EAST WEST UNIVERSITY

Course Title: Statistics and Probability

Course Code: STA102

Topic: Cycling Habit of University Students

Submitted by:
Iffat Tasnim: 2020-2-60-004
Nadia Sultana: 2020-2-60-024
Farhana Moni: 2020-2-60-036
Tasfiha Ahmed: 2020-1-55-009
Ishtiaque Hasan: 2020-1-80-087

Submitted to:
Nazneen Sultana
Senior Lecturer
Department of Mathematical and Physical Sciences

Topic: Cycling habit of university student.

Q1) Do you use cycle in your daily life?

Q2) How frequent do you ride cycle per week?

Q3) Have you ever seen any cycle lane in Bangladesh?

Q4) What is your current body weight?

Q5) Do you think Govt. is supportive about cycling?

Make a questionnaire using google form. You can add option for answer or it can be open ended.

1) Find out the mean, median, mode, standard deviation for Q2 and Q4. Also calculate skewness and kurtosis and interpret the result. Show the data in a suitable graph.

2) Is there any association between Q2 and Q4? Find the result and interpret it. Show the association in a graph.

3) Also find the suitable graphs for Q1, Q3 and Q5.

4) What is your conclusion from the study?

Data Collection Sheet:

Observations	Q1. Do you use cycle in your daily life?	Q2. How frequent do you ride cycle per week? (days)	Q3. Have you ever seen any cycle lane in Bangladesh?	Q4. What is your current body weight?	Q5. Do you think Govt. is supportive about cycling?
1	Yes	1	No	40	No
2	No	2	Yes	95	Yes
3	Yes	3	No	50	Yes
4	Yes	7	Yes	80	Yes
5	No	3	No	55	No
6	No	1	Yes	35	Yes
7	No	1	Yes	60	No
8	Yes	3	No	65	No
9	No	5	Yes	55	No
10	No	6	Yes	55	No
11	Yes	1	No	40	Yes
12	No	4	Yes	60	Yes
13	Yes	4	Yes	60	No
14	Yes	1	No	60	No
15	Yes	1	No	40	No
16	No	1	No	70	No
17	Yes	7	Yes	70	No
18	Yes	1	Yes	50	No
19	Yes	1	No	55	No
20	No	1	No	55	No
21	Yes	4	No	65	No
22	No	3	No	60	No
23	Yes	5	No	65	No
24	Yes	7	No	85	No
25	Yes	5	No	50	No
26	Yes	7	No	75	No
27	No	1	No	40	No
28	Yes	6	No	65	No
29	Yes	5	No	60	No
30	Yes	1	No	50	No
Total	-	98	-	1765	-

Ans-to-the-Q-No-1

Data(Q2): 1, 2 ,3 ,7, 3, 1, 1, 3, 5, 6, 1, 4, 4, 1, 1, 1, 7, 1, 1, 1, 4, 3, 5, 7, 5, 7, 1, 6, 5, 1

Mean:

Formula:

Mean = $\bar{X} = \frac{\Sigma x}{n} = \frac{98}{30} = 3.2667$

Here,

Total number of observations(n) = 30

X is a particular value

$\sum$ indicates the operation of adding

Sum of data(Q2) = 98

Median:

Data(Q2) in ascending order,

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7.

The number of observations of data set, n = 30 (Even)

Formula:

Median, $M_e = \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1})$

$= \frac{1}{2}(X_{\frac{30}{2}} + X_{\frac{30}{2}+1})$

$= \frac{1}{2}(X_{15} + X_{16})$

$= \frac{1}{2}(3 + 3)$

$= 3$

Mode:

Mode is the most appearing set in a data and from data(Q2), mode is 1.

M_o = 1 [Because 1 is the most repeated variable (12 times) in data(Q2)]

Standard Deviation:

Standard deviation of the data can be obtained from the variance of the data. Standard deviation is the square root of the variance. Whereas, variance is obtained by calculating the mean of the squared difference between each data point and the mean of the data. In our circumstance, we have collected data from some of the university students. So, collected data is a sample in that

case and that's why we must use Sample Standard Deviation formula to find Standard Deviation.

