Capstone Project 1:

Project Proposal

I) What is the problem you want to solve?

In 2015 there were lots of flight delays in the United States. The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the status of the flights by providing a summary of the arrival delays, departure delays, scheduled departure, on-time arrival in their monthly report. The data is acquired from Kaggle:https://www.kaggle.com/usdot/flight-delays. The flight data set contains 5819079 rows with 31 columns and 14 different carriers.

Who is my client and why do they care about this problem? In other words, what will my client do or decide based on my analysis that they wouldn't have done otherwise?

My clients are Airlines Passengers, vacationers, the Department of Transportation/National Aviation System.

No one enjoys spending extra time at the airport because flight delays have kept their plane grounded.

The analysis will help passengers to be mindful of the airlines that have the largest/lowest delays and cancellations; what days of the week are the best to fly. And what airports should they avoid? The analysis will also help The Department of Transportation(DOT)/National Aviation System to implement regulations that will reduce delays and provide better protection for consumers.

Many U.S. airlines are now prohibited from allowing the domestic flight to remain on the Tarmac for more than three hours. Carriers are required to notify passengers of known delays and provide accommodations.

What data are you using? How will you acquire the data?
I am using this free dataset from Kaggle: 2015 Flight Delays and Cancellations: https://www.kaggle.com/usdot/flight-delays
This data has 3 sources: airlines.csv, airports.csv, flights.csv.

Briefly outline how you'll solve this problem.
I am going to solve the problem by following these steps

- Process the data (data wrangling):
- Explore the data
- Perform in-depth analysis (machine learning, statistical models, algorithms):
- Communicate results of the analysis:

What are your deliverables? Code & Paper

II)Data Collection & Wrangling Summary

The dataset consists of 5819079 rows and 31 columns. When I took a closer look at the data, I noticed that several features have Null values. I performed **data wrangling** by Renaming, Sorting reordering, duplicating data, addressing missing or invalid data, and Filtering to the desired subset of data.

I used the **IsNull**() function to detect missing values. And The dropna() function to remove missing values.

```
airport.isnull().sum()
```

```
IATA_CODE     0
AIRPORT       0
CITY          0
STATE         0
COUNTRY       0
LATITUDE      3
LONGITUDE     3
dtype: int64
```

```
airport = airport.dropna(subset = ['LATITUDE','LONGITUDE'])
```

```
airport.isnull().sum()
```

```
IATA_CODE     0
AIRPORT       0
CITY          0
STATE         0
COUNTRY       0
LATITUDE      0
LONGITUDE     0
dtype: int64
```

```
In [14]: Data_NULL = data.isnull().sum()*100/data.shape[0]
         Data_NULL
```

```
Out[14]: YEAR                   0.000
         MONTH                  0.000
         DAY                    0.000
         DAY_OF_WEEK            0.000
         AIRLINE                0.000
         FLIGHT_NUMBER          0.000
         TAIL_NUMBER            0.251
         ORIGIN_AIRPORT         0.000
         DESTINATION_AIRPORT    0.000
         SCHEDULED_DEPARTURE    0.000
         DEPARTURE_TIME         1.433
         DEPARTURE_DELAY        1.433
         TAXI_OUT               1.481
         WHEELS_OFF             1.481
         SCHEDULED_TIME         0.000
         ELAPSED_TIME           1.738
         AIR_TIME               1.738
         DISTANCE               0.000
         WHEELS_ON              1.537
         TAXI_IN                1.537
         SCHEDULED_ARRIVAL      0.000
         ARRIVAL_TIME           1.537
         ARRIVAL_DELAY          1.738
         DIVERTED               0.000
         CANCELLED              0.000
         CANCELLATION_REASON   98.503
         AIR_SYSTEM_DELAY      81.702
         SECURITY_DELAY        81.702
         AIRLINE_DELAY         81.702
         LATE_AIRCRAFT_DELAY   81.702
         WEATHER_DELAY         81.702
         dtype: float64
```

We can see that 98% of the values in the Cancellation reason column are null for which it is of less use while predicting

Delays. Some other columns include 81.7% in Air System Delay, Security Delay, Airline Delay, Weather Delay etc. So I am going to create two Datasets which have no null values. First, I am removing all the null values irrespective of different types of Delays. Second, I am going to take the data set with respect to different types of delays. The first Dataset is named as Flights and the other one is named as Flight_Delays.

