Prediction about the U.S.A. Flight's Arrival Delay Departed from Washington State based on its Potential Influential Factors

~ AUTHORS: Albert Yee, Jiening Fu, Yixuan Yao ~

Table of Contents

Summary of Questions and Results

(Q1) From the visualization (flight delay distribution plots), can we discern some specific factors that influence the flight arrival delay? (such as airlines, states, months, or even arrival delays based on departure delays) (The use of Pandas will be a big part)

Result:

We do find some factors that are influential to flight delays from WA state to any airports in the U.S.A., like seasons, airlines, states and regions, departure delay. However, we also surprisingly find that some factors are not influential, like the time duration a plane slides in the origin or destination airport. And days in a week have only little influence on arrival delays based on the graphs we plot. The details will be explained with pictures in the **"Results"** later.

(Q2) Could we set some standards to classify if one airline is "seriously delayed"? Could we use classification methods to predict if a flight will be "seriously delayed" given specific features? How good are the classification performances?

Result:

Yes, based on each flight's arrival time length and its planned flight time length, we could calculate the duration ratio of delay for each flight in the sample dataset. After we calculate the delay-ratio at TOP 25% quantile and the delay-ratio at 1 standard deviation above mean, we calculate the mean value of these two delay-ratio values, which is 15.689133139%. Thus, we classify all the flights whose delay-ration to the planned time greater than 15.689133139% as "seriously delayed" flights. After assigning such labels to the flights in the dataset, we train several classification models, including K-nearest neighbors algorithm (k-NN classification) and Decision Tree classification.

We found that both KNN model and Decision Tree Model did a really good job in classifying whether a flight will be “seriously delayed”. Both of them had an accuracy of over 98%: with big enough depths, the decision tree model could even achieve nearly 100% in terms of accuracy and f1-score. In summary, the use of a decision tree model is very dependable in predicting if a flight will be “seriously delayed”.

(Q3) Could the delay time of flights be predicted? How well could it be predicted? Will the regression prediction be good enough as a reference?

Result:

Yes, the delay time of flights could be well predicted with small errors. We tried three regression models and evaluated their performances to pick up the optimal one for future prediction use. Based on the results comparison and analysis, the decision tree regression model is the best to use. The decision tree shows a great power in predicting the delay time with a very trivial/tiny error (RMSE as small as 5, while MAE is just 2.8). Such errors would not impose significant impacts on the effectiveness of prediction results. In summary, we could use multiple models to predict the delay time of flights well, and the decision tree model proves that it is the best and most reliable tool to use for our stakeholders.

Motivation

With the development of the aviation industry and the business world, flight has been an important part of our life, especially for business people and professionals in many industries. However, flight delays have been troublesome and are affecting flight experiences and consumers’ satisfactions seriously. In addition to hurting the emotions, flight delays could even lead to failures that people want to avoid. Specifically and for example, passengers want to learn about the duration time they’ll spend in the airplane to decide how they could schedule events and activities; family, friends, or people who will pick up the passengers want to know more exact arrival times; airline companies want to know when, where, and how the flight delays would happen and to what extent. Generally, we want to find solutions to all those questions to help people with the decisions and to improve the flight experience.

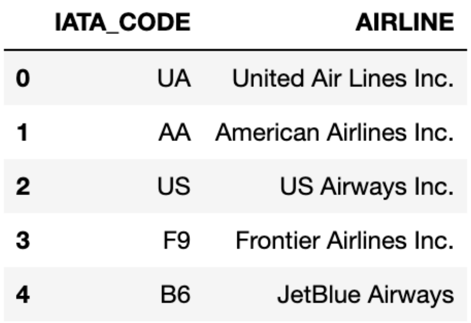
Besides, we are curious if change of seasons or days in weeks (holidays, vacations) influence the light delay in the US, possibly more people will take flights during summer months and December to travel. Are the flight delays correlative to the month/quarter (to be extended to connect with the increase of travel population)? Also, we are interested in whether the delay of flights is correlated to the state, because different state's climate and terrain conditions are different.

With the mind that consumers (the public) and airline companies will be our main target audience, we'll try to analyze the statistical data, find helpful information, and generate business intelligence.

Dataset

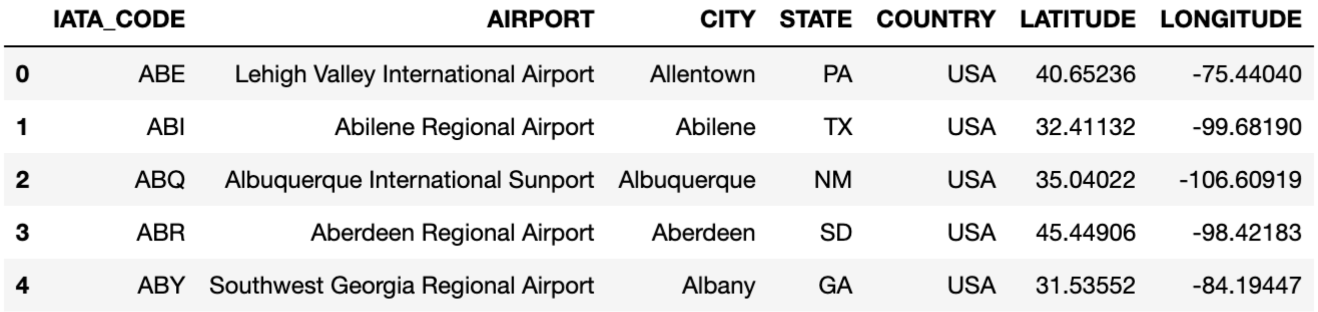
Our main dataset is "2015 Flight Delays and Cancellations" which is about the on-time performance of domestic flights operated by the large air carriers in 2015. The data contains three csv files, airlines.csv, airports.csv, and flights.csv.

- The "airlines.csv" file: it contains the name of the airline and its location identifier.



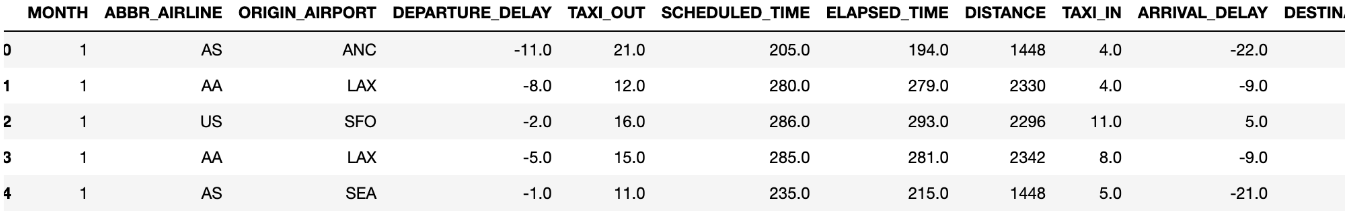
	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways

- The “airports.csv” file: it contains the name of the airports and their detailed location information.



	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
0	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.44040
1	ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.68190
2	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
3	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
4	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447

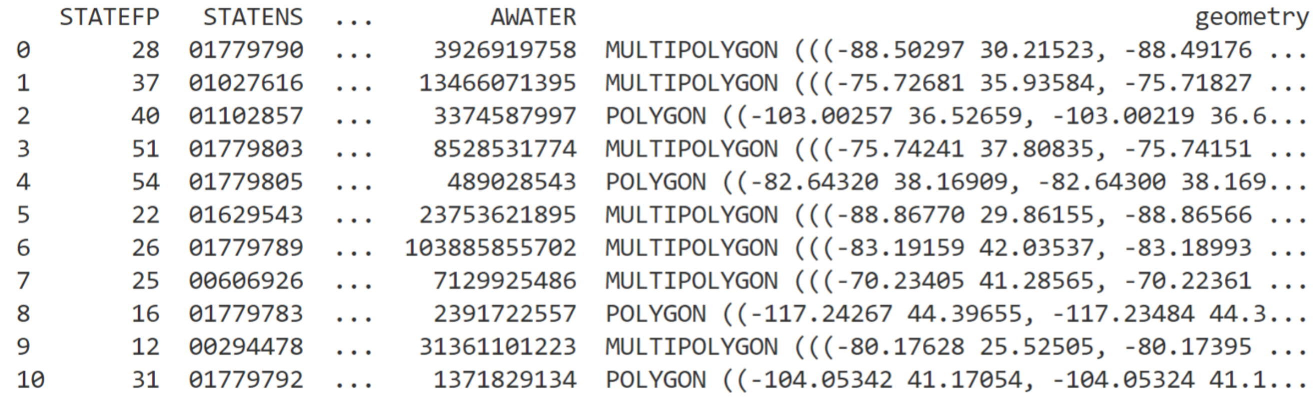
- The “flights.csv” file: it contains the flight information, the departure and arrival time and its delay status, and some other elements related to the flights.



	MONTH	ABBR_AIRLINE	ORIGIN_AIRPORT	DEPARTURE_DELAY	TAXI_OUT	SCHEDULED_TIME	ELAPSED_TIME	DISTANCE	TAXI_IN	ARRIVAL_DELAY	DESTIN.
0	1	AS	ANC	-11.0	21.0	205.0	194.0	1448	4.0	-22.0	
1	1	AA	LAX	-8.0	12.0	280.0	279.0	2330	4.0	-9.0	
2	1	US	SFO	-2.0	16.0	286.0	293.0	2296	11.0	5.0	
3	1	AA	LAX	-5.0	15.0	285.0	281.0	2342	8.0	-9.0	
4	1	AS	SEA	-1.0	11.0	235.0	215.0	1448	5.0	-21.0	

URLs: https://www.kaggle.com/datasets/usdot/flight-delays

Our second dataset is “The cb_2018_us_state_5m.zip " file which is a zip of geospatial dataset files in different formats showing the map of 51 states in the United States. We will use the one with shp file, “The cb_2018_us_state_5m.shp”



```
   STATEFP   STATENS  ...        AWATER                                           geometry
0       28  01779790  ...    3926919758  MULTIPOLYGON (((-88.50297 30.21523, -88.49176 ...
1       37  01027616  ...   13466071395  MULTIPOLYGON (((-75.72681 35.93584, -75.71827 ...
2       40  01102857  ...    3374587997  POLYGON ((-103.00257 36.52659, -103.00219 36.6...
3       51  01779803  ...    8528531774  MULTIPOLYGON (((-75.74241 37.80835, -75.74151 ...
4       54  01779805  ...     489028543  POLYGON ((-82.64320 38.16909, -82.64300 38.169...
5       22  01629543  ...   23753621895  MULTIPOLYGON (((-88.86770 29.86155, -88.86566 ...
6       26  01779789  ...  103885855702  MULTIPOLYGON (((-83.19159 42.03537, -83.18993 ...
7       25  00606926  ...    7129925486  MULTIPOLYGON (((-70.23405 41.28565, -70.22361 ...
8       16  01779783  ...    2391722557  POLYGON ((-117.24267 44.39655, -117.23484 44.3...
9       12  00294478  ...   31361101223  MULTIPOLYGON (((-80.17628 25.52505, -80.17395 ...
10      31  01779792  ...    1371829134  POLYGON ((-104.05342 41.17054, -104.05324 41.1...
```

URL:https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

Methods

(Q1) From the visualization (flight delay distribution plots), can we discern some specific factors that influence the light delay? (such as airlines, states, months, or even arrival delays based on departure delays) (The use of Pandas will be a big part)

Firstly, Pandas library are used to merge 4 data frames each containing different aspects of information into a single one. Also, some functions in Pandas are used to calculate some new columns like "seriously_delay" and "DELAY_RATIO_TIME" and add them to the dataframe. ".loc[ ]" function from Pandas are used to filter out irrelevant columns. Furthermore, groupby function in Pandas is also used to create mean values of flight delay's ratio based on come categorical variable columns such as airlines and states, which assists to making histograms later.

Secondly, Seaborn is used to create histograms and scatter plots, and matplotlib is used to change the behavior of the plots, like adding x, y axises' name and title. More importantly, matplotlib library enables subplots of histogram, which is helpful for visualizing different months and different days in a week's flights' serious delay parallely in a single plane. Moreover, some plots use the matplotlib library to change the range of axes and “sharey = True” to make the plot more readable.

Thirdly, geopandas library is used to change a typical csv file with a column “geometry” into a geoDataFrame file so that it can produce a colorful map with legend when using the ".plot( )” function.

(Q2) Could we set some standards to classify if one airline is “seriously delayed”? Could we use classification methods to predict if a flight will be “seriously delayed” given specific features?

Pandas will be used to extract the columns about the flight times. Then we calculate the delay ratio by (delay_time / total flight time). Then we will discuss and set a standard for a “serious delay”, and add a new column “SERIOUS_DELAY” indicating whether the delay is a serious delay, if the flight is seriously delayed, then mark “1”, if not, then mark “0”.

The standard we set is the sum of half of the 75% quantile of the delay ratio and half of the delay ratio value that is 1 standard deviation above its mean. [0.5 * 75% quantile of delay ratio + (1 standard deviation of delay ratio+ mean of delay ratio)].

After adding the label of “seriously delay”, we make it as our target label y. We set ['MONTH', 'ABBR_AIRLINE', 'DEPARTURE_DELAY', 'ORIGIN_AIRPORT', 'TAXI_OUT', 'SCHEDULED_TIME', 'ELAPSED_TIME', 'DISTANCE', 'TAXI_IN', 'DESTINATION_AIRPORT', 'AIRLINE'] as the features X because we think they may have impacts on whether a flight would be “seriously delayed” or not. Now, given that we have both the label y and the feature variables X, we’re ready to run machine learning classification models.

Generally, we use both KNN and DecisionTree models to do the classification work. To use the KNN algorithm, we import the KNeighborsClassifier from sklearn.neighbors. To measure the “nearby” accurately, knowing that variables with the largest scales would dominate and skew results, we want to make sure that we do a reasonable feature scaling and put all feature variables into the same scale. To normalize the features, we import the StandardScaler from sklearn.preprocessing. After the preparation work, we could then run the KNN model. At the first trial, we set n_neighbors to be 3 to fit the KNN classification model.

To use the Decision Tree Classification model, we import the DecisionTreeClassifier from sklearn.tree to train our model. We will first train our model with a decision tree max depth at 3 and output its confusion matrix and report for analysis and draw the decision tree. And then we will loop over the different max depth to see which one optimized our model.

(Q3) Could the delay time of flights be predicted? How well could it be predicted? Will the regression prediction be good enough as a reference?

We plan to use several Machine Learning models to do this part. We're going to use DecisionTreeRegression, KNN, and Linear Regression to do the regression and predict the exact delay time given the testing samples or new data.

