

Notebook

August 9, 2019

0.0.1 Question 1d

Print a summary of the data selection and cleaning you performed. Your Python code should not include any number literals, but instead should refer to the shape of `all_taxi`, `clean_taxi`, and `manhattan_taxi`.

E.g., 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 the example above are not accurate.)

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

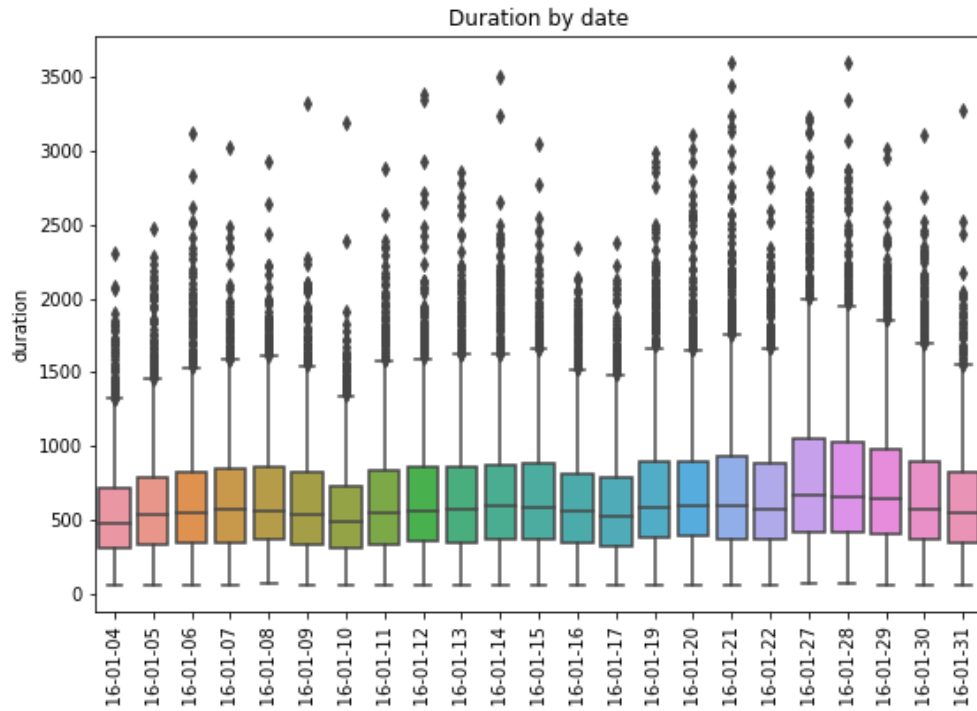
Your response will be scored based on whether you generate an accurate description and do not include any number literals in your Python expression, but instead refer to the dataframes you have created.

```
In [ ]: print('Of the original', len(all_taxi), 'trips,',  
             len(all_taxi)-len(clean_taxi), 'anomolous trips (', round(((len(all_taxi)-len(clean_taxi))  
             '%) were removed through data cleaning, and then the', len(manhattan_taxi),  
             'trips within Manhattan were selected for further analysis.')
```


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.

```
In [ ]: a = manhattan_taxi.groupby(manhattan_taxi['date']).mean().reset_index()
        hb=sns.barplot(x='date',y='duration',data=a)
        hb.set_xticklabels(hb.get_xticklabels(),rotation=90,ha='right')
        plt.tight_layout()
        plt.show()
```