Dropna() Function

The dropna() function is used to remove missing values.

```
In [15]: # Dropping of subset of null values
         data1 = data.dropna(subset = ["TAIL_NUMBER",'DEPARTURE_TIME','DEPARTURE_DELAY','TAXI_OUT','WHEELS_OFF','SCHEDULED_TIME'
                    'ELAPSED_TIME','AIR_TIME','WHEELS_ON','TAXI_IN','ARRIVAL_TIME','ARRIVAL_DELAY'])
```

III) Exploratory data analysis summary (visualization and inferential statistics)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypotheses, and to check assumptions with the help of summary statistics and graphical representations.

For the Exploratory Data Analysis, I used Python coding and Tableau Visualization to get a brief insight and inference from the data.

I created various Plots so I can visualize the Dataset. And to get a glance at the variable affecting the delays of the Airlines.

Departure_Delay is the labeled variable. It plays a significant part in terms of explaining the answer to the project question.

The other subgroups /variables that are significant to the project are :

AIR_SYSTEM_DELAY, SECURITY_DELAY',
AIRLINE_DELAY, LATE_AIRCRAFT_DELAY,
WEATHER_DELAY, YEAR, MONTH, DAY,
DAY_OF_WEEK, TAIL_NUMBER,
SCHEDULED_DEPARTURE, DEPARTURE_TIME,
SCHEDULED_TIME, SCHEDULED_ARRIVAL,
ARRIVAL_TIME, DIVERTED, CANCELLED,
CANCELLATION_REASON, 'FLIGHT_NUMBER,
WHEELS_OFF, WHEELS_ON, AIR_TIME

From the Pearson correlation coefficient, I noticed that there is a strong correlation between Distance and Air_Time; Distance & Elapse_Time; Distance & Time_Scheduled; Schedule_Time & Air_time; Schedule Time & Distance, Departure_Delay & Departure_Delay, etc.

The most appropriate test to use to analyze these relations is the p-value testing.

Summary Statistics

```
In [7]: data.describe()
```

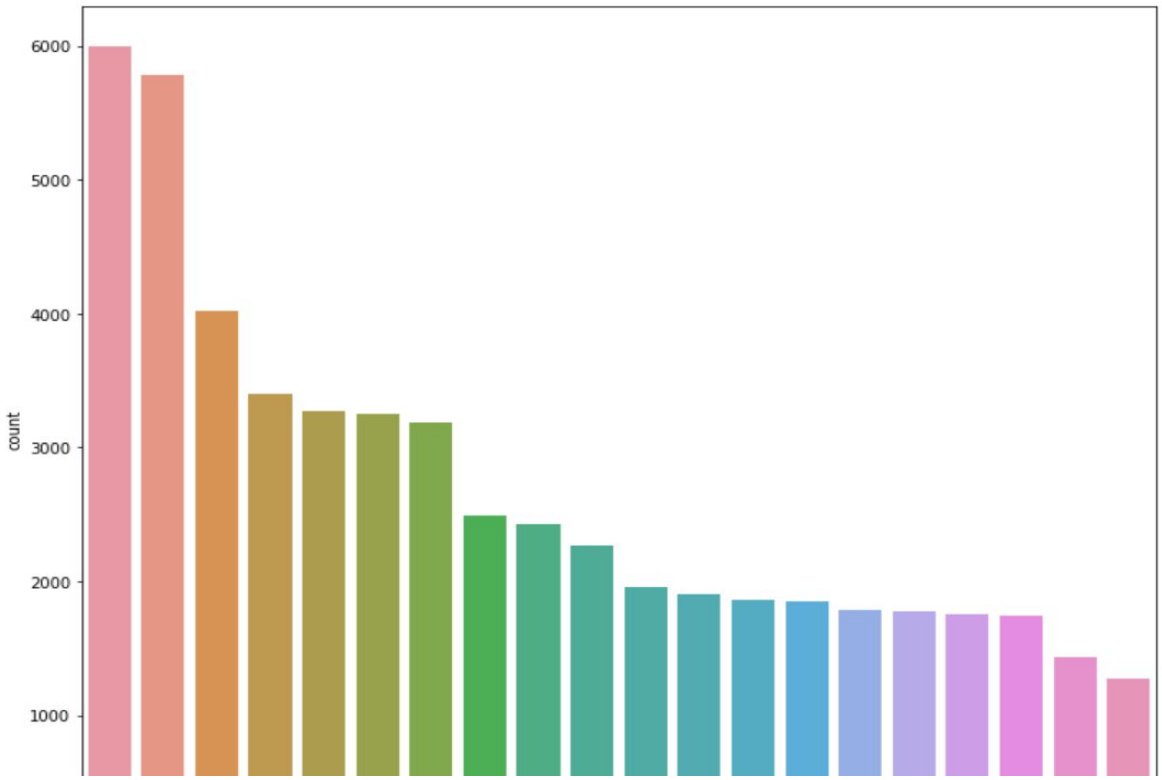
Out[7]:

