Flights Delay Prediction

YAOHUI(ED) WU

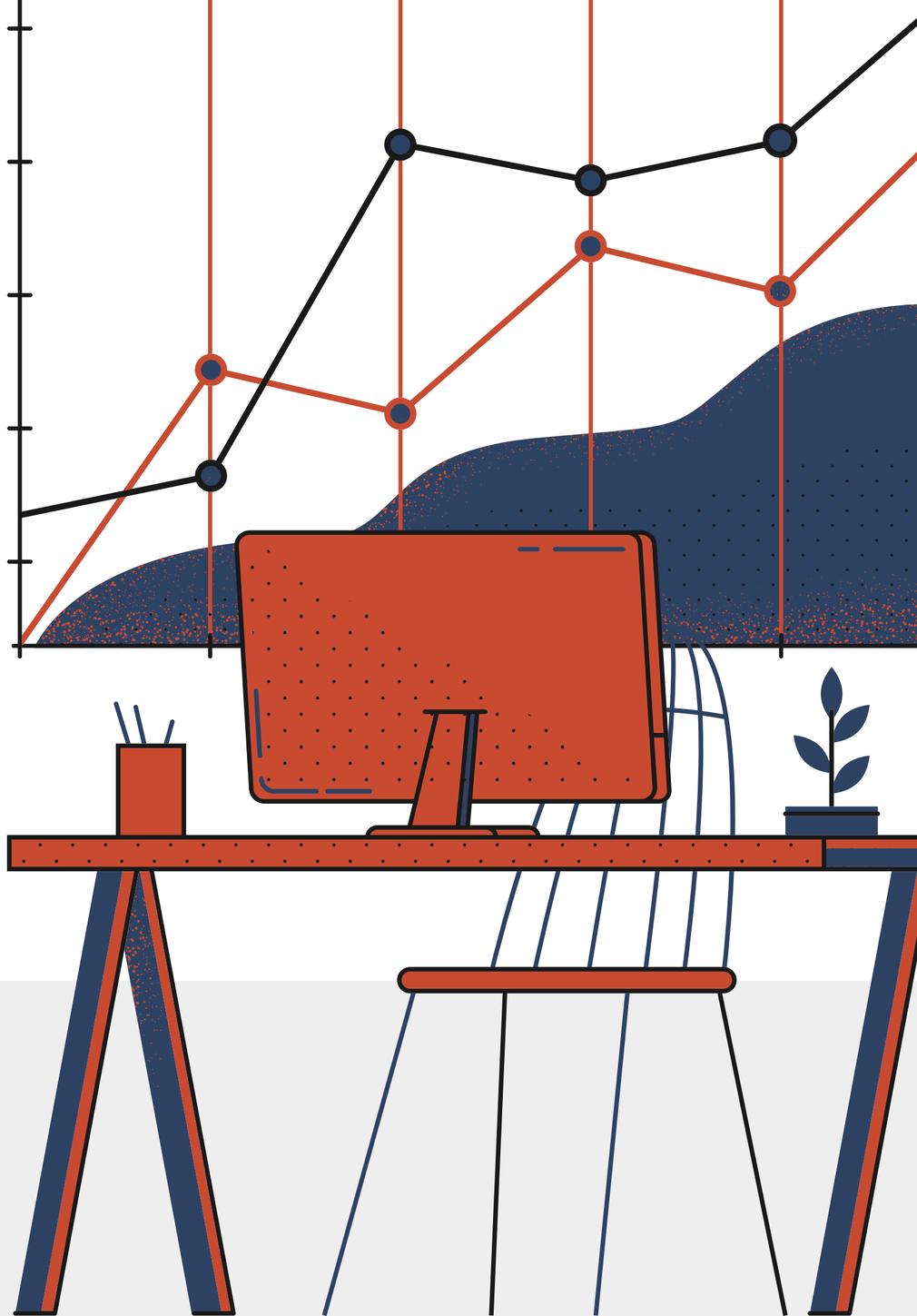


Table of Contents

- Define the problem
  - Classification or Regression
  - What information i have (data checking)
  - Ask questions
- Read Articles about flight delay
  - Analysis angle
  - Factors cause delay(help with data collection)
  - Help with modeling

Weather information

- Where
  - IEM (Iowa Environmental Mesonet) from Iowa State University
- How
  - Web Scraping (Pycharm)
- What
  - Origin, Dest
  - Jan to July each day each hour(more than 600K)
  - 6 Weather features(Air Temperature, Humidity, Wind Speed, Visibility etc)

	station	valid	lon	lat	tmpf	dwpf	relh	drct	sknt	vsby
0	ABQ	2013–01–01 00:00:00	–106.6155	35.0419	35.96	14.00	40.05	250.0	10.0	10.0
1	ABQ	2013–01–01 01:00:00	–106.6155	35.0419	35.06	14.00	41.51	280.0	8.0	10.0
2	ABQ	2013–01–01 02:00:00	–106.6155	35.0419	33.08	17.96	53.36	330.0	5.0	10.0
3	ABQ	2013–01–01 03:00:00	–106.6155	35.0419	32.00	19.04	58.37	310.0	5.0	10.0
4	ABQ	2013–01–01 04:00:00	–106.6155	35.0419	30.02	19.04	63.24	250.0	4.0	10.0

`weather` *Hourly weather data*

Description

Hourly meterological data for LGA, JFK and EWR.

Usage

```
weather
```

Format

A data frame with columns:

origin Weather station. Named `origin` to facilitate merging with `flights` data.

year, month, day, hour Time of recording.

temp, dewp Temperature and dewpoint in F.

humid Relative humidity.

wind_dir, wind_speed, wind_gust Wind direction (in degrees), speed and gust speed (in mph).

precip Precipitation, in inches.

pressure Sea level pressure in millibars.

visib Visibility in miles.

time_hour Date and hour of the recording as a `POSIXct` date.

Source

ASOS download from Iowa Environmental Mesonet, https://mesonet.agron.iastate.edu/request/download.phtml.

