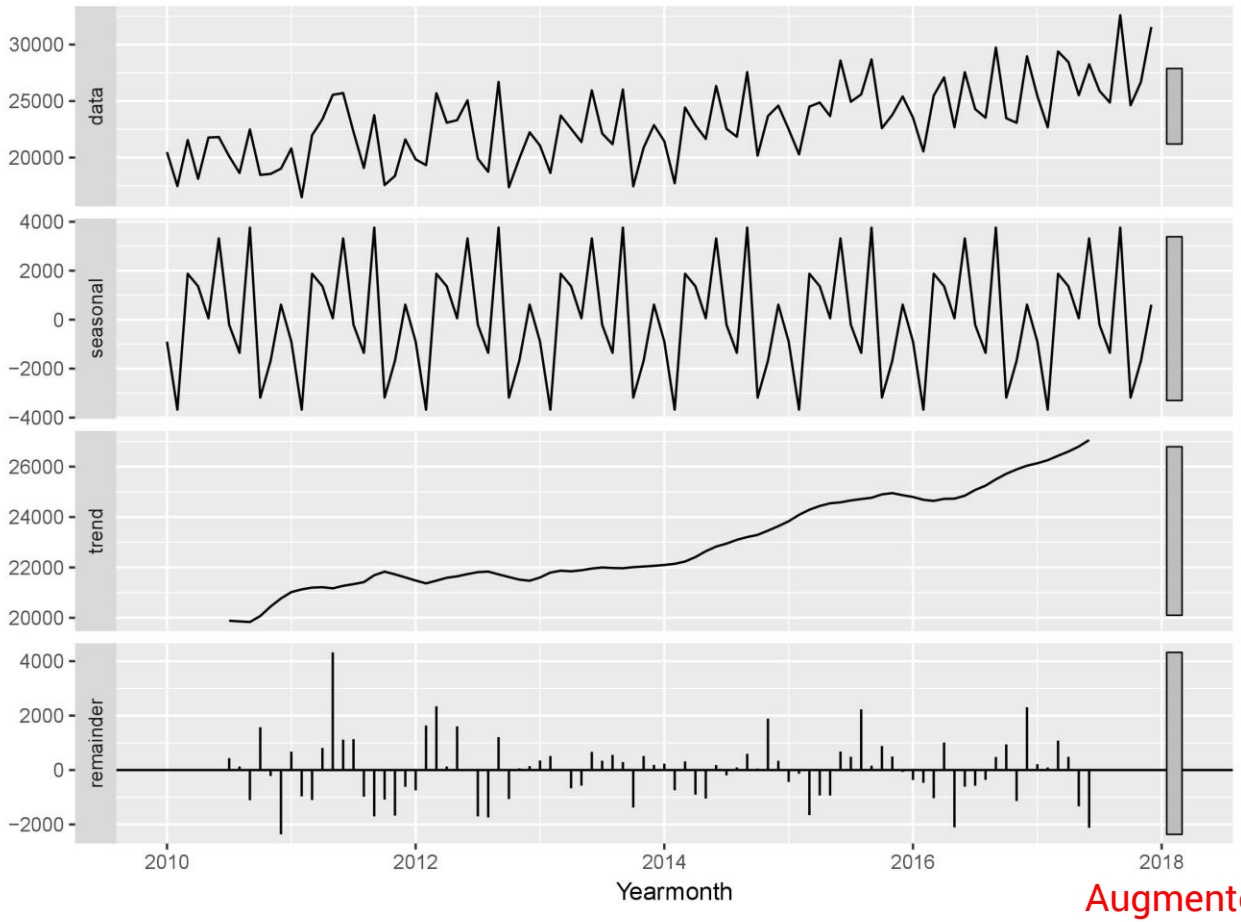

### Augmented Dickey-Fuller Test

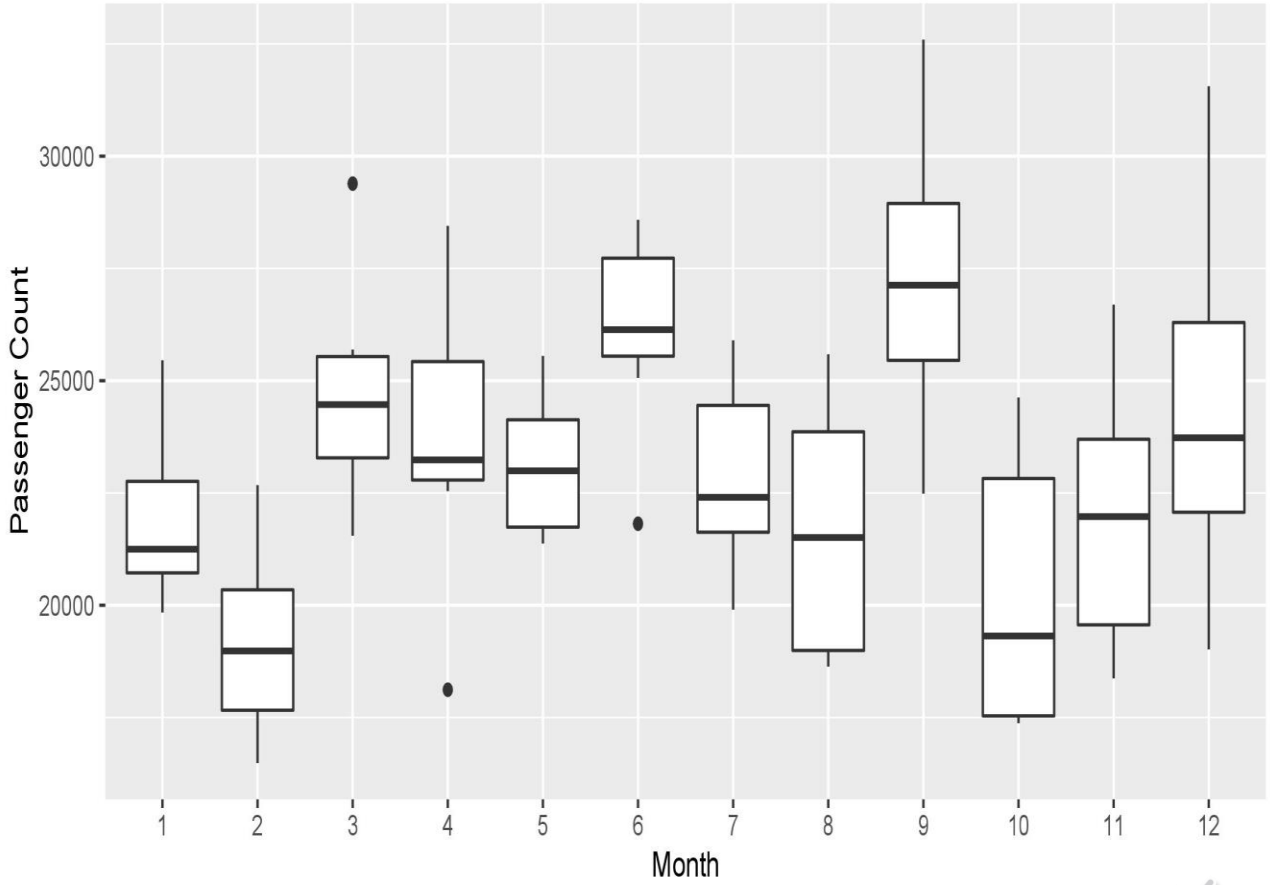
- Dickey-Fuller = -2.9549
- Lag order = 4
- p-value = 0.1818

# Data Analysis

## Long Haul

Ex: *Singapore to London (Air line: 10,886 km)*




### Augmented Dickey-Fuller Test

- Dickey-Fuller = -5.8809