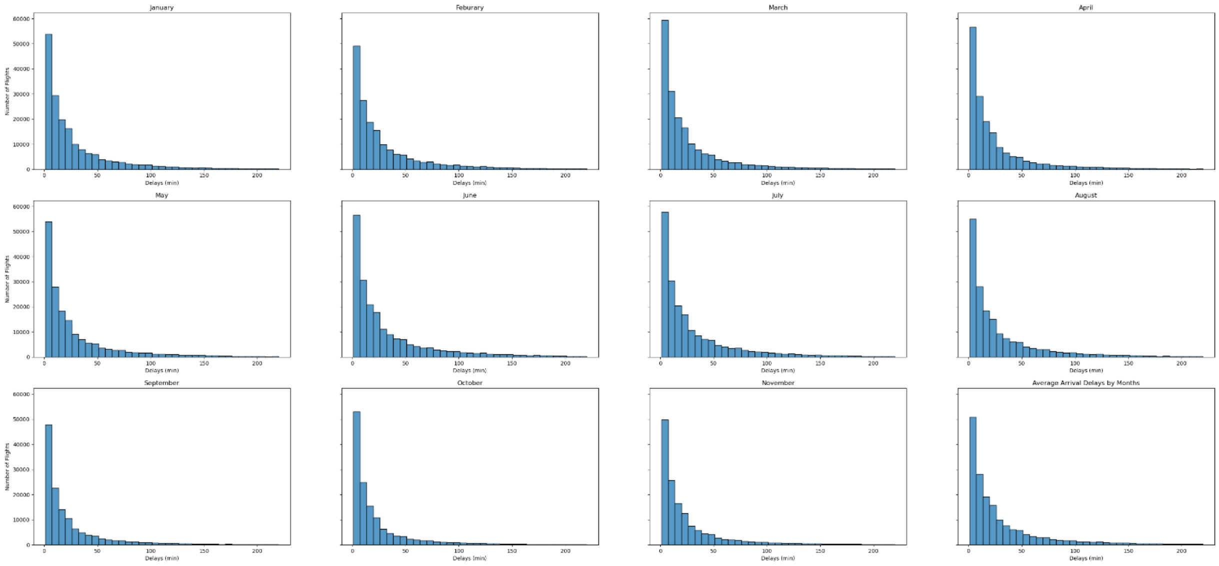
We set ['MONTH', 'ABBR_AIRLINE', 'DEPARTURE_DELAY', 'ORIGIN_AIRPORT', 'TAXI_OUT', 'SCHEDULED_TIME', 'ELAPSED_TIME', 'DISTANCE', 'TAXI_IN', 'DESTINATION_AIRPORT', 'AIRLINE'] as the features X because we think they may have unignorable impacts on the arrival delay time. Now, given that we have both the label y and the feature variables X, we're ready to run machine learning classification models.

We'll also measure the performance of the regressions by using some metrics, such as mean square errors. Specifically, statsmodels will be used to evaluate the regression performance. What's more, we'll use something like for loops to test which depth used in the decision tree will help arrive at a more precise prediction.

Results

(Q1) From the visualization (flight delay distribution plots), can we discern some specific factors that influence the light delay? (such as airlines, states, months, or even arrival delays based on departure delays) (The use of Pandas will be a big part)

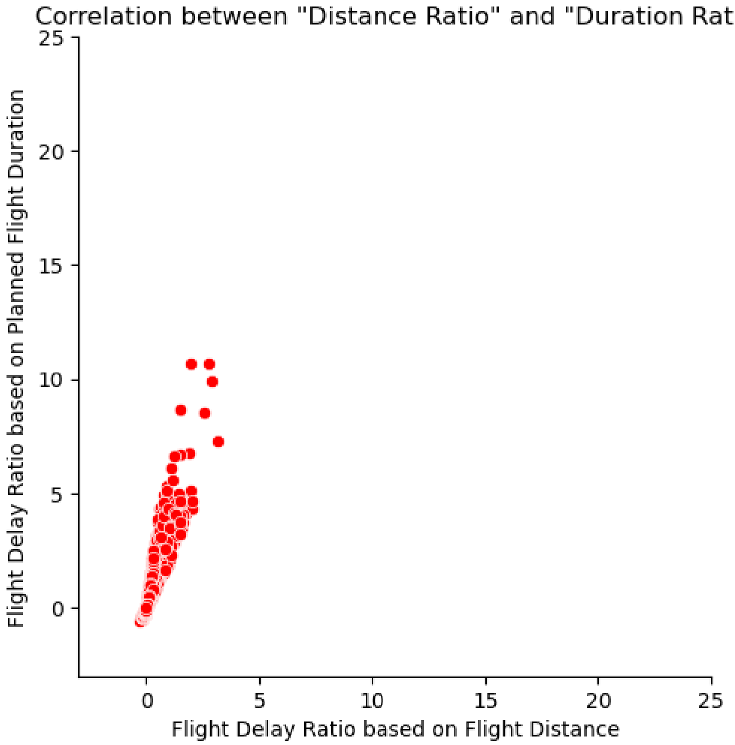
Why omit the data in October:



Based on the distribution of seriously-delayed flights across all airports in the U.S.A. (both departure and arrival airports), we find all of the 12 months have roughly the same pattern and no obvious outliers of some months have much more or less seriously-delayed flights compared to other months'. (This can be discerned by the total volume of the histogram bars, by noticing the y-axis is the same for the 12 subplots.) As a result, dropping the October month will not cause a significant bias for revising the dataset although October has slightly fewer serious delays (in acceptable range).

We don't intentionally delete all flights' information in October, but all October flights' airport information is chaotic: all airports are represented by some random numbers instead of the name. And this will cause hundreds of thousands of new variables in later machine learning's regression model, which will not be understandable. Also, the large number of columns created by the ".get_dummies( )" function will cause very slow computation.

Why choose Delay-Time-Ratio as Standard and not consider Delay-Distance Ratio?

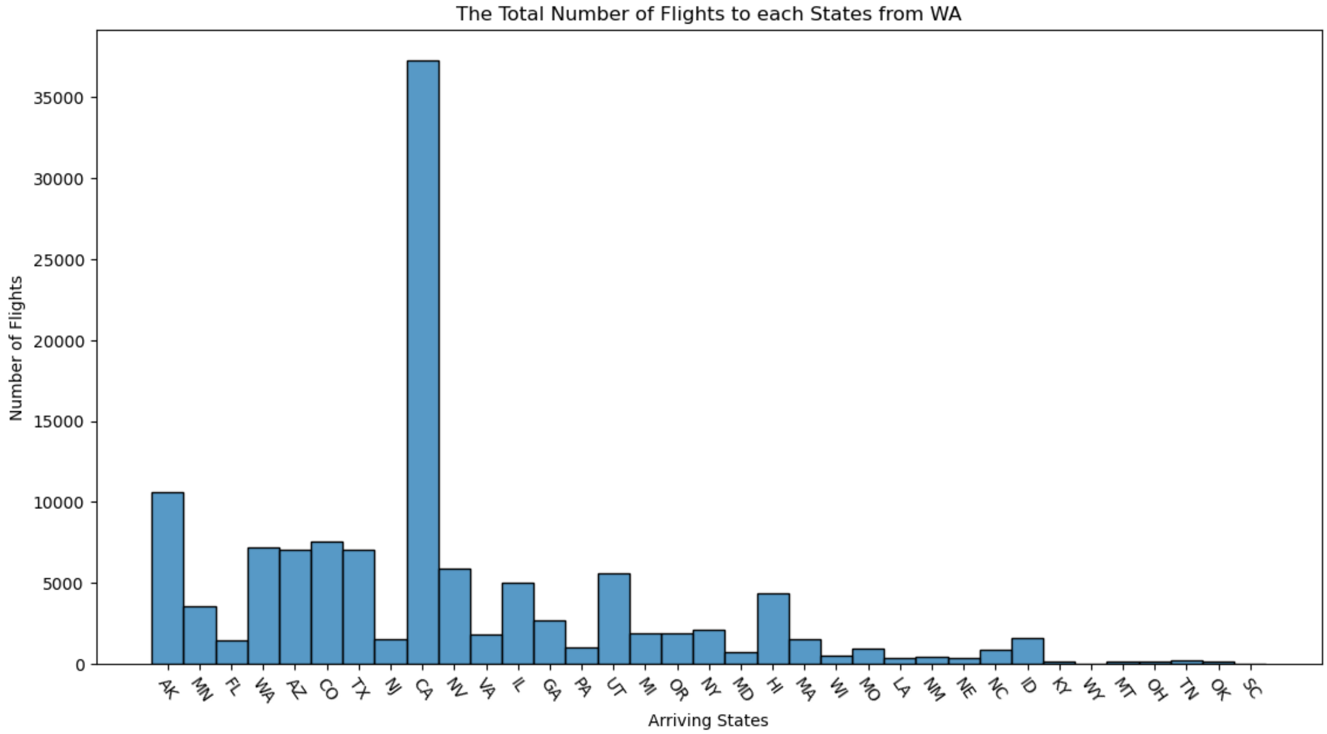


At first, our group think about two ways to calculate the degree of flight delays:

1. divide delayed time length by the planned flight length: Delay-Time-Ratio
2. Divide delayed time length by the distance between two airports: Delay-Distance-Ratio

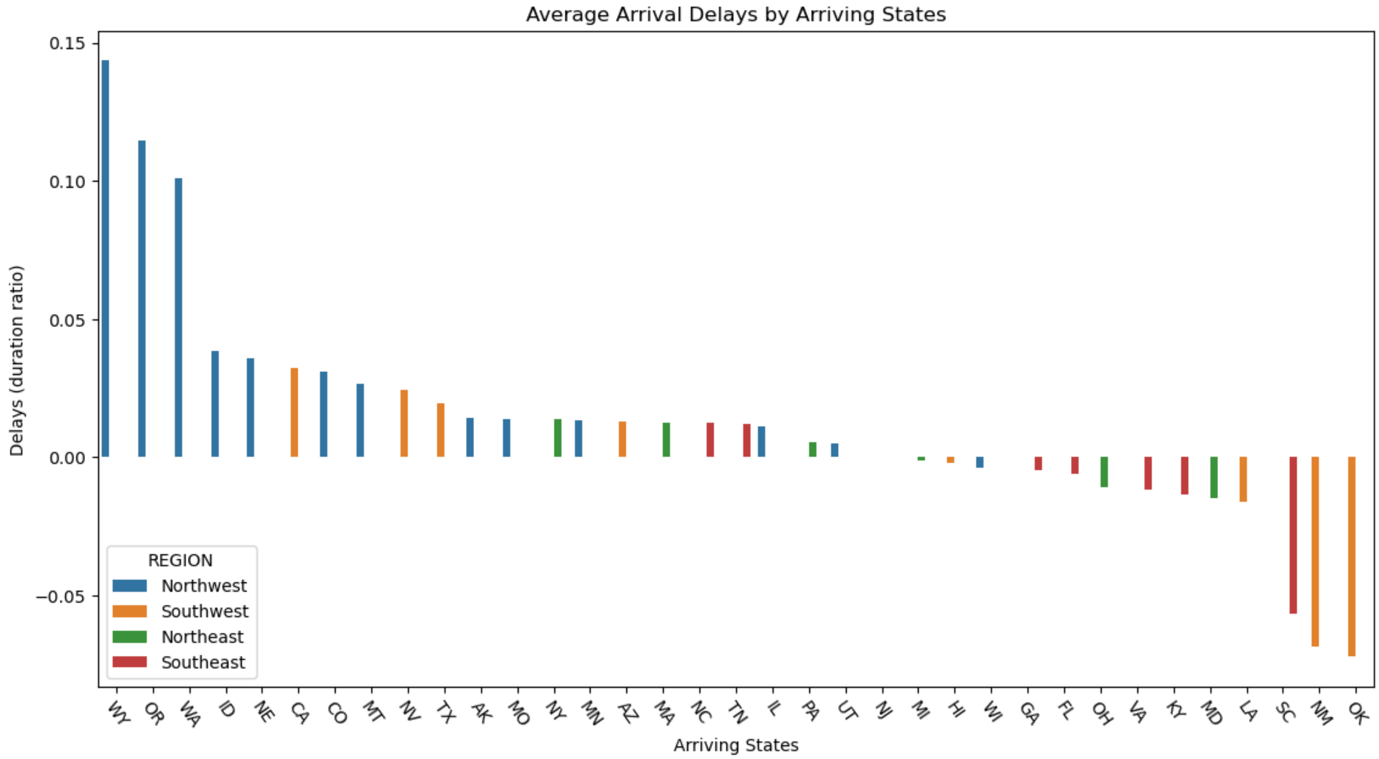
And as we draw a scatter plot with x-coordinate equals to Delay-Distance-Ratio and y-coordinate equals to Delay-Time-Ratio, we could observe that the two measurement values are strongly positively correlated. Thus, selecting any one of the two values as an indicator for degree of flight delays is workable, so we mainly use Delay-Time-Ratio as the indicator for later research topics.

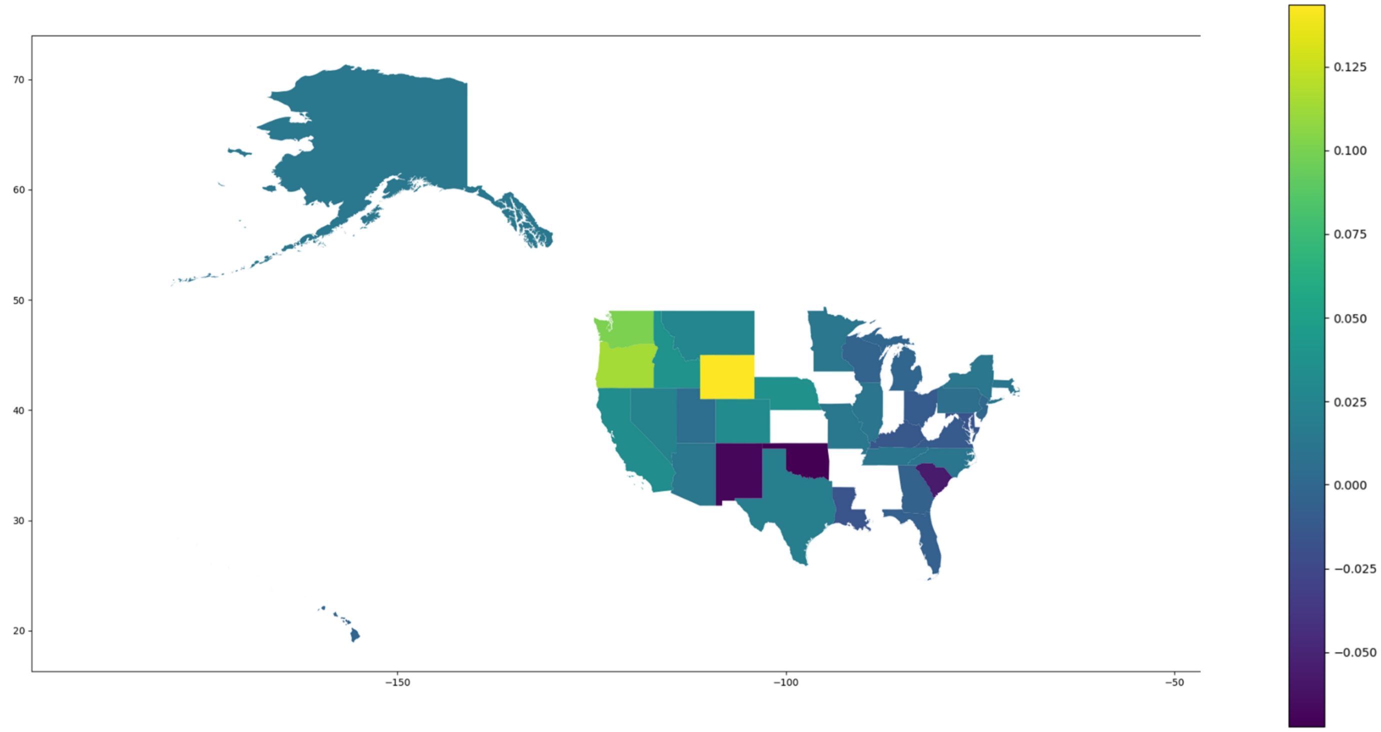
Flights per states:



For the total number of flights in general, the arrival states' flights distribution from Washington state's airports shows that the states that had the most flights with WA in the year 2015 are California, followed by Western states such as Alaska, Colorado, Washington itself, Arizona and Texas. East states usually have poor connection with WA through flights.

