Data analysis – Lab 1-3

Sampling and descriptive statistics

Name: Nguyễn Chí Trung Student ID: ITDSIU19024

Questions:

1. Given the events.csv dataset,
    a. do random simple sampling to obtain 100 data samples, sample size is 2000 viewed items. Each sample contains itemids and the corresponding number of views. A random function should be used to get random samples.
    b. Select 10 samples, draw the bar charts of number of item views for each sample.
    c. Select 10 samples, draw the boxplot of number of item views for each sample.

Notice: please point out what samples are selected in your answers.

2. For each sample in the selected 10 samples,
    a. What proportion of the items had more than the mean number of views?
    b. For what proportion of the items was the number of views more than one standard deviation greater than the mean?
    c. For what proportion of the items was the number of views within one standard deviation of the mean?

3. Present sampling distribution of $\bar{x}$. $\bar{x}$ is sample mean number of views from the above 10 data samples.
    a. Draw a relative frequency histogram of $\bar{x}$ values.
    b. Evaluate the sampling distribution.

Question 1:

First, imported csv file in python by using read_csv(...)

```
dt = pd.read_csv('lab 1-2-
3/events.csv',sep=',',header=0,usecols=[2,3],dtype={2:str, 3:np.int32})
```

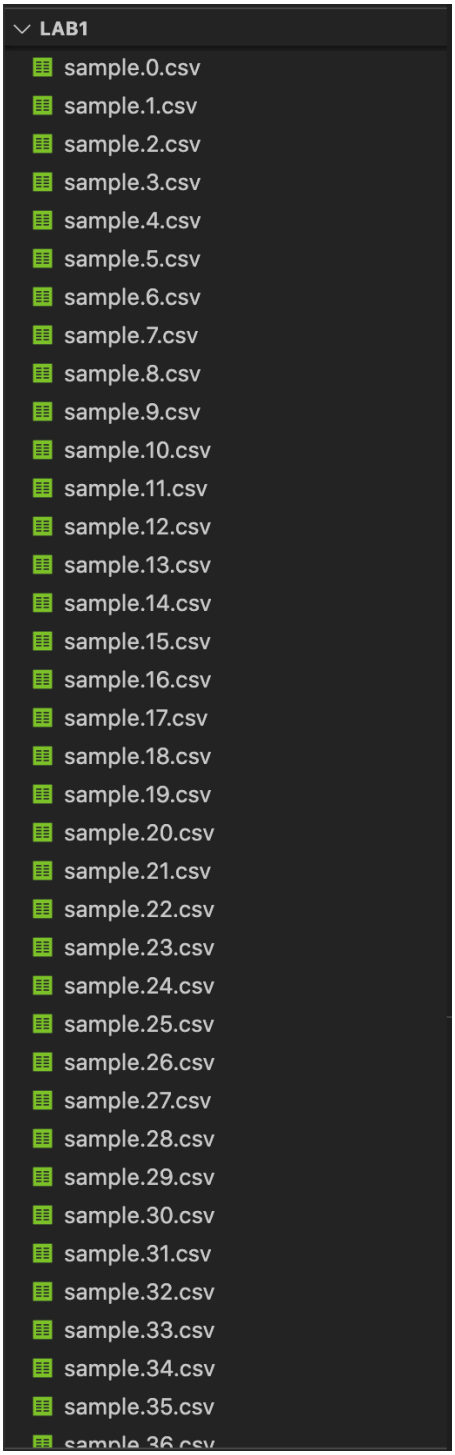
Then, make a quick adjust to shorten our data because we just need two columns "ItemID" and "View".

```
dt.columns = ['Type','ItemId']
dt = dt[dt.Type == 'view']
dt= dt.groupby('ItemId').count().reset_index()
dt.columns = ['ItemId','View']
```

Next step is to use loop to split the original csv file into 100 samples and each sample has 2000 viewed items

```
for n in range (100):
    dt.sample(2000).to_csv('lab 1-2-3/'+'sample.'+str(n)+'.csv',sep = ',',index =
False)
```

