# Lab Mid Exam

**Course:** Data Science Section 8A
**Instructor:** Maemoona Kayani
**Date:** 11th April 2022

## Empirical rule

The **Empirical Rule** or the **68-95-99.7 Rule** states that for if a frequency distribution of a set of sample data is **normally distributed** then

- Approximately 68% of the data falls within 1 standard deviation of the mean i.e. within $\bar{x} \pm s$.
- Approximately 95% of the data falls within 2 standard deviations of the mean i.e. within $\bar{x} \pm 2s$.
- Approximately 99.7% of the data falls within 3 standard deviations of the mean i.e. within $\bar{x} \pm 3s$.

## Dataset "faithful"

To test the data set "the length of time of eruptions of the Old Faithful Geyser in Yellowstone" that it satisfies the Empirical Rule. In python,

load the data,

look at the eruption times now in variable eruptions.

Plot histogram of eruptions.

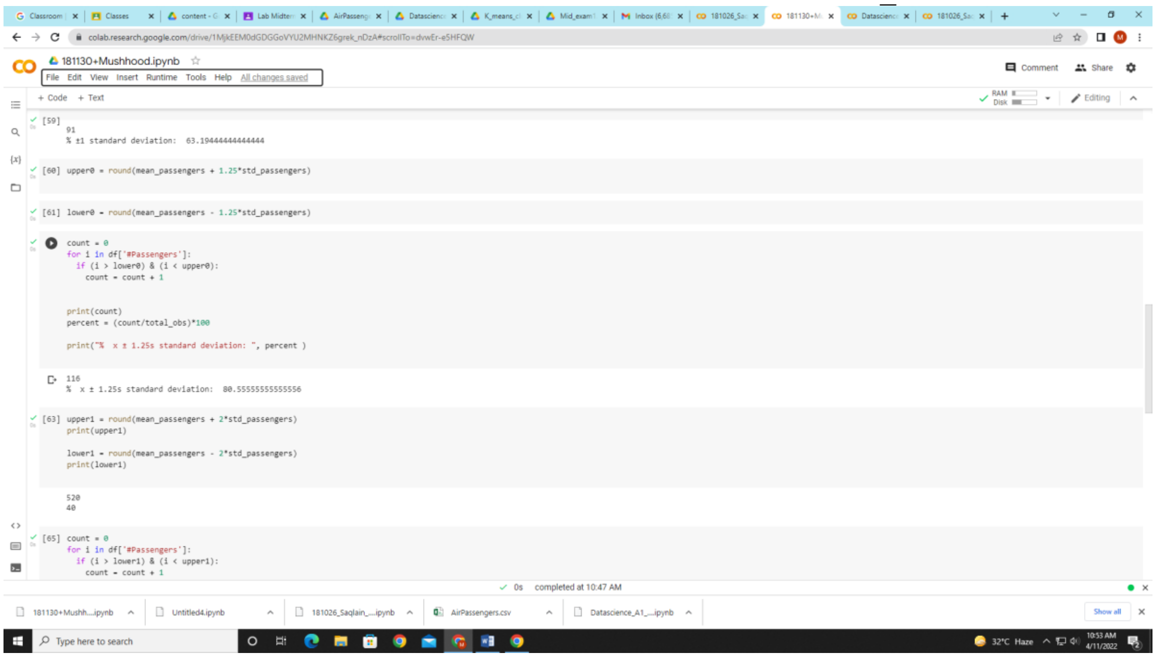
To see what percentage of the data is within one, two and three standard deviations, compute the mean and the standard deviation and save the numbers.

Find the number of observations within ±1 standard deviation of the mean and calculate percentage by dividing by the number of observations.

The data set eruptions DOES NOT satisfy the Empirical rule, which in turn means that the data set is NOT normally distributed. This can also be seen from the histogram.