0.03 Question 3a

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

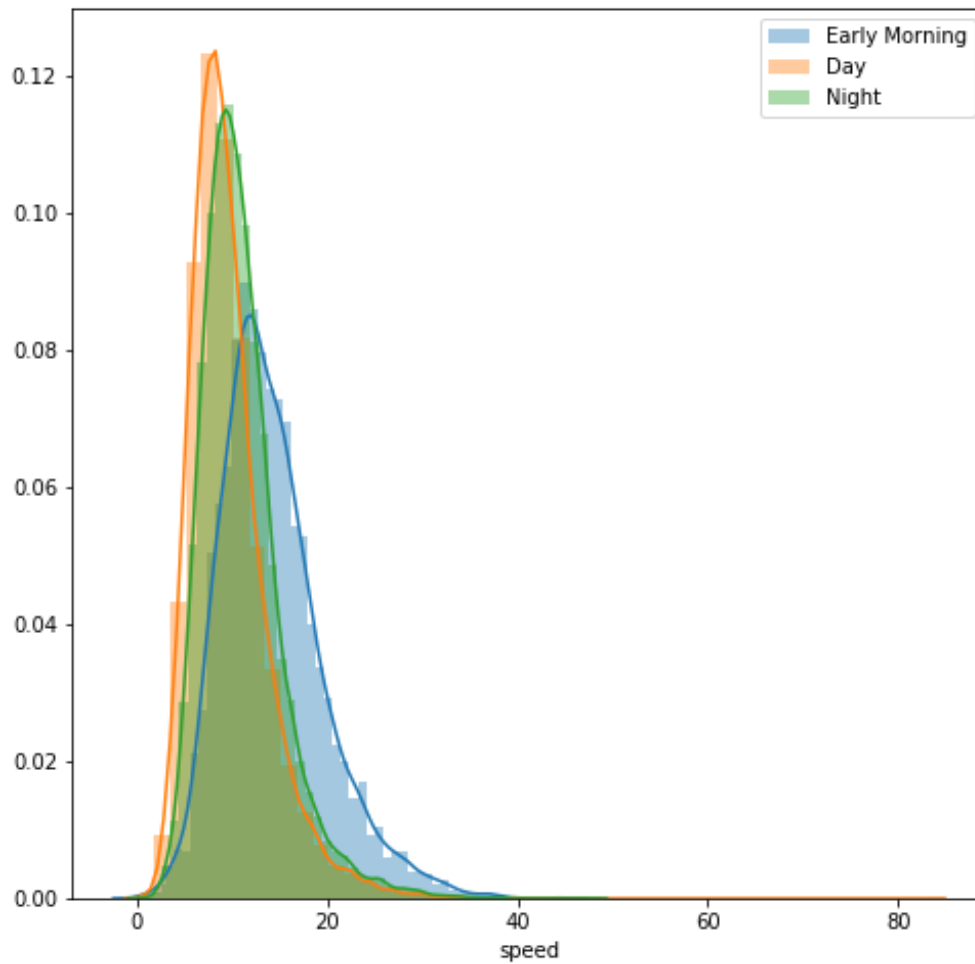
```
In [ ]: new_train=train.sort_values(by='date')
dd=sns.boxplot(x='date',y='duration',data=new_train)
plt.title('Duration by Date')
dd.set_xticklabels(dd.get_xticklabels(),rotation=90,ha='right')
plt.tight_layout()
plt.show()
```


0.0.4 Question 3b

In one or two sentences, describe the association between the day of the week and the duration of a taxi trip.

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

The mean of the duration of taxi trips in weekdays is usually higher than the mean of the duration of taxi trips in the weekends.



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 [ ]: sns.distplot(train[train['hour']<=6]['speed'],label='Early Morning')
        sns.distplot(train[(train['hour']<=18) & (train['hour']>=6)]['speed'],label='Day')
        sns.distplot(train[train['hour']>=18]['speed'],label='Night')
        plt.legend()
```


0.0.6 Question 4e

In one or two sentences, explain how the period regression model could possibly outperform linear regression when the design matrix for linear regression already includes one feature for each possible hour, which can be combined linearly to determine the period value.

The period regression model is a simple form of decision tree regression, which does well in capturing the non-linearity in the data by dividing the data into smaller subset depending on the periods. However, the linear regression works well when the data has a linear shape, but when the data has a non-linear shape, then it cannot capture all the non-linear features.