Formula:

Sample Standard Deviation, $S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Here,

Number of observations, n= 30

$\bar{x}$ is the mean of data(Q2) = 3.2667

Now,

X	$(\mathbf{x}-\bar{\mathbf{x}})^{\mathbf{2}}$
1	5.13792889
2	1.60452889
3	0.07112889
7	13.93752889
3	0.07112889
1	5.13792889
1	5.13792889
3	0.07112889
5	3.00432889
6	7.47092889
1	5.13792889
4	0.53772889
4	0.53772889
1	5.13792889
1	5.13792889
1	5.13792889
7	13.93752889
1	5.13792889
1	5.13792889
1	5.13792889
4	0.53772889
3	0.07112889
5	3.00432889
7	13.93752889

5	3.00432889
7	13.93752889
1	5.13792889
6	7.47092889
5	3.00432889
1	5.13792889
Total	$\sum(x-\bar{x})^2$= 147.866667

Sample Standard Deviation,

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{147.866667}{30-1}}=\sqrt{5.09885057}=2.2581$$

Skewness:

Skewness, $sk=\frac{3(Mean-Median)}{Standard\ Deviation}$

$$=\frac{3(3.2667-3)}{2.2581}$$

$$=0.354$$

Interpretation: If the value of sk > 0 means the distribution is positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left.

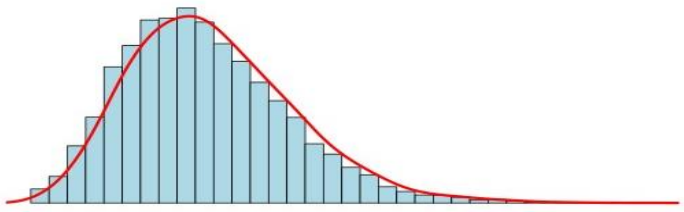


Figure: Positively Skewed distribution

Kurtosis:

The kurtosis value of data(Q2) is -1.30592 (From excel sheet)

Interpretation: The value of the kurtosis is less than 3. So, it is a platykurtic kurtosis. The tails of platykurtic kurtosis are very thin compared to the normal distribution.

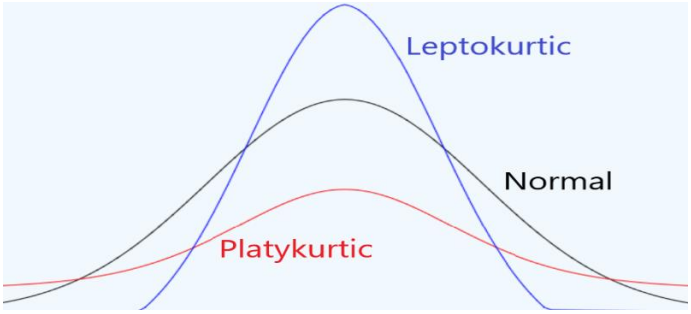


Figure: Kurtosis

Suitable graph for data(Q2):

Number of classes:

$2^k > 30$

⇨ $2^5 > 30$

⇨ k ⩾ 30 [So, there should be at least 5 classes]

Here,

k = Number of classes

n = Total observation

Class interval:

$i \geqslant \frac{H-L}{k}$

⇨ $i \geqslant \frac{7-1}{5}$

⇨ i ⩾ 1.2 [We will take exact 1.2]

Here,

H= Highest observation of the dataset

L= Lowest observation of the dataset

i = Class interval

Class	Frequency
1-2.2	13
2.2-3.4	4
3.4-4.6	3
4.6-5.8	4
5.8-7	2
7-8.2	4
	Total = 30

