Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

From the categorical variables, we can infer that there are no chances of outliers in this as outliers tend to come only in regression problems.

The categorical variable like holiday was a significant feature for the model. If there is a holiday the bike count increases.

As the year increased demand increased.

For the later months during September, October, November, the holidays are more hence also the bike demand increases.

Weekends have a greater number of bookings.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   If it is not given the same column is treated and it will be an extra variable which has same information which will lead to overfitting of the model. Hence it is recommended to delete it.

   For example, if we have multiple seasons, then the first variable is dropped which contains the same information.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Feature “temp” has highest correlation.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - No Multicollinearity
   - Homoscedasticity (For example variable ‘register’ was heteroscedastic)
   - Linear relationship between independent and target variable.
   - Error terms are normally distributed
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Temp, year, holiday

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Linear regression elucidates about linear relationship between independent variable with target variable.

   There are two types of regression:
   - Simple linear regression
   - Multiple linear regression

   Simple linear regression:

   It describes relationship between one dependant and one independent variable. If the value keeps increasing, then there is a positive correlation. If it is decreasing, it has negative correlation.

   Multiple linear regression:

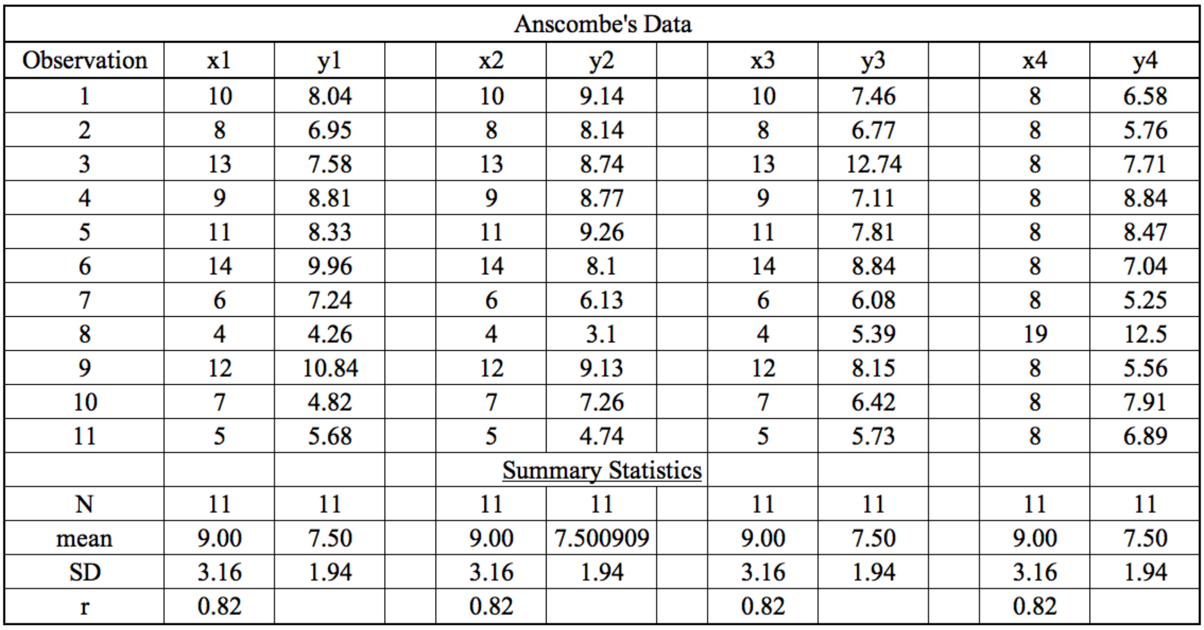
   It describes relationship between one many independent variables.

The equation is **y = mx + c**
y = dependant variable
m = slope
c = y intercept constant.

2. Explain the Anscombe's quartet in detail. (3 marks)

Francis Anscombe had developed Anscombe's quartet after his name in 1973 . It uses descriptive statistics for grouping together four similar datasets. It gives weightage on the visualization of the dataset first and then using algorithms. It says that linear regression is the best algorithm for determining the linear relationship between the variables. But when the model is plotted using scatter plot, it has difference even when the datasets are similar.