This is the result of this code



For question b) and c) we need to import 10 samples that has been created, use the same method as first step

```
dt1 = pd.read_csv('lab 1-2-3/sample.1.csv')
dt2 = pd.read_csv('lab 1-2-3/sample.2.csv')
dt3 = pd.read_csv('lab 1-2-3/sample.3.csv')
dt4 = pd.read_csv('lab 1-2-3/sample.4.csv')
dt5 = pd.read_csv('lab 1-2-3/sample.5.csv')
dt6 = pd.read_csv('lab 1-2-3/sample.6.csv')
dt7 = pd.read_csv('lab 1-2-3/sample.7.csv')
dt8 = pd.read_csv('lab 1-2-3/sample.8.csv')
```

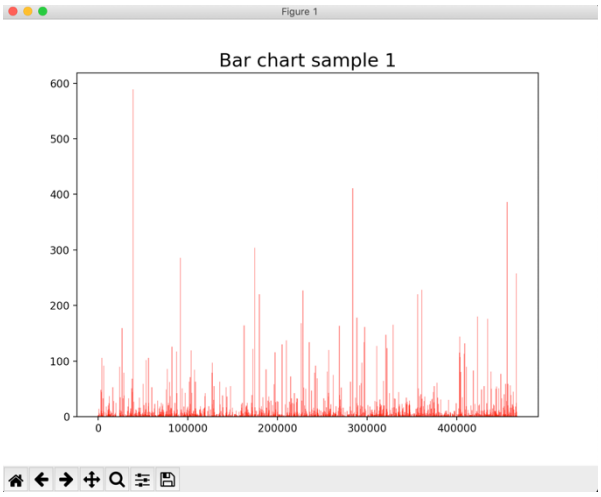
```
dt9 = pd.read_csv('lab 1-2-3/sample.9.csv')
dt10 = pd.read_csv('lab 1-2-3/sample.10.csv')
```

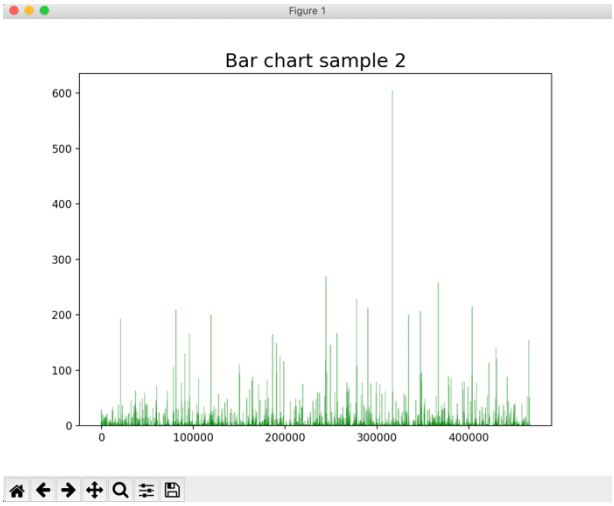
Now just take turn to draw bar chart

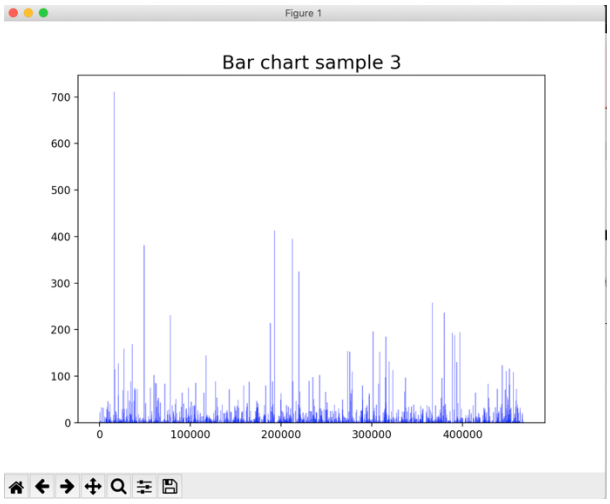
```
plt.figure(figsize=(8,6))
plt.bar(dt1['ItemId'],dt1['View'],color = 'red',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 1',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt2['ItemId'],dt2['View'],color = 'green',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 2',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt3['ItemId'],dt3['View'],color = 'blue',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 3',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt4['ItemId'],dt4['View'],color = 'yellow',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 4',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt5['ItemId'],dt5['View'],color = 'black',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 5',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt6['ItemId'],dt6['View'],color = 'c',width=1000,align='center',alpha= 0.5)
plt.title('Bar chart sample 6',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt7['ItemId'],dt7['View'],color = 'pink',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 7',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt8['ItemId'],dt8['View'],color = 'red',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 8',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
```