- Lag order = 4
- p-value = 0.01

# Algorithms

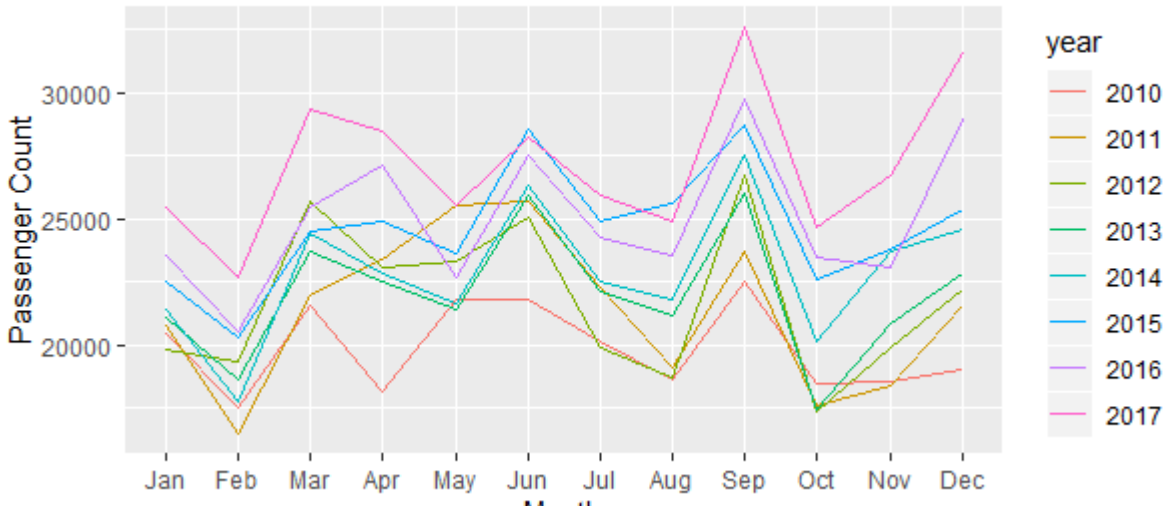
1. Seasonal Naïve Method
2. ARIMA
3. Gradient Boosting based Regression

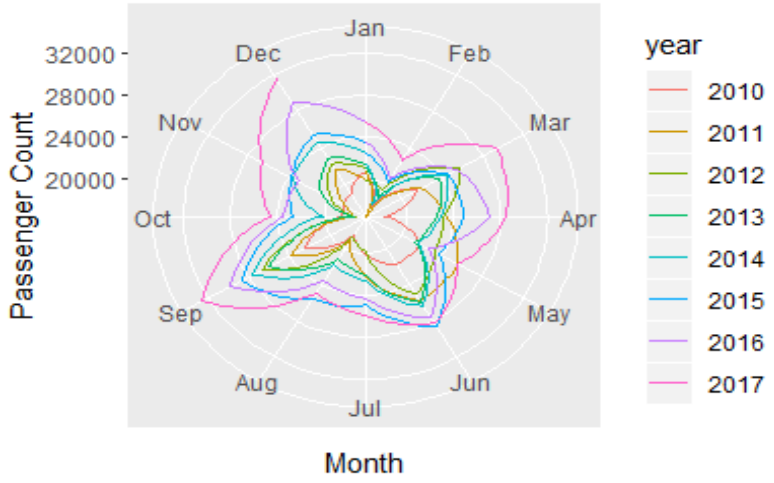
# Seasonal Naïve Method

- For Naïve forecast, current month's passenger count is set to previous month's passenger count.

