Notebook

December 9, 2019

0.0.1 Question 1d

In the following cell, print a summary of the data selection and cleaning you performed. For example, you should print something like: ”Of the original 1000 trips, 21 anomolous trips (2.1%) were removed through data cleaning, and then the 600 trips within Manhattan were selected for further analysis.” (Note that the numbers in this example are not accurate.)

Your Python code should not include any number literals, but instead should refer to the shape of `all_taxi`, `clean_taxi`, and `manhattan_taxi`. Your response will be scored based on whether you generate an accurate description and do not include any number literals in your Python expression, but instead refer to the dataframes you have created.

One way to do this is with Python’s f-strings. For instance,

```
name = "Joshua"
print(f"Hi {name}, how are you?")
```

prints `Hi Joshua, how are you?`.

Please ensure that your Python code does not contain any very long lines, or we can’t grade it.

```
In [12]: all_to_clean = (all_taxi.shape[0] - clean_taxi.shape[0]) / all_taxi.shape[0] * 100
         clean_to_manhattan = (clean_taxi.shape[0] - manhattan_taxi.shape[0]) / clean_taxi.shape[0] * 1
         print(f"In the original {all_taxi.shape[0]} trips, {all_to_clean}% of them were removed during
         print(f"During the second stage of cleaning, which removes those trips that are not in Manhatta
         print(f"After this two-step cleaning process, {all_taxi.shape[0] - manhattan_taxi.shape[0]} tri
```

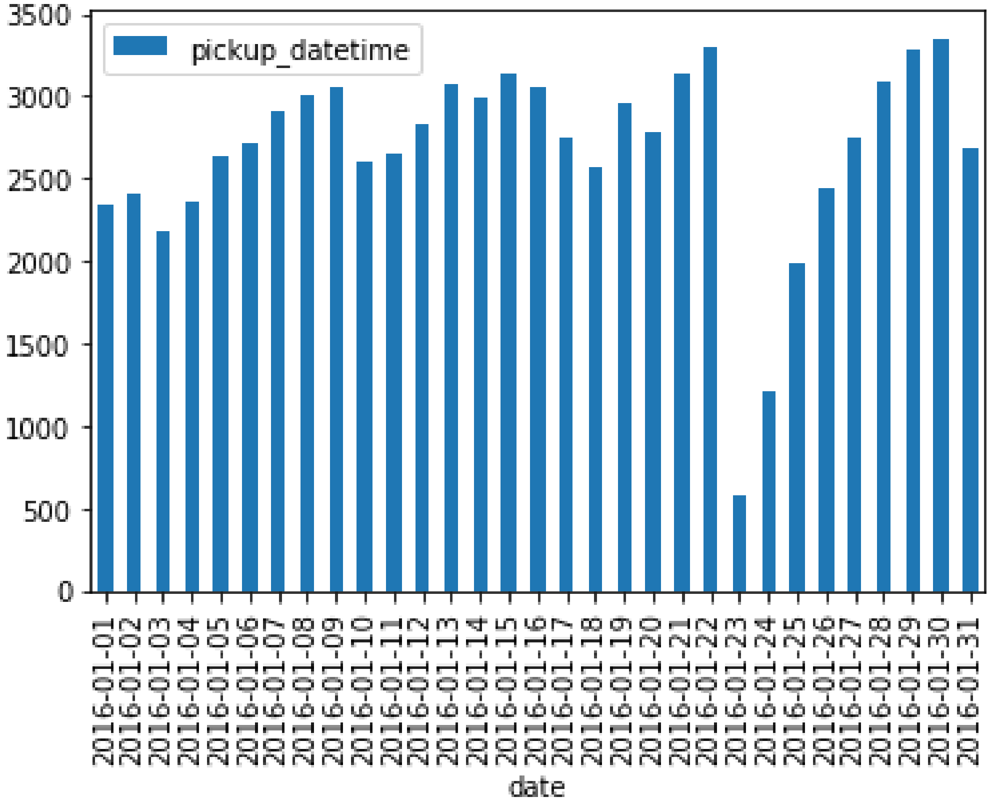
```
In the original 97692 trips, 1.276460713262089% of them were removed during the first stage of cleaning

During the second stage of cleaning, which removes those trips that are not in Manhattan(by not in Manha

After this two-step cleaning process, 14892 trips were removed from the original data and 82800 trips r
```

0.0.2 Question 2b

Create a data visualization that allows you to identify which dates were affected by the historic blizzard of January 2016. Make sure that the visualization type is appropriate for the visualized data.

