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)

After the analysis on the categorical variables, their effect on the dependent variable is as follows:

i. The bike rental popularity increased in 2019 as compared to the previous year (2018)
ii. The bike rental also increased every month. The maximum bikes were rented between March and October.
iii. The bike rentals are maximum during Saturday and Sunday.
iv. The bike rental is little high during the working days as compared to holidays.
v. Bike rental is high during the clear weather.
vi. Out of four seasons, bike rental is more popular during 'fall' season.

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

Machine Learning cannot work with non-numerical data, so if we have categorical variables, we need to convert those variables into numerical data.

If there are 'n' number of distinct values in a categorical variable, those values can be converted into numerical data by creating dummy variables. The idea of dummy variable creation is to create 'n-1' variables and these dummy variables are created against each distinct value into the categorical variable. To create 'n-1' dummy variables, drop_first=True is used which automatically drop the first dummy variable during creation. If we don't drop the first variable, our dummy variables will be correlated. This may affect some models adversely and the affect is stronger when the cardinality is smaller.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

'temp' and 'atemp' both variables are equally correlated with the target variable. The correlation is 0.63.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

We have validated the assumptions of Linear Regression as follows:

Little or No Autocorrelation: The summary of our model shows Durbin-Watson value as 2.040 which is between 2 and 4. As this value is very close to 2 which seems to be very close to the ideal case and hence have no autocorrelation.

Little or no Multicollinearity: During our analysis we found that 'temp' and 'atemp' variables are highly correlated to each other, so we dropped 'atemp' before the model building to avoid

multicollinearity. After building our model the VIF values came up less than 5 which means that there is no extreme multicollinearity between selected independent variables.

Normally distributed Residuals: After building the model, we have found that the error terms are centered around 0 which tells us that residuals 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)**

The following features contributing significantly towards the demand of bike rentals:

i. Sunday
ii. Temperature
iii. Mist weather

General Subjective Questions

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

Linear regression is a simple statistical regression method used for predictive analysis which shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).

There are two types of Linear regressions, Simple Linear Regression (SLR) and Multiple Linear Regression (MLR).
Simple Linear Regression: If there is a single independent variable, such linear regression is called simple linear regression.
Multiple Linear Regression: If there are more than one independent variables, such linear regression is called multiple linear regression.

The above graph shows the linear relationship between dependent variable and independent variable. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.
A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. If the dependent variable increases on the Y-axis as the independent variable progress on X-axis, then

such a relationship is called as a Positive linear relationship. If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a Negative linear relationship.

To calculate the best-fit line, linear regression uses following formula:

$y = mx + b$ => $y = a_0 + a_1 x$

y = Dependent variable
x = Independent variable
a_0 = Intercept of the line
a_1 = Linear regression coefficient

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

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 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

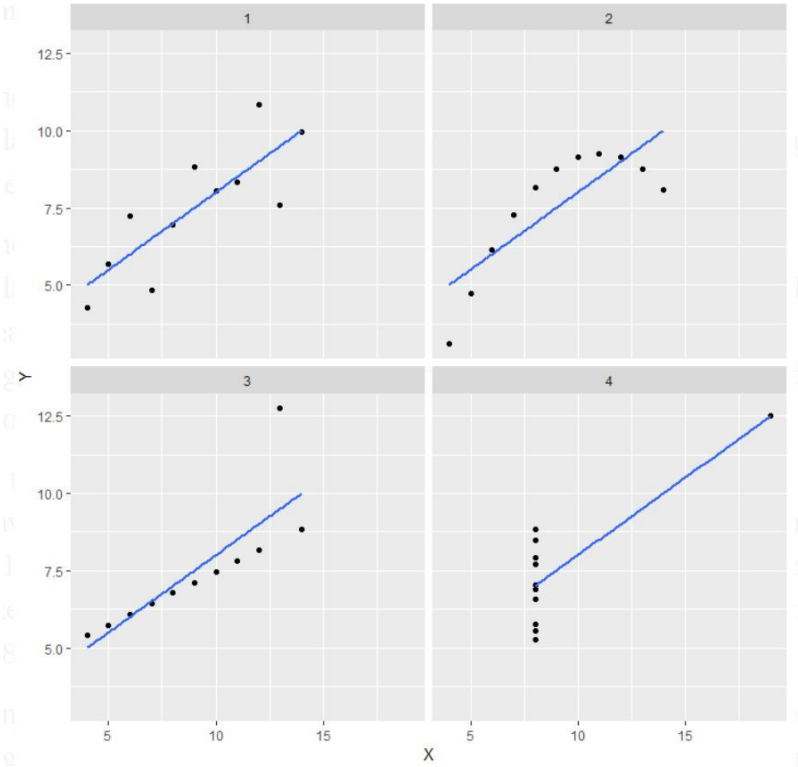
It was called as Anscombe's quartet. The definition of Anscombe's quartet comprises of four data sets that have nearly identical simple statistical properties yet appear very different when graphed.

The council analyzed these datasets using only descriptive statistics and found the mean, standard deviation and correlation between x and y.