	YEAR	MONTH	DAY	DAY_OF_WEEK	FLIGHT_NUMBER	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	TAXI_
count	100000.0	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	98567.000000	98567.000000	98519.00
mean	2015.0	6.518270	15.752430	3.928900	2181.275680	1330.654150	1335.983940	9.388061	16.13
std	0.0	3.411504	8.791039	2.001436	1759.683847	484.068943	496.750957	36.531606	9.07
min	2015.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	-40.000000	1.00
25%	2015.0	4.000000	8.000000	2.000000	735.000000	918.000000	922.000000	-5.000000	11.00
50%	2015.0	7.000000	16.000000	4.000000	1702.000000	1326.000000	1332.000000	-2.000000	14.00
75%	2015.0	9.000000	23.000000	6.000000	3247.000000	1730.000000	1740.000000	7.000000	19.00
max	2015.0	12.000000	31.000000	7.000000	7438.000000	2359.000000	2400.000000	1076.000000	180.00

8 rows × 26 columns

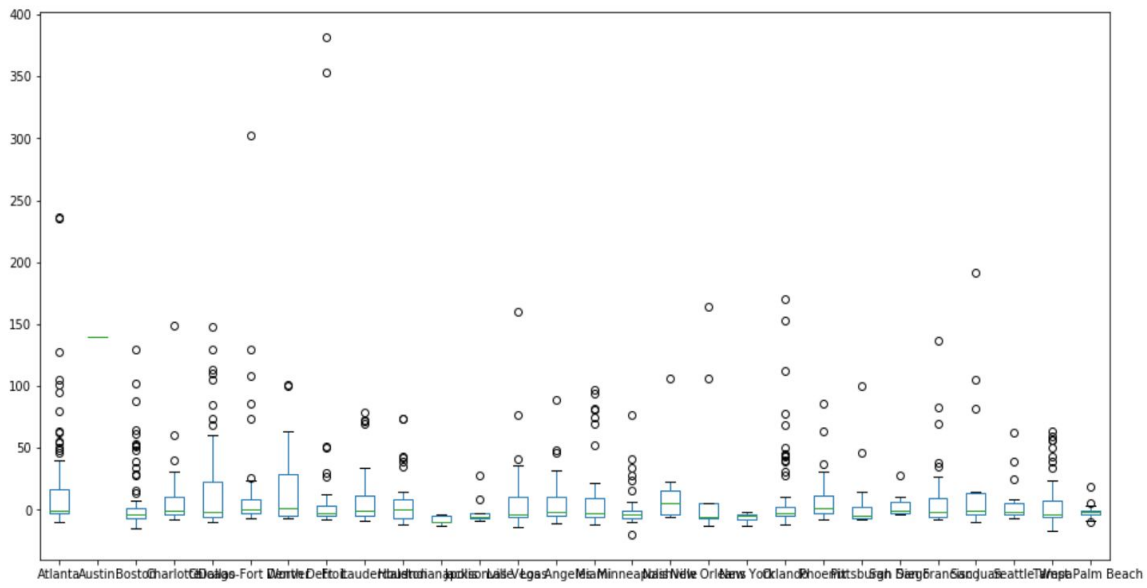
```
]: # Dropping of subset of null values
data1 = data.dropna(subset = ["TAIL_NUMBER",'DEPARTURE_TIME','DEPARTURE_DELAY','TAXI_OUT','WHEELS_OFF','SCHEDULED_TIME'
            'ELAPSED_TIME','AIR_TIME','WHEELS_ON','TAXI_IN','ARRIVAL_TIME','ARRIVAL_DELAY'])
```

```
]: plt.figure(figsize=(10, 10))
axis = sns.countplot(x=Flights['Origin_city'], data = Flights,
              order=Flights['Origin_city'].value_counts().iloc[:20].index)
axis.set_xticklabels(axis.get_xticklabels(), rotation=90, ha="right")
plt.tight_layout()
plt.show()
```



```
]:
hi_volume_airports_pivots = Flights.pivot_table(index='Date', columns=Flights['Origin_city'].iloc[:1000], values='DEPAR
hi_volume_airports_pivots.plot(kind='box', figsize=[16,8])
```

