Project_1

September 26, 2021

#

Project 1

1 Dataset Exploration

For this project you will be given a dataset and an associated problem. Over the course of the day, you will explore the dataset and train the best model you can in order to solve the problem. At the end of the day, you will give a short presentation about your model and solution.

1.0.1 Deliverables

1. A **copy of this Colab notebook** containing your code and responses to the ethical considerations below.
2. At the end of the day, we will ask you and your group to stand in front of the class and give a brief **presentation about what you have done**.

1.1 Team

Please enter your team members' names in the placeholders in this text area:

- *Wren Priest*
- *Jose Martinez*
- *Maria Quintero-Pena*

2 Exercises

2.1 Exercise 1: Coding

Kaggle hosts a dataset containing US airline on-time statistics and delay data from the US Department of Transportation's Bureau of Transportation Statistics (BTS). In this project, we will **use flight statistics data to gain insights into US airports' and airlines' flights in 2008.**

You are free to use any toolkit we've covered in class to solve the problem (e.g. Pandas, Matplotlib, Seaborn).

Demonstrations of competency: 1. Get the data into a Python object. 1. Inspect the data for each column's data type and summary statistics. 1. Explore the data programmatically and visually. 1. Produce an answer and visualization, where applicable, for at least three questions from the list below, and discuss any relevant insights. Feel free to generate and answer some of your own questions.

- Which U.S. airport is the busiest airport? You can decide how you'd like to measure "business" (e.g., annually, monthly, daily).
- Of the 2008 flights that are *actually delayed*, think about:
  - Which 10 U.S. airlines have the most delays?
  - Which 10 U.S. airlines have the longest average delay time?
  - Which 10 U.S. airports have the most delays?
  - Which 10 U.S. airports have the longest average delay time?
- More analysis:
  - Are there patterns on how flight delays are distributed across different hours of the day?
  - How about across months or seasons? Can you think of any reasons for these seasonal delays?
  - If you look at average delay time or number of delays by airport, does the data show linearity? Does any subset of the data show linearity?
  - Add reason for delay to your delay analysis above.
  - Examine flight frequencies, delays, time of day or year, etc. for a specific airport, airline or origin-arrival airport pair.

2.1.1 Student Solution

Imports

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import zipfile
```

1. Get the data into a Python object.

[ ]:
```
# Use Kaggle API to download airlinedelaycauses.zip
! cp ~/.kaggle/kaggle.json .
! chmod 600 kaggle.json && (ls ~/.kaggle 2>/dev/null || mkdir ~/.kaggle) && mv␣
 ↪kaggle.json ~/.kaggle/ && echo 'Done'
! kaggle datasets download giovamata/airlinedelaycauses
! ls
```

```
kaggle.json
Done
401 - Unauthorized
airlinedelaycauses.zip  sample_data
```

[ ]:
```
# unzip airlinedelaycauses.zip
zf = zipfile.ZipFile('airlinedelaycauses.zip')

# read into df
df = pd.read_csv(zf.open('DelayedFlights.csv'))
df
```

[ ]:
```
          Unnamed: 0  Year  Month  …  NASDelay  SecurityDelay
LateAircraftDelay
0                  0  2008      1  …       NaN            NaN
```

```
NaN
1                  1  2008      1  …       NaN            NaN
NaN
2                  2  2008      1  …       NaN            NaN
NaN
3                  4  2008      1  …       0.0            0.0
32.0
4                  5  2008      1  …       NaN            NaN
NaN
…                ...  ...    ...  …       …                …
…
1936753      7009710  2008     12  …       0.0            0.0
22.0
1936754      7009717  2008     12  …      18.0            0.0
0.0
1936755      7009718  2008     12  …      19.0            0.0
79.0
1936756      7009726  2008     12  …       NaN            NaN
NaN
1936757      7009727  2008     12  …       NaN            NaN
NaN

[1936758 rows x 30 columns]
```

2. Inspect the data for each column's data type and summary statistics.

[ ]:
```
# print datatype of each column
print(df.dtypes)

# describe
df.describe()
```

```
Unnamed: 0             int64
Year                   int64
Month                  int64
DayofMonth             int64
DayOfWeek              int64
DepTime              float64
CRSDepTime             int64
ArrTime              float64
CRSArrTime             int64
UniqueCarrier         object
FlightNum              int64
TailNum               object
ActualElapsedTime    float64
CRSElapsedTime       float64
AirTime              float64
ArrDelay             float64
```