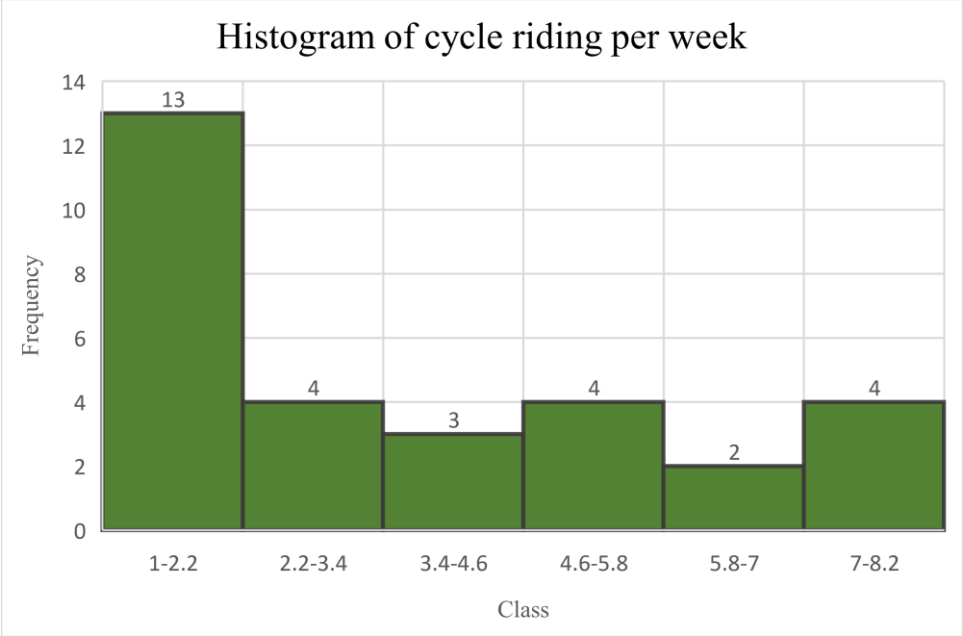


Figure: Histogram

Data(Q4): 40, 95, 50, 80, 55, 35, 60, 65, 55, 55, 40, 60, 60, 60, 40, 70, 70, 50, 55, 55, 65, 60, 65, 85, 50, 75, 40, 65, 60, 50

Mean:

Formula:

Mean, $\bar{x} = \frac{\Sigma x}{n} = \frac{1765}{30} = 58.8333$

Here,

n is the total number of observations = 30

From the data, total sum of data set = 1765

Median:

Data(Q4) in ascending order,

35, 40, 40, 40, 40, 50, 50, 50, 50, 55, 55, 55, 55, 55, 60, 60, 60, 60, 60, 60, 65, 65, 65, 65, 70, 70, 75, 80, 85, 95.

The number of observations of data set, n = 30 (Even)

Formula:

Median, $M_e = \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1})$

$= \frac{1}{2}(X_{\frac{30}{2}} + X_{\frac{30}{2}+1})$

$= \frac{1}{2}(X_{15} + X_{16})$

$= \frac{1}{2}(60 + 60)$

$= 60$

Mode:

Mode is the most appearing set in a data and from data(Q4), mode is 60.

M_o = 60 [Because 60 is the most repeated variable (6 times) in data(Q4)]

Standard Deviation:

Standard deviation of the data can be obtained from the variance of the data. Standard deviation is the square root of the variance. Whereas, variance is obtained by calculating the mean of the squared difference between each data point and the mean of the data. In our circumstance, we have collected data from some of the university students. So, collected data is the sample in that case and that's why we must use Sample Standard Deviation formula to find Standard Deviation.

Formula:

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Here,

Number of Observation, n = 30

$\bar{x}$ is the mean of the dataset = 58.8333

Now,

X	$(\mathbf{x}-\bar{\mathbf{x}})^2$
40	354.6931889
95	1308.030189
50	78.02718889
80	448.0291889
55	14.69418889
35	568.0261889
60	1.36118889

65	38.02818889
55	14.69418889
55	14.69418889
40	354.6931889
60	1.36118889
60	1.36118889
60	1.36118889
40	354.6931889
70	124.6951889
70	124.6951889
50	78.02718889
55	14.69418889
55	14.69418889
65	38.02818889
60	1.36118889
65	38.02818889
85	684,6961889
50	78.02718889
75	261.3621889
40	354.6931889
65	38.02818889
60	1.36118889
50	78.02718889
Total	$\sum(x - \bar{x})^2$= 5484.16667

Now,

So, Sample Standard Deviation,

$$S= \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{5484.16667}{30-1}} = \sqrt{189.109195} = 13.7517$$

Skewness:

$$\text{Skewness, sk} = \frac{3(\text{Mean}-\text{Median})}{\text{Standard Deviation}}$$

$$= \frac{3(\,58.8333-60\,)}{13.7517}$$

$$= -\,0.25$$

Interpretation: If the value of sk < 0 means the distribution isnegatively skewed or skewed left, meaning that the left tail of the distribution is longer than the right.