Planes and Airports information

- Where
  - R nycflights13 package
- What
  - Airports information
  - Planes information

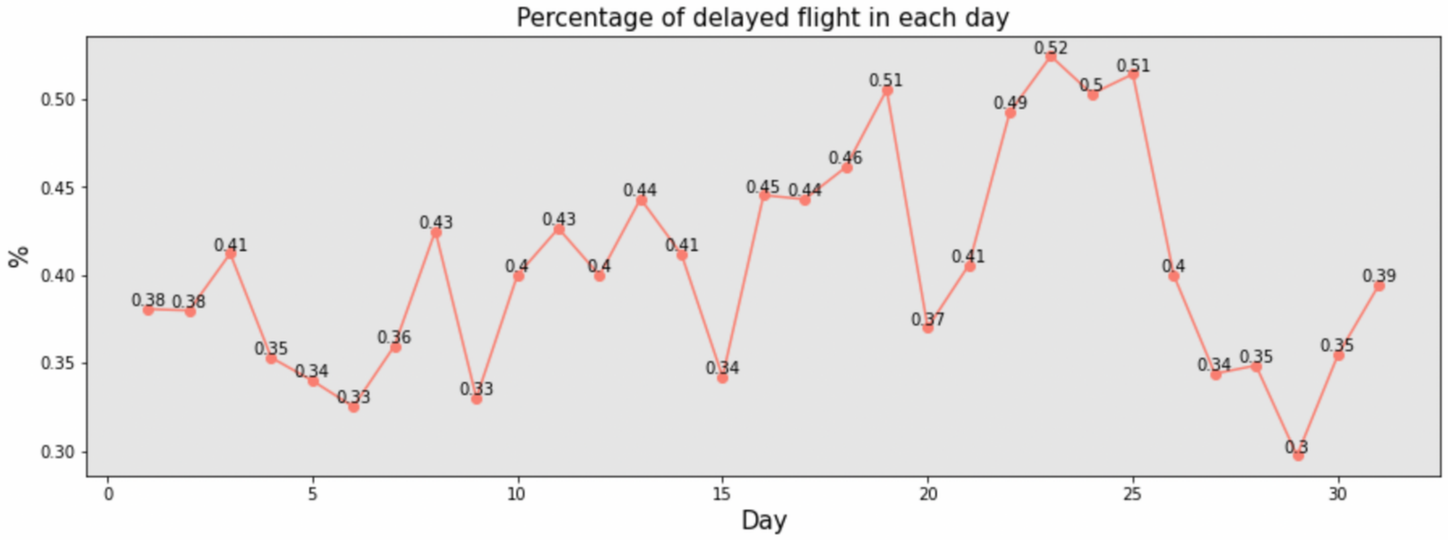
```
> planes
# A tibble: 3,322 × 9
   tailnum  year type                  manufacturer     model     engines seats speed engine
   <chr>   <int> <chr>                 <chr>            <chr>       <int> <int> <int> <chr>
 1 N10156   2004 Fixed wing multi engine EMBRAER          EMB-145XR       2    55    NA Turbo-fan
 2 N102UW   1998 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
 3 N103US   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
 4 N104UW   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
 5 N10575   2002 Fixed wing multi engine EMBRAER          EMB-145LR       2    55    NA Turbo-fan
 6 N105UW   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
 7 N107US   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
 8 N108UW   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
 9 N109UW   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
10 N110UW   1999 Fixed wing multi engine AIRBUS INDUSTRIE A320-214       2   182    NA Turbo-fan
# … with 3,312 more rows
```

```
> airports
# A tibble: 1,458 × 8
   faa   name                           lat    lon   alt    tz dst   tzone
   <chr> <chr>                        <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
 1 04G   Lansdowne Airport             41.1  -80.6  1044    -5 A     America/New_York
 2 06A   Moton Field Municipal Airport 32.5  -85.7   264    -6 A     America/Chicago
 3 06C   Schaumburg Regional           42.0  -88.1   801    -6 A     America/Chicago
 4 06N   Randall Airport               41.4  -74.4   523    -5 A     America/New_York
 5 09J   Jekyll Island Airport         31.1  -81.4    11    -5 A     America/New_York
 6 0A9   Elizabethton Municipal Airport 36.4 -82.2  1593    -5 A     America/New_York
 7 0G6   Williams County Airport       41.5  -84.5   730    -5 A     America/New_York
 8 0G7   Finger Lakes Regional Airport 42.9  -76.8   492    -5 A     America/New_York
 9 0P2   Shoestring Aviation Airfield  39.8  -76.6  1000    -5 U     America/New_York
10 0S9   Jefferson County Intl         48.1 -123.    108    -8 A     America/Los_Angeles
# … with 1,448 more rows
```

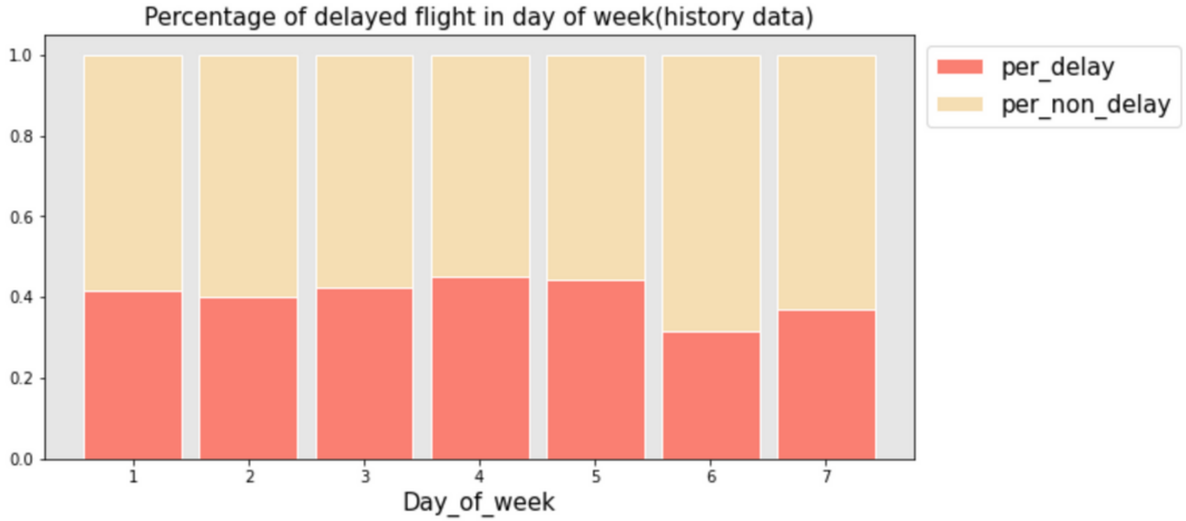
EDA (Exploratory Data Analysis)