$$y_{T+h|H} = y_T$$

- Airline passenger traffic data is seasonal (`freq=12`, for monthly passenger count).
- We set the forecasted passenger count to passenger count observed from the same month in the last year.
- $y_{T+h|H} = y_{T+h-12}$



# Auto-Regressive Integrated Moving Average (ARIMA)

- $ARIMA(p, q, d)$ model,

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

- Where,
  - $\boldsymbol{\varepsilon_t}$ = white noise;
  - $p$ = order of the autoregressive part;
  - $d$ = degree of first differencing involved;
  - $q$ = order of the moving average part;

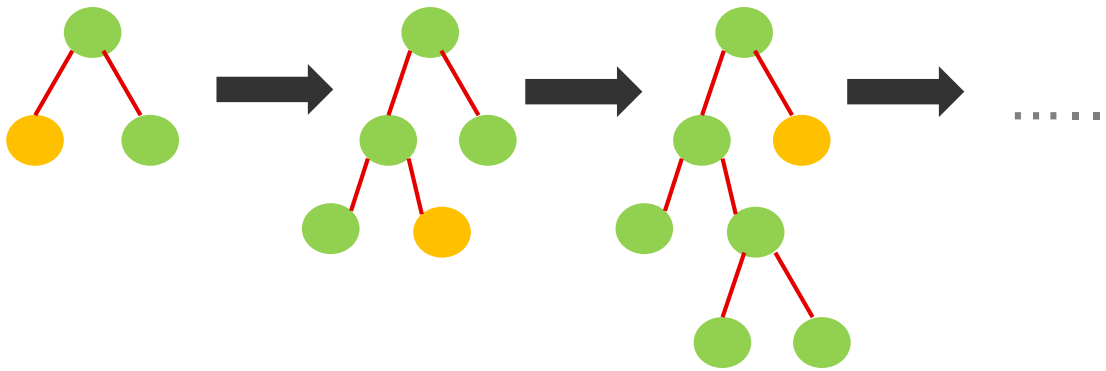
- We used `auto.arima()` [1] from the `R` package `forecast` which returns best ARIMA model according to either AIC, AICc or BIC value. The function conducts a search over possible model within the order constraints provided.

Source:

[1] Hyndman, R., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software, 27*(3), 1 - 22. doi:http://dx.doi.org/10.18637/jss.v027.i03

# eXtreme Gradient Boosting (XGBoost)

- Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made.
- Gradient boosting is a boosting approach where gradient descent algorithm is used to minimize the loss when adding new models.
- The XGBoost library [2] implements the gradient boosting decision tree algorithm.



Real value (label) known from the training data-set

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Can be seen as f(x + Δx) where x = $\hat{y}_i^{(t-1)}$

Source:

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

# Evaluation Metrics

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

- Weighted Mean Absolute Percentage Error (WMAPE)

$$WMAPE = \frac{\left( \sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t} \right) * A_t}{\sum_{t=1}^{n} A_t} * \frac{100\%}{n}$$

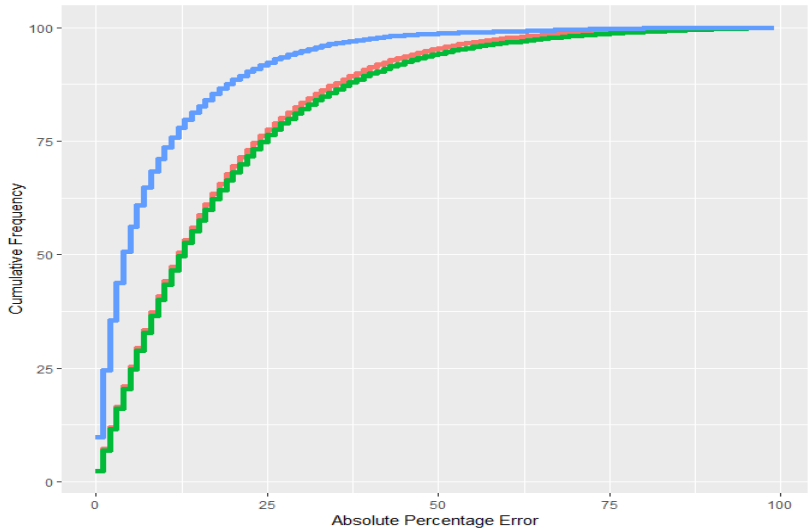
- Root Mean Squared Percentage Error (RMSPE)

$$RMSPE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left( \frac{A_t - F_t}{A_t} \right)^2}$$

