Assignment-based Subjective Questions and Answers:

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

As per my analysis, the fall season has high demand for bikes. In spring season, the demand was less when compared to the other seasons. In addition to this, bike demand has increased in 2019 when compare to 2018. During the beginning of the year and towards the ending the bike demand was less. It could due to the bad weather condition on those days.

Demand is continuously increasing each month till June. On September month has highest demand. Weekdays did not give conclusive evidence on demand.

On holidays the demand is high as expected. As expected during good weather days, the demand was high.

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

During dummy variables creation there is a chance for additional columns get created. To avoid this, it is important to use **drop_first=True.** This way we can reduce the correlations created among dummy variables.

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

Based on the heatmap, the variable **'temp'** has the highest correlation with target variable **'cnt'**.

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

I would you use the following parameters to validate the assumptions.

Linearity: Validate if the relationship between X and the mean of Y is linear.
Homoscedasticity: Check if the variance of residual is the same for any value of X.
Independence: Validate and ensure the Observations are independent of each other.
Normality: For any fixed value of X, Y is normally distributed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Season, workingday and weather are three important features that contributes to the demand of the shared bikes.

General Subjective Questions and Answers:

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as temperature, price, age etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

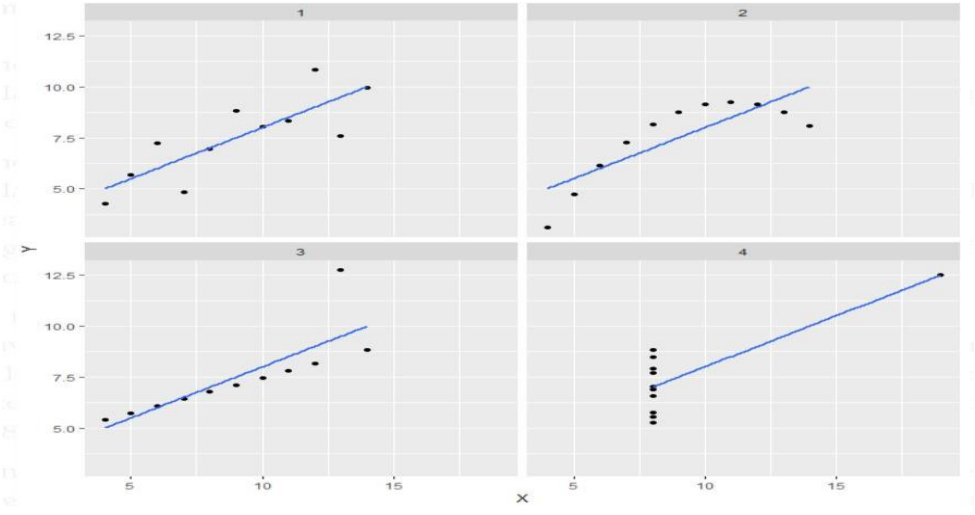
2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Therefore, the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

Below four similar looking sample data sets given.

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

Let's plot the above data in to a scatter plot.



Explanation on the above scatter plot:

1. 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.
2. 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.
3. In the third one (bottom left) you can say when 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.
4. Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

3. What is Pearson's R?

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation also called Pearson's R, this is a correlation coefficient commonly used in linear regression.

The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

The PPMC is not able to tell the difference between dependent variables and independent variables. For example, if we are trying to find the correlation between a high calorie diet and diabetes, we might find a high correlation of .8. However, we could also get the same result with the variables switched around. In other words, we could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, we have to be aware of the data that we are plugging in. In addition, the PPMC will not give any information about the slope of the line; it just shows, whether there is a relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example-centered around 0 or in the range (0,1) depending on the scaling technique.

Reason For Scaling: An unscaled data can adversely impact a model's ability to make accurate predictions. It is because, the real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

There are two types of scaling:

1) Normalisation scaling.
2) Standardisation scaling.

Difference between Normalized scaling and Standardized scaling: The choice between normalisation and standardisation really comes down to the application.

Standardisation is generally preferred over normalisation in most machine learning context as it is especially important when comparing the similarities between features based on certain distance measures. This is most prominent in Principal Component Analysis (PCA), a dimensionality reduction algorithm, where we are interested in the components that maximise the variance in the data.

Normalisation, on the other hand, also offers many practical applications particularly in computer vision and image processing where pixel intensities have to be normalised in order to fit within the RGB colour range between 0 and 255. Moreover, neural network algorithms typically require data to be normalised to a 0 to 1 scale before model training.

At the end of the day, there is no definitive answer as to whether you should normalise or standardise your data. One can always apply both techniques and compare the model performance under each approach for the best result.

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

If there is 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 lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

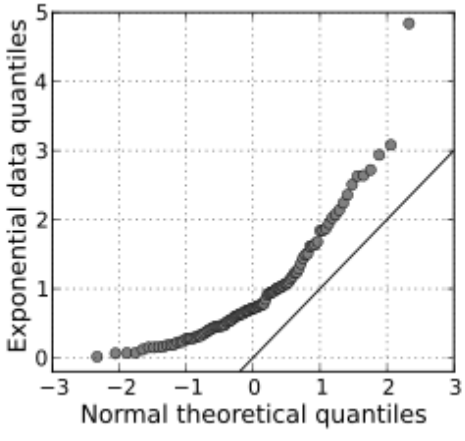
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

A Q-Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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.