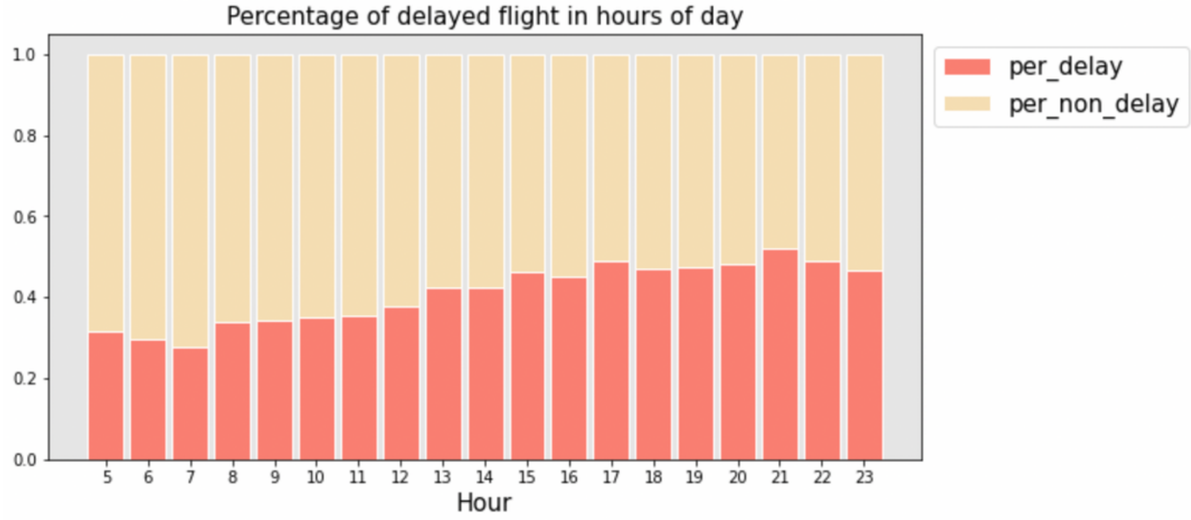
- Benefits from EDA
  - New features
  - Feature selection
  - Extract insight
- Angle
  - Time angle
    - Day of month, day of week, hour of day
  - Location
    - Departure airport, destination airport
  - Plane
    - Age, plane type, manufacture, Seats, Speed, carrier
  - Weather
    - Temperature, Humidity, Wind Speed, Visibility etc

EDA-Time Angle

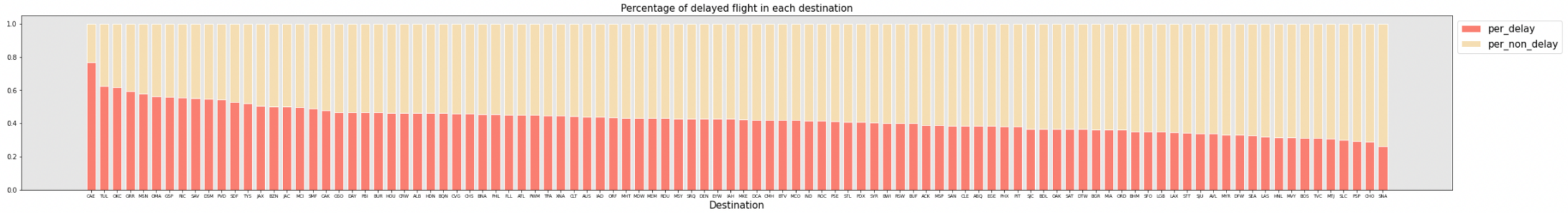


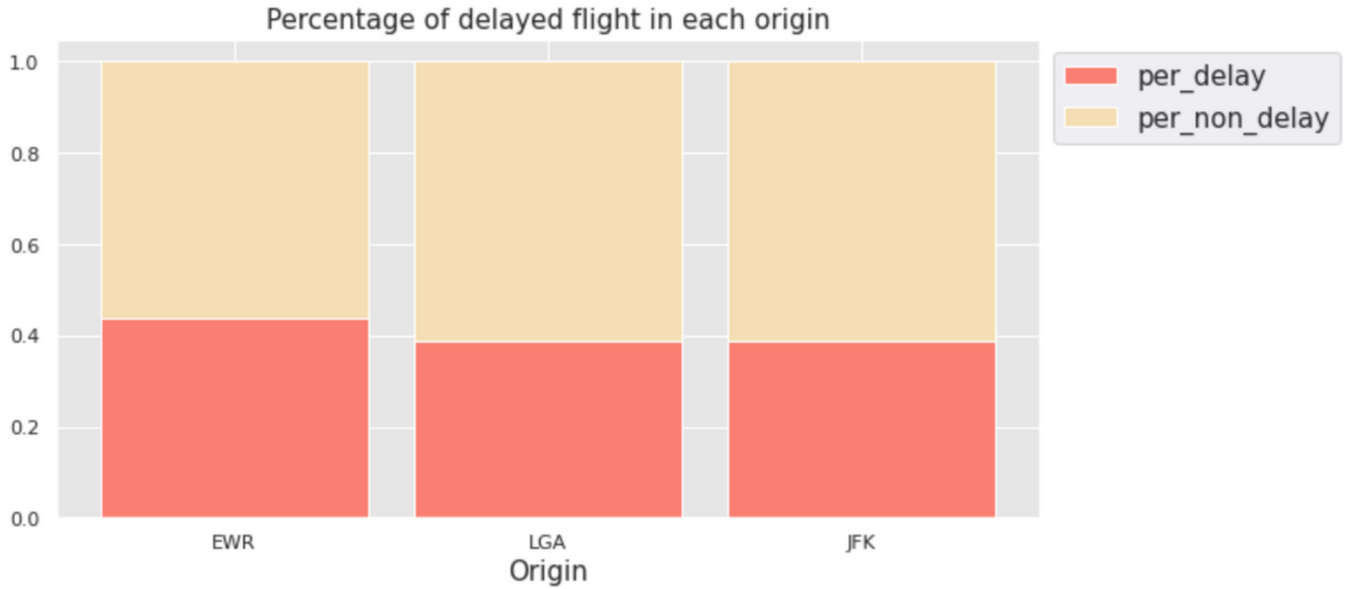
- No obvious insight from day of month
- Saturday has obvious lower delay percentage
- An increasing trend from morning to night