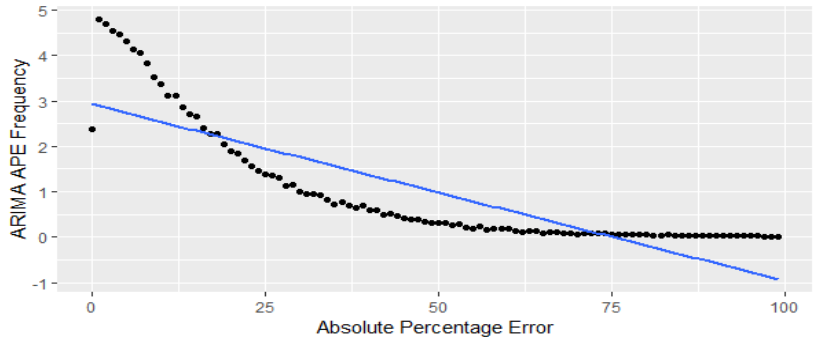
Where,
$A_t$ – Actual Passenger Count
$F_t$ – Forecasted Passenger Count

# Evaluation Metrics (Continued)

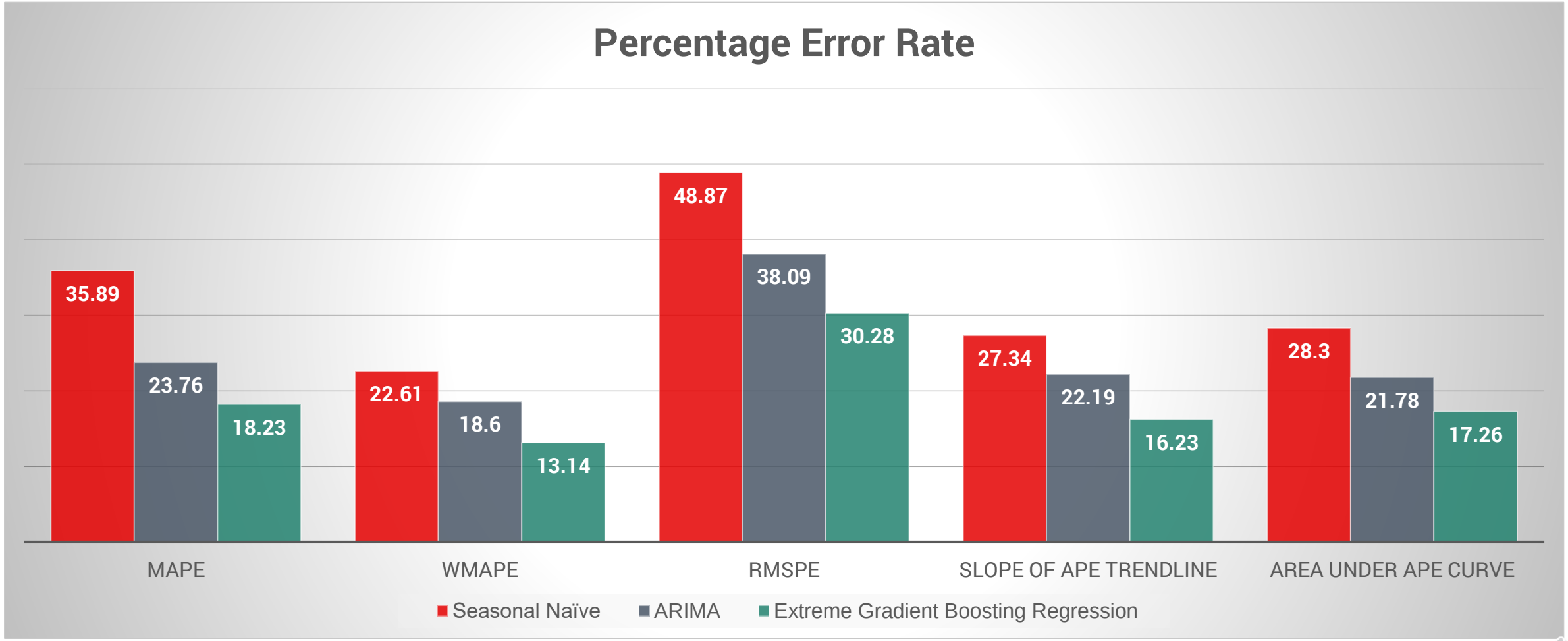
- Area Under the **A**bsolute **P**ercentage **E**rror(APE) Curve
- $$APE = 100\% \left| \frac{A_t - F_t}{A_t} \right|$$
  - Cumulative frequency of APE bins
  - $APE > 100$ are ignored in this curve
  - More the area under the curve, accuracy is better
  - Trapezoidal rule to approximate the area under the APE curve
  - Normalized to match other error metrics