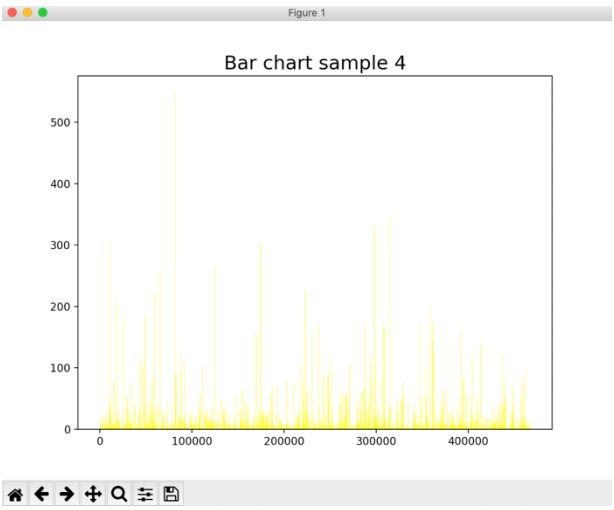
```
plt.bar(dt9['ItemId'],dt9['View'],color = 'green',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 9',fontsize = 18)
plt.grid(False)
plt.show()
plt.figure(figsize=(8,6))
plt.bar(dt10['ItemId'],dt10['View'],color = 'blue',width=1000,align='center',alpha=
0.5)
plt.title('Bar chart sample 10',fontsize = 18)
plt.grid(False)
plt.show()
```

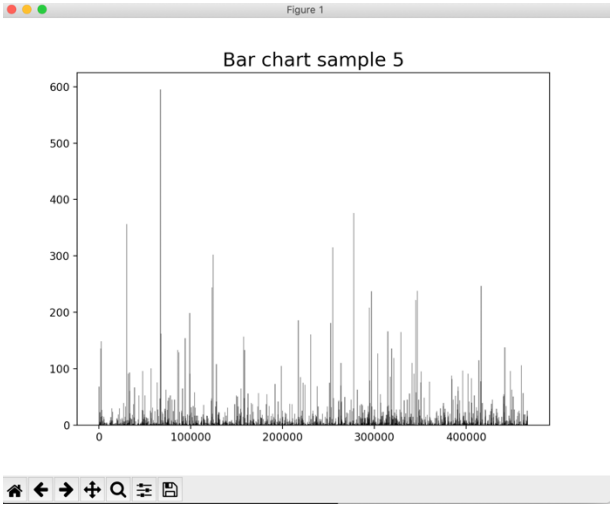
Results:

