Assignment-based Subjective Questions

Q.1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Ans -> From the basis of analysis I got following points -

- Fall season (season 3) has highest demand for rental bikes.
- The demand is increasing till September and after that its decreasing.
- There is not much clarity in weekday variable with target.
- When the weather is good the Demand is more and when weather is bad the demand is less.
- I can see that the demand for next year will increase.
- When there is holiday demand is slightly decreasing.

==

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

Ans ->

When we are creating dummy variables, so it is creating different columns based on number of values in that column.

If we used drop_first=True so it is creating columns and removing first value of the column so the data will optimize and less data to handle as well as for Linear Regression.

E.g.

The variable `furnishingstatus` has three levels. We need to convert these levels into integer as well. For this, we will use something called `dummy variables.`

Now, you don't need three columns. You can drop the `furnished` column, as the type of furnishing can be identified with just the last two columns where —

- `00` will correspond to `furnished.`
- `01` will correspond to `unfurnished.`
- `10` will correspond to `semi-furnished.`

==

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

Ans -> **temp and atemp** are highly correlated with the target variable cnt.

==

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

Ans -> I has checked the following assumptions -

- Error terms are normally distributed with mean 0.
- Error Terms do not follow any pattern.
- Multicollinearity check using VIF.
- Linearity check.
- Ensured the overfitting by looking the R2 value and Adjusted R2.

==

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

Ans -> top 3 features contributing significantly are -

1. Temp
2. Windspeed
3. Season

==

General Subjective Questions

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

Ans ->

Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.

Linear regression is of the 2 types:

1. **Simple Linear Regression**: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$$Y=\beta0+\beta1X1 +\epsilon$$

2. **Multiple Linear Regression**: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation

that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.

Formula for the Multiple Linear Regression:

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$$

The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by the following two methods:

· Differentiation

· Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression.

===

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

Ans ->

Definition: **Anscombe’s quartet** comprises a set of four datasets, having identical descriptive statistical properties in terms of **means, variance, R-squared, correlations, and linear regression** lines but having different representations when we scatter plots on a graph.

Purpose: The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. **Anscombe’s quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Advantages: Reveals limitations of summary statistics, emphasizing the need for visual exploration to detect nuances, outliers, and diverse relationships in datasets.

Steps:

- Find mean for x and y for all four datasets.
- Find standard deviations for x and y for all four datasets.
- Find correlations with their corresponding pair of each dataset.
- Find slope and intercept for each dataset.
- Find R-square for each dataset.
- To find R-square first find residual sum of square error and Total sum of square error.
- Create a statistical summary by using all these data and print it.
- Plot the scatter plot and linear regression line for each dataset.

===

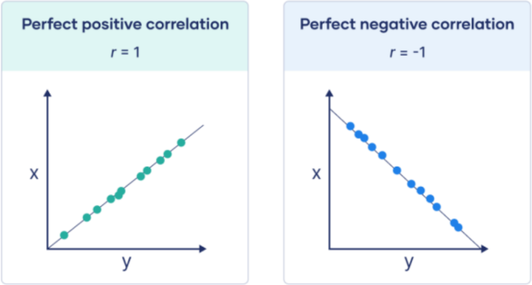
Q.3. What is Pearson's R? (3 marks)

Ans -> The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Between 0 and 1 = positive correlation

Between 0 and -1 = negative correlation

0 = no correlation



When to use it –

- Both the variables are quantitative (numeric)
- The variables are normally distributed.
- The data have no outliers.
- The relationship is linear.

Formula for calculating r

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

==

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

Ans ->

Scaling: Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

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

==

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

(3 marks)

Ans -> The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

==

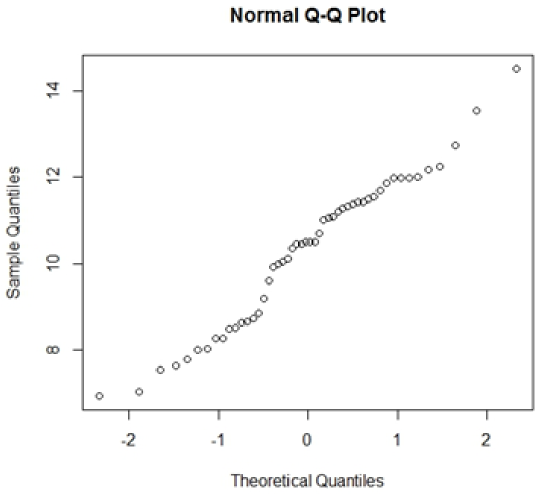
Q.6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans ->

Q-Q plot –

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.

===