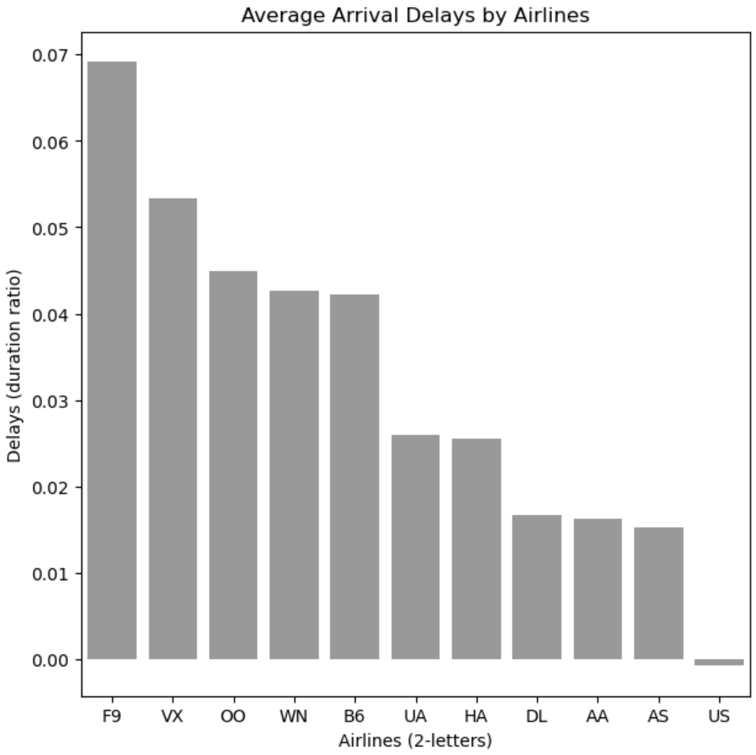
Average Delay per States：





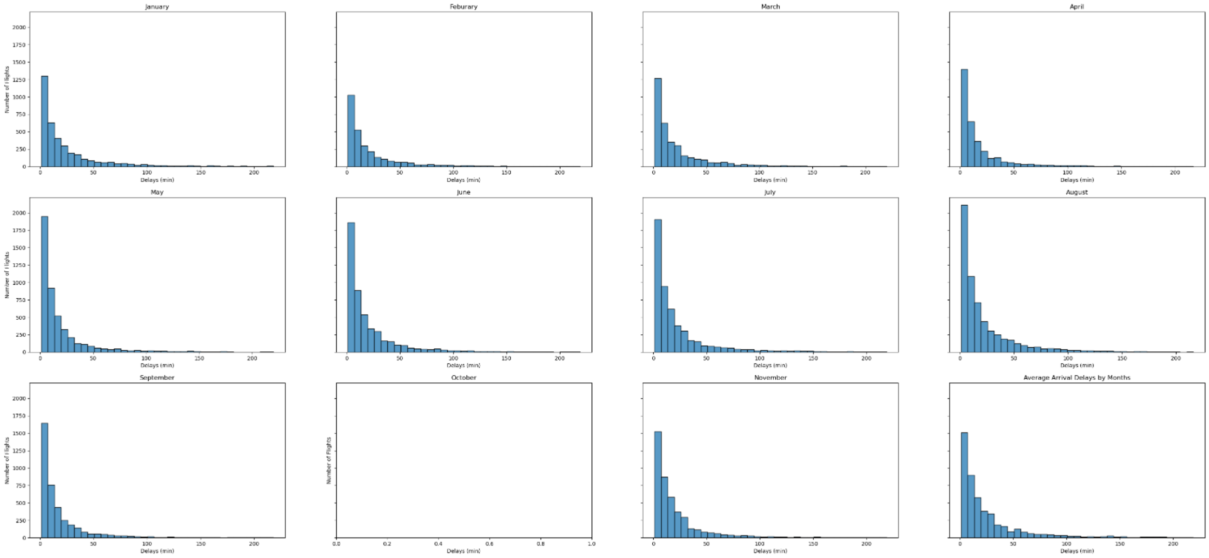
For arrival states, most Northwestern state's arrival delay is more serious on average than other regions', and this is clearly from the color pattern in the map. For example, Wyoming's average flight-time-delay ratio is over 14%, with Oregon and Washington's ratio both above 10%: the top 3 most seriously-delayed arriving states are all in the Northwest, as observed from the bar chart above. Perhaps, this is because it has a relatively shorter total flight time since the departure airport is in WA: if delay-time-length is the same, the delay-time-ratio is higher. Another potential reason is that the northwest has a lot of mountain regions, which could increase the landing difficulties and thus increase the landing-preparing duration in the sky.

Average delay per Airlines:



For airlines, different airlines' flight delays vary a lot. For example, Frontier Airlines's flight's arrival delay is the most serious: its average flight-time-delay ratio is about 7%; however, some other airlines' flight delays are little, like US Airways's average flight-time-delay ratio is even negative, which means their flights averagely arrives earlier than expected. Also, Alaska Airlines, American Airlines and Delta Airlines average delay ratios are all below 2%, which is in the reasonable error range.

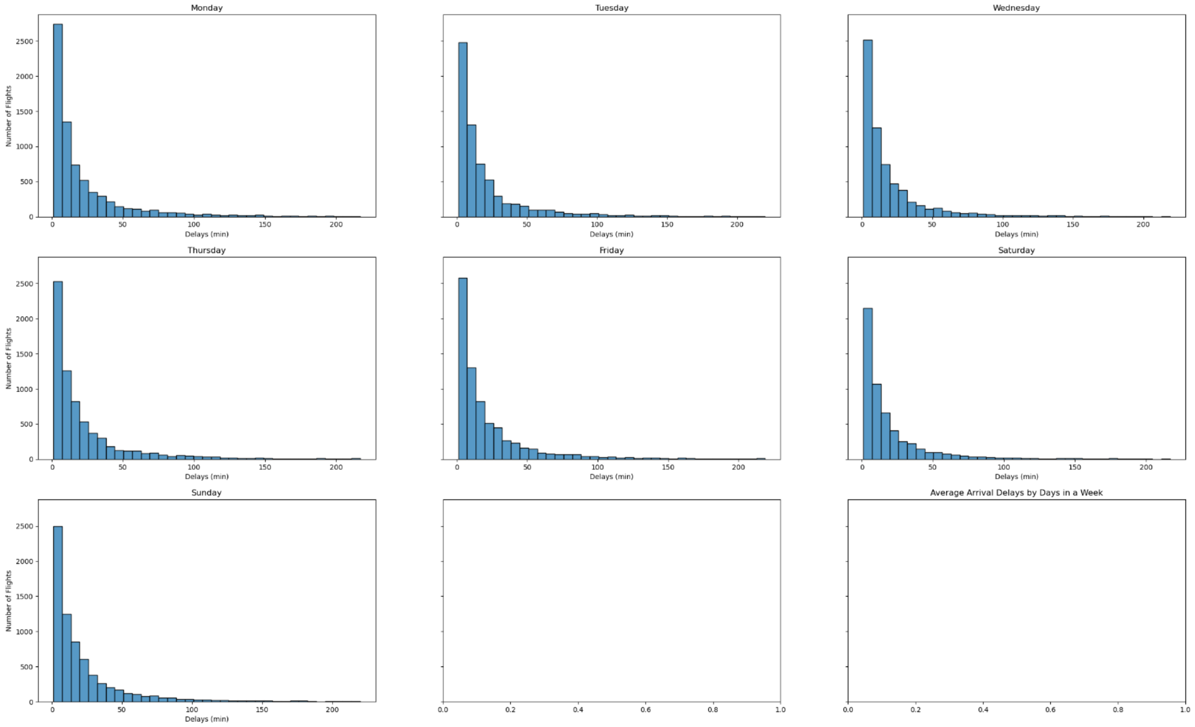
Delay-Time-Ratio per Month:



For all 11 months except for October (missing valid data in October), the summer months: May, June, July, August's flights flying from WA state delay slightly more seriously than other seasons' flights. Because their total seriously-delayed flights' distribution histogram volume is comparatively larger. Except for summer months, December's histogram also has a comparatively longer "right-tail", which means more flights in December are delayed for a very long time (like the Delay-Time-Ratio is even more than 100%).

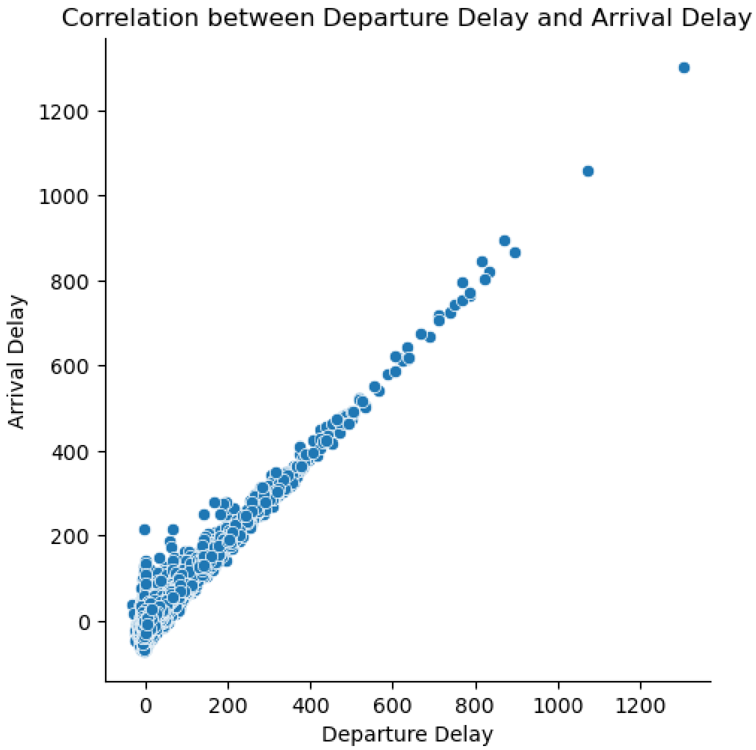
This phenomenon is possible due to more people taking planes during holidays (summer vacation and Christmas) so that it might take a longer time for passengers to go in and out from the plane while the plane parks stationary, thus causing the plane to move later than expected to the next destination.

Delay-Time-Ratio per Week:



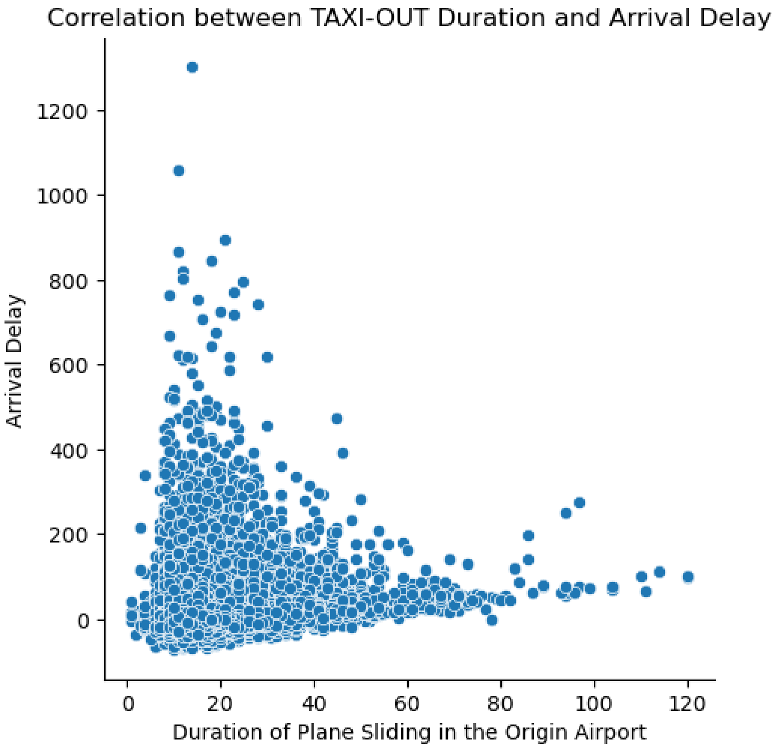
For all 7 days in a week, Saturday flights' delay is less serious compared to other 6 days': it has a shorter "right-tail" and the lowest bars for nearly every Delay-Time-Ratio interval.

Departure Delay and Arrival Delay:



For departure delays, it is strongly-positively correlated with arrival delays, which means longer departure delays are highly likely to lead to longer arrival delays. This is obviously reasonable because a longer departure delay means the flight has already wasted a longer time before flying to the destination. Then, if the flying speed keeps the same, it will arrive later by the later starting time.

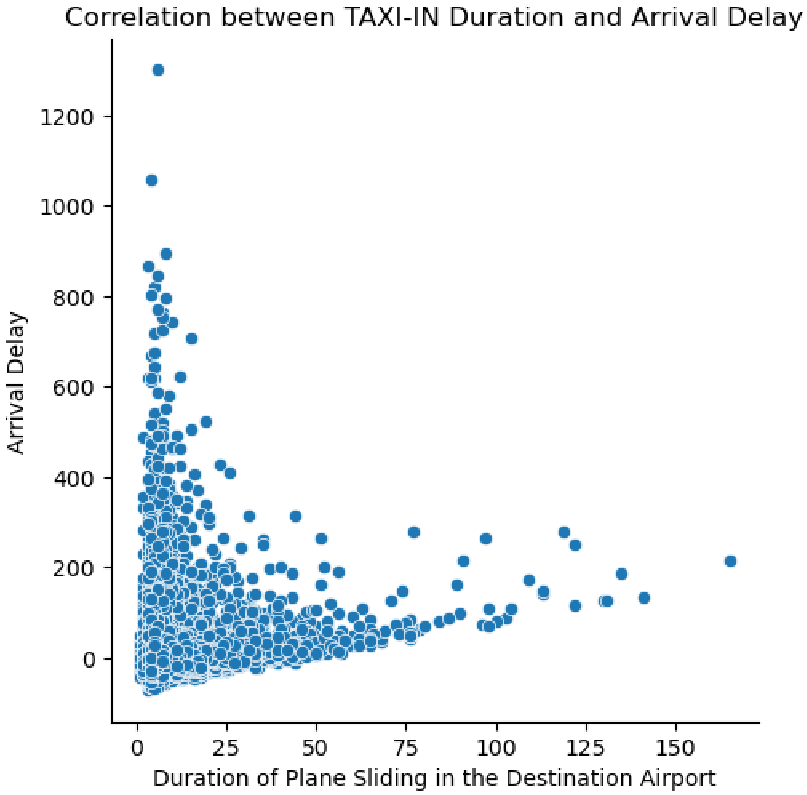
Taxi-Out Duration and Arrival:



For the duration of Plane Sliding in the Origin Airport (namely Taxi Out Duration), we get a surprising result from the scatter plot. There seems to exist a weak inverse proportional relationship between Arrival Delay and Taxi Out Duration, which means the longer a plane delays, the shorter a plane slides in the origin airport before flying. This is not what we expected because a longer sliding in the airport will cause a longer arrival delay in logic.

Our possible explanation is that the graph is misleading. Because the points that construct this inverse proportional pattern are just some outliers, and if we remove those outliers out, there will only be some chaotic random points in the middle circle which means there’s no obvious relationship.

Taxi-In Duration and Arrival:



Also, we find the same relationship between the destination airport's sliding duration and arrival delay length, and if we drop all outliers from the (n > 100000) size sample, there's no obvious relation between these two variables.

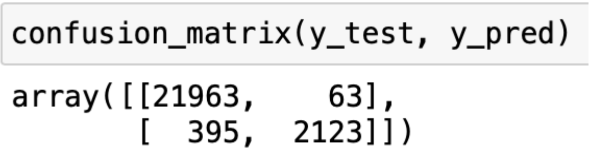
For Question II (Could we set some standards to classify if one airline is "seriously delayed"? Could we use classification methods to predict if a flight will be "seriously delayed" given specific features? How good are the classification performances?)

Yes, we've developed a systematic rule and set a standard about "seriously delay". We tried KNN Classification and Decision Tree Classification to predict if a flight will be "seriously delayed".

