Nama : Emil Hardiansyah
NIM : D082221005

Tugas 2 Case Study – Cycle Sharing Scheme

Mengimport Library

```
In [18]: %matplotlib inline

         # Mengimport Library
         %matplotlib inline
         import random
         import datetime
         import pandas as pd
         import matplotlib.pyplot as plt
         import statistics
         import numpy as np
         import scipy
         from scipy import stats
         import seaborn
```

Mengimport file dataset

```
In [3]: data = pd.read_csv('D:/Unhas/Big Data Analysis/Case Study 1/trip.csv') #Membaca file dataset
```

Menjabarkan tanggal ke dalam variabel tersendiri

```
In [5]: #Mengkonversi String ke datetime, dan memperoleh fitur baru
        list_of_starttime = list(data['starttime'])

        list_of_starttime = [datetime.datetime.strptime(x, '%m/%d/%Y %H:%M') for x in list_of_starttime]
        data['starttime_mod'] = pd.Series(list_of_starttime,index=data.index)
        data['starttime_date'] = pd.Series([x.date() for x in list_of_starttime],index=data.index)
        data['starttime_year'] = pd.Series([x.year for x in list_of_starttime],index=data.index)
        data['starttime_month'] = pd.Series([x.month for x in list_of_starttime],index=data.index)
        data['starttime_day'] = pd.Series([x.day for x in list_of_starttime],index=data.index)
        data['starttime_hour'] = pd.Series([x.hour for x in list_of_starttime],index=data.index)
```

a. Find the mean, median, and mode of the trip duration of gender type male.

Answer:

```
In [12]: # filter data khusus "Male"
         data_male = data[data['gender'] == 'Male']
         data_male
```

Out[12]:

	trip_id	starttime	stoptime	bikeid	tripduration	from_station_name	to_station_name	from_station_id	to_station_id	usertype	gender	birthyear
0	431	10/13/2014 10:31	10/13/2014 10:48	SEA00298	985.935	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washing...	CBD-06	PS-04	Member	Male	1960.0
1	432	10/13/2014 10:32	10/13/2014 10:48	SEA00195	926.375	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washing...	CBD-06	PS-04	Member	Male	1970.0
4	435	10/13/2014 10:34	10/13/2014 10:49	SEA00202	923.923	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washing...	CBD-06	PS-04	Member	Male	1971.0
5	436	10/13/2014 10:34	10/13/2014 10:47	SEA00337	808.805	2nd Ave & Spring St	Occidental Park / Occidental Ave S & S Washing...	CBD-06	PS-04	Member	Male	1974.0
6	437	10/13/2014 11:35	10/13/2014 11:45	SEA00202	596.715	Occidental Park / Occidental Ave S & S Washing...	King Street Station Plaza / 2nd Ave Extension ...	PS-04	PS-05	Member	Male	1978.0
...	...	...	...	...	...	...	...	...	...	...	...	...
236049	255230	8/31/2016 21:27	8/31/2016 21:30	SEA00056	198.324	Republican St & Westlake Ave N	Dexter Ave N & Aloha St	SLU-04	SLU-02	Member	Male	1981.0
236051	255232	8/31/2016 21:59	8/31/2016 22:04	SEA00499	308.484	E Harrison St & Broadway Ave E	Bellevue Ave & E Pine St	CH-02	CH-12	Member	Male	1989.0
236053	255234	8/31/2016 22:02	8/31/2016 22:17	SEA00448	879.160	Key Arena / 1st Ave N & Harrison St	Pier 66 / Alaskan Way & Bell St	SLU-19	WF-03	Member	Male	1981.0
236055	255236	8/31/2016 22:13	8/31/2016 22:25	SEA00254	674.993	3rd Ave & Broad St	Occidental Park / Occidental Ave S & S Washing...	BT-01	PS-04	Member	Male	1984.0
236056	255237	8/31/2016 22:37	8/31/2016 22:39	SEA00330	144.477	Summit Ave & E Denny Way	Summit Ave E & E Republican St	CH-01	CH-03	Member	Male	1990.0

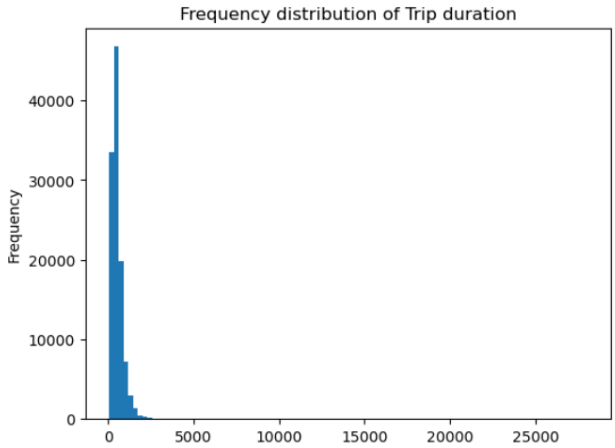
```
In [13]: trip_duration = list(data_male['tripduration'])
         station_from = list(data_male['from_station_name'])
         print ('Mean of trip duration: %f' %statistics.mean(trip_duration))
         print ('Median of trip duration: %f' %statistics.median(trip_duration))
         print ('Mode of the trip duration: %s' %statistics.mode(trip_duration))

         Mean of trip duration: 563.402797
         Median of trip duration: 458.451500
         Mode of the trip duration: 466.174
```