Hint: How do you expect taxi usage to differ on blizzard days?

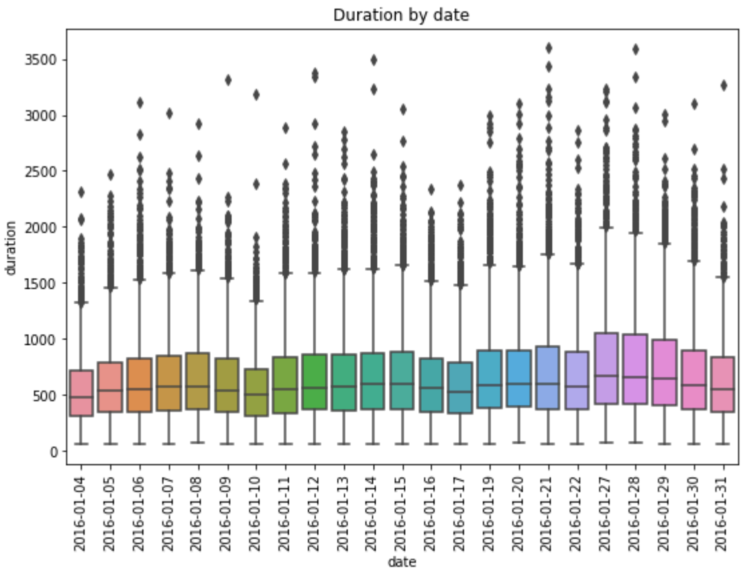
```
In [15]: plt.figure()
         plt_table = manhattan_taxi.groupby("date").count().reset_index().loc[:, ["date", "pickup_datet
         plt_table.plot.bar(x = "date", y = "pickup_datetime")

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc7fb724f98>

<Figure size 432x288 with 0 Axes>
```



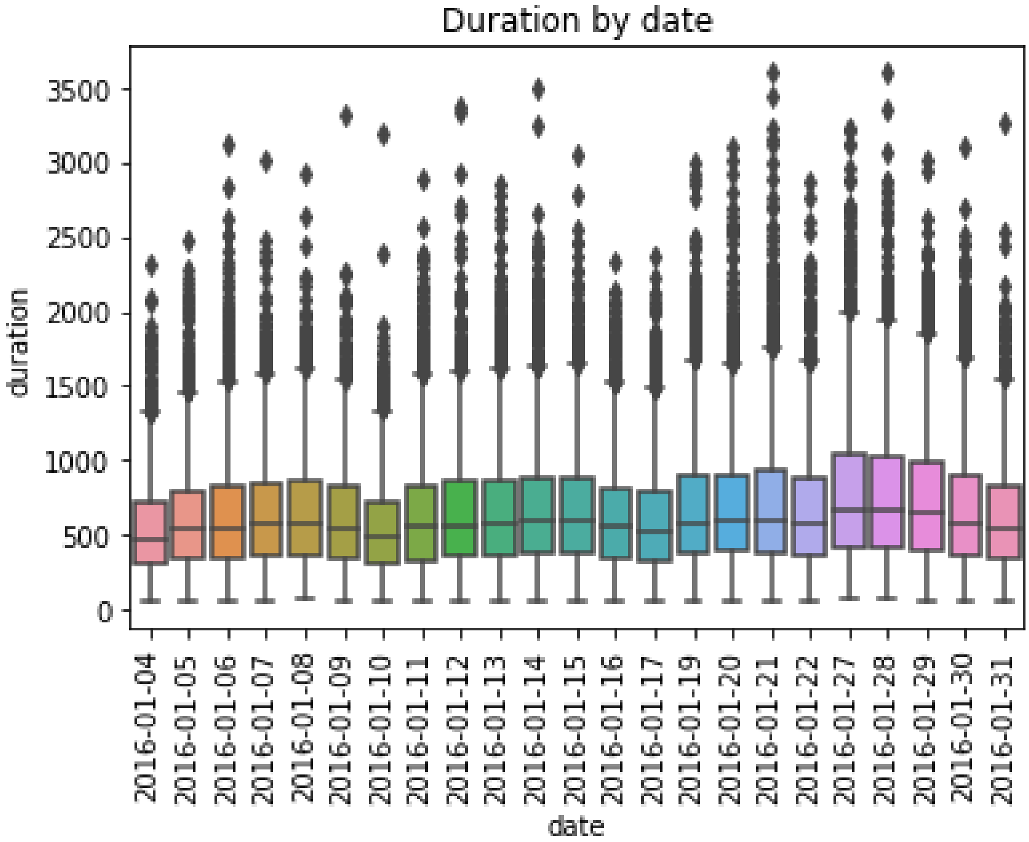
0.0.3 Question 3a

Create a box plot that compares the distributions of taxi trip durations for each day **using `train` only**. Individual dates shoud appear on the horizontal axis, and duration values should appear on the vertical axis. Your plot should look like the following.