Classification: to predict whether a flight would be "Serious Delayed" or not

KNN Classification (confusion matrix and classification report clarification included):

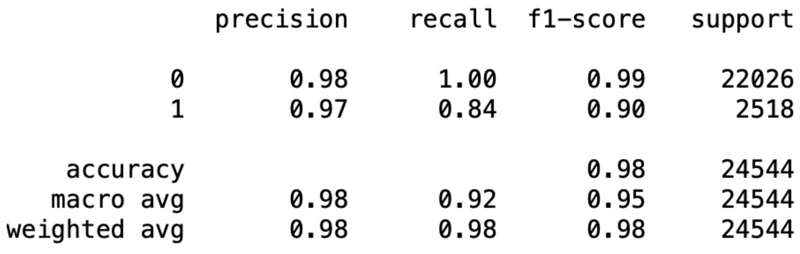
The KNN Classifier shows a good performance on the classification of “seriously delayed” problems. At first, we set n_neighbors = 3 to fit the model. To evaluate the performance, we choose to use and print the confusion matrix and the classification report. Before the interpretation, we’d like to clarify that C_11 (upper left) in the matrix stands for “True Negative”, which means that both the Predicted label and the Actual label are 0; X_12 (upper right) in the matrix stands for “False Positive”, meaning that the Predicted label is 1 while the Actual label is 0; X_21 (lower left) stands for “False Negative”, meaning that the Predicted label is 0 while the Actual label is 1; X_22 (lower right) stands for “True Positive”. When n_neighbors = 3, the confusion matrix is shown below.



```
confusion_matrix(y_test, y_pred)

array([[21963,    63],
       [  395,  2123]])
```

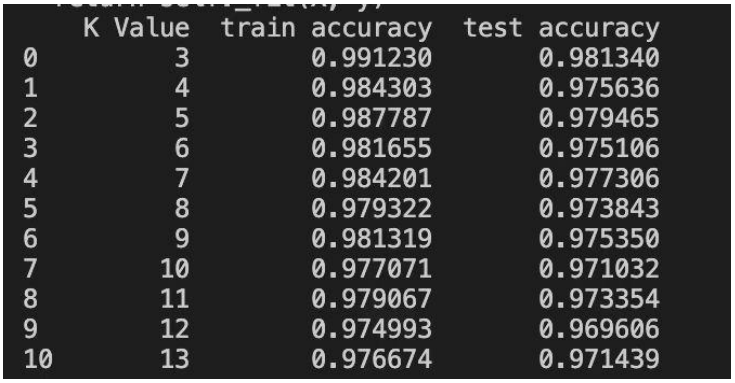
The results show that when we use the trained KNN model to test new data, there are 63 “False Positive” and 395 “False Negative”. This indicates that KNN classification does a good job in limiting the number of “False Positive” but a worse performance in controlling the number of “False Negative”. This situation indicates that when the KNN model predicts one flight to be “seriously delayed”, the possibility that it is wrong is lower than that of the scenario when the KNN model predicts that it would not be “seriously delayed”.

To explore and expand such findings, we introduce the classification report.



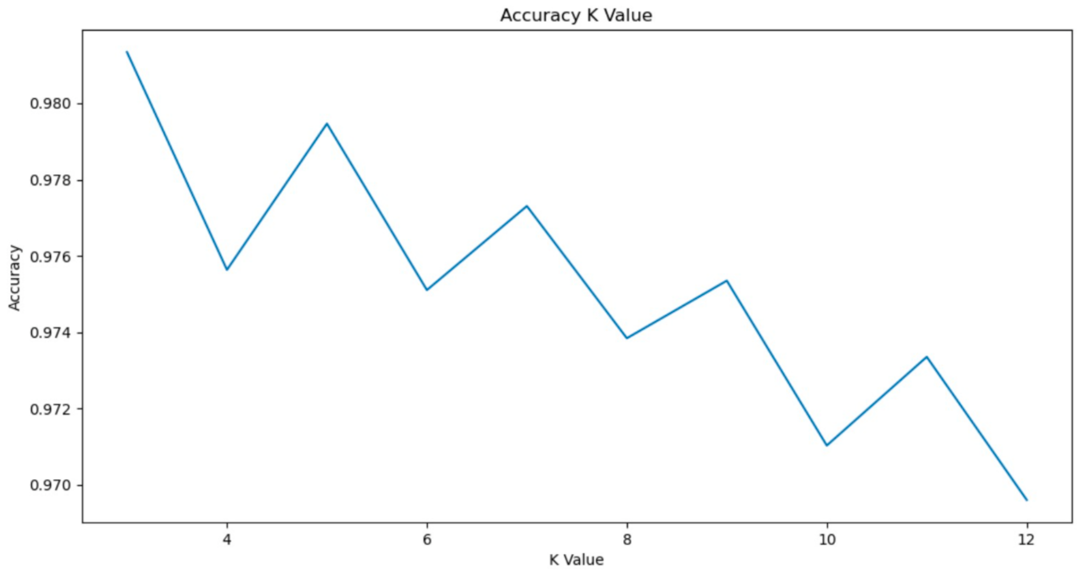
```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     22026
           1       0.97      0.84      0.90      2518

    accuracy                           0.98     24544
   macro avg       0.98      0.92      0.95     24544
weighted avg       0.98      0.98      0.98     24544
```

According to the report, when n_neighbors = 3, the positive precision is 0.97. Precision is equal to (TruePositive) / (TruePositive + FalsePositive). In other words, out of all the results which were predicted as “seriously delay”, the percentage of those actually “seriously delay” is 97%. The philosophy behind improving precision is that we don’t want to predict one as “seriously delayed” arbitrarily and wrongly. The high precision value indicates that our KNN model performs carefully when making the judgment about “seriously delay”: though it may miss some “seriously delayed” scenarios, the “seriously delay” prediction it makes is dependable. However, in terms of the positive recall, its performance is poorer. Recall is defined to be (TruePositive) / (TruePositive + FalseNegative). In this case, in other words, out of all the results which were actually “seriously delayed”, the percentage of those predicted as “seriously delayed” is 84%. Such value is not bad, though it’s not comparable to the high precision. Another good news is that the accuracy of this model (n_neighbors = 3) is 98% and the f1-score is 90%.



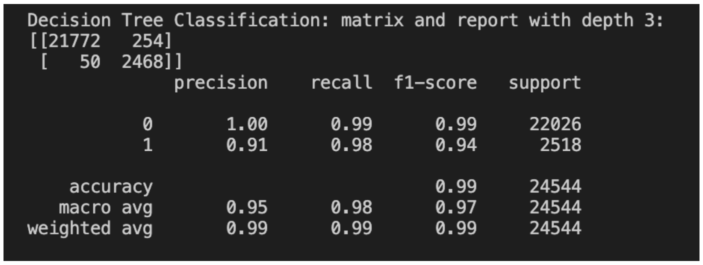
```
    K Value  train accuracy  test accuracy
0         3        0.991230       0.981340
1         4        0.984303       0.975636
2         5        0.987787       0.979465
3         6        0.981655       0.975106
4         7        0.984201       0.977306
5         8        0.979322       0.973843
6         9        0.981319       0.975350
7        10        0.977071       0.971032
8        11        0.979067       0.973354
9        12        0.974993       0.969606
10       13        0.976674       0.971439
```

However, we’d like to explore more and try to find the best KNN model to run the classification. Thus, we did a for loop and tried to find the best n_neighbors. We aimed to find a K value in the range of (3, 14): we don’t want to consider K = 1 or 2 since it may be too over-fitting; we set the max to be 13 (odd number) since it could help the model to vote. According to the graph and the table, we found that n_neighbors = 3 is actually the best choice we could have when choosing the KNN model. Thus, the evaluations above for n_neighbors = 3 is the final result we’ll consider when using KNN classification. Generally, the use of KNN models to predict whether one flight would be seriously delayed or not is good and presents high enough accuracy.

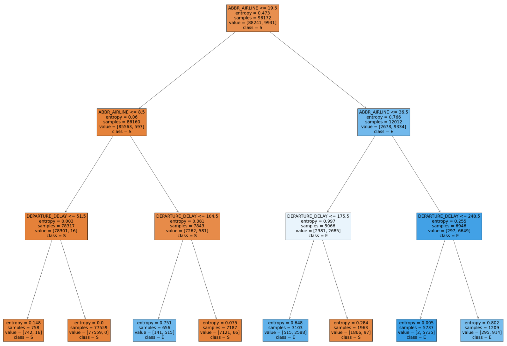


Decision Tree Classification:

Similar to the KNN Classification model, we firstly trained our Decision Tree Classification model with max depth of 3, and output its confusion matrix and classification report.



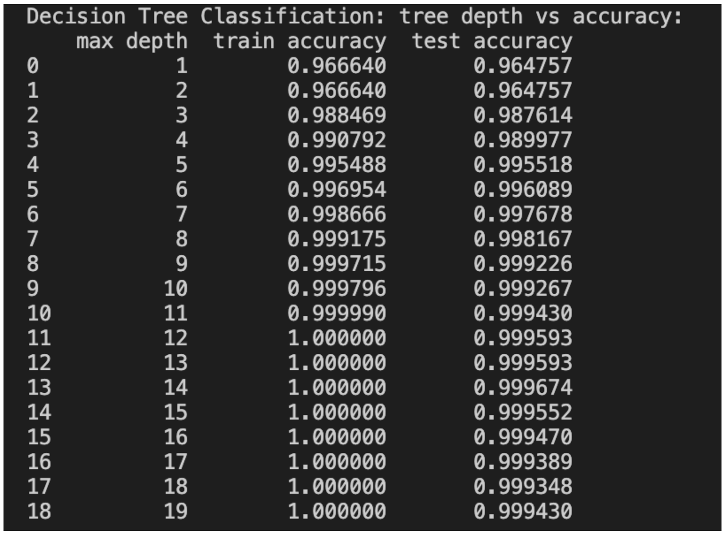
```
Decision Tree Classification: matrix and report with depth 3:
[[21772   254]
 [   50  2468]]
              precision    recall  f1-score   support

           0       1.00      0.99      0.99     22026
           1       0.91      0.98      0.94      2518

    accuracy                           0.99     24544
   macro avg       0.95      0.98      0.97     24544
weighted avg       0.99      0.99      0.99     24544
```



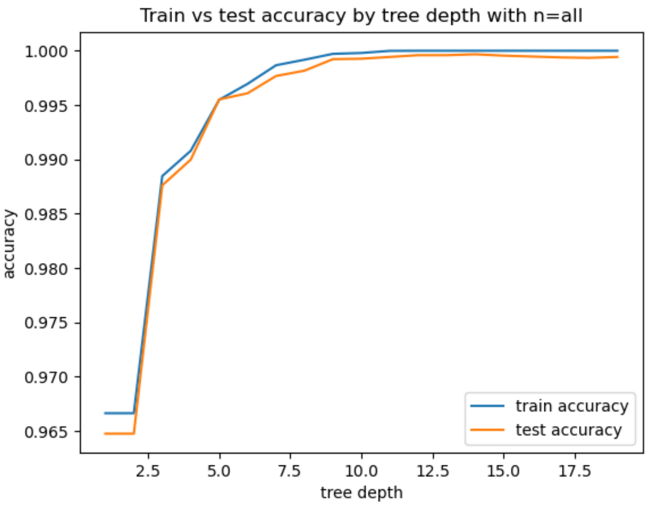
As explained in KNN Classification results, Decision Tree Classification models have 254 "False Positive" out of 22026 of "Positive" (1.15%) and 50 "False Negative" out of 2518 "Negative" (1.99%). This means the probability of predicting flight to be seriously delayed while the truth is not is higher than the probability of predicting it to be not seriously delayed but it actually is.

Also, by looking at the classification report, the precision score for 1 (marking as seriously delay) is 0.91. In other words, out of all the results which were predicted as “seriously delay”, the percentage of those actually “seriously delay” is 91%. This means that it is making a good prediction on seriously delay, but compared to the KNN model described above, the Decision Tree Classification model is not as accurate as KNN.

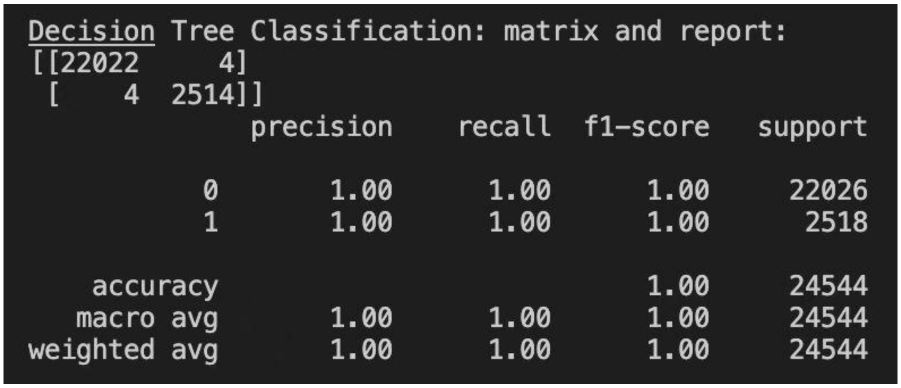
The recall score is 0.98 which is much higher than its precision score. This means that among all the flights which are actually “seriously delayed”, 98% of them are predicted as “seriously delayed” by our model.



```
Decision Tree Classification: tree depth vs accuracy:
    max depth  train accuracy  test accuracy
0           1        0.966640       0.964757
1           2        0.966640       0.964757
2           3        0.988469       0.987614
3           4        0.990792       0.989977
4           5        0.995488       0.995518
5           6        0.996954       0.996089
6           7        0.998666       0.997678
7           8        0.999175       0.998167
8           9        0.999715       0.999226
9          10        0.999796       0.999267
10         11        0.999990       0.999430
11         12        1.000000       0.999593
12         13        1.000000       0.999593
13         14        1.000000       0.999674
14         15        1.000000       0.999552
15         16        1.000000       0.999470
16         17        1.000000       0.999389
17         18        1.000000       0.999348
18         19        1.000000       0.999430
```



This model also has a very high accuracy score, 99%, which is enough for us to make the prediction. However, we also explore our accuracy trying to find the best decision tree max depth to optimize our measures. Similar to KNN model, we use a for loop to print the accuracies with different decision tree depth. We loop over the decision tree max depth from 1 to 19, we find out that the accuracy is highest when max depth equals to 13 with an accuracy of 0.999593. The confusion matrix and the report with max depth of 13 also shows that our model has an incredible performance on prediction with all the scores equal to 1.



```
Decision Tree Classification: matrix and report:
[[22022     4]
 [    4  2514]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     22026
           1       1.00      1.00      1.00      2518

    accuracy                           1.00     24544
   macro avg       1.00      1.00      1.00     24544
weighted avg       1.00      1.00      1.00     24544
```

Comprehensive Evaluations and the Choice of Optimal Classification Models:

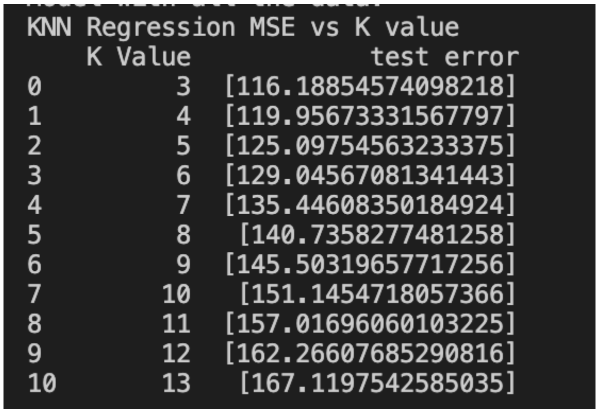
According to the results and analysis before, the performance of the decision tree classifier dominates this section with a wonderful classification report: the Precisions, Recalls, f1-scores, and accuracy are all 1.

(Q3) Could the delay time of flights be predicted? How well could it be predicted? Will the regression prediction be good enough as a reference?

Regression: to predict the arrival delay time of a flight, we tried three different models: KNN, Decision Tree, and Linear Regression.