```
DepDelay             float64
Origin                object
Dest                  object
Distance               int64
TaxiIn               float64
TaxiOut              float64
Cancelled              int64
CancellationCode      object
Diverted               int64
CarrierDelay         float64
WeatherDelay         float64
NASDelay             float64
SecurityDelay        float64
LateAircraftDelay    float64
dtype: object
```

[ ]:

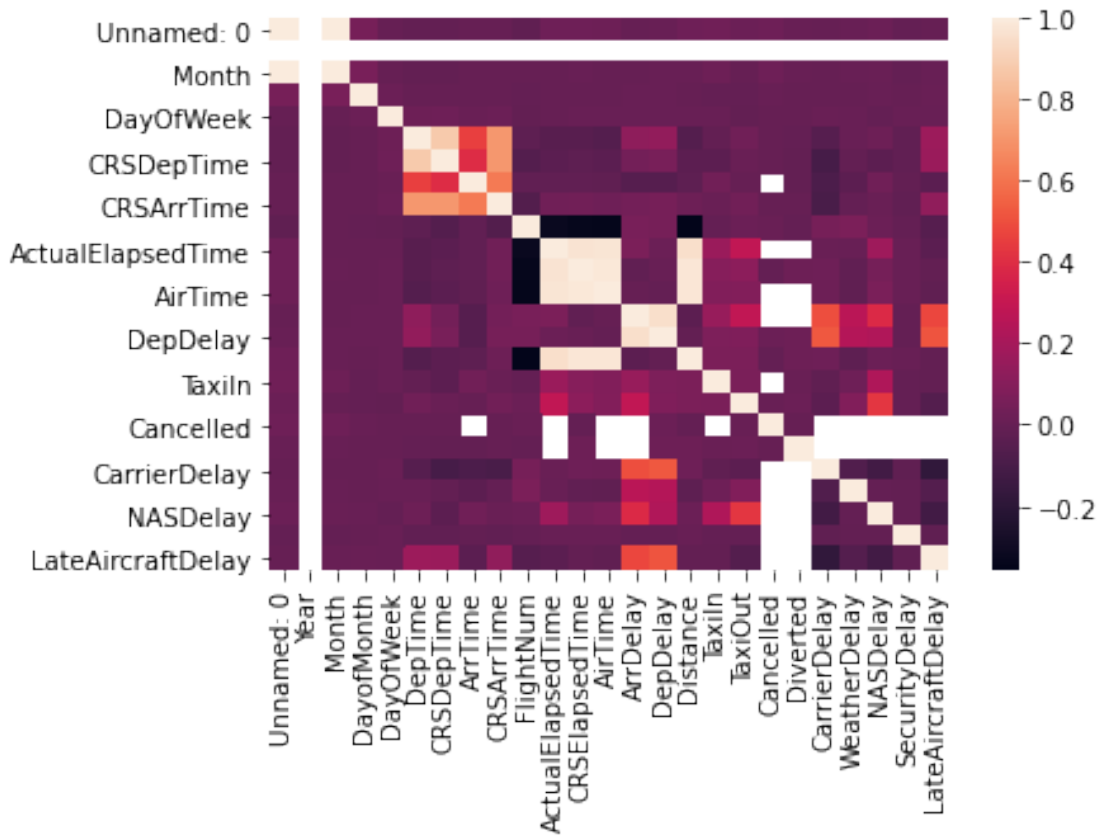
	Unnamed: 0	Year	…	SecurityDelay	LateAircraftDelay
count	1.936758e+06	1936758.0	…	1.247488e+06	1.247488e+06
mean	3.341651e+06	2008.0	…	9.013714e-02	2.529647e+01
std	2.066065e+06	0.0	…	2.022714e+00	4.205486e+01
min	0.000000e+00	2008.0	…	0.000000e+00	0.000000e+00
25%	1.517452e+06	2008.0	…	0.000000e+00	0.000000e+00
50%	3.242558e+06	2008.0	…	0.000000e+00	8.000000e+00
75%	4.972467e+06	2008.0	…	0.000000e+00	3.300000e+01
max	7.009727e+06	2008.0	…	3.920000e+02	1.316000e+03

[8 rows x 25 columns]

3. Explore the data programmatically and visually.

[ ]:
```
sns.heatmap(df.corr())
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f73db755890>
```



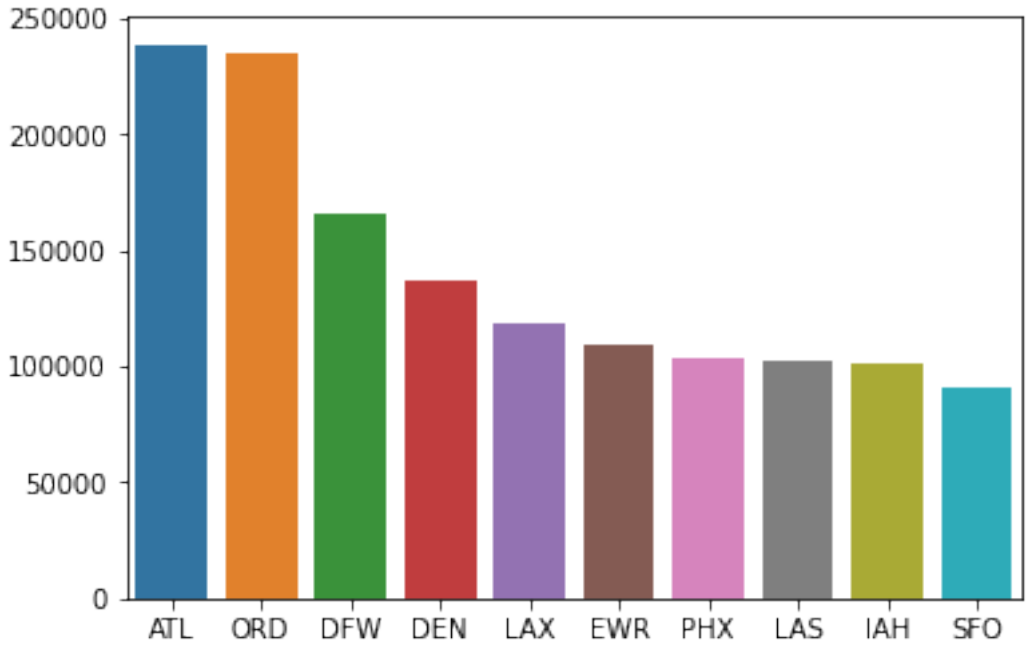
Series used to find the busiest airport

[ ]:
```
#Using origin and destination to determine how busy an airport is
origin = df['Origin'].value_counts()
dest= df['Dest'].value_counts()