Hint: Use `sns.boxplot`.



```
In [19]: sns.boxplot(x = "date", y = "duration", data = train.sort_values(by = "date"))
         plt.xticks(rotation = 90)
         plt.title("Duration by date")
         plt.xlabel("date")
         plt.ylabel("duration")

         # x axis datetime unit

Out[19]: Text(0, 0.5, 'duration')
```


Duration by date
duration
0
500
1000
1500
2000
2500
3000
3500
date
2016-01-04
2016-01-05
2016-01-06
2016-01-07
2016-01-08
2016-01-09
2016-01-10
2016-01-11
2016-01-12
2016-01-13
2016-01-14
2016-01-15
2016-01-16
2016-01-17
2016-01-19
2016-01-20
2016-01-21
2016-01-22
2016-01-27
2016-01-28
2016-01-29
2016-01-30
2016-01-31

0.0.4 Question 3b

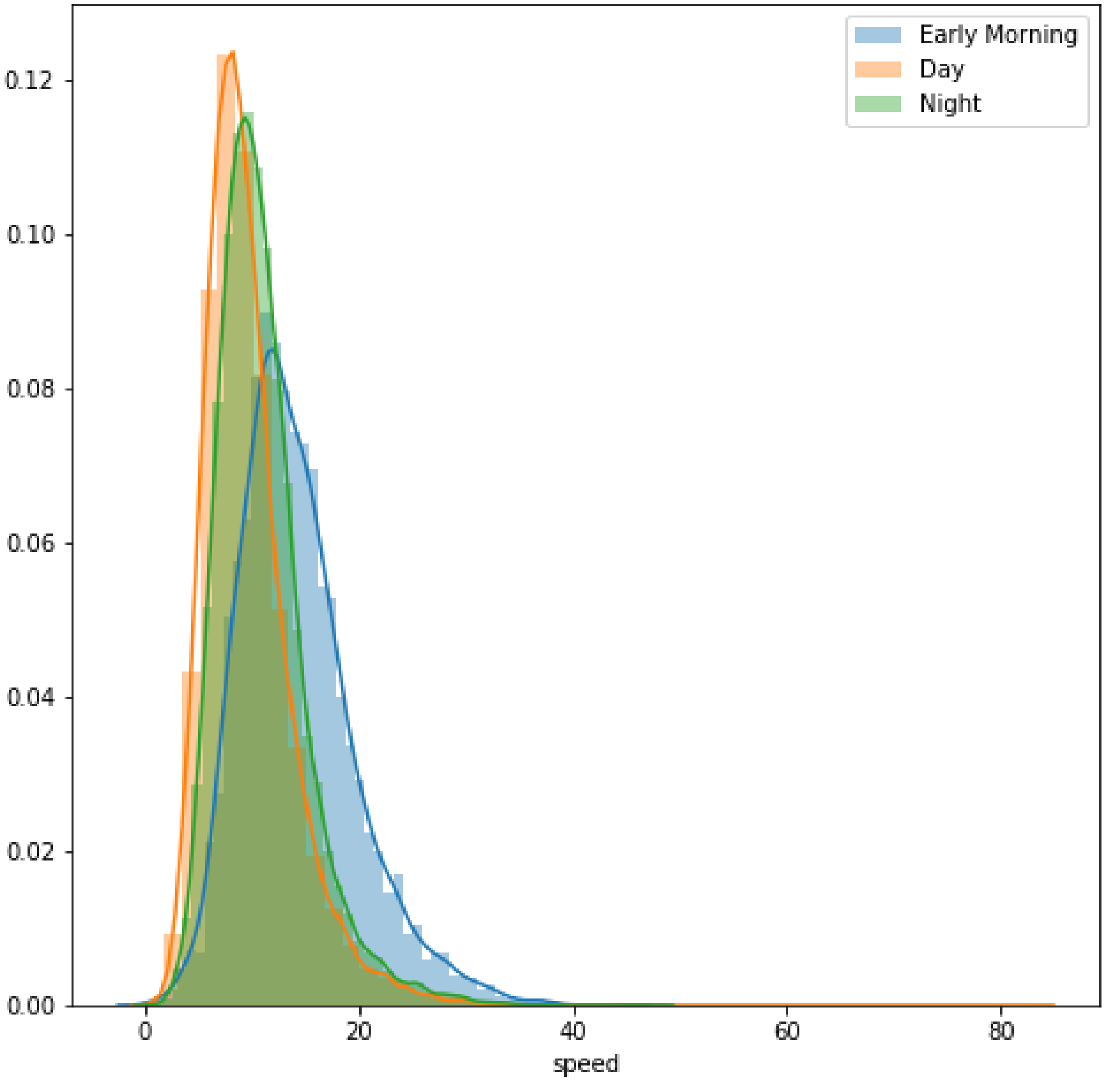
In one or two sentences, describe the assocation between the day of the week and the duration of a taxi trip. This question will be graded on whether your answer is justified by your boxplot and if it is at least somewhat meaningful.