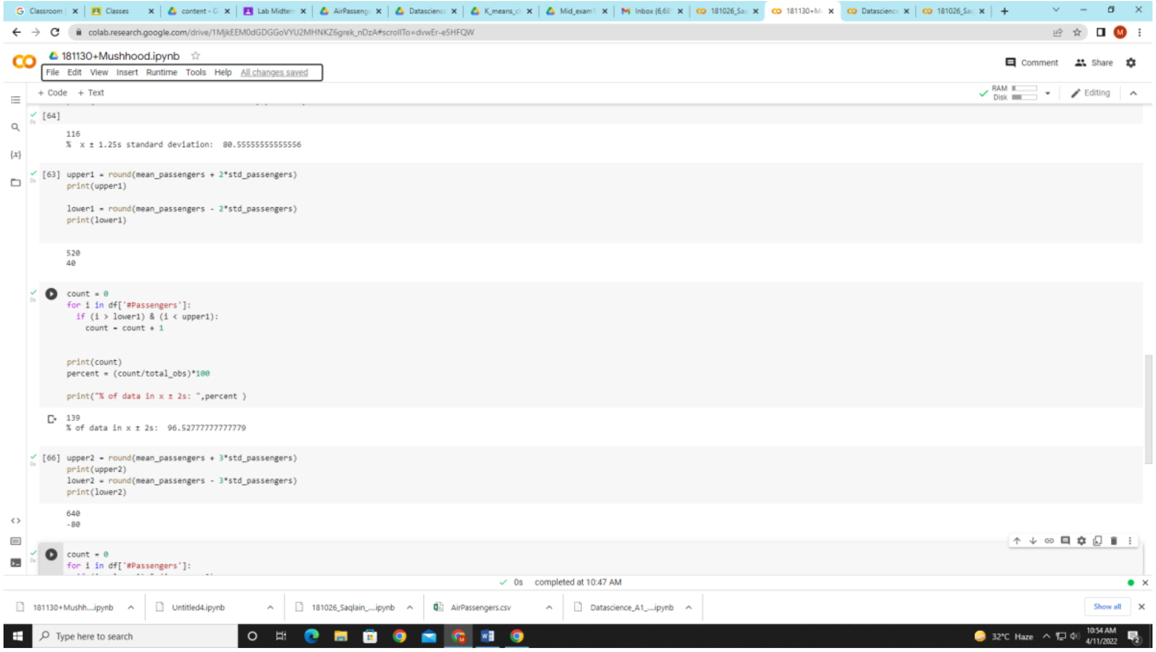
# Questions

1. For the above data set, what percentage of data falls in the range of $x \pm \overline{1}.25s$ ?

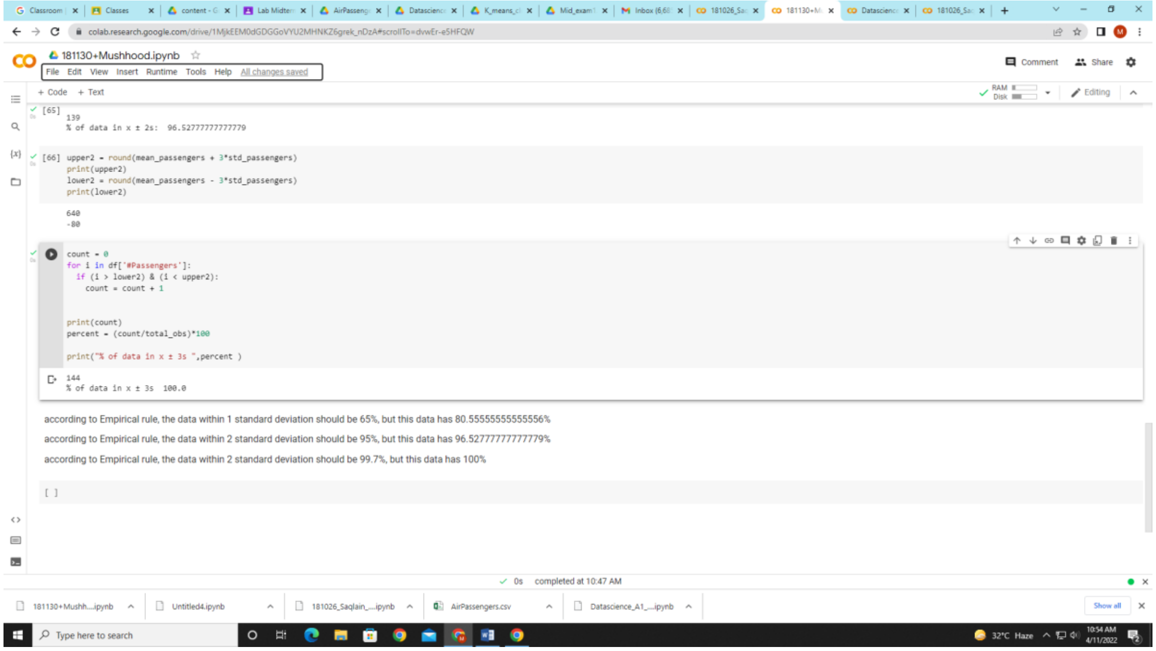


```
[59]
     91
     % ±1 standard deviation:  63.19444444444444

[60] upper0 = round(mean_passengers + 1.25*std_passengers)

[61] lower0 = round(mean_passengers - 1.25*std_passengers)

     count = 0
     for i in df['#Passengers']:
       if (i > lower0) & (i < upper0):
         count = count + 1


     print(count)
     percent = (count/total_obs)*100

     print("%  x ± 1.25s standard deviation: ", percent )

     116
     %  x ± 1.25s standard deviation:  80.55555555555556

[63] upper1 = round(mean_passengers + 2*std_passengers)
     print(upper1)

     lower1 = round(mean_passengers - 2*std_passengers)
     print(lower1)

     520
     40

[65] count = 0
     for i in df['#Passengers']:
       if (i > lower1) & (i < upper1):
         count = count + 1
```

