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)

Answer:

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non-working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall.

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces 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? (1 mark)

Answer: By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Answer:

- Linear relationship: By creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.
- Multicollinearity: If VIF<=5 implies no multicollinearity, whereas VIF>=10 implies serious multicollinearity.
- Homoscedasticity: Scatter plot that shows residual vs fitted value. If the data points are spread across equally, it means the residuals have constant variance (homoscedasticity). Otherwise, if a funnel-shaped pattern is seen, it means the residuals are not distributed equally and depicts heteroscedasticity.
- Normal distribution of error terms: By checking the assumption using a Q-Q (Quantile-Quantile) plot. If the data points on the graph form a straight diagonal line, the assumption is met.

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: Based on final model top three features contributing significantly towards explaining the demand are:

1. Temperature - (0.5398)
2. Weathersit LightSnow - (0.3017)
3. year -(0.2312)

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**.** Regression model is a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Hypothesis function for Linear Regression :

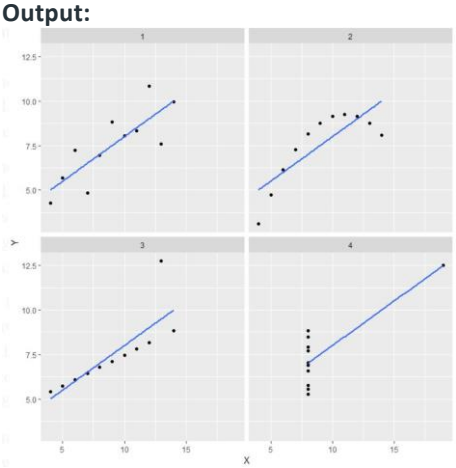
$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. It is quite interesting to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Output:



It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed.

Explanation of this output:

- In the first one (top left) if you look at the scatter plot we will find that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure we can conclude that there is a non-linear relationship between x and y.
- 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.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

Answer: The Pearson correlation coefficient (r) is used to measure a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

Below is the formula to calculate the correlation coefficient(r):

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

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

Answer: Scaling is one of the most important data pre-processing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.
There are some feature scaling techniques such as Normalization and Standardization that are the most popular and at the same time, the most confusing ones.

1.Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1]. Normalization is useful when there are no outliers as it cannot cope up with them.

2.Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

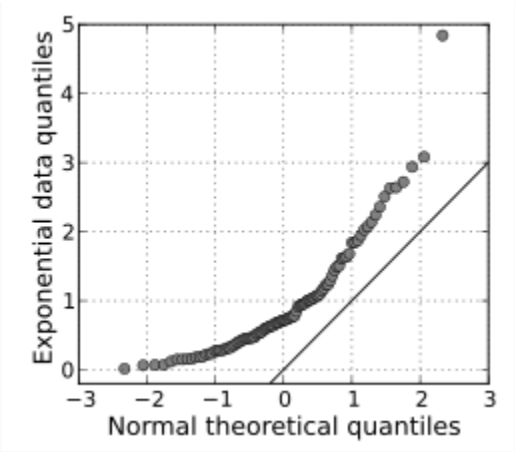
Answer: If VIF -> infinity it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-Square =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. 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. (3 marks)

Answer:

Q-Q Plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. 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 angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

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.