Berdasarkan gambar diatas nilai mean 563.402797, nilai median 458.451500 dan nilai mode 466.174000

b. By looking at the numbers obtained earlier, in your opinion is the distribution symmetric or skewed? If skewed, then is is it positively skewed or negatively skewed?
Answer:
Berdasarkan rata-rata yang lebih besar dibandingkan nilai tengah atau Mean > Median ada kemungkinan distribusinya tidak simetris karena jarak antara mean dan median jauh, dan distribusinya kemungkinan miring. Untuk kemiringan distribusi tidak bisa ditebak, karena positively skewed dan negatively skewed sama-sama bisa terjadi karena mean > median begitupun sebaliknya.

c. Plot a frequency distribution of trip duration for trips availed by gender type male. Does it validate your inference as you did so in the previous question?
Answer:

```
In [15]: data_male['tripduration'].plot.hist(bins=100, title='Frequency distribution of Trip duration')
         plt.show()
```

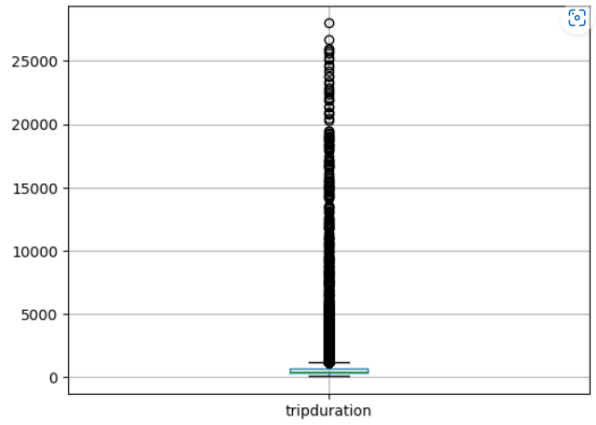


Iya, plot distribusi miring dari kiri ke kanan.

d. Plot a box plot of the trip duration of trips taken by males. Do you think any outliers exist?

Answer:

```
In [16]: box = data_male.boxplot(column=['tripduration'], showfliers = True)
         plt.show()
```



Berdasarkan box plot diaats dapat dilihat kalau terdapat banyak outliers dari data.

e. Apply the formula in Listing 1-19 to determine the percentage of observations for which outliers exists.

Answer:

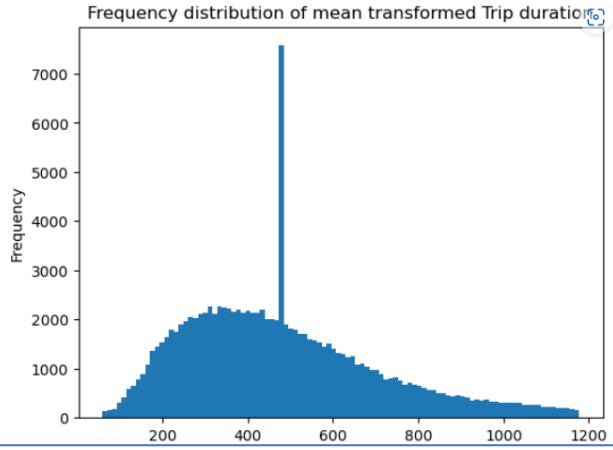
```
In [19]: q75, q25 = np.percentile(trip_duration, [75 ,25])
         iqr = q75 - q25
         print ('Proportion of values as outlier: %f percent'
                %((len(data_male) - len([x for x in trip_duration
                                         if q75+(1.5*iqr)>=x>= q25-(1.5*iqr)]))*100/float(len(data_male))))

         Proportion of values as outlier: 5.030104 percent
```

Berdasarkan perhitungan nilai proporsi pada data sebagai outliers sebesar 5.030104 persen.

f. Perform the treatment of outliers by incorporating one of the methods we discussed earlier for the treatment of outliers (lihat penjelasan di halaman 28). Untuk STB Ganjil gunakan metode "replace outliers with mean", STB Genap gunakan metode "transform the outlier values to upper boundary atau to lower boundary).

Answer:

```
In [20]: mean_trip_duration = np.mean([x for x in trip_duration if q75+(1.5*iqr) >=x>= q25-(1.5*iqr)])
         upper_whisker = q75+(1.5*iqr)

         def transform_tripduration(x):
             if x > upper_whisker:
                 return mean_trip_duration
             return x

         data['tripduration_mean'] = data_male['tripduration'].apply( lambda x: transform_tripduration(x))
         data['tripduration_mean'].plot.hist( bins=100, title='Frequency distribution of mean transformed Trip duration')
         plt.show()
```



Dengan mengganti outliers dengan mean maka boxplot-nya lebih bersih dari outliers dan distribusi data lebih meluas.