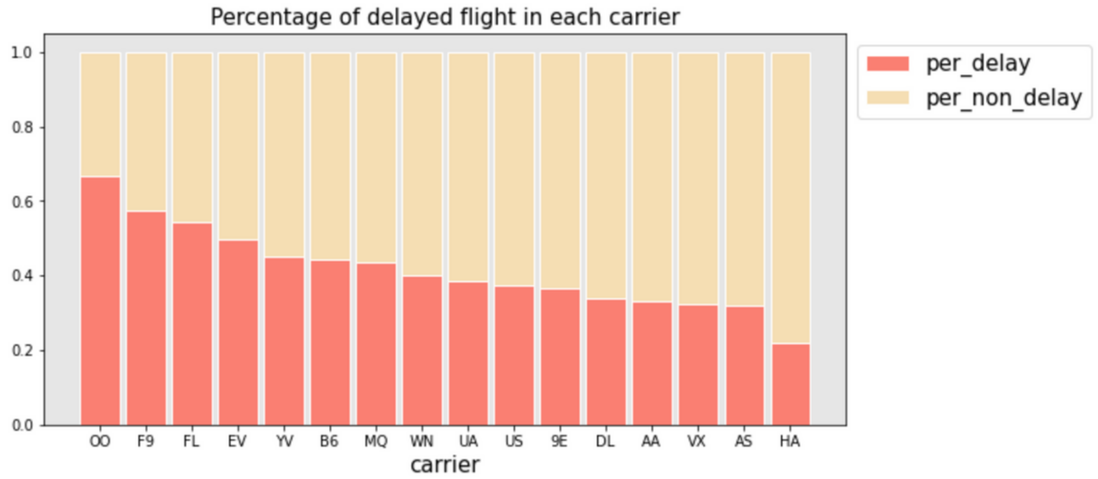




3 EDA-Location Angle & Planes Angle

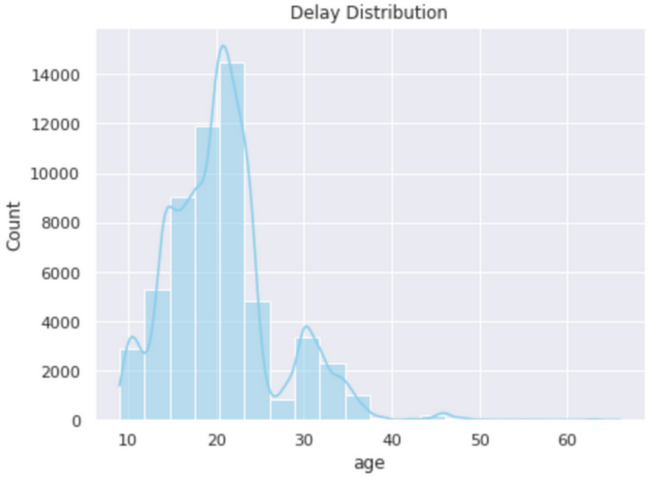


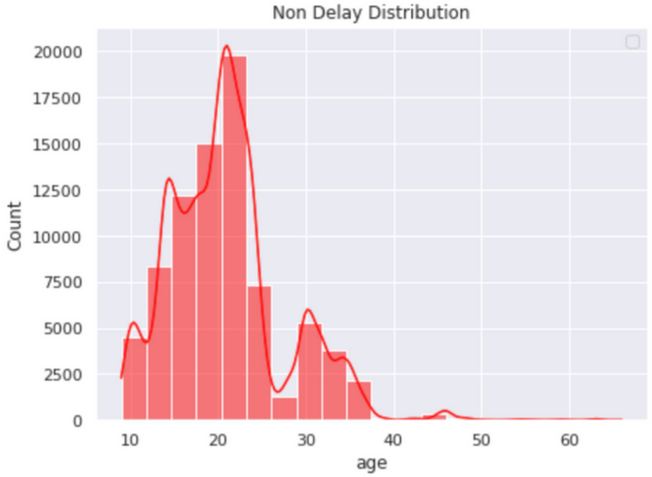


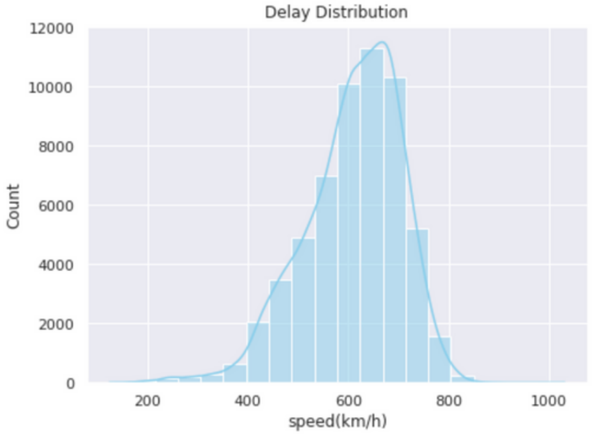


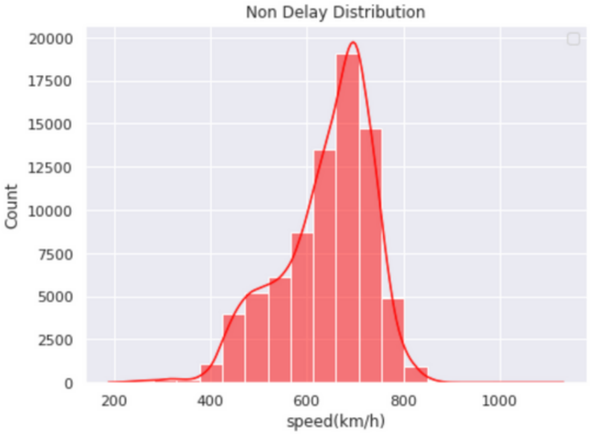
- Obvious difference in destination
- EWR has a relative higher delay percentage
- Different across carriers