#combining origin and destination into one series
origin_plus_dest = df.Origin.append(df.Dest)
origin_dest_count = origin_plus_dest.value_counts()

#x-axis
x = origin_dest_count.head(10).index[0:10]#top 10 busiest aiports

#y-axis
y =origin_dest_count.values[0:10]#amount of origin + desination flights

#Adding Seaborn graph
sns.barplot(x,y)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
```

```
Pass the following variables as keyword args: x, y. From version 0.12, the only
valid positional argument will be `data`, and passing other arguments without an
explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

[ ]: <matplotlib.axes._subplots.AxesSubplot at 0x7f73db633890>



Answering Question II

[ ]:
```
carrier_delay = df.groupby('UniqueCarrier').agg(
    count_ArrDelay = ('ArrDelay', lambda x: x[x>0].count()),
    mean_ArrDelay = ('ArrDelay', lambda x: x[x>0].mean())
)


carrier_delay.sort_values('count_ArrDelay', ascending=False).head(10)
```

[ ]:

UniqueCarrier	count_ArrDelay	mean_ArrDelay
WN	324717.0	35.752200
AA	172197.0	52.308693
MQ	130647.0	49.323276
UA	123989.0	55.247086
OO	121942.0	49.362172
DL	100923.0	45.786501
XE	94313.0	55.424629

```
CO                   83646.0      49.729324
US                   83262.0      44.175278
EV                   75170.0      52.153146
```

Answering Question III

[ ]:
```
carrier_delay.sort_values('mean_ArrDelay', ascending=False).head(10)
```

[ ]:
```
               count_ArrDelay  mean_ArrDelay
UniqueCarrier
B6                    48177.0      63.947859
YV                    63289.0      58.563131
XE                    94313.0      55.424629
UA                   123989.0      55.247086
OH                    49104.0      54.915261
AA                   172197.0      52.308693
EV                    75170.0      52.153146
9E                    46896.0      52.097343
CO                    83646.0      49.729324
OO                   121942.0      49.362172
```

Answering Question IV

[ ]:
```
airport_delay = df.groupby('Origin').agg(
    count_DepDelay = ('DepDelay', lambda x: x[x>0].count()),
    mean_DepDelay = ('DepDelay', lambda x: x[x>0].mean())
)

airport_delay.sort_values('count_DepDelay', ascending=False).head(10)
```

[ ]:
```
        count_DepDelay  mean_DepDelay
Origin
ATL           131613.0      40.893240
ORD           125979.0      50.531168
DFW            95414.0      38.340610
DEN            74323.0      37.698868
LAX            58772.0      38.174998
IAH            56847.0      37.735219
PHX            55720.0      35.242067
LAS            53710.0      38.819643
EWR            52925.0      50.430099
DTW            43923.0      40.493637
```