2. For the above data set, what percentage of data falls in the range of $x \pm 2s$, $x \pm \overline{3}s$ ?



```
[64]
     116
     %  x ± 1.25s standard deviation:  80.55555555555556

[63] upper1 = round(mean_passengers + 2*std_passengers)
     print(upper1)

     lower1 = round(mean_passengers - 2*std_passengers)
     print(lower1)

     520
     40

     count = 0
     for i in df['#Passengers']:
       if (i > lower1) & (i < upper1):
         count = count + 1


     print(count)
     percent = (count/total_obs)*100

     print("% of data in x ± 2s: ",percent )

     139
     % of data in x ± 2s:  96.52777777777779

[66] upper2 = round(mean_passengers + 3*std_passengers)
     print(upper2)
     lower2 = round(mean_passengers - 3*std_passengers)
     print(lower2)

     640
     -80

     count = 0
     for i in df['#Passengers']:
```



```
[65]
     139
     % of data in x ± 2s:  96.52777777777779
```

```
[66] upper2 = round(mean_passengers + 3*std_passengers)
     print(upper2)
     lower2 = round(mean_passengers - 3*std_passengers)
     print(lower2)

     640
     -80
```

```
count = 0
for i in df['#Passengers']:
  if (i > lower2) & (i < upper2):
    count = count + 1


print(count)
percent = (count/total_obs)*100

print("% of data in x ± 3s ",percent )
```

```
144
% of data in x ± 3s  100.0
```

according to Empirical rule, the data within 1 standard deviation should be 65%, but this data has 80.55555555555556%

according to Empirical rule, the data within 2 standard deviation should be 95%, but this data has 96.52777777777779%

according to Empirical rule, the data within 2 standard deviation should be 99.7%, but this data has 100%

3. How does it compare with the Empirical Rule ?

   according to Empirical rule, the data within 1 standard deviation should be 65%, but this data has 80.55555555555556%

   according to Empirical rule, the data within 2 standard deviation should be 95%, but this data has 96.52777777777779%
   according to Empirical rule, the data within 2 standard deviation should be 99.7%, but this data has 100%

4. Find percentages of data which falls in the range $\bar{x} \pm s$, $\bar{x} \pm 1.25s$, $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$ for the data set called "AirPassengers" and "LakeHuron". Compare your results with Empirical Rule.

   Already done.

---

## To hand in:

1. 3 Histograms for the data sets Faithful, AirPassengers or LakeHuron (according to your name), mean and standard deviation typed out.

2. Type out the answers to Questions for the data sets Faithful, AirPassengers or LakeHuron (according to your name) in a MS word.

---

| Roll no | Name | Dataset |
|---|---|---|
| 170246 | Shanila Abid | Faithful |
| 170744 | Areej Sajjad | Airpassenger |
| 180954 | Muhammad Muneeb | LakeHuron |
| 180962 | Areej Zafar | Faithful |
| 180974 | Asra Imtiaz | Airpassenger |
| 180978 | Maryam Munir | LakeHuron |
| 180986 | Mirza Hammad Baig | Faithful |
| 180990 | Muhammad Farrukh Shahid | Airpassenger |
| 180994 | Ahmad Yar Khan | LakeHuron |
| 180998 | Baseerat Lazawal | Faithful |
| 181002 | Haseeb Tariq | Airpassenger |
| 181026 | Saqlain Umer | LakeHuron |
| 181030 | Qurat-ul-Ain | Faithful |
| 181034 | Noman Aziz khan | Airpassenger |
| 181074 | Muhammad Mubashar Saleem | LakeHuron |
| 181094 | Ayesha Jabeen Malik | Faithful |
| 181098 | Hassam Shafique Cheema | Airpassenger |
| 181114 | Hamza Kayani | LakeHuron |
| 181126 | Hammad Rustum | Faithful |
| 181130 | Mir Mashood Afsar | Airpassenger |
| 181138 | Muhammad Haseeb Rafique | LakeHuron |
| 183244 | Maimoona Nawaz Khan | Faithful |
| 180991 | Zainab Noor | Airpassenger |
| 181019 | Muhammad Shahzaib Khan | LakeHuron |
| 181135 | Mubashir Ahmad | Faithful |
| 181913 | Aksam Javed | Airpassenger |