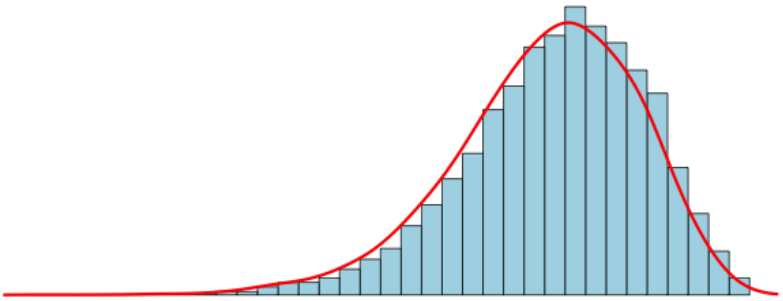


Figure: Negatively Skewed

Kurtosis:

The kurtosis value of data(Q4) is 0.62086(From excel sheet).
Interpretation: The value of the kurtosis is less than 3. So, it is a platykurtic kurtosis. The tails of platykurtic kurtosis are very thin compared to the normal distribution.

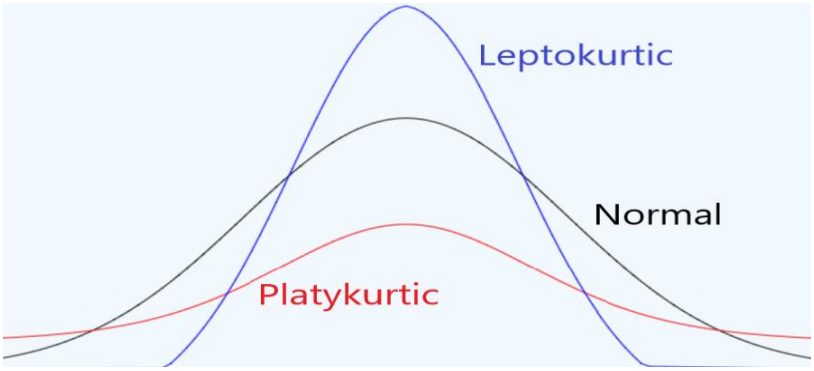


Figure: Kurtosis

Suitable graph for data(Q4):

Number of classes:

$2^k > 30$

⇨ $2^5 > 30$

⇨ $k \geqslant 30$ [So, there should be at least 5 classes]

Here,

k = Number of classes

n = Total observation

Class interval:

$i \geqslant \frac{H-L}{k}$

⇨ $i \geqslant \frac{7-1}{5}$

⇨ $i \geqslant 1.2$ [We will take exact 1.2]

Here,

H= Highest observation of the dataset

L= Lowest observation of the dataset

i = Class interval

Class	Frequency
35-47	5
47-59	9
59-71	12
71-83	2
83-95	1
95-107	1
	Total = 30

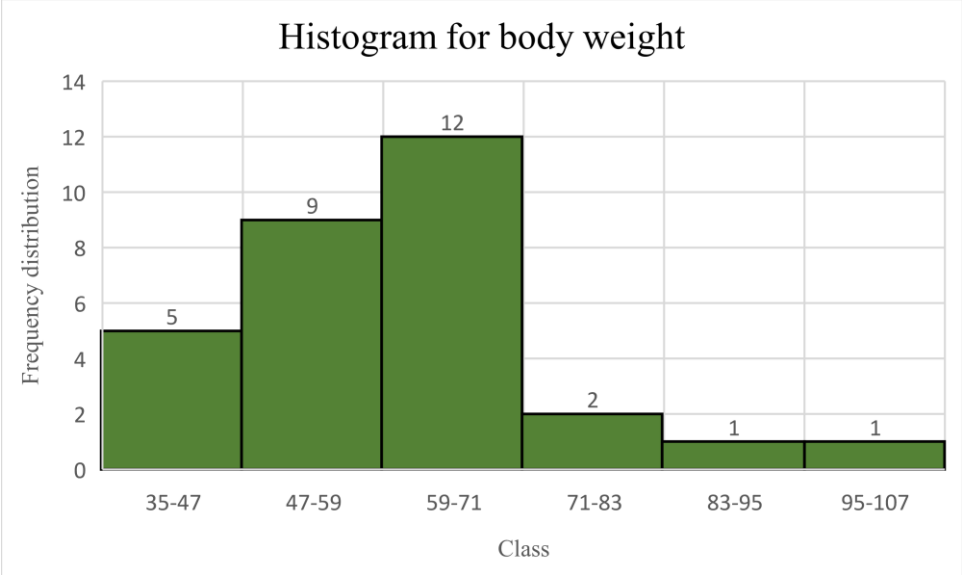


Figure: Histogram

Ans-to-the-Q-No-2

The is an association between data(Q2) and data(Q4).