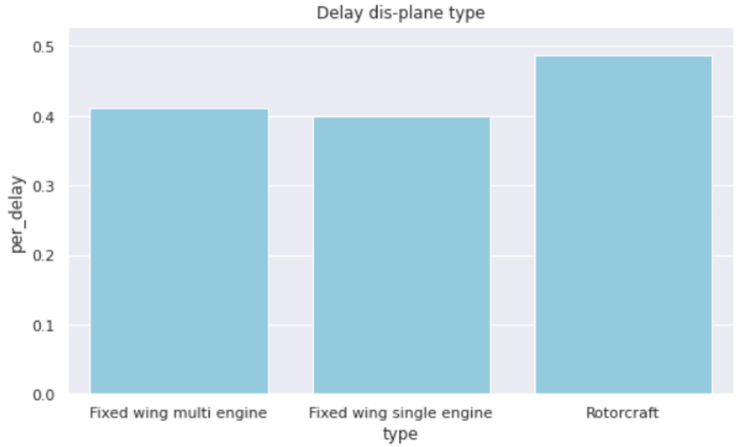
3 EDA-Plane Angle

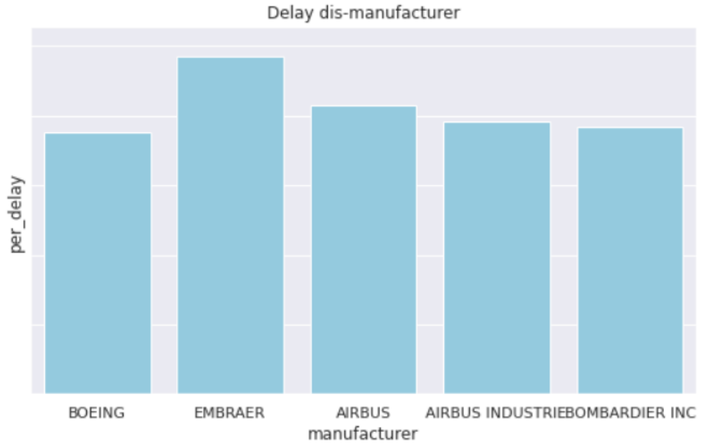


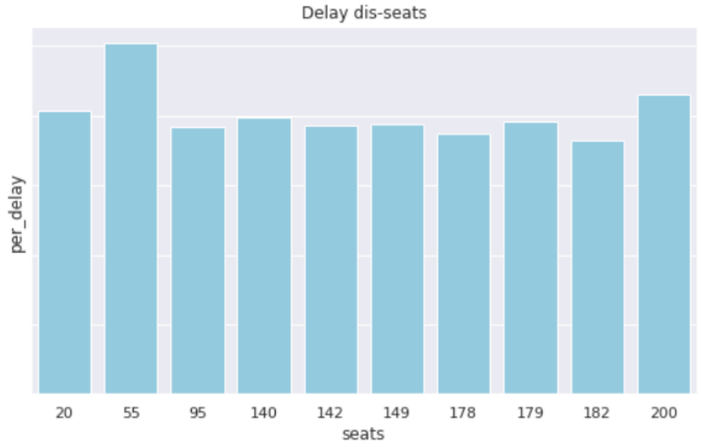






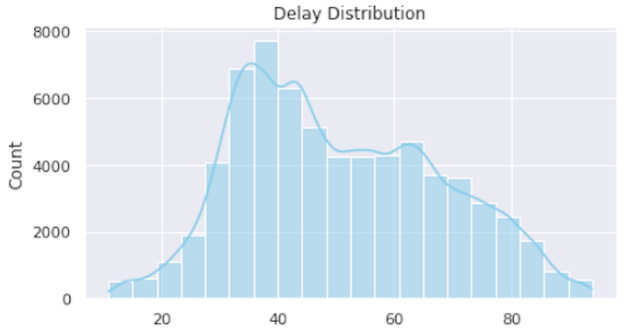


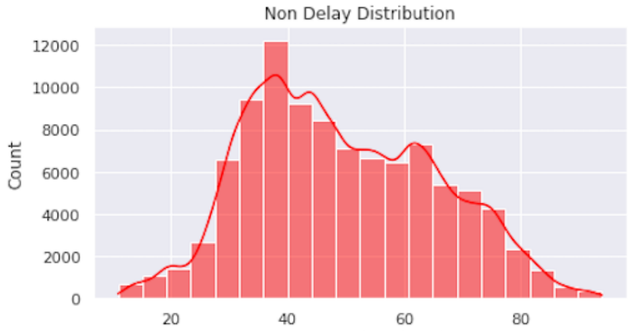


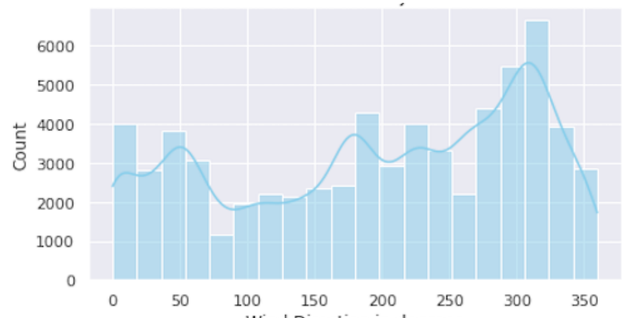


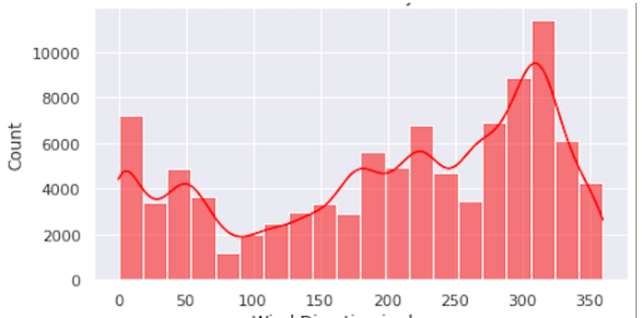
- No obvious difference from age
- Distribution of speed stayed more around 600 and left of 600