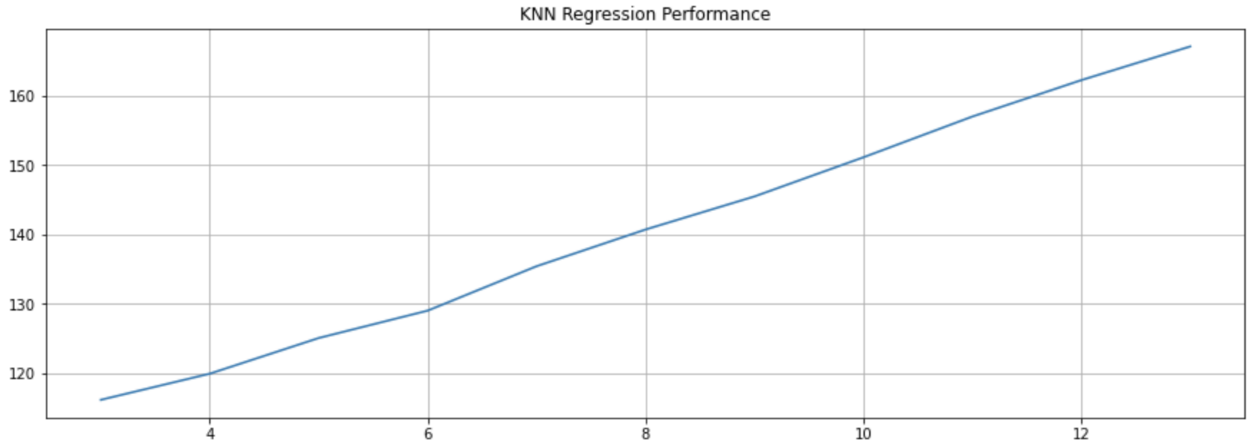
KNN Regression:

Given the good performance of KNN models in the classification problems, we built up a KNN regression model to predict the delay time of one random flight. Generally, the KNN model works and validates that users could predict the delay time of a flight without big errors.



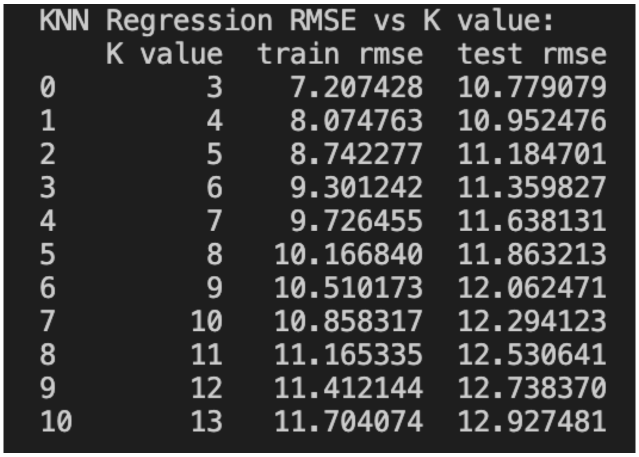
```
KNN Regression MSE vs K value
    K Value            test error
0         3  [116.18854574098218]
1         4  [119.95673331567797]
2         5  [125.09754563233375]
3         6  [129.04567081341443]
4         7  [135.44608350184924]
5         8   [140.7358277481258]
6         9  [145.50319657717256]
7        10   [151.1454718057366]
8        11  [157.01696060103225]
9        12  [162.26607685290816]
10       13   [167.1197542585035]
```

For n_neighbors = 3, the mean square error (MSE) of the KNN model is 116.19. During our exploration for k = range(3,14), we found that n_neighbors = 3 is actually the best one we should use. Thus, we'll stick with n_neighbors = 3 when we apply KNN methods. From the KNN Regression Performance graph, we know that a lower n_neighbors value always means a lower error in this case.

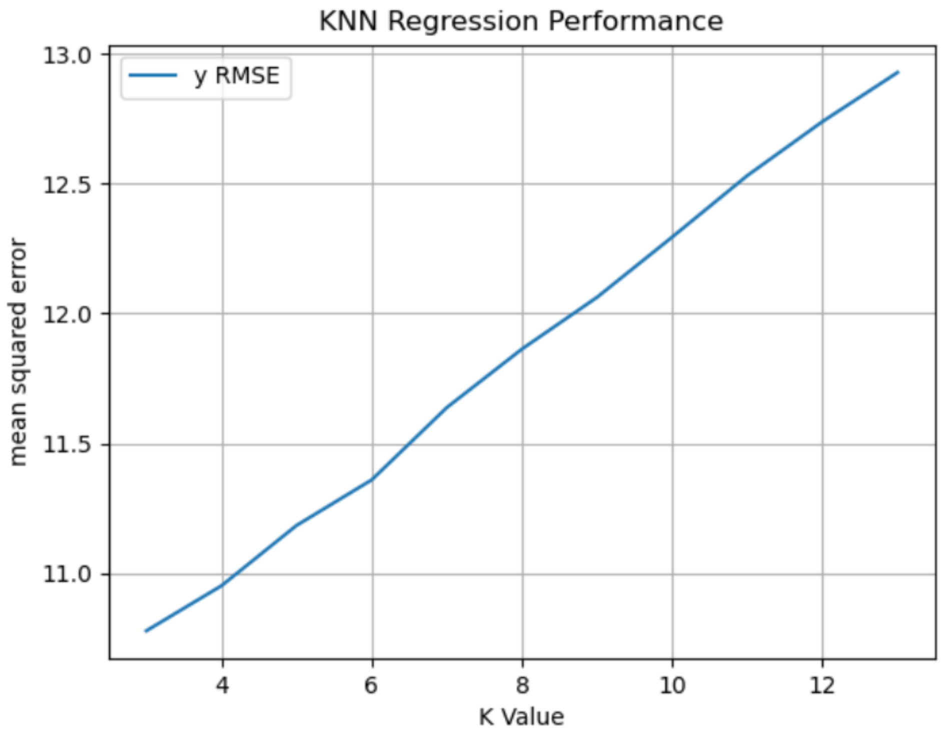


Since we'd like to make the metrics more readable and understandable, it is more appropriate to look at the root-mean-square error (RMSE).

Based on the evaluation results, the RMSE of the KNN model when the n_neighbors = 3 is 10.78. For n_neighbors from 4 to 13, the RMSEs are higher and higher. Thus, we decide to use K value = 3 for the KNN Regression.



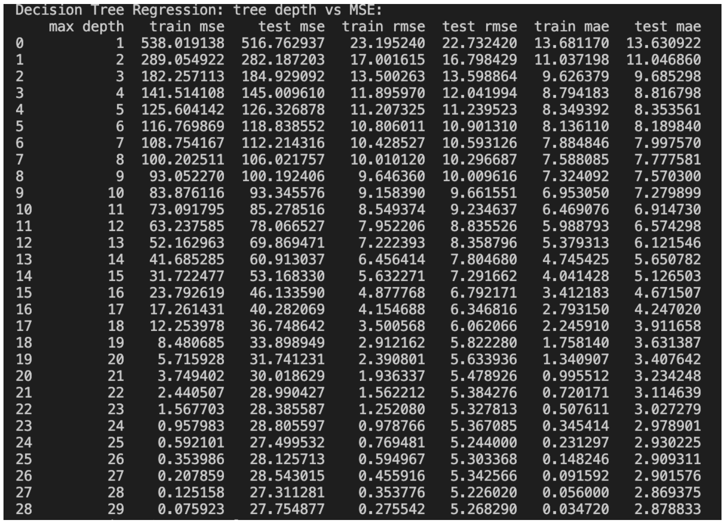
```
KNN Regression RMSE vs K value:
    K value  train rmse  test rmse
0         3    7.207428  10.779079
1         4    8.074763  10.952476
2         5    8.742277  11.184701
3         6    9.301242  11.359827
4         7    9.726455  11.638131
5         8   10.166840  11.863213
6         9   10.510173  12.062471
7        10   10.858317  12.294123
8        11   11.165335  12.530641
9        12   11.412144  12.738370
10       13   11.704074  12.927481
```



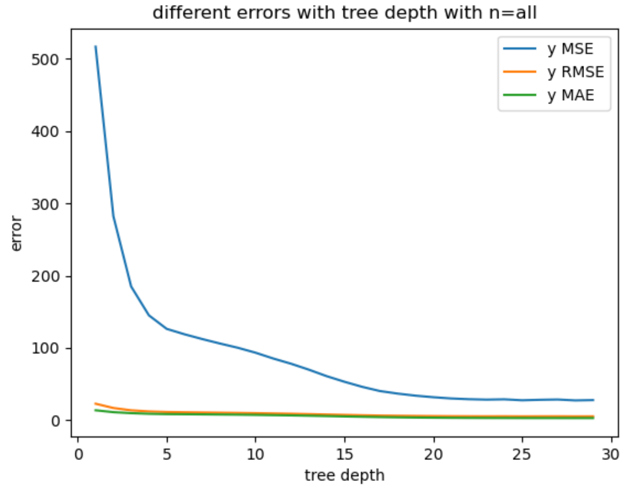
Generally, the RMSE of the KNN model with K value = 3 is not very serious and acceptable. It proves that the prediction of delay time is workable.

Decision Tree Regression:

At first, we trained our model with decision trees max depth at 3, and we calculated its mean squared error, 184.217081. This validates that we could use Decision Tree Regression model to predict the delay time of a flight, but we still want to explore more to find out that with which max depth, our model performs best in predicting delay time, namely with relatively small mean squared error.



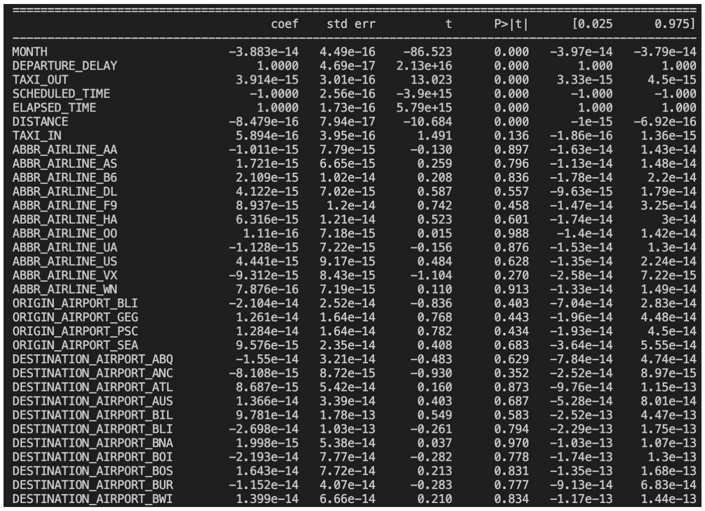
```
Decision Tree Regression: tree depth vs MSE:
    max depth   train mse    test mse  train rmse  test rmse  train mae   test mae
0           1  538.019138  516.762937   23.195240  22.732420  13.681170  13.630922
1           2  289.054922  282.187203   17.001615  16.798429  11.037198  11.046860
2           3  182.257113  184.929092   13.500263  13.598864   9.626379   9.685298
3           4  141.514108  145.009610   11.895970  12.041994   8.794183   8.816798
4           5  125.604142  126.326878   11.207325  11.239523   8.349392   8.353561
5           6  116.769869  118.838552   10.806011  10.901310   8.136110   8.189840
6           7  108.754167  112.214316   10.428527  10.593126   7.884846   7.997570
7           8  100.202511  106.021757   10.010120  10.296687   7.588085   7.777581
8           9   93.052270  100.192406    9.646360  10.009616   7.324092   7.570300
9          10   83.876116   93.345576    9.158390   9.661551   6.953050   7.279899
10         11   73.091795   85.278516    8.549374   9.234637   6.469076   6.914730
11         12   63.237585   78.066527    7.952206   8.835526   5.988793   6.574298
12         13   52.162963   69.869471    7.222393   8.358796   5.379313   6.121546
13         14   41.685285   60.913037    6.456414   7.804680   4.745425   5.650782
14         15   31.722477   53.168330    5.632271   7.291662   4.041428   5.126503
15         16   23.792619   46.133590    4.877768   6.792171   3.412183   4.671507
16         17   17.261431   40.282069    4.154688   6.346816   2.793150   4.247020
17         18   12.253978   36.748642    3.500568   6.062066   2.245910   3.911658
18         19    8.480685   33.898949    2.912162   5.822280   1.758140   3.631387
19         20    5.715928   31.741231    2.390801   5.633936   1.340907   3.407642
20         21    3.749402   30.018629    1.936337   5.478926   0.995512   3.234248
21         22    2.440507   28.990427    1.562212   5.384276   0.720171   3.114639
22         23    1.567703   28.385587    1.252080   5.327813   0.507611   3.027279
23         24    0.957983   28.805597    0.978766   5.367085   0.345414   2.978901
24         25    0.592101   27.499532    0.769481   5.244000   0.231297   2.930225
25         26    0.353986   28.125713    0.594967   5.303368   0.148246   2.909311
26         27    0.207859   28.543015    0.455916   5.342566   0.091592   2.901576
27         28    0.125158   27.311281    0.353776   5.226020   0.056000   2.869375
28         29    0.075923   27.754877    0.275542   5.268290   0.034720   2.878833
```



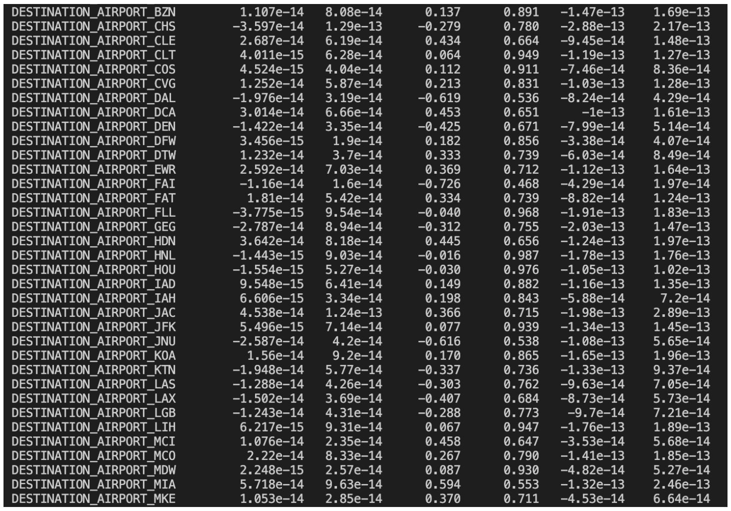
So we loop over the different max depth from 1 to 30 to compare its mean squared error and we find out that the more depth we have the smaller mean squared error our prediction produced. From the table, we can see that, in general, the mean squared error is decreasing as the depth increases. Also, we print the RMSE and MAE to extend our analysis, and see the same trend as the MSE. Thus, if we want to make a very precise result, we could train a model with a large max depth to perfect our prediction.

Linear Regression:

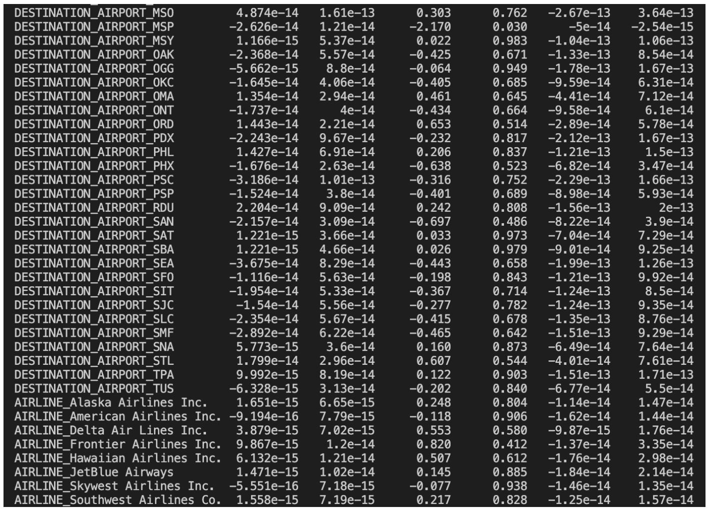
We also trained a linear regression model to predict the delay time and we print out the summary of the model. By looking at the p value of each conditions we find that although the p value for the MONTH, DEPARTURE_DELAY, TAXI_OUT, SCHEDULED_TIME, ELAPSED_TIME, DISTANCE are 0 which is less than 0.05 meaning the results are statistically significant. However, we we looking at all the airports and airline factors, they have very large p values which is larger than 0.05 and this shows that our data used for this model is not statistically significant and has no effects on other observations. Thus we conclude that the Linear Regression Model is not suitable for our prediction.