Note: The end of Part 2 showed a calendar for these dates and their corresponding days of the week.

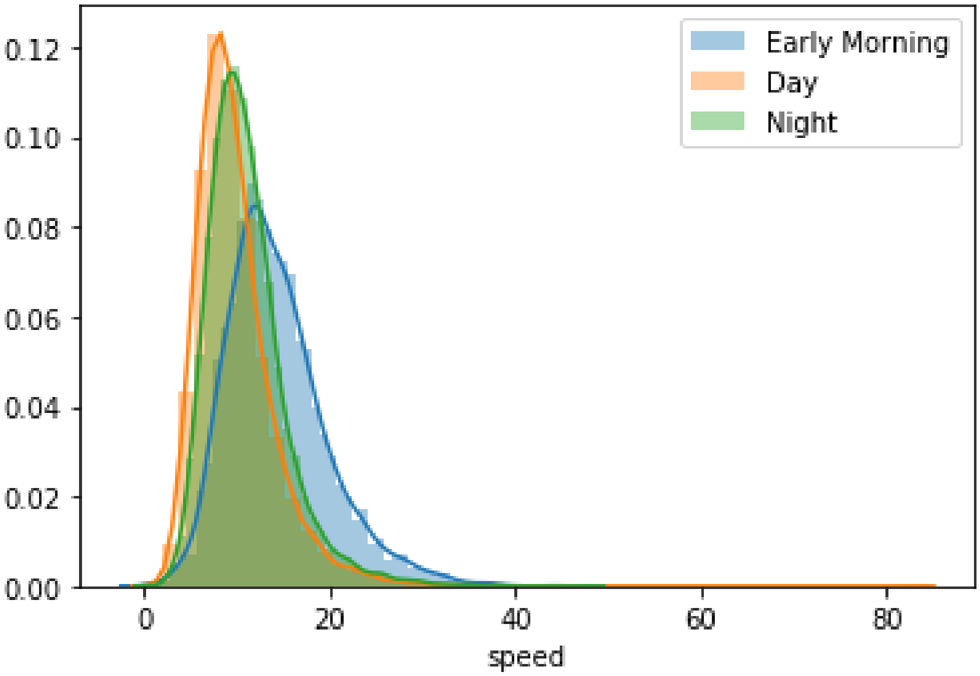
By using the calendar at the end of Part 2, we can see that 2016-01-04 is a Monday, and 2016-01-11 is another. 2016-01-10 is a Sunday, and 2016-01-17 is another. From the barplot above, we can see that the start and the end of a period (Monday and Sunday) has a relatively low median comparing with other days in the period. These two days can be used to seperate periods. For the other days in the week, from Monday to Sunday, the median of the duration generally grows and then decreases.

0.0.5 Question 3c

Use `sns.distplot` to create an overlaid histogram comparing the distribution of average speeds for taxi rides that start in the early morning (12am-6am), day (6am-6pm; 12 hours), and night (6pm-12am; 6 hours). Your plot should look like this:



```
In [21]: sns.distplot(train[train["period"] == 1]["speed"], label = "Early Morning")
         sns.distplot(train[train["period"] == 2]["speed"], label = "Day")
         sns.distplot(train[train["period"] == 3]["speed"], label = "Night")
         plt.legend()

Out[21]: <matplotlib.legend.Legend at 0x7fc7fb82e080>
```


0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
speed
Early Morning
Day
Night

0.0.6 Question 4e

In one or two sentences, explain how the `period` regression model could possibly outperform linear regression model, even when the design matrix of the latter includes one feature for each possible hour.

If for different period, the trips' duration varies a lot, then the period regression model might outperform linear regression model.