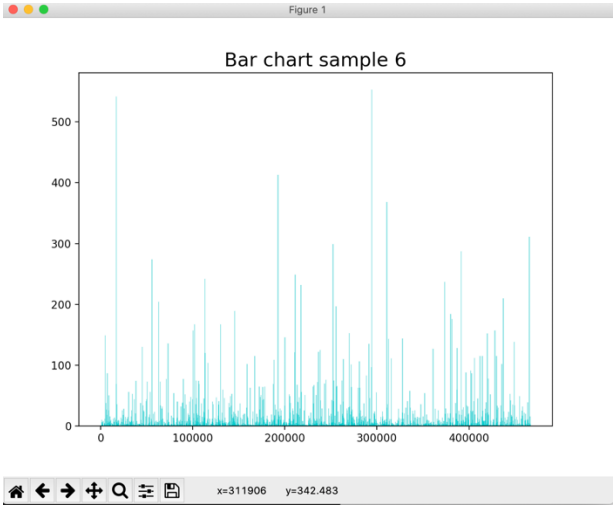


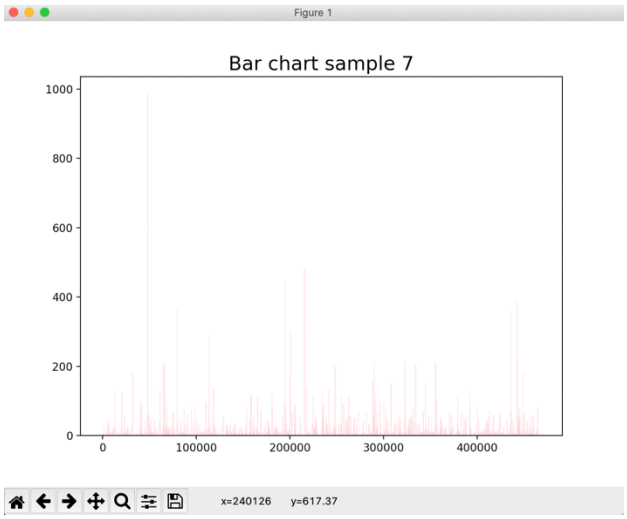


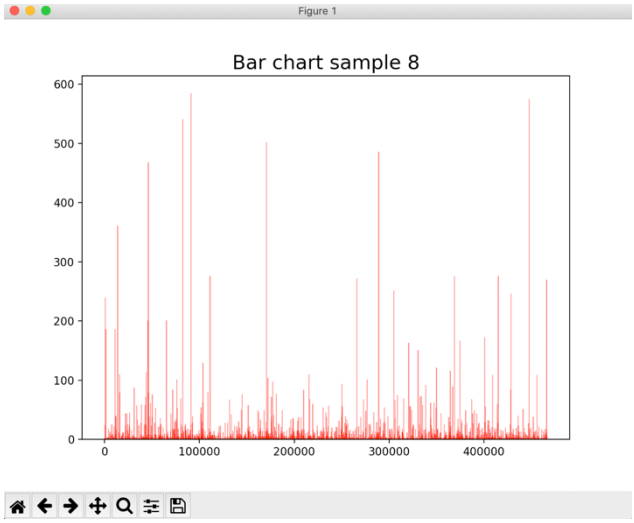


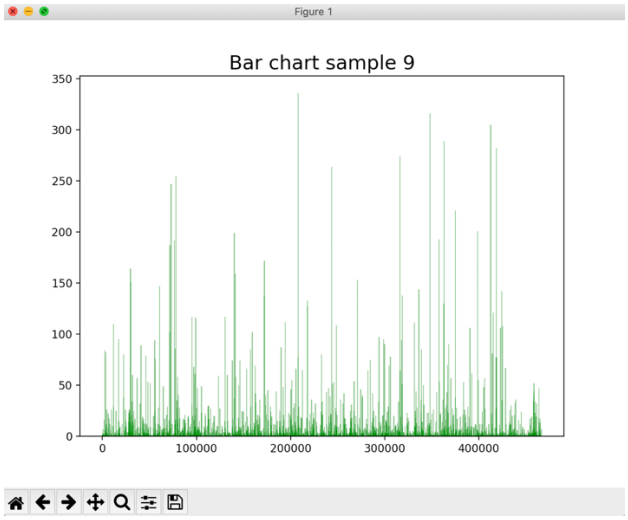


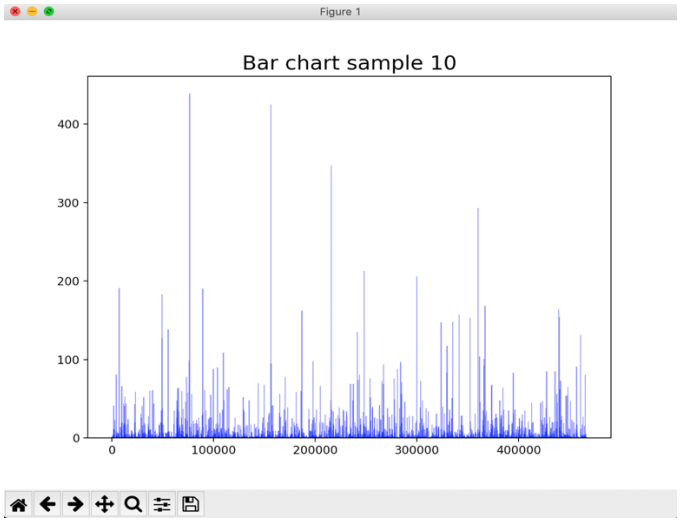








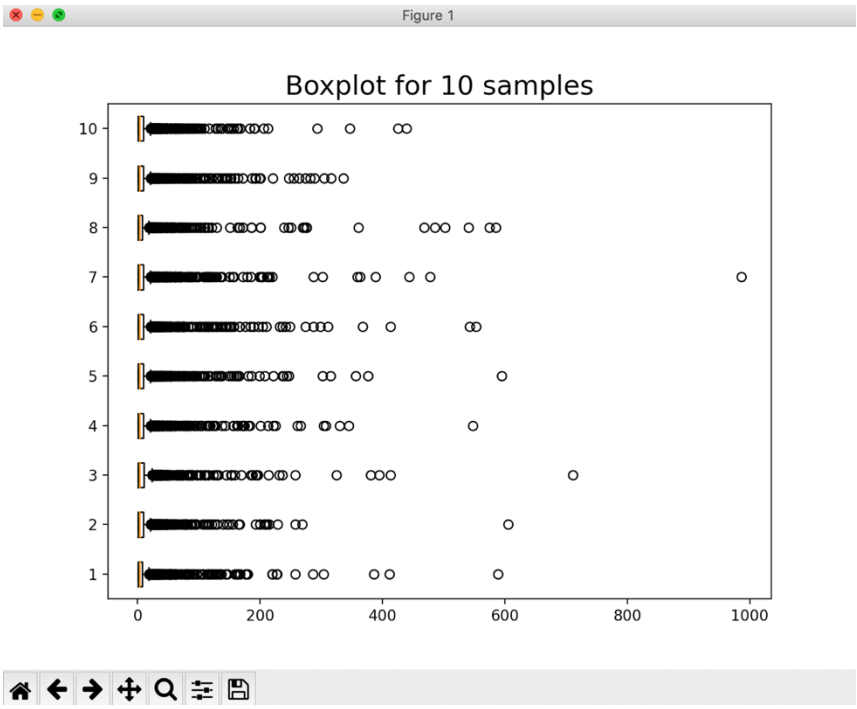




Final step of question 1 is to draw boxplot of 10 samples, we put 10 samples into a place and draw by using this code

```
plt.boxplot([dt1['View'],dt2['View'],dt3['View'],dt4['View'],dt5['View'],dt6['View'
],dt7['View'],dt8['View'],dt9['View'],dt10['View']],notch=False,vert=False)
plt.title('Boxplot for 10 samples',fontsize=18)
plt.grid(False)
plt.show()
```

Result:



Question 2:
I use mean() function to calculate mean, fill out what "View" is satisfied with that condition and print the answer out

```
item1 = dt1[dt1['View'] > dt1['View'].mean()]
itemcount1 = sum(1 for row in item1['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 1
=',round((itemcount1/2000)*100,2),'%')
item2 = dt2[dt2['View'] > dt2['View'].mean()]
itemcount2 = sum(1 for row in item2['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 2
=',round((itemcount2/2000)*100,2),'%')
item3 = dt3[dt3['View'] > dt3['View'].mean()]
itemcount3 = sum(1 for row in item3['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 3
=',round((itemcount3/2000)*100,2),'%')