Cycle riding(X)	Body Weight(Y)	XY	X^2	Y^2
1	40	40	1	1600
2	95	190	4	9025
3	50	150	9	2500
7	80	560	49	6400
3	55	165	9	3025
1	35	35	1	1225
1	60	60	1	3600
3	65	195	9	4225
5	55	275	25	3025
6	55	330	36	3025
1	40	40	1	1600
4	60	240	16	3600
4	60	240	16	3600
1	60	60	1	3600
1	40	40	1	1600
1	70	70	1	4900
7	70	490	49	4900
1	50	50	1	2500
1	55	55	1	3025
1	55	55	1	3025
4	65	260	16	4225

3	60	180	9	3600
5	65	325	25	4225
7	85	595	49	7225
5	50	250	25	2500
7	75	525	49	5625
1	40	40	1	1600
6	65	390	36	4225
5	60	300	25	3600
1	50	50	1	2500
Total = 98	Total = 1765	Total = 6255	Total = 468	Total = 109325

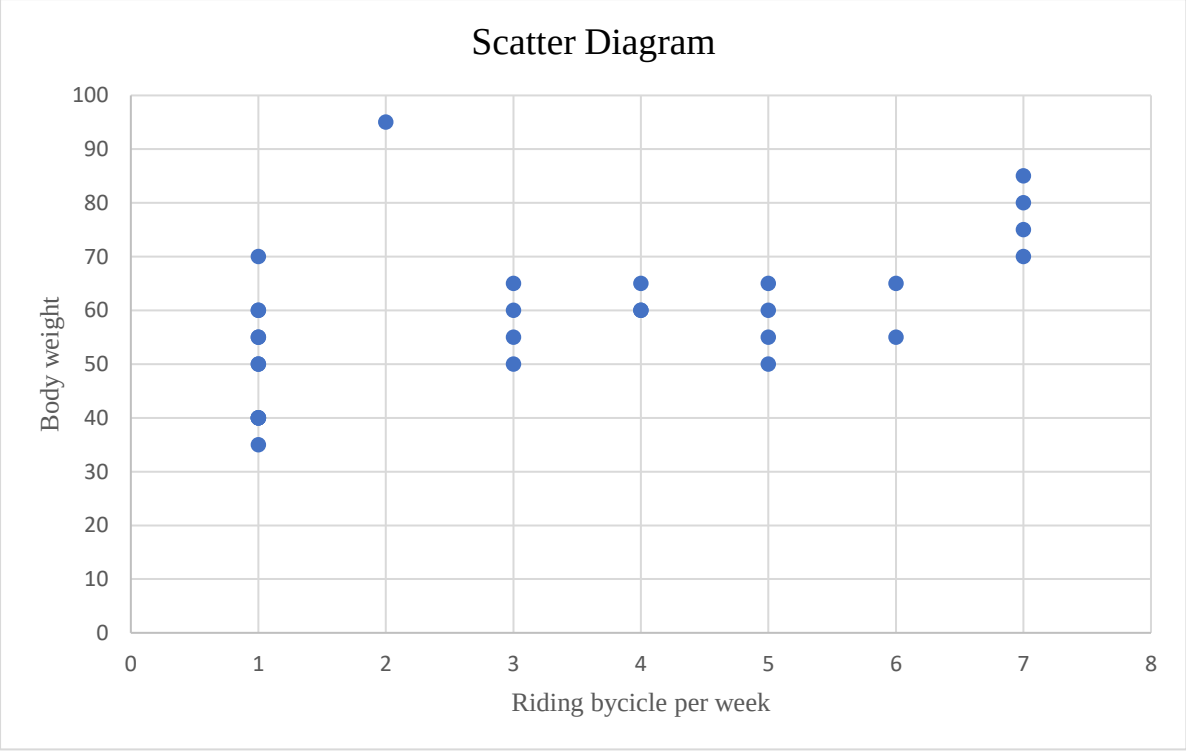
We know,

Correlation coefficient, $r = \frac{n(\sum XY)-(\sum X)(\sum Y)}{\sqrt{[n\sum X^2-(\sum X)^2][n\sum Y^2-(\sum Y)^2]}}$

$= \frac{30\times6255-98\times1765}{\sqrt{[30\times468-98^2][30\times109325-1765^2]}}$

$= 0.54$

We know, the value of r is [0.5,0.7), it is called positively moderate correlation. So, we can say there is an association between data(Q2) and data(Q4).