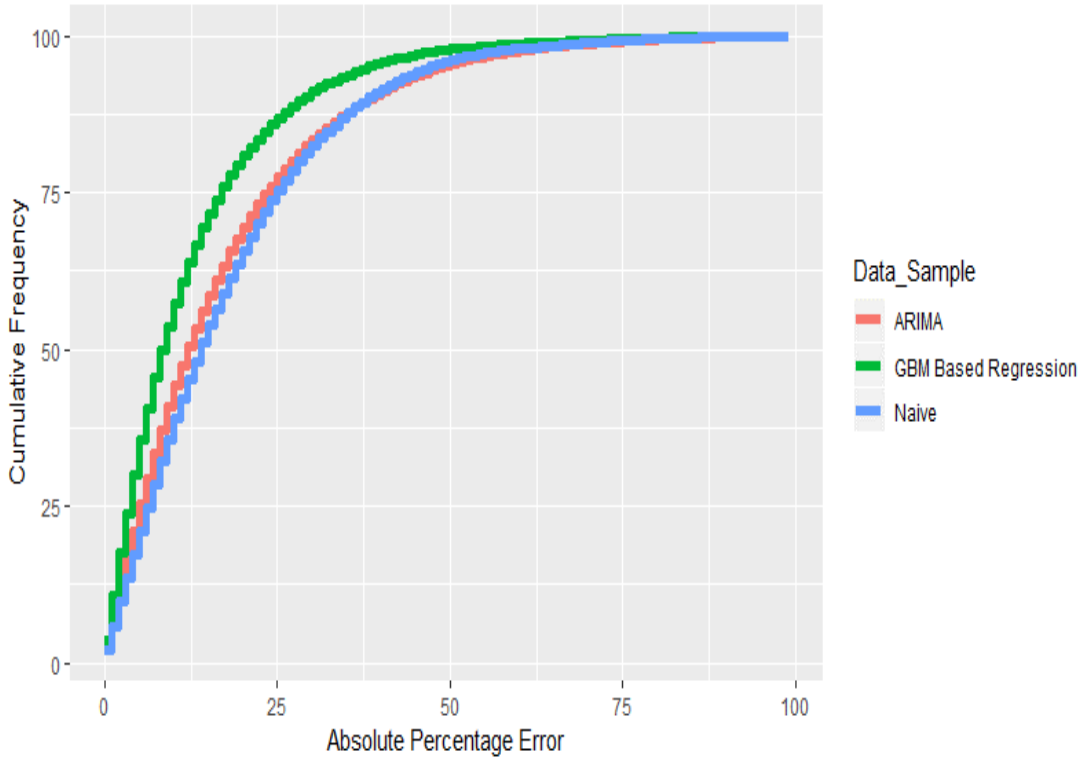

- Slope of Trendline for Absolute Percentage Error Curve
  - Trendline of APE frequency scatter plot
  - Slope coefficient of the trendline
  - Normalized to match other error metrics