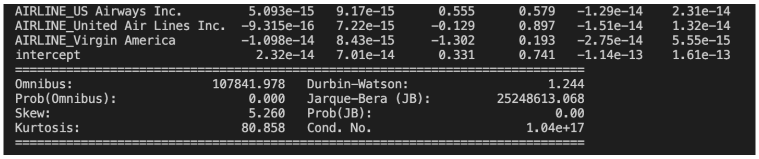
```
==================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
MONTH                       -3.883e-14   4.49e-16    -86.523      0.000   -3.97e-14   -3.79e-14
DEPARTURE_DELAY                 1.0000   4.69e-17   2.13e+16      0.000       1.000       1.000
TAXI_OUT                     3.914e-15   3.01e-16     13.023      0.000    3.33e-15     4.5e-15
SCHEDULED_TIME                 -1.0000   2.56e-16   -3.9e+15      0.000      -1.000      -1.000
ELAPSED_TIME                    1.0000   1.73e-16   5.79e+15      0.000       1.000       1.000
DISTANCE                    -8.479e-16   7.94e-17    -10.684      0.000      -1e-15   -6.92e-16
TAXI_IN                      5.894e-16   3.95e-16      1.491      0.136   -1.86e-16    1.36e-15
ABBR_AIRLINE_AA             -1.011e-15   7.79e-15     -0.130      0.897   -1.63e-14    1.43e-14
ABBR_AIRLINE_AS              1.721e-15   6.65e-15      0.259      0.796   -1.13e-14    1.48e-14
ABBR_AIRLINE_B6              2.109e-15   1.02e-14      0.208      0.836   -1.78e-14     2.2e-14
ABBR_AIRLINE_DL              4.122e-15   7.02e-15      0.587      0.557   -9.63e-15    1.79e-14
ABBR_AIRLINE_F9              8.937e-15    1.2e-14      0.742      0.458   -1.47e-14    3.25e-14
ABBR_AIRLINE_HA              6.316e-15   1.21e-14      0.523      0.601   -1.74e-14       3e-14
ABBR_AIRLINE_OO               1.11e-16   7.18e-15      0.015      0.988    -1.4e-14    1.42e-14
ABBR_AIRLINE_UA             -1.128e-15   7.22e-15     -0.156      0.876   -1.53e-14     1.3e-14
ABBR_AIRLINE_US              4.441e-15   9.17e-15      0.484      0.628   -1.35e-14    2.24e-14
ABBR_AIRLINE_VX             -9.312e-15   8.43e-15     -1.104      0.270   -2.58e-14    7.22e-15
ABBR_AIRLINE_WN              7.876e-16   7.19e-15      0.110      0.913   -1.33e-14    1.49e-14
ORIGIN_AIRPORT_BLI          -2.104e-14   2.52e-14     -0.836      0.403   -7.04e-14    2.83e-14
ORIGIN_AIRPORT_GEG           1.261e-14   1.64e-14      0.768      0.443   -1.96e-14    4.48e-14
ORIGIN_AIRPORT_PSC           1.284e-14   1.64e-14      0.782      0.434   -1.93e-14     4.5e-14
ORIGIN_AIRPORT_SEA           9.576e-15   2.35e-14      0.408      0.683   -3.64e-14    5.55e-14
DESTINATION_AIRPORT_ABQ      -1.55e-14   3.21e-14     -0.483      0.629   -7.84e-14    4.74e-14
DESTINATION_AIRPORT_ANC     -8.108e-15   8.72e-15     -0.930      0.352   -2.52e-14    8.97e-15
DESTINATION_AIRPORT_ATL      8.687e-15   5.42e-14      0.160      0.873   -9.76e-14    1.15e-13
DESTINATION_AIRPORT_AUS      1.366e-14   3.39e-14      0.403      0.687   -5.28e-14    8.01e-14
DESTINATION_AIRPORT_BIL      9.781e-14   1.78e-13      0.549      0.583   -2.52e-13    4.47e-13
DESTINATION_AIRPORT_BLI     -2.698e-14   1.03e-13     -0.261      0.794   -2.29e-13    1.75e-13
DESTINATION_AIRPORT_BNA      1.998e-15   5.38e-14      0.037      0.970   -1.03e-13    1.07e-13
DESTINATION_AIRPORT_BOI     -2.193e-14   7.77e-14     -0.282      0.778   -1.74e-13     1.3e-13
DESTINATION_AIRPORT_BOS      1.643e-14   7.72e-14      0.213      0.831   -1.35e-13    1.68e-13
DESTINATION_AIRPORT_BUR     -1.152e-14   4.07e-14     -0.283      0.777   -9.13e-14    6.83e-14
DESTINATION_AIRPORT_BWI      1.399e-14   6.66e-14      0.210      0.834   -1.17e-13    1.44e-13
```



```
DESTINATION_AIRPORT_BZN      1.107e-14   8.08e-14      0.137      0.891   -1.47e-13    1.69e-13
DESTINATION_AIRPORT_CHS     -3.597e-14   1.29e-13     -0.279      0.780   -2.88e-13    2.17e-13
DESTINATION_AIRPORT_CLE      2.687e-14   6.19e-14      0.434      0.664   -9.45e-14    1.48e-13
DESTINATION_AIRPORT_CLT      4.011e-15   6.28e-14      0.064      0.949   -1.19e-13    1.27e-13
DESTINATION_AIRPORT_COS      4.524e-15   4.04e-14      0.112      0.911   -7.46e-14    8.36e-14
DESTINATION_AIRPORT_CVG      1.252e-14   5.87e-14      0.213      0.831   -1.03e-13    1.28e-13
DESTINATION_AIRPORT_DAL     -1.976e-14   3.19e-14     -0.619      0.536   -8.24e-14    4.29e-14
DESTINATION_AIRPORT_DCA      3.014e-14   6.66e-14      0.453      0.651      -1e-13    1.61e-13
DESTINATION_AIRPORT_DEN     -1.422e-14   3.35e-14     -0.425      0.671   -7.99e-14    5.14e-14
DESTINATION_AIRPORT_DFW      3.456e-15    1.9e-14      0.182      0.856   -3.38e-14    4.07e-14
DESTINATION_AIRPORT_DTW      1.232e-14    3.7e-14      0.333      0.739   -6.03e-14    8.49e-14
DESTINATION_AIRPORT_EWR      2.592e-14   7.03e-14      0.369      0.712   -1.12e-13    1.64e-13
DESTINATION_AIRPORT_FAI      -1.16e-14    1.6e-14     -0.726      0.468   -4.29e-14    1.97e-14
DESTINATION_AIRPORT_FAT       1.81e-14   5.42e-14      0.334      0.739   -8.82e-14    1.24e-13
DESTINATION_AIRPORT_FLL     -3.775e-15   9.54e-14     -0.040      0.968   -1.91e-13    1.83e-13
DESTINATION_AIRPORT_GEG     -2.787e-14   8.94e-14     -0.312      0.755   -2.03e-13    1.47e-13
DESTINATION_AIRPORT_HDN      3.642e-14   8.18e-14      0.445      0.656   -1.24e-13    1.97e-13
DESTINATION_AIRPORT_HNL     -1.443e-15   9.03e-14     -0.016      0.987   -1.78e-13    1.76e-13
DESTINATION_AIRPORT_HOU     -1.554e-15   5.27e-14     -0.030      0.976   -1.05e-13    1.02e-13
DESTINATION_AIRPORT_IAD      9.548e-15   6.41e-14      0.149      0.882   -1.16e-13    1.35e-13
DESTINATION_AIRPORT_IAH      6.606e-15   3.34e-14      0.198      0.843   -5.88e-14     7.2e-14
DESTINATION_AIRPORT_JAC      4.538e-14   1.24e-13      0.366      0.715   -1.98e-13    2.89e-13
DESTINATION_AIRPORT_JFK      5.496e-15   7.14e-14      0.077      0.939   -1.34e-13    1.45e-13
DESTINATION_AIRPORT_JNU     -2.587e-14    4.2e-14     -0.616      0.538   -1.08e-13    5.65e-14
DESTINATION_AIRPORT_KOA       1.56e-14    9.2e-14      0.170      0.865   -1.65e-13    1.96e-13
DESTINATION_AIRPORT_KTN     -1.948e-14   5.77e-14     -0.337      0.736   -1.33e-13    9.37e-14
DESTINATION_AIRPORT_LAS     -1.288e-14   4.26e-14     -0.303      0.762   -9.63e-14    7.05e-14
DESTINATION_AIRPORT_LAX     -1.502e-14   3.69e-14     -0.407      0.684   -8.73e-14    5.73e-14
DESTINATION_AIRPORT_LGB     -1.243e-14   4.31e-14     -0.288      0.773    -9.7e-14    7.21e-14
DESTINATION_AIRPORT_LIH      6.217e-15   9.31e-14      0.067      0.947   -1.76e-13    1.89e-13
DESTINATION_AIRPORT_MCI      1.076e-14   2.35e-14      0.458      0.647   -3.53e-14    5.68e-14
DESTINATION_AIRPORT_MCO       2.22e-14   8.33e-14      0.267      0.790   -1.41e-13    1.85e-13
DESTINATION_AIRPORT_MDW      2.248e-15   2.57e-14      0.087      0.930   -4.82e-14    5.27e-14
DESTINATION_AIRPORT_MIA      5.718e-14   9.63e-14      0.594      0.553   -1.32e-13    2.46e-13
DESTINATION_AIRPORT_MKE      1.053e-14   2.85e-14      0.370      0.711   -4.53e-14    6.64e-14
```



```
DESTINATION_AIRPORT_MSO        4.874e-14   1.61e-13      0.303      0.762   -2.67e-13    3.64e-13
DESTINATION_AIRPORT_MSP       -2.626e-14   1.21e-14     -2.170      0.030       -5e-14   -2.54e-15
DESTINATION_AIRPORT_MSY        1.166e-15   5.37e-14      0.022      0.983   -1.04e-13    1.06e-13
DESTINATION_AIRPORT_OAK       -2.368e-14   5.57e-14     -0.425      0.671   -1.33e-13    8.54e-14
DESTINATION_AIRPORT_OGG       -5.662e-15    8.8e-14     -0.064      0.949   -1.78e-13    1.67e-13
DESTINATION_AIRPORT_OKC       -1.645e-14   4.06e-14     -0.405      0.685   -9.59e-14    6.31e-14
DESTINATION_AIRPORT_OMA        1.354e-14   2.94e-14      0.461      0.645   -4.41e-14    7.12e-14
DESTINATION_AIRPORT_ONT       -1.737e-14      4e-14     -0.434      0.664   -9.58e-14     6.1e-14
DESTINATION_AIRPORT_ORD        1.443e-14   2.21e-14      0.653      0.514   -2.89e-14    5.78e-14
DESTINATION_AIRPORT_PDX       -2.243e-14   9.67e-14     -0.232      0.817   -2.12e-13    1.67e-13
DESTINATION_AIRPORT_PHL        1.427e-14   6.91e-14      0.206      0.837   -1.21e-13     1.5e-13
DESTINATION_AIRPORT_PHX       -1.676e-14   2.63e-14     -0.638      0.523   -6.82e-14    3.47e-14
DESTINATION_AIRPORT_PSC       -3.186e-14   1.01e-13     -0.316      0.752   -2.29e-13    1.66e-13
DESTINATION_AIRPORT_PSP       -1.524e-14    3.8e-14     -0.401      0.689   -8.98e-14    5.93e-14
DESTINATION_AIRPORT_RDU        2.204e-14   9.09e-14      0.242      0.808   -1.56e-13       2e-13
DESTINATION_AIRPORT_SAN       -2.157e-14   3.09e-14     -0.697      0.486   -8.22e-14     3.9e-14
DESTINATION_AIRPORT_SAT        1.221e-15   3.66e-14      0.033      0.973   -7.04e-14    7.29e-14
DESTINATION_AIRPORT_SBA        1.221e-15   4.66e-14      0.026      0.979   -9.01e-14    9.25e-14
DESTINATION_AIRPORT_SEA       -3.675e-14   8.29e-14     -0.443      0.658   -1.99e-13    1.26e-13
DESTINATION_AIRPORT_SFO       -1.116e-14   5.63e-14     -0.198      0.843   -1.21e-13    9.92e-14
DESTINATION_AIRPORT_SIT       -1.954e-14   5.33e-14     -0.367      0.714   -1.24e-13     8.5e-14
DESTINATION_AIRPORT_SJC        -1.54e-14   5.56e-14     -0.277      0.782   -1.24e-13    9.35e-14
DESTINATION_AIRPORT_SLC       -2.354e-14   5.67e-14     -0.415      0.678   -1.35e-13    8.76e-14
DESTINATION_AIRPORT_SMF       -2.892e-14   6.22e-14     -0.465      0.642   -1.51e-13    9.29e-14
DESTINATION_AIRPORT_SNA        5.773e-15    3.6e-14      0.160      0.873   -6.49e-14    7.64e-14
DESTINATION_AIRPORT_STL        1.799e-14   2.96e-14      0.607      0.544   -4.01e-14    7.61e-14
DESTINATION_AIRPORT_TPA        9.992e-15   8.19e-14      0.122      0.903   -1.51e-13    1.71e-13
DESTINATION_AIRPORT_TUS       -6.328e-15   3.13e-14     -0.202      0.840   -6.77e-14     5.5e-14
AIRLINE_Alaska Airlines Inc.   1.651e-15   6.65e-15      0.248      0.804   -1.14e-14    1.47e-14
AIRLINE_American Airlines Inc. -9.194e-16   7.79e-15     -0.118      0.906   -1.62e-14    1.44e-14
AIRLINE_Delta Air Lines Inc.   3.879e-15   7.02e-15      0.553      0.580   -9.87e-15    1.76e-14
AIRLINE_Frontier Airlines Inc. 9.867e-15    1.2e-14      0.820      0.412   -1.37e-14    3.35e-14
AIRLINE_Hawaiian Airlines Inc. 6.132e-15   1.21e-14      0.507      0.612   -1.76e-14    2.98e-14
AIRLINE_JetBlue Airways        1.471e-15   1.02e-14      0.145      0.885   -1.84e-14    2.14e-14
AIRLINE_Skywest Airlines Inc. -5.551e-16   7.18e-15     -0.077      0.938   -1.46e-14    1.35e-14
AIRLINE_Southwest Airlines Co. 1.558e-15   7.19e-15      0.217      0.828   -1.25e-14    1.57e-14
```



```
AIRLINE_US Airways Inc.          5.093e-15   9.17e-15      0.555      0.579   -1.29e-14    2.31e-14
AIRLINE_United Air Lines Inc.   -9.315e-16   7.22e-15     -0.129      0.897   -1.51e-14    1.32e-14
AIRLINE_Virgin America          -1.098e-14   8.43e-15     -1.302      0.193   -2.75e-14    5.55e-15
intercept                          2.32e-14   7.01e-14      0.331      0.741   -1.14e-13    1.61e-13
==============================================================================
Omnibus:                   107841.978   Durbin-Watson:                   1.244
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         25248613.068
Skew:                           5.260   Prob(JB):                         0.00
Kurtosis:                      80.858   Cond. No.                     1.04e+17
==============================================================================
```

Impact and Limitations

Impact：

This comprehensive project aims to serve multiple groups of people as the audience. The stakeholders benefited from this project include flight passengers/consumers, family/friends/related people of the passengers, airline companies, taxi/Uber drivers, Aviation Administration, and the federal government.