Below is the data



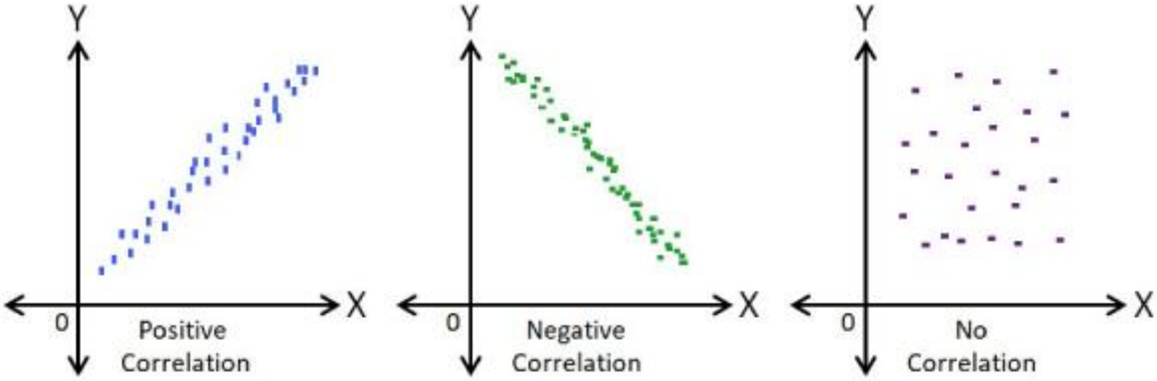
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

The summary statistics determine that x and y are similar for the datasets.

3. What is Pearson's R? (3 marks)

It determines the strength of linearity between the variables. It is used to determine the correlation between the variables.
If the values of variables go upwards, there is a positive correlation
If the values of variables go downwards, there is a negative correlation
If the values of variables doesn't change, there is no correlation
The range is -1, 0, +1
negative correlation = -1
no correlation = 1
positive correlation = 1

For multiple linear regression, it helps in detecting multicollinearity beforehand so that the model avoids overfitting.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   The purpose is to provide a fixed range to the independent data with deviating values. Scaling provides an automatic solution to outliers. If there is a huge deviation from one value to another value. Scaling helps to order between 0 and 1 for a better learning for the model. Min max scaling is most importantly done for numerical features. If scaling is prior, then there will be huge coefficient of weights and the learning will be improper.

   i. Normalized Scaling:

      Scale values are either between (-1,1) or (0,1)

      Outlier detection is not that good in normalized scaling compared to Min max.

   ii. Standardized scaling:

      It doesn’t have ranges. Mean is 0 and standard deviation is always 1. Best for outliers.

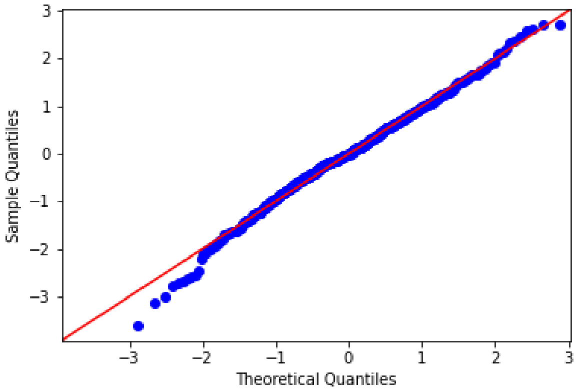
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

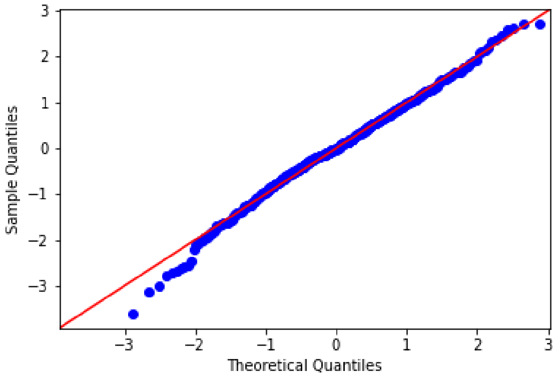
   If there is a perfect correlation between the variables. VIF is used.
   The rule for VIF is:

   - Greater than 10: Variable should be eliminated.
   - Greater than 5: Medium.
   - Less than 5: Variable needn’t be eliminated.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

   The qq plot says if two values or a set of values are having a common distribution between one another. It is a plot between theoretical quantiles and Sample quantiles.
   This elucidates whether the plot distributions of the two variables are equal/similar or not.
   Below is an example of Q-Q plot demand unit of predicted and actual values of a Bike sharing dataset.


Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Sample Quantiles
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3