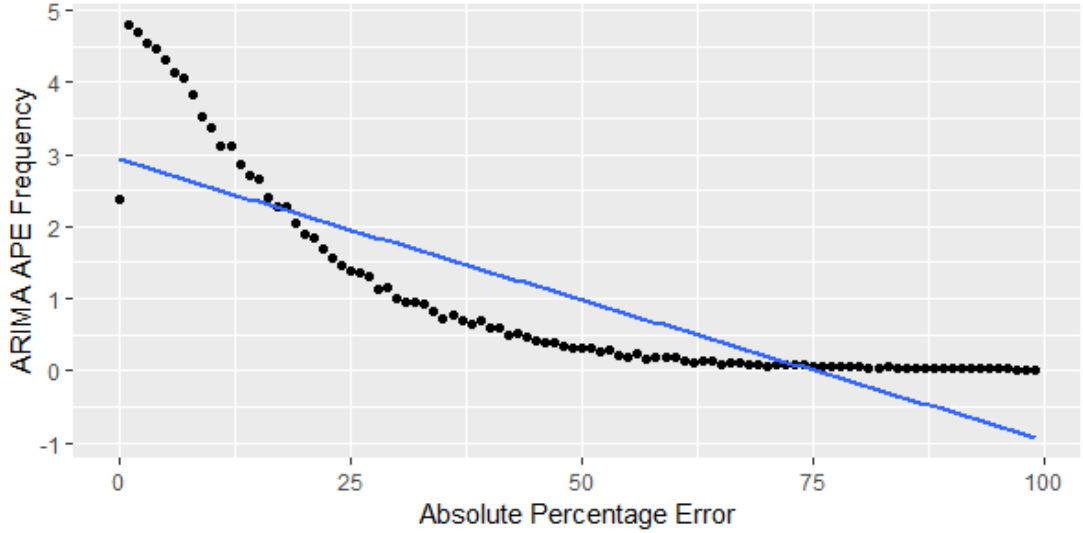


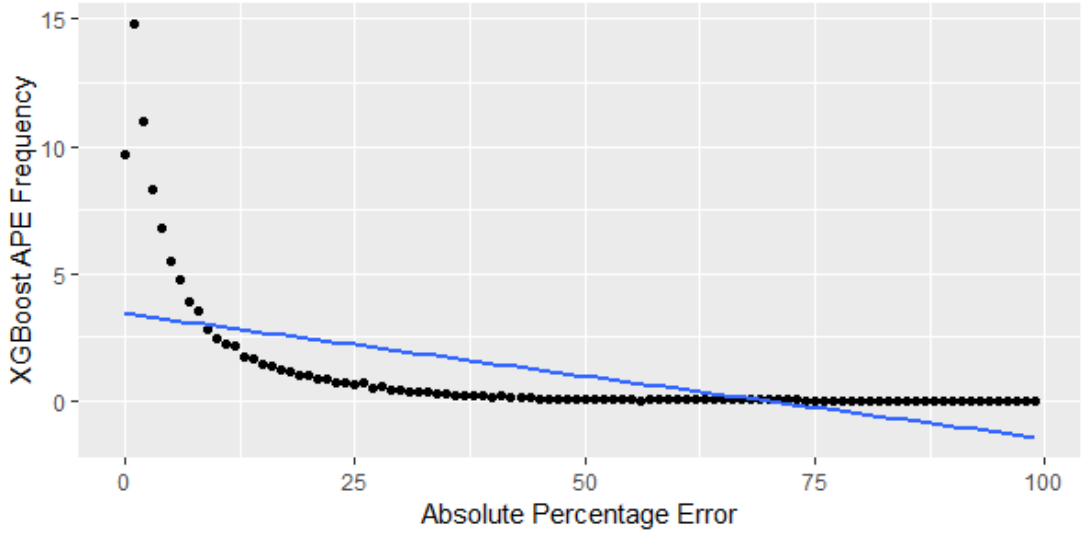
# Results


Percentage Error Rate
35.89
23.76
18.23
22.61
18.6
13.14
48.87
38.09
30.28
27.34
22.19
16.23
28.3
21.78
17.26
MAPE
WMAPE
RMSPE
SLOPE OF APE TRENDLINE
AREA UNDER APE CURVE
Seasonal Naïve
ARIMA
Extreme Gradient Boosting Regression

# Results

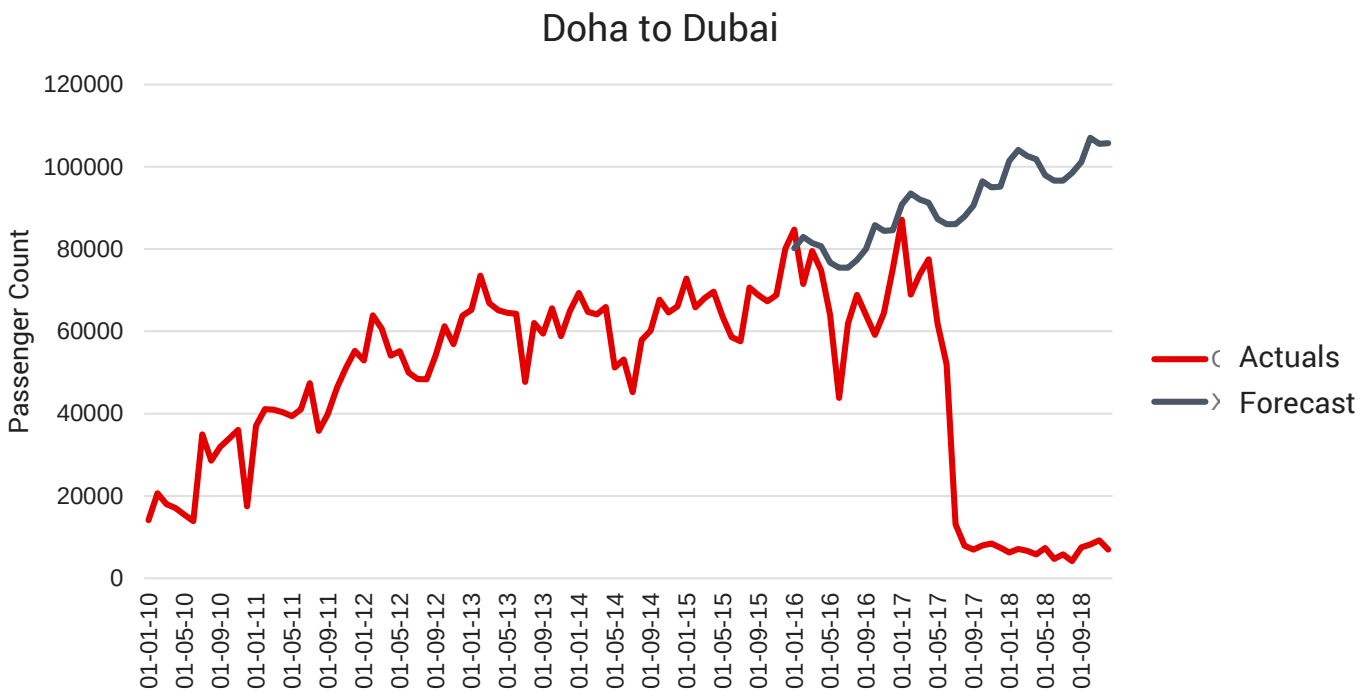

Cumulative Frequency
100
75
50
25
0
0
25
50
75
100
Absolute Percentage Error
Data_Sample
ARIMA
GBM Based Regression
Naive