item4 = dt4[dt4['View'] > dt4['View'].mean()]
itemcount4 = sum(1 for row in item4['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 4
=',round((itemcount4/2000)*100,2),'%')
item5 = dt5[dt5['View'] > dt5['View'].mean()]
itemcount5 = sum(1 for row in item5['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 5
=',round((itemcount5/2000)*100,2),'%')
item6 = dt6[dt6['View'] > dt6['View'].mean()]
itemcount6 = sum(1 for row in item6['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 6
=',round((itemcount6/2000)*100,2),'%')
item7 = dt7[dt7['View'] > dt7['View'].mean()]
itemcount7 = sum(1 for row in item7['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 7
=',round((itemcount7/2000)*100,2),'%')
item8 = dt8[dt8['View'] > dt8['View'].mean()]
itemcount8 = sum(1 for row in item8['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 8
=',round((itemcount8/2000)*100,2),'%')
item9 = dt9[dt9['View'] > dt9['View'].mean()]
itemcount9 = sum(1 for row in item9['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 9
=',round((itemcount9/2000)*100,2),'%')
item10 = dt10[dt10['View'] > dt10['View'].mean()]
itemcount10 = sum(1 for row in item10['ItemId'])
print('Proportion of the items had more than the mean number of views of sample 10
=',round((itemcount10/2000)*100,2),'%')
```