Answering Question V

[ ]:
```
airport_delay.sort_values('mean_DepDelay', ascending=False).head(10)
```

[ ]:
```
        count_DepDelay  mean_DepDelay
Origin
```

CMX	34.0	116.147059
PLN	21.0	93.761905
SPI	357.0	83.848739
ALO	31.0	82.225806
MQT	203.0	79.556650
ACY	29.0	79.310345
MOT	133.0	78.661654
HHH	183.0	76.530055
EGE	861.0	74.128920
BGM	103.0	73.155340

4. Questions and Answers 1.) Which U.S. airport is the busiest airport?

Based on the code and visualizations done above,when consiering arrivals and departures as parameters for how busy an aiport is, the busiest aiport was ATL

2.) Which 10 U.S. airlines have the most delays?

WN, AA, MQ, UA, OO, DL, XE, CO,and EV were the the 10 airlines that had the most delays

3.) Which 10 U.S. airlines have the longest average delay time?

Based on the DataFrame created, the airline that had the longest average delay timme were B6, YV,XE,UA,OH,AA,EV,9E,CO, and OO

4.) Which 10 U.S. airports have the most delays?

ATL, ORG, DFW, DEN, LAX, IAH, PHX, LAS, EWR and DTW

5.) Which 10 U.S. airlines have the longest average delay time?

Based on the frame created, CMX, PLN, SPI, ALO,MQT, ACY, MOT, EGE, AND BGM .

2.2 Exercise 2: Ethical Implications

Even the most basic of data manipulations has the potential to affect segments of the population in different ways. It is important to consider how your code might positively and negatively affect different types of users.

In this section of the project, you will reflect on the ethical implications of your analysis.

2.2.1 Student Solution

Positive Impact

Your analysis is trying to solve a problem. Think about who will benefit if the problem is solved, and write a brief narrative about how the model will help.

[ ]:

Travelers will benefit because they will be able to buy tickets from more efficient airlines. Airlines with not many delays or long delays will benefit because they will be able to

sell more tickets. Smaller airports may be positively impacted assuming they have less delays.

Negative Impact

Solutions usually don't have a universal benefit. Think about who might be negatively impacted by your analysis. This person or persons might not be directly considered in the analysis, but they might be impacted indirectly.

> Airplane customers may be negatively impacted because the prices for airlines with less delays may rise. Airline companys will be negatively impacted if they have more delays or longer delays because they may have less sales. The busier airports may be negatively affected assuming they have more delays.

Bias

Data analysis can be biased for many reasons. The bias can come from the data itself (e.g. sampling, data collection methods, available sources), and from the interpretation of the analysis outcome.

Think of at least two ways that bias might have been introduced to your analysis and explain them below.

> Sampling bias can occur since larger airports tend to have more flights, therefore more delays. Larger airlines will also have a sampling bias since they will also have more flights. There may also be coverage bias since different regions also tend to have more frequent or less frequent activity that can affect airline delays. There may also be a Reporting bias due to Holidays which tend to have more flight and so can have more delays. Airlines in regions with harsh weather may be adversely affected since weather would affect the number of delays in that region.

Changing the Dataset to Mitigate Bias

The most common way that an analysis is biased is when the dataset itself is biased. Look back at the input data that you used for your analysis. Think about how you might change something about the data to reduce bias in your model.

What changes could you make to make your dataset less biased? Consider the data that you have, how and where that data was collected, and what other sources of data might be used to reduce bias.

Write a summary of the changes that could be made to your input data.

> Since the data has potential bias sampling, we could adjust the models to account for the number and size of airports. Since the nnumber of delays will vary with time of year we adjust the data to account for seasonal delays. We could also add a column to take into account weather.

Changing the Analysis Questions to Mitigate Bias

Are there any ways to reduce bias by changing the analysis itself? This could include modifying the choice of questions you ask, the approach you take to answer the questions, etc.

Write a brief summary of any changes that you could make to help reduce bias in your analysis.

> Since the analysis has potential bias sampling, we can weigh the data based on airport size. For example we could ask what percent of flights are delayed instead of the number

of flights that are delayed to avoid penalizing airports that are really busy. We could also adjust the time and location that data is evaluated.

Mitigating Bias Downstream

While analysis can point to suggestions, it is people who make decisions based on them. What processes and/or rules should be in place for people and systems interpreting and acting on the results of your analysis to reduce the bias? Describe these below.

Since the analysis has potential bias reporting and sampling, we can implement newer data. Since the data is from 2008 it is not a good representation of recent flights. The data could also be organized depending on region and airline size to provide more individualized data.