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:

   a. The demand of rent bikes is low in the month of **`spring`** compared to other seasons
   b. The demand of rent bikes increased in the year **2019** compared to year **2018**.
   c. Number of Ride Count drastically increases between May to October which are comparatively Summer & Fall Season
   d. Rent bikes demand is less in holidays in comparison to not being holiday.
   e. The demand of rent bikes shows no measure variations in weekday.
   f. No measure variations in rent bike count in working day
   g. The rent bikes demand is high when weather is `clear and Few clouds` however demand is less in case of `Lightsnow and light rainfall`. We do not have any data for `Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog`, so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of rent bikes.

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:` `temp` has highest positive 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:

   -After building model, we cannot finalise until we prove the residual analysis wherein, we check whether the distribution of Error is around 0 or not

   - From the plotted graph it is evident that Error Distribution Is Normally Distributed Across 0, which indicates that our model has handled the assumption of Error Normal Distribution properly

   - From the regplot, we see that there is almost no relation between Residual & Predicted Value. This is what we had expected from our model to have no specific pattern.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared rent bikess? (2 marks)

   Answer: Temp, season, Month are the top 3 features contributing significantly towards explaining the demand of the shared rent bikess

General Subjective Questions

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

   Answer: Linear regression of machine learning where we train a model to predict the behaviour of your data based on one or more variables. In the case of linear regression, linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

   We can write a linear regression equation as:

   y = a + bx

   Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

   Here, x and y are two variables on the regression line.

   b = Slope of the line

   a = y-intercept of the line

   x = Independent variable from dataset

   y = Dependent variable from dataset

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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

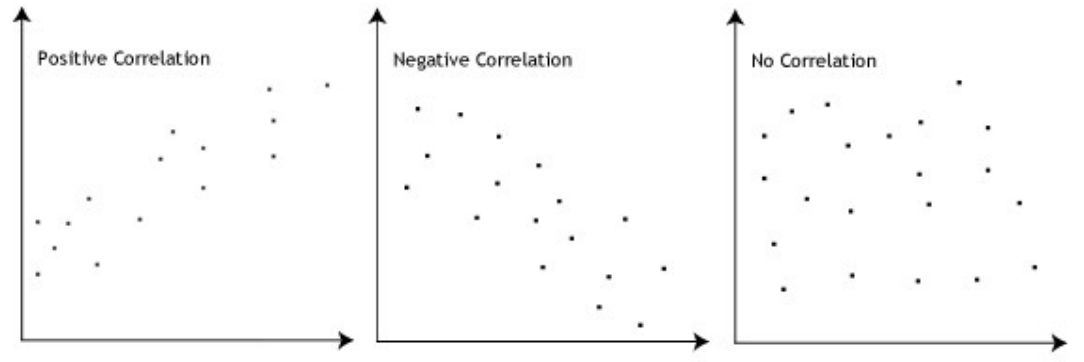
3. What is Pearson's R? (3 marks)

Answer:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

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

Answer:

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

   Answer:
   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.

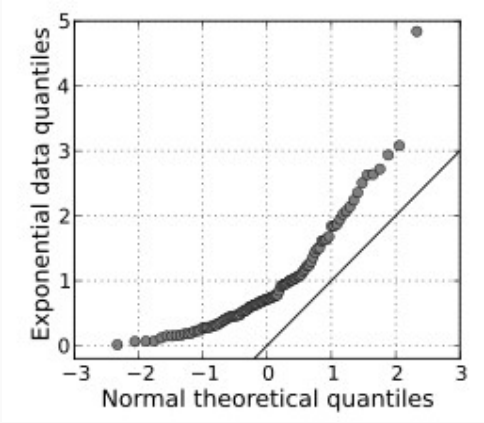
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

   Answer:

   Q-Q Plots (Quantile-Quantile 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.

   Q Q plot showing the 45-degree reference line:

   A 

   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.