Question b is quiet the same, that is find which is greater than sum of mean and standard deviation, we use std() to calculate the standard deviation of samples, next two steps is same with question a)

```
item1b = dt1[dt1['View'] > (dt1['View'].mean() + dt1['View'].std())]
itemcount1b = sum(1 for row in item1b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 1 =',round((itemcount1b/2000)*100,2),'%')
```

```
item2b = dt2[dt2['View'] > (dt2['View'].mean() + dt2['View'].std())]
itemcount2b = sum(1 for row in item2b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 2 =',round((itemcount2b/2000)*100,2),'%')
item3b = dt3[dt3['View'] > (dt3['View'].mean() + dt3['View'].std())]
itemcount3b = sum(1 for row in item3b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 3 =',round((itemcount3b/2000)*100,2),'%')
item4b = dt4[dt4['View'] > (dt4['View'].mean() + dt4['View'].std())]
itemcount4b = sum(1 for row in item4b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 4 =',round((itemcount4b/2000)*100,2),'%')
item5b = dt5[dt5['View'] > (dt5['View'].mean() + dt5['View'].std())]
itemcount5b = sum(1 for row in item5b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 5 =',round((itemcount5b/2000)*100,2),'%')
item6b = dt6[dt6['View'] > (dt6['View'].mean() + dt6['View'].std())]
itemcount6b = sum(1 for row in item6b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 6 =',round((itemcount6b/2000)*100,2),'%')
item7b = dt7[dt7['View'] > (dt7['View'].mean() + dt7['View'].std())]
itemcount7b = sum(1 for row in item7b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 7 =',round((itemcount7b/2000)*100,2),'%')
item8b = dt8[dt8['View'] > (dt8['View'].mean() + dt8['View'].std())]
itemcount8b = sum(1 for row in item8b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 8 =',round((itemcount8b/2000)*100,2),'%')
item9b = dt9[dt9['View'] > (dt9['View'].mean() + dt9['View'].std())]
itemcount9b = sum(1 for row in item9b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 9 =',round((itemcount9b/2000)*100,2),'%')
item10b = dt10[dt10['View'] > (dt10['View'].mean() + dt10['View'].std())]
itemcount10b = sum(1 for row in item10b['ItemId'])
print('Proportion of the items that the number of views more than one standard
deviation greater than the mean of sample 10
=',round((itemcount10b/2000)*100,2),'%')
```

For question c, the condition is to find the ID that fall between mean - standard deviation and mean + standard deviation, we use the same method

```
item1c = dt1[(dt1['View'] >= (dt1['View'].mean() - dt1['View'].std())) &
(dt1['View'] < (dt1['View'].mean() + dt1['View'].std()))]
itemcount1c = sum(1 for row in item1c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 1 =',round((itemcount1c/2000)*100,2),'%')
item2c = dt2[(dt2['View'] >= (dt2['View'].mean() - dt2['View'].std())) &
(dt2['View'] < (dt2['View'].mean() + dt2['View'].std()))]
itemcount2c = sum(1 for row in item2c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 2 =',round((itemcount2c/2000)*100,2),'%')
item3c = dt3[(dt3['View'] >= (dt3['View'].mean() - dt3['View'].std())) &
(dt3['View'] < (dt3['View'].mean() + dt3['View'].std()))]
```

```
itemcount3c = sum(1 for row in item3c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 3 =',round((itemcount3c/2000)*100,2),'%')
item4c = dt4[(dt4['View'] >= (dt4['View'].mean() - dt4['View'].std())) &
(dt4['View'] < (dt4['View'].mean() + dt4['View'].std()))]
itemcount4c = sum(1 for row in item4c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 4 =',round((itemcount4c/2000)*100,2),'%')
item5c = dt5[(dt5['View'] >= (dt5['View'].mean() - dt5['View'].std())) &
(dt5['View'] < (dt5['View'].mean() + dt5['View'].std()))]
itemcount5c = sum(1 for row in item5c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 5 =',round((itemcount5c/2000)*100,2),'%')
item6c = dt6[(dt6['View'] >= (dt6['View'].mean() - dt6['View'].std())) &
(dt6['View'] < (dt6['View'].mean() + dt6['View'].std()))]
itemcount6c = sum(1 for row in item6c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 6 =',round((itemcount6c/2000)*100,2),'%')
item7c = dt7[(dt7['View'] >= (dt7['View'].mean() - dt7['View'].std())) &
(dt7['View'] < (dt7['View'].mean() + dt7['View'].std()))]
itemcount7c = sum(1 for row in item7c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 7 =',round((itemcount7c/2000)*100,2),'%')
item8c = dt8[(dt8['View'] >= (dt8['View'].mean() - dt8['View'].std())) &
(dt8['View'] < (dt8['View'].mean() + dt8['View'].std()))]
itemcount8c = sum(1 for row in item8c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 8 =',round((itemcount8c/2000)*100,2),'%')
item9c = dt9[(dt9['View'] >= (dt9['View'].mean() - dt9['View'].std())) &
(dt9['View'] < (dt9['View'].mean() + dt9['View'].std()))]
itemcount9c = sum(1 for row in item9c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 9 =',round((itemcount9c/2000)*100,2),'%')
item10c = dt10[(dt10['View'] >= (dt10['View'].mean() - dt10['View'].std())) &
(dt10['View'] < (dt10['View'].mean() + dt10['View'].std()))]
itemcount10c = sum(1 for row in item10c['ItemId'])
print('Proportion of the items that the number of views within one standard
deviation of the mean of sample 10 =',round((itemcount10c/2000)*100,2),'%')
```