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x1a5f5201d0>
```



The count plot depicts the origin city on the x-axis and the departure delay on the y-axis. The count plot also shows that Chicago has the highest count of flight from origin city, follows by Atlanta.

IV: In-Depth Analysis using machine learning

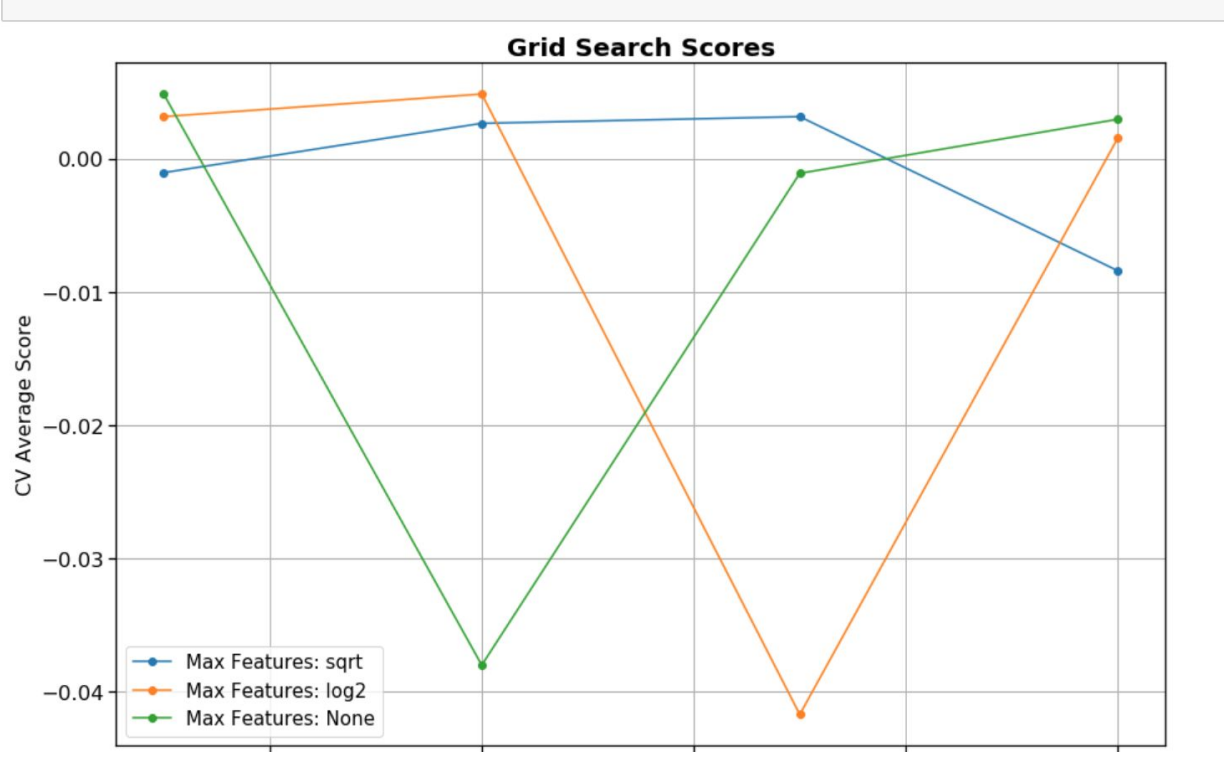
I use 100000 samples to predict how much a flight will delay. Before applying machine learning, I performed One-hot coding to convert the categorical values into numerical values.

For the prediction, I separated the dataset into training and testing. Training data helps recognize patterns in the dataset. And Testing data is used to evaluate the performance of the model.

Also, From sci-kit-learn, I used Linear regression, Random forest, and Decision Tree for model comparison.

I also used scaling, a method used to normalize the range of independent variables or features of data.

To determine the optimal values and accuracy of the model, I used Grid Search and cross-validation



Model fitting and results