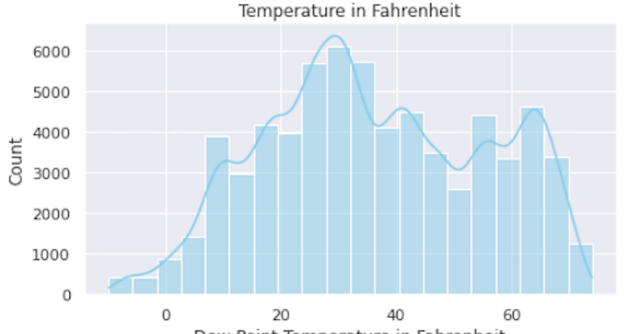
3 EDA-Weather Angle

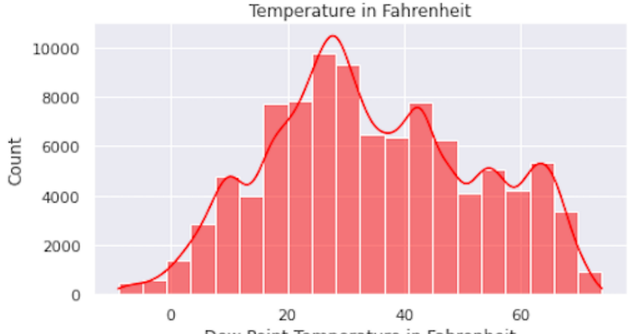


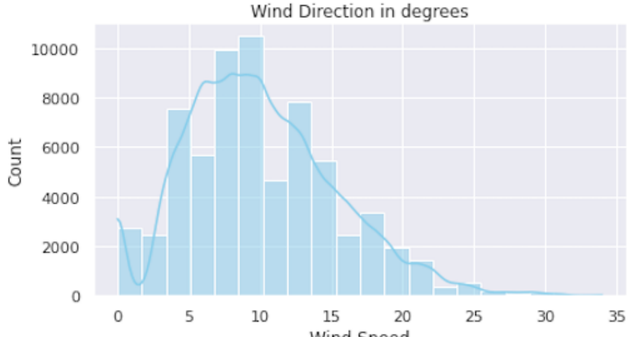


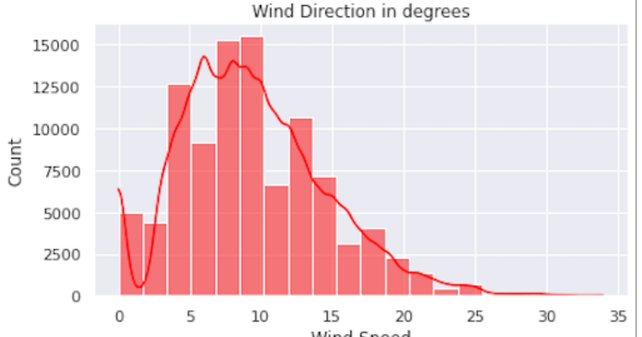


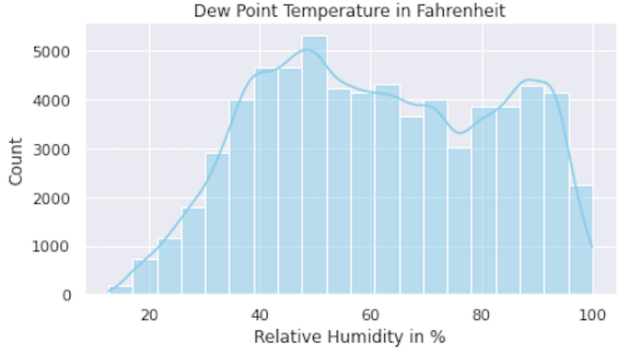


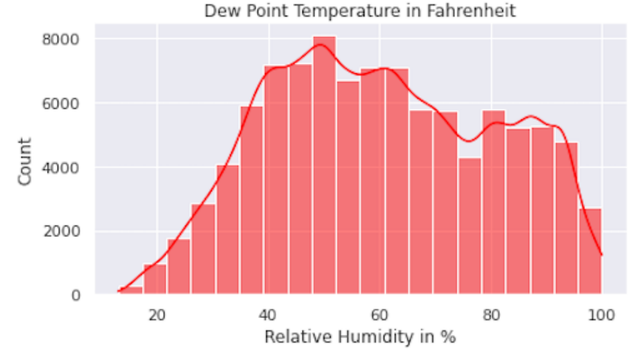


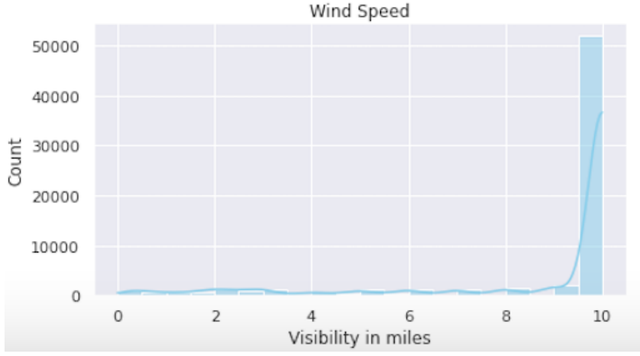


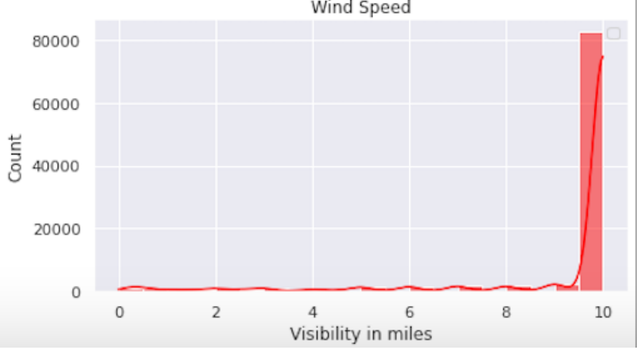








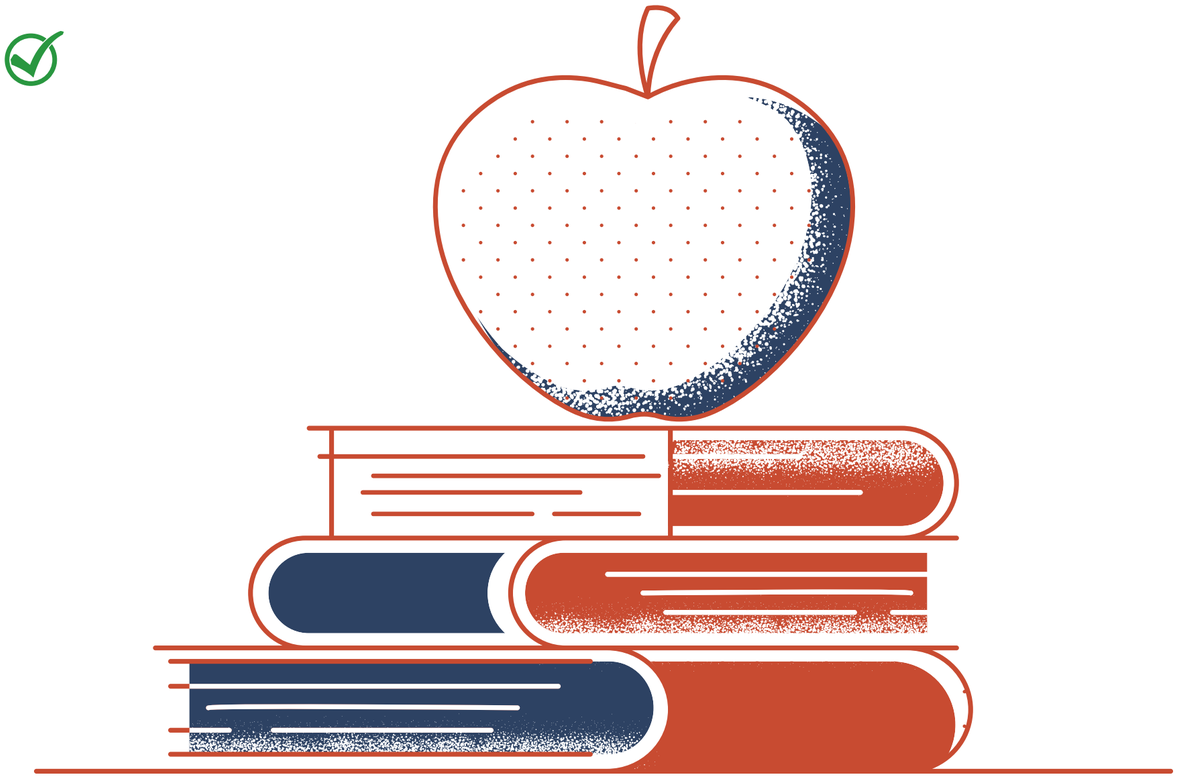




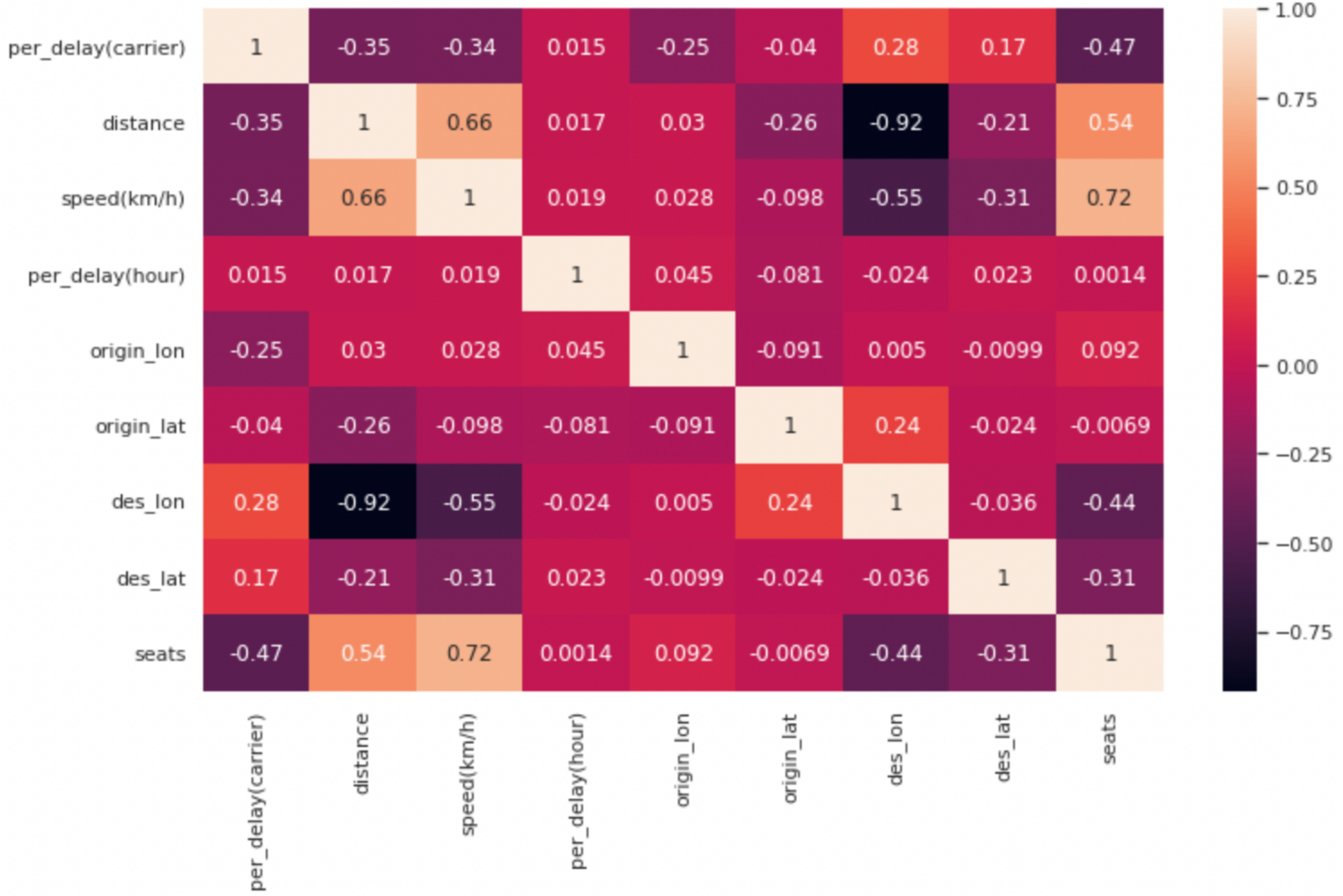
4 Data preparation(Feature Creation and Selection)

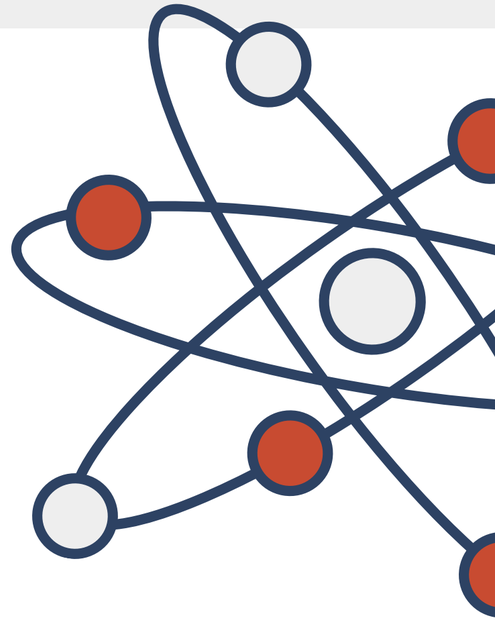
- Flight history data
  - % of delayed flight in each day (from day1 to day30)
  - % of delayed flight in day of week (from Monday to Sunday)
  - % of delayed flight in each hour (from 5:00 am to 23:00 pm) ✓
  - % of delayed flight for each carrier ✓
  - % of delayed flight in each origin
  - % of delayed flight in each destination
  - Speed of each plane (in km/h) ✓
  - Distance between origin airport and destination airport ✓
- R 'nycflights13' package
  - Age of each plane
  - Seats of each plane
  - Manufacturer of each plane…
- Weather
  - tmpf (Air Temperature in Fahrenheit)
  - relh (Relative Humidity in %)
  - drct (Wind Direction in degrees)
  - sknt (Wind Speed)
  - vsby (Visibility in miles)
  - longitude and latitude of departure airport ✓
  - longitude and latitude of destination airport ✓