Interpretation: From the graph, we can see the dot pattern is upward. The dots are staying near to each other. As the spreadness is low, the relationship of the data is strong. So, the diagram will be strong positive correlation.

Ans-to-the-Q-No-3

Suitable graph for data(Q1):

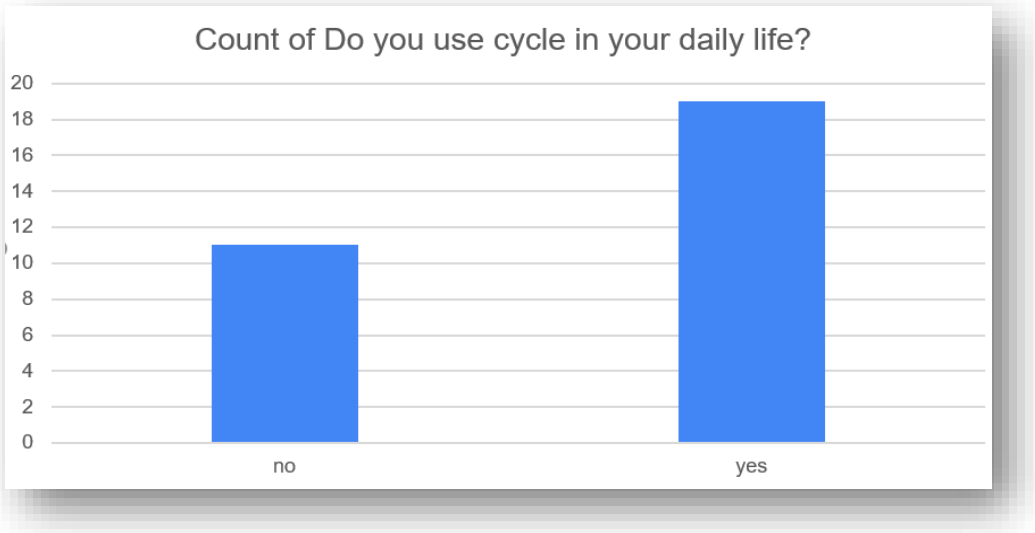


Figure: Bar chart for data(Q1)

Here, in x-axis we have what people reacted and in y-axis we have how many people reacted.

Suitable graph for data(Q3):

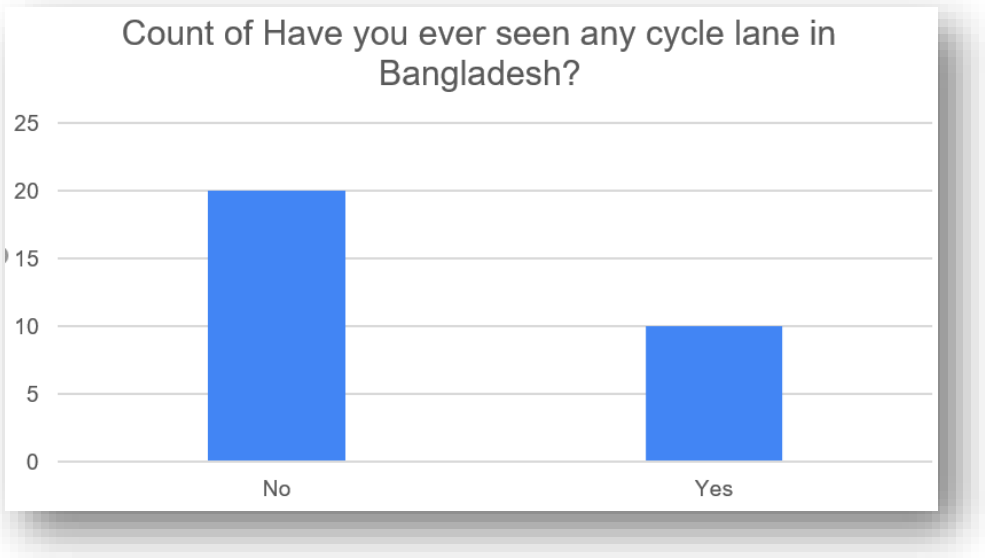


Figure: Bar chart for data(Q3)

Here, in x-axis we have what people reacted and in y-axis we have how many people reacted.