Most passengers will be benefited if they see the result of this report. When passengers have several choices of flights, such as which days in a week to fly, which airlines to choose, or which airport to arrive (if a city has more than one airport), they can minimize their probability of experiencing a serious delay by looking at our plots and models. Also, the results might help people to decide which months or seasons are wise for travel to avoid delayed flights if they want to avoid it. More importantly, the machine learning models we trained, tested, and evaluated could tell them if their flights will be “seriously delayed” or not and tell them the potential arrival delay time with very tiny errors. That function should be very important to many passengers, especially to the businessmen. They always want to figure out the delay time so they won’t miss any important meetings and visits. If they know their estimated arrival time after considering potential delays, they could deal with other events better and more confidently. Given that our models, especially the decision tree models with large depths, present a wonderful accuracy and great error controls, flight passengers could trust the prediction results most of the time. Businessmen usually hope that the classification models could have a high **Recall**, since they do not want to dismiss/ignore any potential“seriously delayed” case that could be catched: they do not want to let the people/companies they plan to visit and meet wait for them - that will hurt their business. Thus, they wish our models to catch the “seriously delay” case as much as possible: by contrast, they’re willing to take the corresponding cost that our model predicts “seriously delayed” but actually not. In this way, the decision tree model (Recall = nearly 1) serves the best, instead of the KNN model (Recall = 0.84).

Taxi/Uber/Lyft drivers, or any people coming to pick up the flight passengers, could also be greatly benefited from this project. They are usually bothered and confused by flight delays. Many drivers had complained about the unstable and volatile arrival time provided by the airline companies. The introduction of our regression and classification models help them a lot and save their time. Besides, they want to make sure that they can pick their passengers/friends/family up in time without letting them wait. Therefore, they may want the models with excellent **Precision**: they could accept that our models miss some “seriously delay” cases, but they want that our model will not judge the normal one to be “seriously delay”, otherwise they will be late to arrive. Both KNN model and the decision tree model can work for them very well. The KNN model with k value = 3 could have a 0.97 for the positive precision, while the decision tree model’s Precision can be nearly 1.

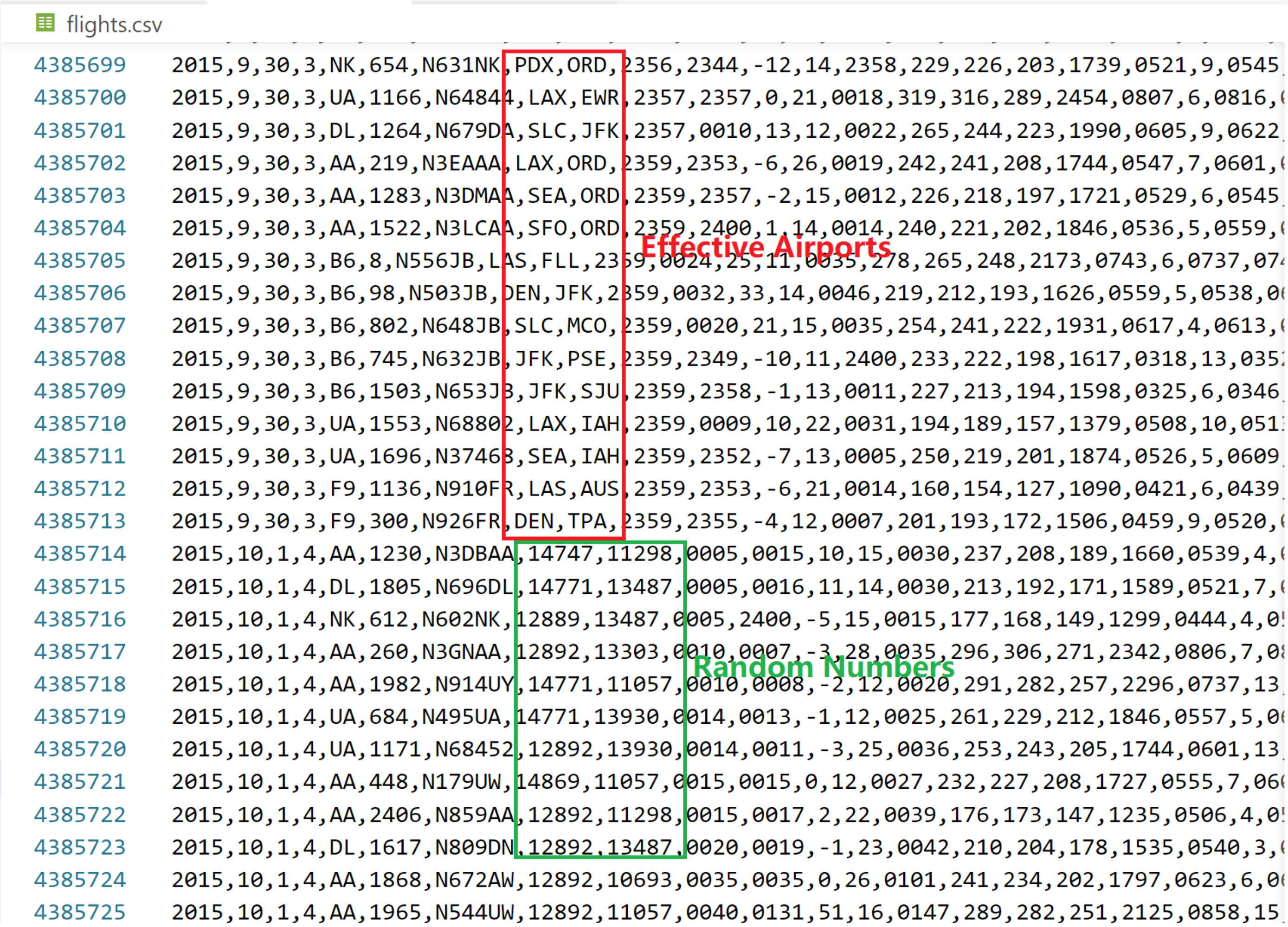
Airline companies and airports may find our project very helpful as well. They could utilize our plotting and mapping results to optimize their operations, logistics management, and staff training. Airline companies may appreciate the introduction of our training models since the accurate estimations of delay time can help them in many fields. Airline companies can use our prediction models and results to enhance customer experiences and satisfactions.

What’s more, the aviation administration and government can use the Mapping we created and the machine learning models to improve their management and investment. The official agencies may want to increase the funding and support for the areas and airports where flight delays happen frequently and seriously.

In contrast, the airlines that have comparatively high average Delay-Time-Ratio on the plot will be harmed by these results because passengers who want to avoid delays will intentionally avoid taking those airlines. So, this report in public might let those airlines sell fewer air tickets in the future and earn less money than if this report is not published.

Limitation:

First and most importantly, the original dataset we found online contains some mistaken information about airports, which means some rows of the airport’s content are meaningless. We found this mistake while running the linear regression model above, there are about 500000 variables in our fractionated dataframe’s model (even though the rows are reduced to 1% of the original size, it still takes a very long time to run), and as we observe the variables, we found the majority is some airports with random numbers. This leads us to check the original data frame again. As we examine the flights in October, both the origin and destination airports are numbers. As a result, this leads us to eliminate all the flights in October because we want to find the relationship between flight delay and its corresponding airports.



```
2015,9,30,3,NK,654,N631NK,PDX,ORD,2356,2344,-12,14,2358,229,226,203,1739,0521,9,0545
2015,9,30,3,UA,1166,N64844,LAX,EWR,2357,2357,0,21,0018,319,316,289,2454,0807,6,0816,
2015,9,30,3,DL,1264,N679DA,SLC,JFK,2357,0010,13,12,0022,265,244,223,1990,0605,9,0622
2015,9,30,3,AA,219,N3EAAA,LAX,ORD,2359,2353,-6,26,0019,242,241,208,1744,0547,7,0601,
2015,9,30,3,AA,1283,N3DMAA,SEA,ORD,2359,2357,-2,15,0012,226,218,197,1721,0529,6,0545
2015,9,30,3,AA,1522,N3LCAA,SFO,ORD,2359,2400,1,14,0014,240,221,202,1846,0536,5,0559,
2015,9,30,3,B6,8,N556JB,LAS,FLL,2359,0024,25,11,0035,278,265,248,2173,0743,6,0737,07
2015,9,30,3,B6,98,N503JB,DEN,JFK,2359,0032,33,14,0046,219,212,193,1626,0559,5,0538,0
2015,9,30,3,B6,802,N648JB,SLC,MCO,2359,0020,21,15,0035,254,241,222,1931,0617,4,0613,
2015,9,30,3,B6,745,N632JB,JFK,PSE,2359,2349,-10,11,2400,233,222,198,1617,0318,13,035
2015,9,30,3,B6,1503,N653JB,JFK,SJU,2359,2358,-1,13,0011,227,213,194,1598,0325,6,0346
2015,9,30,3,UA,1553,N68802,LAX,IAH,2359,0009,10,22,0031,194,189,157,1379,0508,10,051
2015,9,30,3,UA,1696,N37468,SEA,IAH,2359,2352,-7,13,0005,250,219,201,1874,0526,5,0609
2015,9,30,3,F9,1136,N910FR,LAS,AUS,2359,2353,-6,21,0014,160,154,127,1090,0421,6,0439
2015,9,30,3,F9,300,N926FR,DEN,TPA,2359,2355,-4,12,0007,201,193,172,1506,0459,9,0520,
2015,10,1,4,AA,1230,N3DBAA,14747,11298,0005,0015,10,15,0030,237,208,189,1660,0539,4,
2015,10,1,4,DL,1805,N696DL,14771,13487,0005,0016,11,14,0030,213,192,171,1589,0521,7,
2015,10,1,4,NK,612,N602NK,12889,13487,0005,2400,-5,15,0015,177,168,149,1299,0444,4,0
2015,10,1,4,AA,260,N3GNAA,12892,13303,0010,0007,-3,28,0035,296,306,271,2342,0806,7,0
2015,10,1,4,AA,1982,N914UY,14771,11057,0010,0008,-2,12,0020,291,282,257,2296,0737,13
2015,10,1,4,UA,684,N495UA,14771,13930,0014,0013,-1,12,0025,261,229,212,1846,0557,5,0
2015,10,1,4,UA,1171,N68452,12892,13930,0014,0011,-3,25,0036,253,243,205,1744,0601,13
2015,10,1,4,AA,448,N179UW,14869,11057,0015,0015,0,12,0027,232,227,208,1727,0555,7,06
2015,10,1,4,AA,2406,N859AA,12892,11298,0015,0017,2,22,0039,176,173,147,1235,0506,4,0
2015,10,1,4,DL,1617,N809DN,12892,13487,0020,0019,-1,23,0042,210,204,178,1535,0540,3,
2015,10,1,4,AA,1868,N672AW,12892,10693,0035,0035,0,26,0101,241,234,202,1797,0623,6,0
2015,10,1,4,AA,1965,N544UW,12892,11057,0040,0131,51,16,0147,289,282,251,2125,0858,15
```

However, eliminating a whole month of flights is not the same as random sampling: there is a weak pattern about flight delays across different months in the result from "Serious Delay Distribution per Month", October has slightly fewer flight delays compared to other months. Moreover, the data constructed from the histogram is "flights from all airports in the U.S.A. in 2015" while this project's limited topic is "flights from WA state's airports in 2015". This is because we are unable to construct a histogram both with October's flight data and the origin airports are limited in WA state: it is impossible to find flights in the 4 airports in WA state in October because all of October's airports are numbers (This is defined as a **Flight Paradox**!) Therefore, omitting the entire October's flight data will increase bias, reduce the accuracy of predicting afterwards, because the flight range of research is the whole year.

The second major limitation is about mapping. Because there exist some states that are not a destination of any flights from Washington state at all, we miss the average Delay-Time-Ratio in those states, as left some states as white color on the map. As a result, the pattern of Northwestern states' average delay is generally longer might be misleading, if the white state exhibits the opposite trend if there are some flights to those states in the future.

Challenging Goals

1) Multiple Datasets:

In our project, we need to visualize and analyze the relationship between the flight delay and the destination states, and even the corresponding airlines the flight belongs to. Since flights.csv does not contain the detailed location information for each airport, we need to combine the flights.csv with airports.csv to find the corresponding destination states for each flight by a join type based on the airports' names in both files. For the same logic, combining flights.csv with airlines.csv by a join type based on airlines information in each file, too. Also, we need to combine the csv file with the cb_2018_us_state_5m.shp file together by a join type based on the state name in order to draw a geospatial map showing the flight delay contrast between each state. But unfortunately, the state names represented in airports' csv file are in the abbreviation form, but the states in the "shape" file are represented in full-name. Therefore, a transformation from abbreviation to full-name for each state must be constructed in order to make the two data frames share a matchable column to merge. So, this is a challenging goal in this program.

2) Machine Learning:

In this project, we'll use a series of methods based on machine learning to solve the conundrums of flight delay. We have two main goals in terms of machine learning application. The first one is that we want to use regressions to predict the arrival delay time given any random flight. We plan to use data in the dataset to train our model. Decision Tree Regression, Linear Regression, and KNN could be applied in this part. We want to evaluate each model via different metrics, including Mean Square Error and RMSE. Finally, we wish that we could determine which model is better for the prediction. The second goal is to classify if a flight will be seriously delayed. We will (and have) developed a way and figure out the standard to see if a labeled flight is "seriously delayed". With the training of the models, we'll use classification models to see if a random/new/testing flight will be seriously delayed. We consider various models, such as Decision Tree Classification, Logistic Regression, and KNN. After that, we would like to evaluate each model, using accuracy Precision, Recall, and F-1 scores, to see which model is better for classifications.

3) Mappings:

After we merged the shp file with other csv files, the resulting file is default to be in csv type, so that we are unable to plot a map from it although it contains a column "geometry" with each state's polygon or multipolygon information. So we need to find a way to convert it into a shp file again.

Also, the background map of this shp file we downloaded is a map with longitude from 180° E to 180° W, and we tried 4 different websites to download at least 7 versions of the U.S. state's shapefile, but all of them have the same longitude range. A possible explanation is that the graph contains some Pacific Islands (which is small but still considered to be the U.S.A. 's territory) so the overall background map looks very long, with vague boundaries between states and not reader-friendly. Therefore, we need to search for some functions that limit the longitude range of a map plot, and also make the legend on the right to be smaller. (not a similar size to the entire map)

Work Plan Evaluation

Task 1: Update the columns of the original dataset, combine airlines.csv, airport.csv and flights.csv together by merge functions, also remove some useless columns, like flight number. Then, add three new columns: 1. Delay_ratio_flytime (arrival_delay / scheduled_time), 2. Delay_ratio_distance (arrival_delay / distance), 3. Seriously_delay

- Expected Time: 6 hours
- Actual Time: 20 hours

This task takes way longer than our expectation since as we explore our data, we find that the data is too large and it takes longer than we expected to load and clean the data. Also, as described in the limitation, we meet some issues when we try to clean our data and do machine learning with a large dataset. We spend most of our time solving our problems.

Task 2: Use Panda library and plotting function to draw distribution histograms of flight delay (Delay_ratio_flytime) based on different factors: month, days in a week, airline, state, tax in & tax out time length, and departure delay.

- Expected Time: 4.5 hours
- Actual Time: 16 hours

We spend a lot of time planning and arranging our plots for the data, also we met some issues in mapping the geospatial graph when we try to show the contrast among the states. Also, due to the change of data in the middle of the process, we need to redraw all the plots and maps for our analysis, and there’s some variation in the code that we need to fix for the new data.

Task 3: Write a logistic regression model to classify which flights are seriously delayed (the results are either ‘1’ or ‘0’), which is the prerequisite for adding the Seriously_delay column in Task 1.

