Assignment-based Subjective Questions

1. From you analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
   Answer:
   After analysis of the categorical variables against the dependent variable:
   - Fall has the highest demand for bike.
   - The demand for bikes has gone up significantly from 2018 to 2019.
   - The demand for bikes steadily increases from January to June, the demand then shows no significant change till September. From October the demand for bikes drops down again.
   - The demand for bikes decreases when there is a holiday.
   - There is no significant change in demand of the bikes for each day of the week
   - The demand of for bikes is highest when the weather is clear, and starts to drop down when the weather is cloudy/misty or there is light rain/snow, reaching the lowest with light rain/snow

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 marks)
   Answer:
   Setting **drop_first=True** while creating dummy variables helps in reducing the number of dummy variables created by '1', as doing so gets rid of the redundant column as this column can be represented with the help of the remaining dummy variables.

3. Looking at the pair-plot among the numeric variables, which one has the highest correlation with the target variable? (1 mark)
   Answer:
   The variable '**temp**' shows the highest correlation with the target variable '**cnt**'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer:
   The following steps were taken to ensure that the assumptions were valid:
   - Multicollinearity check using VIF
   - Checking for overfitting by looking at the R-squared and Adjusted R-squared value
   - Plotted a histogram to ensure the error terms are normally distributed with mean 0
   - Plotted a scatterplot to ensure that the error terms do follow any pattern and have constant variance

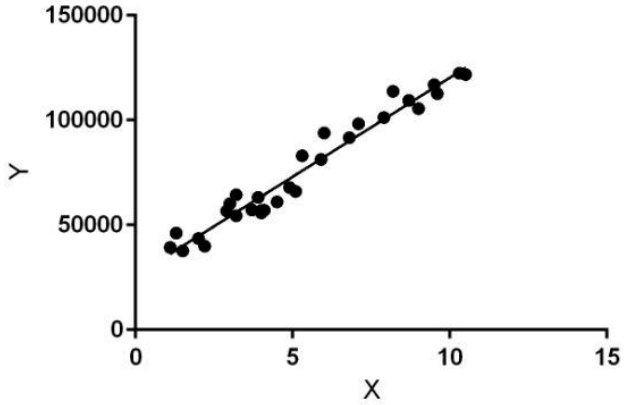
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer:
   The top 3 features that contribute significantly towards explaining the demand of shared bikes are '**temp**', '**weather** (light rain/snow)' and '**season** (spring)'.

General Subjective Questions

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

   Answer:

   Linear Regression is a machine learning algorithm based on **supervised learning.** It performs a regression task. Regression models a target prediction value based on the independent variables. It is mostly used for finding relationship between variables and forecasting. Different regression models differ based on the kind of relationship between the dependent and independent variables they are considering and the number of independent variables being used.



   Linear regression performs the task of predicting a dependent variable value (y) based on the given independent variable (X).

   Mathematically the equation for linear regression can be written as:

$$y = mx + c$$

   This expression can also be expressed as $y = \beta_0 + \beta_1 x$

   Where, β_0 is the intercept and β_1 is the slope and x is the independent variable

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

   Answer:

   Anscombe's quartet comprises of four datasets that have nearly identical simple statistical properties, yet appear different when graphed. Each dataset consists of eleven (x, y) points. Four sets of 11 data-points are given below:

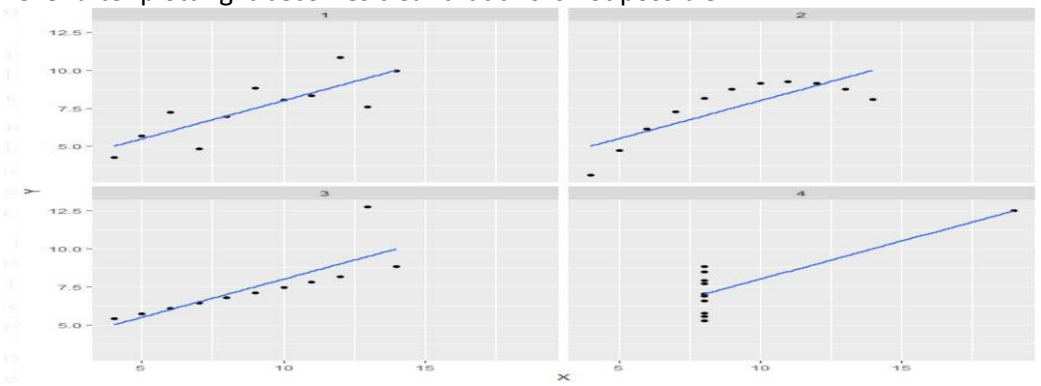
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

   The table below shows the mean, standard deviation and correlation between X and y for the 4 sets.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Going solely by these statistics, one could think that a line could be fitted on these datasets, however after plotting it becomes clear that this is not possible



From the scatterplots we can see that:

- For the first set, there seems to be a linear relationship between X and y
- For the second set, there is non-linear relationship between X and y
- For the third set, there is a perfect linear relationship between X and y for all the data points except one which seems to be an outlier value
- The fourth set, shows an example when one high-leverage point is enough to produce high correlation coefficient.

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

Answer:

The Pearson's Correlation Coefficient is also known as Pearson's R is statistic used to measure the linear correlation between two variables. The numeric value of the Pearson's R lies between -1 and 1.

If data lies on a perfect straight with positive slope the value of r will equal to '1' or approximately equal to '1', suggesting high positive correlation i.e. when X increases y will also increase, inversely if the slope is negative, then r will be equal to '-1' or approximately equal to '-1' suggesting high negative correlation, i.e. when X increases y will decrease.

An r value which is closer to '0' suggests that X and y do not share a linear relationship.

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

Answer:

Scaling is a used to normalize the range of the independent variables.

It brings all the independent variables on the same scale for regression to work properly. If Scaling is not performed then the linear regression algorithm will treat values with a higher number to be higher irrespective of the scale (e.g. grams, kg's, etc.)

An example of this would be if a dataset contains the weight of product in both grams and kilograms, in this case the value of 100 g which should be smaller than a value 5 kg however the model will treat 100 to be greater 5, which would be correct had both values been on the same scale, however as that is not the case the prediction will go wrong.

There are two types of scaling that can be done:

1. Normalization: This will scale the variables in a range of zero and one
2. Standardization: This transforms the data such that it has a mean of zero and a standard of one

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

Answer:

If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 = 1, which leads to 1/(1 – R2) to be infinity

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Advantages of q-q plot:

- The sample sizes do not need to be equal
- Many distributional such as shifts in scale, shifts in location, changes in symmetry and the presence of outliers can be tested simultaneously.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

The q-q plot is used to answer the following question:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justifiable. The q-q plot can provide more insight into the nature of the difference of the two data sets.