- Around 20 newly created feature



4 Data preparation(Feature Creation and Selection)





- Regular classification model
  - Logistic Regression
  - KNN
- Ensemble learning
  - Random Forest
  - Adaboost
- Dataset splitting(Around 6000 in total)(around 60% training and 40% testing)
  - Training is 4000
  - Testing is 2000
- Metric for evaluation
  - Confusion matrix
  - Accuracy
  - Stratified Accuracy

	lateflight	count
0	0.0	3088
1	1.0	2698

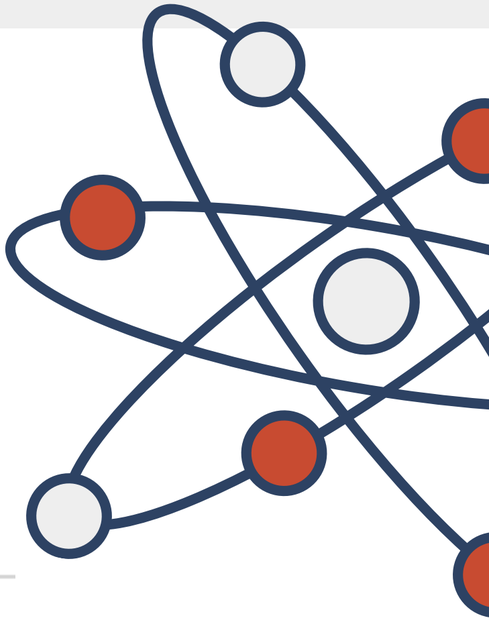
	Training Accuracy	Stratified	Testing Accuracy	Stratified accuracy
Logistic Regression	63%	Class 0: 64% Class 1: 63%	63%	Class 0: 64% Class 1: 62%
KNN	67%	Class 0: 75% Class 1: 59%	57%	Class 0: 64% Class 1: 49%
Random Forest	69%	Class 0: 73% Class 1: 65%	64%	Class 0: 69% Class 1: 59%
Adaboost	65%	Class 0: 68% Class 1: 61%	62%	Class 0: 66% Class 1: 58%

```
[[592 361]
 [293 540]]
Accuracy: 0.6338185890257558
              precision    recall  f1-score   support

         0.0       0.67      0.62      0.64       953
         1.0       0.60      0.65      0.62       833
```

Modeling & Optimization

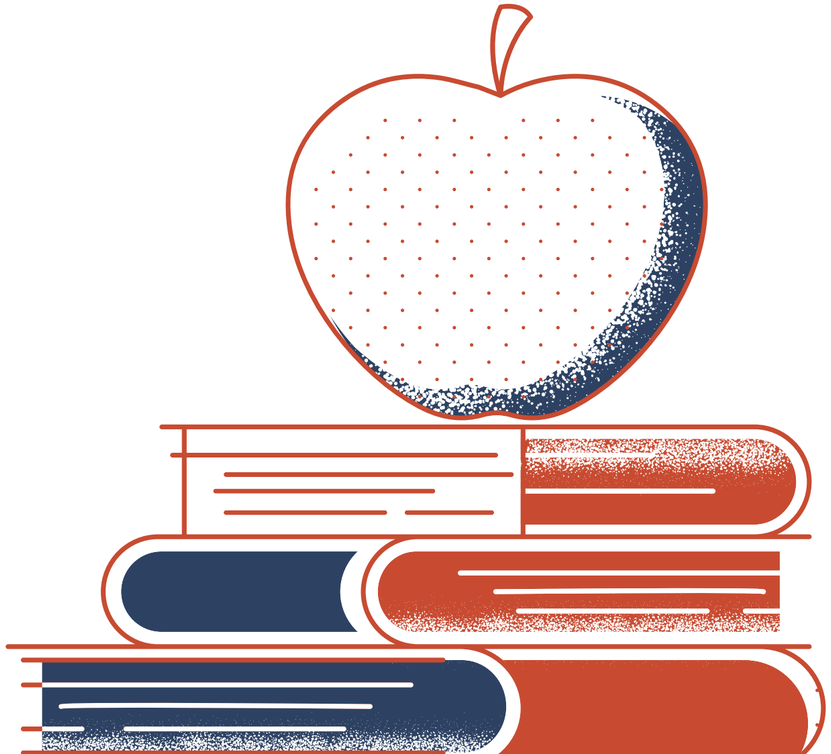
- Add additional Features
- Tune parameters for each model
- Split dataset into training and testing with different proportion
- Feature scaling



	Training Accuracy	Stratified	Testing Accuracy	Stratified accuracy
Logistic Regression	65%	Class 0: 71% Class 1: 57%	65%	Class 0: 70% Class 1: 60%
KNN	73%	Class 0: 76% Class 1: 70%	58%	Class 0: 62% Class 1: 53%
Random Forest	77%	Class 0: 84% Class 1: 69%	67%	Class 0: 71% Class 1: 62%
Adaboost	70%	Class 0: 74% Class 1: 64%	67%	Class 0: 69% Class 1: 64%

Conclusion & Further Analysis

- Why some of the results from EDA is not as expected
  - Weather feature choice
  - Angle problem(plane itself)
- Problem about modeling
  - Delay is influenced by many factors
- From data aspect
  - Additional features like demand of an airport, whether it's a connecting flight or not, passenger's comments.
  - Whether there is any connecting passenger/bags or not.
  - Air traffic control.
  - Actual arrival time
  - Other weather features like rainy, heavy rainy, sunny, storm etc.
  - Acquire more observations.
- From model aspect
  - Employ new models.
  - Combine the result of different models



The End

Thank you for listening

Q & A