*Use round to make sure the result has two decimals, here the results

```
trungnguyen@Trungs-MBP lab1 % /usr/local/bin/python3 "/Users/trungnguyen/Downloads/Sem1-year2/Data Analys/lab1/ques2.py"
Proportion of the items had more than the mean number of views of sample 1 = 19.9 %
Proportion of the items had more than the mean number of views of sample 2 = 21.2 %
Proportion of the items had more than the mean number of views of sample 3 = 18.85 %
Proportion of the items had more than the mean number of views of sample 4 = 17.45 %
Proportion of the items had more than the mean number of views of sample 5 = 19.6 %
Proportion of the items had more than the mean number of views of sample 6 = 19.7 %
Proportion of the items had more than the mean number of views of sample 7 = 21.0 %
Proportion of the items had more than the mean number of views of sample 8 = 19.75 %
Proportion of the items had more than the mean number of views of sample 9 = 20.3 %
Proportion of the items had more than the mean number of views of sample 10 = 19.5 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 1 = 5.0 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 2 = 5.8 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 3 = 4.35 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 4 = 2.8 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 5 = 4.75 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 6 = 5.1 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 7 = 6.5 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 8 = 4.6 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 9 = 5.65 %
Proportion of the items that the number of views more than one standard deviation greater than the mean of sample 10 = 5.15 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 1 = 95.0 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 2 = 94.2 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 3 = 95.65 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 4 = 97.2 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 5 = 95.25 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 6 = 94.9 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 7 = 93.5 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 8 = 95.4 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 9 = 94.35 %
Proportion of the items that the number of views within one standard deviation of the mean of sample 10 = 94.85 %
trungnguyen@Trungs-MBP lab1 %
```

Question 3:

First calculate mean of 10 samples

```
x1 = dt1['View'].mean()
x2 = dt2['View'].mean()
x3 = dt3['View'].mean()
x4 = dt4['View'].mean()
x5 = dt5['View'].mean()
x6 = dt6['View'].mean()
x7 = dt7['View'].mean()
x8 = dt8['View'].mean()
x9 = dt9['View'].mean()
x10 = dt10['View'].mean()
```