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

The graphical representation and explanation of these four datasets is as follows:



i. In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

ii. In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

iii. In the third one (bottom left) you can say where there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

iv. Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's R, the Pearson product moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and 1.0. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient R. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula is given below:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

N = The number of pairs of scores.
Σxy = The sum of the products of paired scores.
Σx = The sum of x scores.
Σy = The sum of y scores.
Σx2 = The sum of squared x scores.
Σy2 = The sum of squared y scores.

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

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. Scaling is performed because of two following reasons:

i. Ease of interpretation.
ii. Faster convergence for gradient descent methods.

It is important to note that scaling just affects the coefficients and none of other parameters like t-statistics, F-statistics, p-values, R-squared etc.
There are two types of populate scaling which are as follows:

Normalized/Min-Max Scaling: In this, the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. The formula is given below:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized Scaling: Standardization brings all the data into a standard normal distribution which has mean zero and standard deviation one. The formula for it is given below:

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

A VIF value can be infinite when there is a multicollinearity in our model. Multicollinearity happens when there are two or more independent variables are highly correlated to each other. This can lead to wider confidence intervals that produce less reliable probabilities in terms of the effect of independent variables in a model.

As a result of perfect correlation, the VIF value shows infinity and we get R2 = 1, which lead to 1 / (1 – R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing the perform multicollinearity.

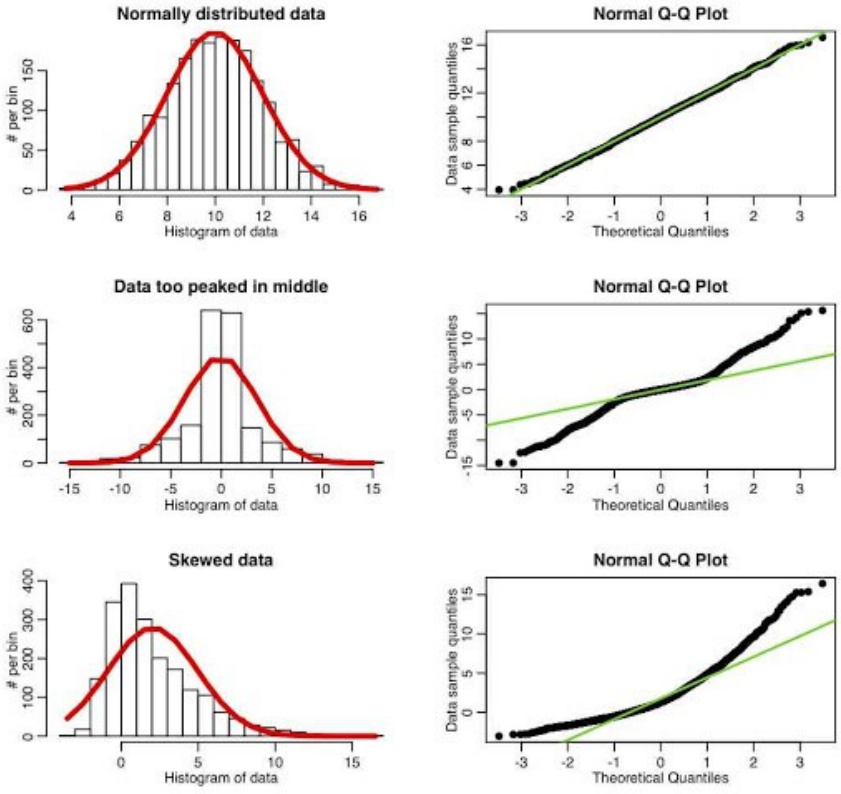
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plot (Quantile-Quantile plot) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

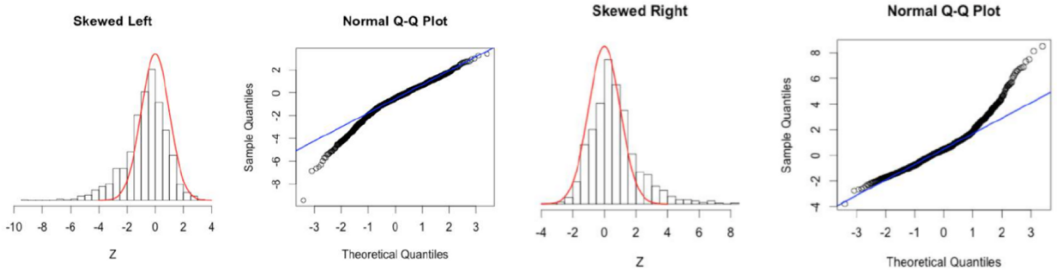
A q-q plot is a plot of the quantiles of the first data set against the quantiles of second data set. By a quantile, we mean the fraction (or percent) of points below the given value. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line, and we can say that this

distribution is Normally distribution. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.



Skewed Q-Q plots: Q-Q plots are also used to find the skewness of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution, we want to know on the y-axis then we see a very peculiar shape of a Normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to it's left or simply it is left-skewed (or negatively skewed) but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower end follows a straight line then the curve has a longer tail to its right and it is rights-skewed (or positively skewed).



In the end we can say that a Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.