ARIMA APE Frequency
5
4
3
2
1
0
-1
0
25
50
75
100
Absolute Percentage Error


XGBoost APE Frequency
15
10
5
0
0
25
50
75
100
Absolute Percentage Error

# Example: Where forecasting fails

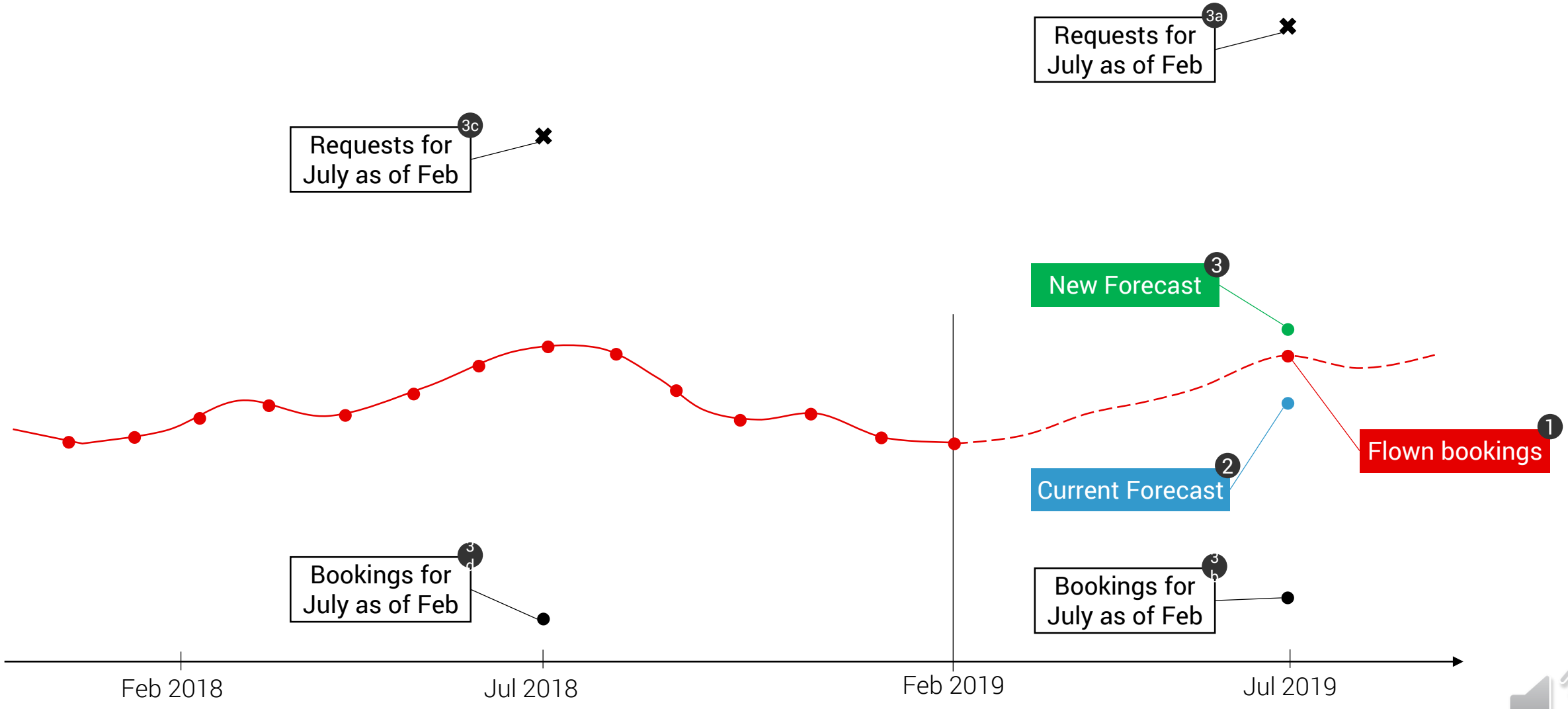
- In 2017, Doha to Dubai market's passenger count dropped by big margin due to diplomatic crisis between UAE and Qatar.



# SUPPLEMENTARY DATA

Shopping data to aid in forecasting demand

# Shopping Data for Network Planning



# The Data

- Base historical demand data
  - Consists of historical global demand reported by market
- Exogenous historical Shopping Data
  - Consists of historical shopping data across markets

## Sample Shopping Data

| ORG | DST | YEARMONTH | DEP_YEARMONTH | REQ | BOOKED |
|---|---|---|---|---|---|
| ATL | BOS | 201408 | 201408 | 33779 | 3717 |
| ATL | BOS | 201408 | 201410 | 12622 | 909 |
| ATL | BOS | 201408 | 201505 | 287 | 1 |
| ATL | BOS | 201408 | 201504 | 306 | 2 |
| ATL | BOS | 201408 | 201409 | 27474 | 3409 |
| ATL | BOS | 201408 | 201506 | 196 | 0 |

## Sample Demand Data

| ORG | DST | PAX | YEARMONTH |
|---|---|---|---|
| ATL | BOS | 29045.15 | 201408 |
| ATL | CHI | 48327.07 | 201408 |
| ATL | DAL | 31440.27 | 201408 |
| ATL | DEN | 30472.01 | 201408 |
| ATL | FLL | 32653.29 | 201408 |
| ATL | LAS | 22212.86 | 201408 |

# The Approach: Data Preprocessing

- Join shopping data to historical demand data keeping track of
  - Markets shopped, Markets booked, Departure Date, Shopped Dated, Booked Date
- Create lagged input variables for all variables
  - Resulting data set has 89 columns