Then add it into a list

```
ls = {x1,x2,x3,x4,x5,x6,x7,x8,x9,x10}
```

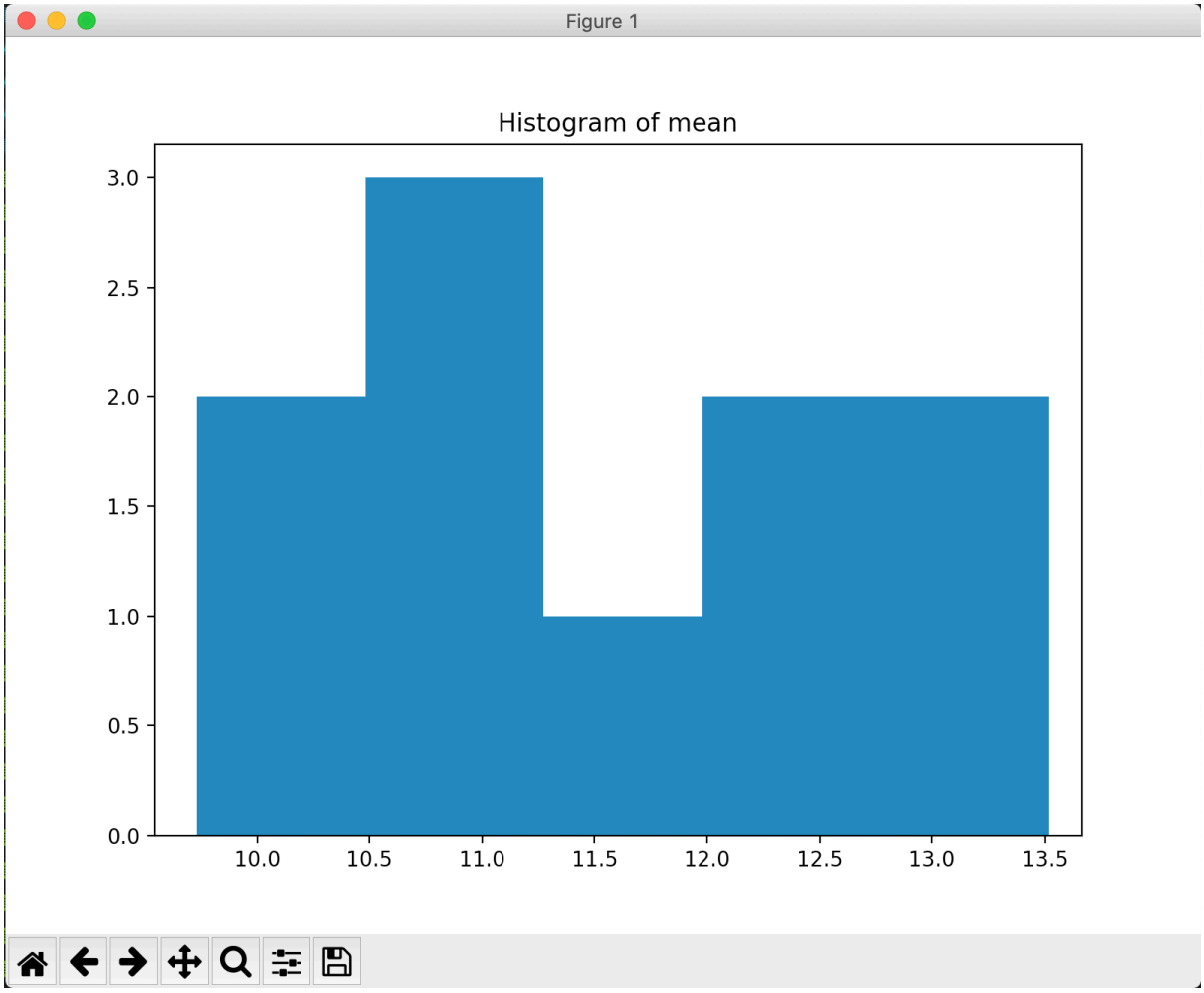
Next set it as a data frame

```
q3 = pd.DataFrame(ls, columns= ['mean'])
```

Now draw histogram

```
plt.figure(figsize=(8,6))
plt.hist(q3['mean'], width = 0.79, bins= 'auto')
plt.title('Histogram of mean')
plt.show()
```

Result:



From this histogram, we can easily conclude that this is not a normal distribution

Here is the range, variance and standard deviation of this 10 samples to make sure we have all information about this data

Range:

```
q3b = [x1,x2,x3,x4,x5,x6,x7,x8,x9,x10]
rg = (min(q3b),max(q3b))
print(rg)
```

Variance:

```
x1c = dt1['View'].var()
x2c = dt2['View'].var()
x3c = dt3['View'].var()
x4c = dt4['View'].var()
x5c = dt5['View'].var()
x6c = dt6['View'].var()
x7c = dt7['View'].var()
x8c = dt8['View'].var()
x9c = dt9['View'].var()
x10c = dt10['View'].var()
```

Standard deviation:

```
x1b = dt1['View'].std()
```

```
x2b = dt2['View'].std()
x3b = dt3['View'].std()
x4b = dt4['View'].std()
x5b = dt5['View'].std()
x6b = dt6['View'].std()
x7b = dt7['View'].std()
x8b = dt8['View'].std()
x9b = dt9['View'].std()
x10b = dt10['View'].std()
```