- Expected Time: 3.5 hours
- Actual Time: 5 hours

We spent most of the time training a logistic regression model at the beginning, but we failed to create one due to some errors. So, after our discussion, we decided to use other methods to set our serious delay standard, which is a relatively simple one and we figured it out quickly.

Task 4: Use machine learning models (decision tree, KNN, and Linear Regression) to predict the delay time of some flights in the future. Also, based on the Performance metrics of OLS, such as “mean squared error”, conclude which models’ prediction is good enough to be used as references for future lights.

- Expected Time: 13 hour
- Actual Time: 30 hours

Since we have a lot of data at the beginning when we want to do the flight for the whole nation, training a machine takes a longer time than we expected, about 20 minutes each time, and sometimes will cause an error about running out of computer memory. Even after we shrunk our data to the scope of WA, it is still large for some machine learning models such as KNN and linear regression. Hence, we still need to wait for a relatively long time to get our results.

Notice: Each of the group members will develop codes in Task 1, 2, 3 separately, and collaborate on Task 4 together. However, during the revising period, 3 of us will work together through task 1 to task 4. Additionally, if one is stuck with merge and join data frames in Task 1, which is potentially confusing for us, the person who works Task 2 and

3 should help develop ideas together, especially for testing what different join types will produce.

Testing

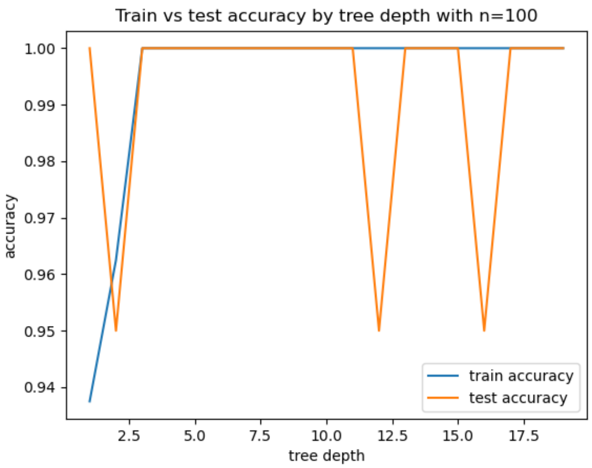
Visualization:

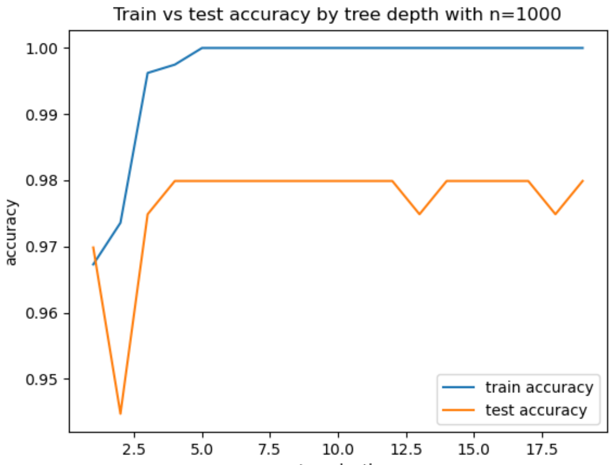
For the visualization part, we create a test_df which contains only 3 rows and all columns as our smaller testing data frame. When we only run the test.py, it will produce some plots with the same logic in plot.py but with test_df, so there are only at most 3 symbols on each graph. In the main function in test.py, we also print the test_df, so we can calculate whether each symbol in the graph is in the correct position. If so, the test for visualization is passed.

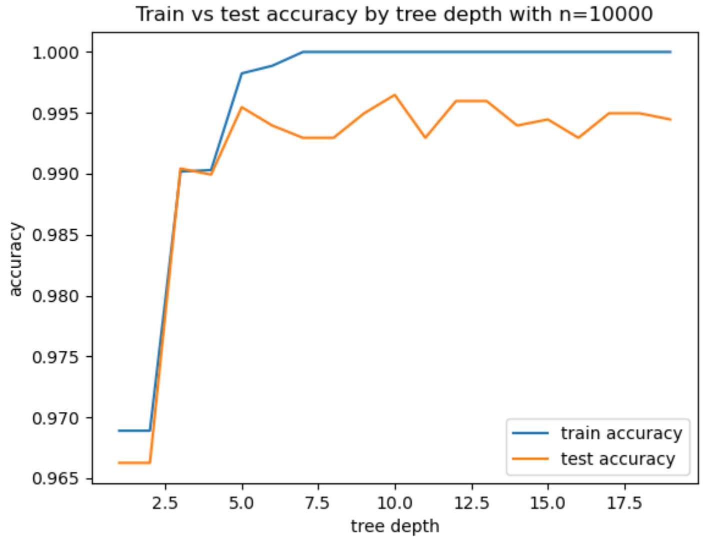
Machine learning:

For the machine learning part, we create a test_ML.py file for our testing. And we take three different sizes of subset from our main data for our testing which are data with 100 rows, 1000 rows, and 10000 rows. We analyze the graph and accuracy results of our test sets and we think all of our test data results and accuracies follows the general trend of our train data. Also the result produced by the subset data makes sense to us, so we pass the test.

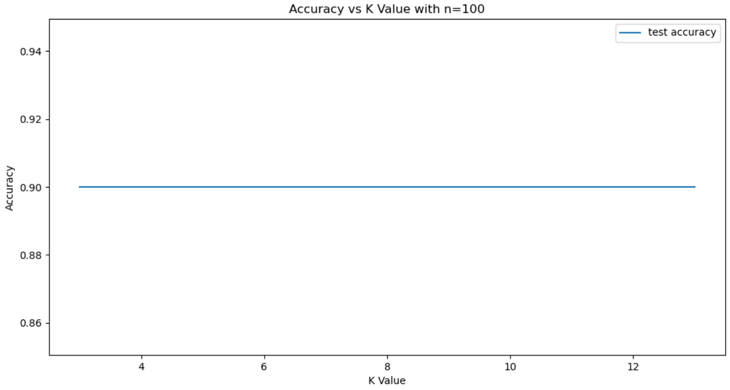
- Decision Tree Classification:

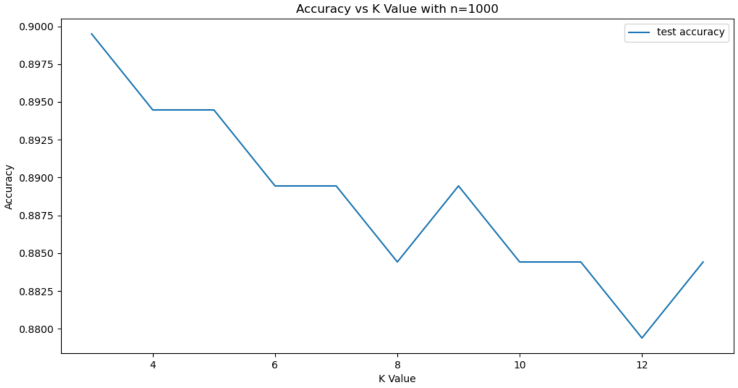


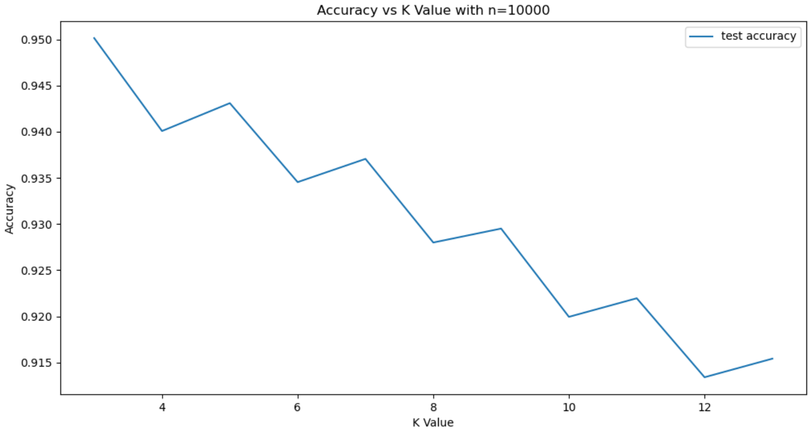




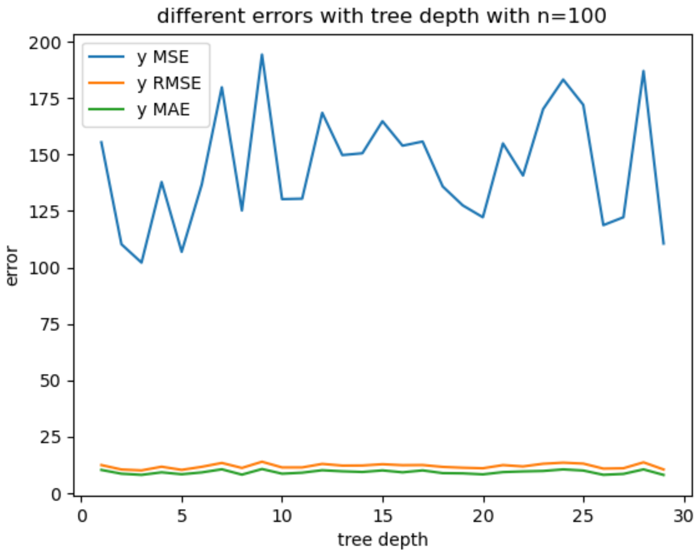
- KNN Classification

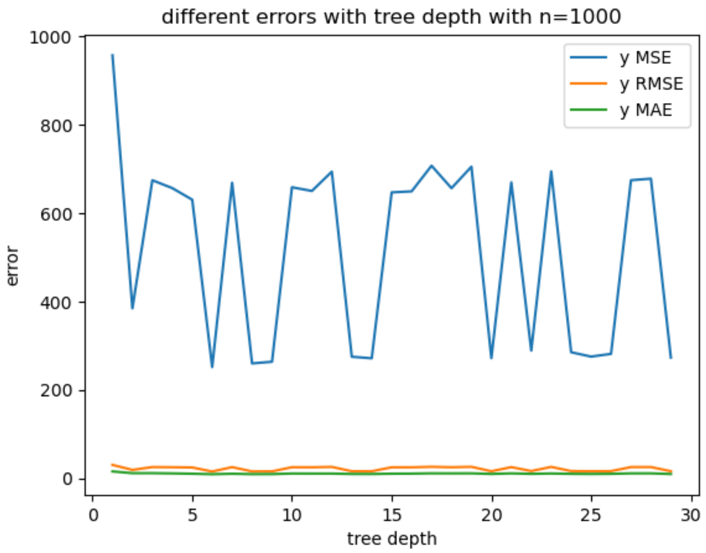


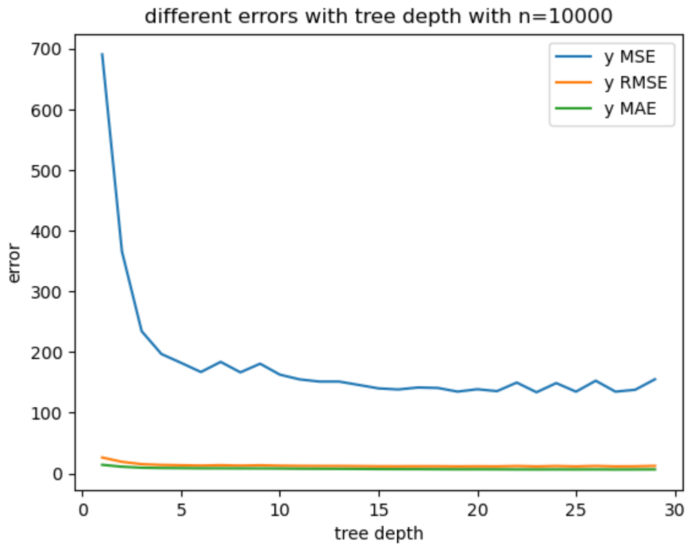




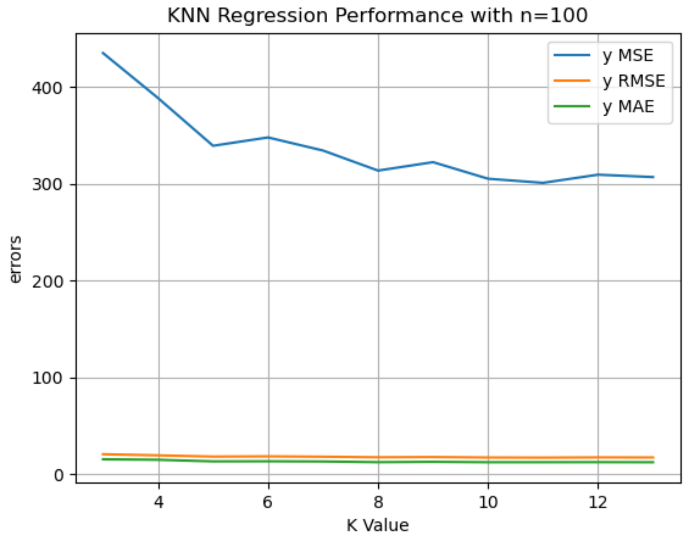
-
- Decision Tree Regression

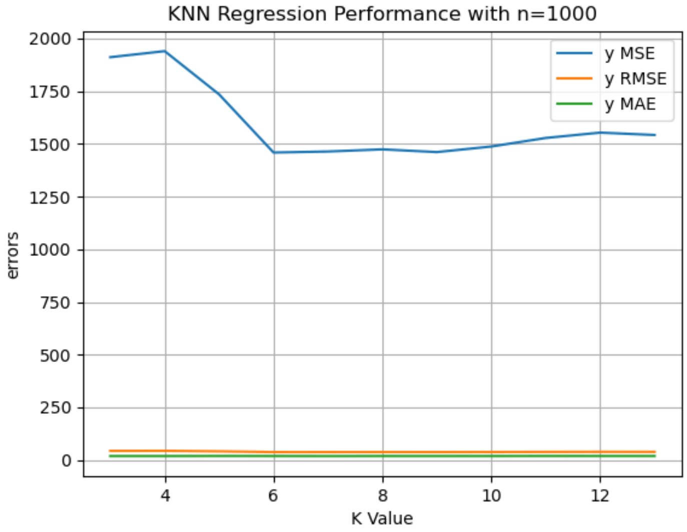


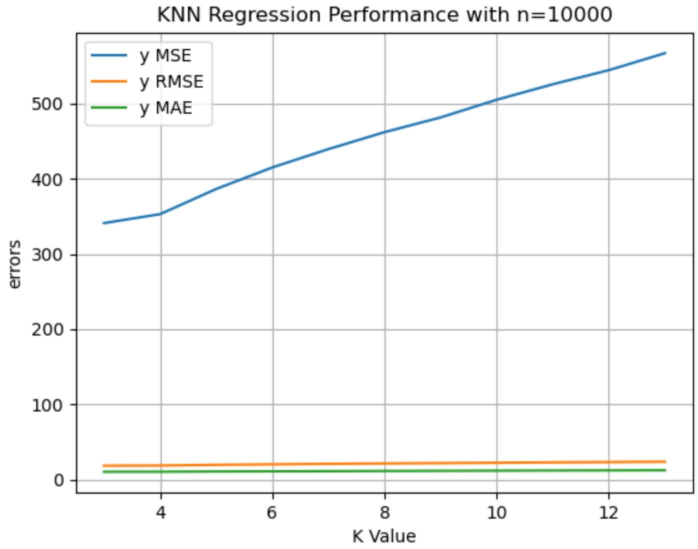




- KNN Regression:


KNN Regression Performance with n=100


KNN Regression Performance with n=1000


KNN Regression Performance with n=10000

Collaboration

We asked TA for some technical issues about setting up environments in Office Hour. The mentor gave us kind support and good flexibility when we have questions. For any other parts, we are doing entirely by the three of us.