```
]: for model, name in zip([LinR,Rfc,Dtc],
       ['Linear Regression','Random forest Regressor',"DecisionTreeRegressor"]):
      model1 = model.fit(X_train_sc,y_train)
      Y_predict=model1.predict(X_test_sc)
      print(name)
      print('Mean Absolute Error:', mean_absolute_error(y_test, Y_predict))
      print('Mean Squared Error:', mean_squared_error(y_test, Y_predict))
      print('Root Mean Squared Error:', np.sqrt(mean_squared_error(y_test, Y_predict)))
      print('R2 : ',r2_score(y_test, Y_predict))
      print()
```

```
Linear Regression
Mean Absolute Error: 18.39922563412102
Mean Squared Error: 1277.0927484702281
Root Mean Squared Error: 35.73643446778411
R2 :  0.028298113427712046

Random forest Regressor
Mean Absolute Error: 17.29092119928667
Mean Squared Error: 1425.9005813698175
Root Mean Squared Error: 37.761098783931295
R2 :  -0.08492534049798817

DecisionTreeRegressor
Mean Absolute Error: 21.48367142220241
Mean Squared Error: 2389.095575122604
Root Mean Squared Error: 48.878375332273514
R2 :  -0.8177917620540451
```

Results

name	R2	Root Mean Squared Error
Linear Regression	0.028298113427712046	35.73643446778411
Random forest Regressor	-0.08492534049798817	37.761098783931295
Decision Tree Regressor	-0.8177917620540451	48.878375332273514

Advantage & Disadvantage of using Linear Regression, Decision Tree and Random Forest

Advantage:

Linear Regression is that it helps to find the relationships between independent and dependent variables.

Disadvantage:

Linear regression only looks at linear relationships between dependent and independent variables.

Advantage

decision tree can handle both numerical and categorical data. And it is extremely fast to run.

Disadvantage

The model is prone to overfitting, especially when a tree is particularly deep.

Advantages:

1)Random Forest has the ability to handle multiple input features without the need for features deletion.

2. Random Forest Works well with missing data. Ant it still gives better predictive accuracy.

Disadvantage:

Random Forest is Not easily interpretable

DISCUSSION

The models did not work. I only compared three models: Linear Regression, Random Forest, and Decision Tree. For future reference, I would use more models to make my comparison. I can also try some deep learning techniques.

Feedback from my mentor.

Hi Vertuile,

Yes, we can keep working with modifying the features and see if we can get better results. I am not sure it will improve the model a lot.

Keep in mind that the important factors are that you perform the visualization of the statistical tests and the model and you know that the data is not sufficient. Add a discussion that will describe which other data might improve the results. or what analysis you think might improve the results.

I think that we should continue the next project. If we will have time after we finish the next project we can have a look at this analysis again.

Please let me know what you think.

Best

Shmuel