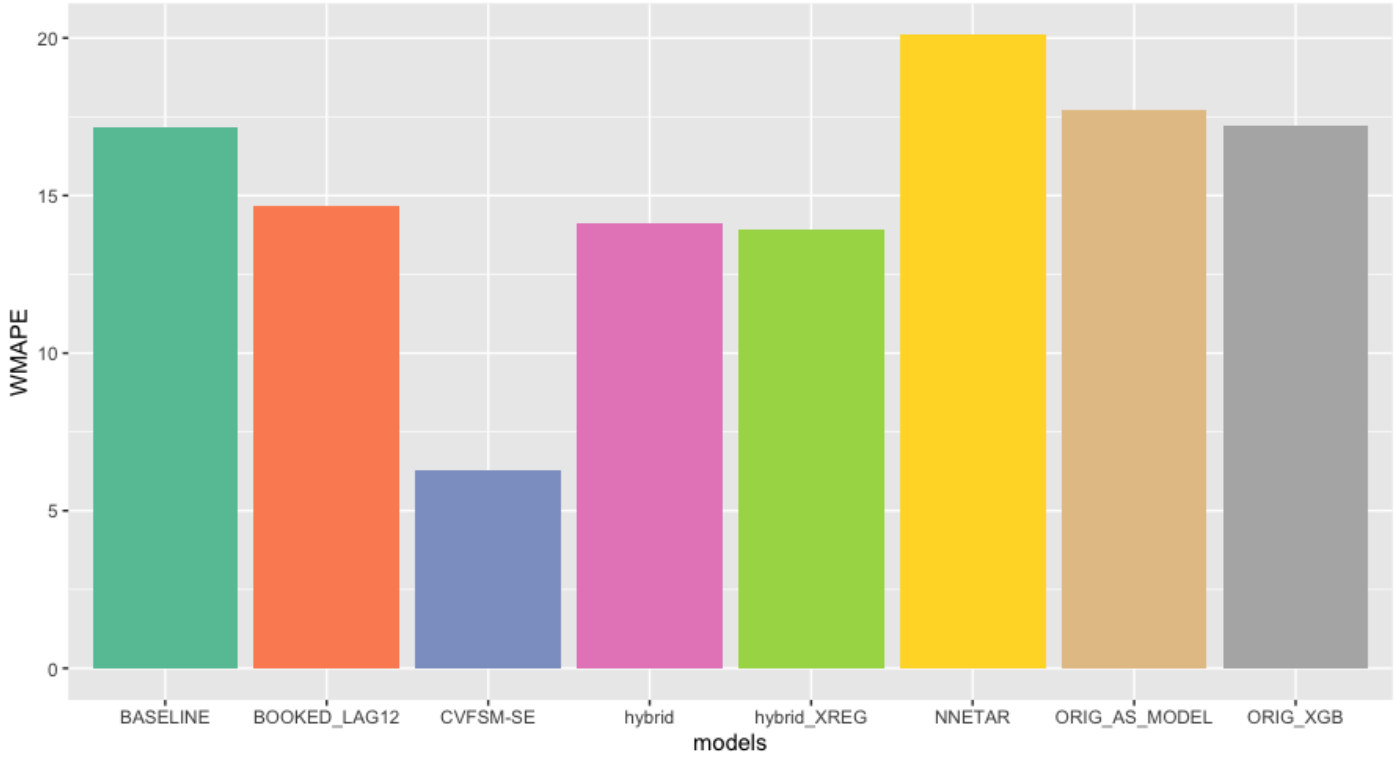
| DEPARTURE_YEARMONTH | SHOP_YEARMONTH | REQUESTED_PAX | BOOKED_PAX | req_prior_month_1 | req_prior_month_2 | req_prior_month_3 | req_prior_month_4 |
|---|---|---|---|---|---|---|---|
| 201408 | 201408 | 33779 | 3717 | NA | NA | NA | NA |
| 201409 | 201409 | 25850 | 4834 | 27474 | NA | NA | NA |
| 201410 | 201410 | 26294 | 4324 | 28607 | 12622 | NA | NA |
| 201411 | 201411 | 19610 | 2763 | 23212 | 9198 | 3711 | NA |
| 201412 | 201412 | 23349 | 2548 | 19865 | 8123 | 3613 | 1899 |
| 201501 | 201501 | 22090 | 3415 | 14391 | 3630 | 1266 | 706 |
| 201502 | 201502 | 19546 | 3347 | 16029 | 3383 | 1198 | 601 |
| 201503 | 201503 | 21456 | 4575 | 22721 | 7594 | 2622 | 1135 |
| 201504 | 201504 | 20570 | 4048 | 21389 | 8768 | 4692 | 1932 |
| 201505 | 201505 | 26592 | 3703 | 27545 | 10895 | 6969 | 4105 |

# The Approach: Time Series & ML Strategies

- For each market build series of univariate time series model and combine these models using a simple average.  (Call this combination model SE- Strong Ensemble)
- Fuse exhaustive search feature selection wrapper with xgboost model that uses as input the shopping data and demand data at different lags, say lag5, lag6, lag7,..,lag12.
- For each market using the wrapper defined above perform Cross Validation on training data to get a model that uses shopping data to predict demand data 5 months ahead. (Call this CVFSM- Cross Validated-Feature Selection)
- Combine the CVFSM and SE models using simple average.

# Results

- We predict market size 5 months in advance for 250 top NA markets
- HIST_DATA_MODEL model only uses historical data and has 17.7% error
- New CVFSM-SE model uses both historical and shopping data and has 6.3% error




| Model | WMAPE |
|---|---|
| HIST_DATA_MODEL | 17.7% |
| BASELINE | 17.2% |
| BOOKED_LAG12 | 14.7% |
| CVFSM-SE | 6.3% |
| hybrid | 14.1% |
| hybrid_XREG | 13.9% |
| NNETAR | 20.1% |
| ORIG_XGB | 17.2% |

# Future Work

- Use additional data sources
  - Econometric data like GDP
  - Advance booking data to better predict spikes/dips in demand
- Other algorithms like neural networks and random forests

Sabre®