Suitable graph for data(Q5):

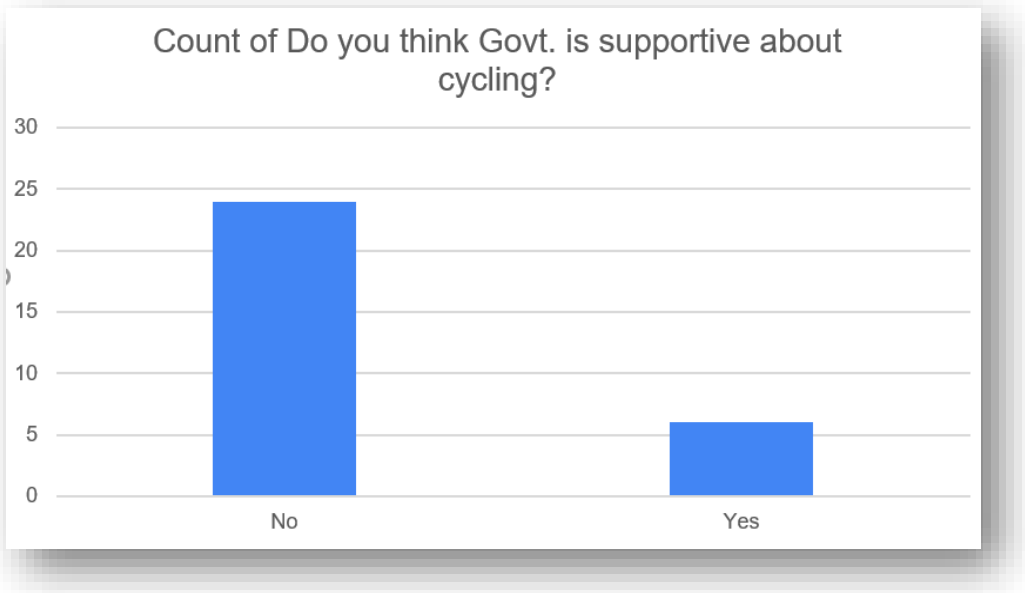


Figure: Bar chart for data(Q5)

Here, in x-axis we have what people reacted and in y-axis we have how many people reacted.

Ans-to-the-Q-No-4

Conclusion: In this project, we have predicted "Cycling habit of University student". We have observed 30 student and asked them some question about their daily routine of riding cycle and health condition from this observation. We used Google form to collect that information.

From the form we got, 63.3% students ride cycle in their daily life where 36.7% do not (from Q1). 40% students ride cycle 1 day, 3.3% ride cycle 2 days, 10% ride cycle 3 days, 13.3% ride 4 days, 13.3% ride 5 days, 6.7% ride 6 days, 13.3% ride 7 days (from the Q2). Then 55.7% students did not see any cycle lane in Bangladesh, on the other hand 33.3% saw (from Q3). Body weights: 35(3.3%), 40(13.3%), 45(0%), 50(13.3%), 55(16.7%), 60(20%), 65(13.3%), 70(6.7%), 75(3.3%), 80(3.3%), 85(3.3%), 90(3.3%), 95(3.3%) (from Q4). 20% gave their opinion yes about governments' support for cycling, but 80% are against the opinion (from Q5).

Then we used the data to find mean, median, mode, standard deviation for Q2 and Q4 and also calculated skewness and kurtosis and interpreted the result. We also made suitable graph for those. From the mean we get measure of central tendency that incorporates the score from

every subject in the research study. From the median we get statistical measure that determines the middle value of a dataset listed in ascending order. From mode we get the central tendency when examining the data. From standard deviation we get a measure that can tell how measurements for a group are spread out from the average. We got the measurement of the lack of symmetry of the distribution from skewness. From kurtosis we get a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution, mainly we can see is there any extreme value in the tail of the distribution.

There is an association between Q2 and Q4 and we have showed this association in scatter diagram. It shows the measure of the strength of the relationship between two variables.

Again, we have represented the data od Q1, Q3, Q5 in bar chart. It shows a visual representation of the responses of the students.

From this project, we realized all students are not careful about their health. Also, government do not do enough effort to make consciousness. But the main limitation of this project is this dataset is too short. If we could observe more students, the prediction would be more subtle.