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:

1. Bike rental total count increased from 2018 to 2019.
2. During no holidays, the bike rental counts is highest, compared to during holidays for different seasons.
3. There is an increase in the bike rental count in spring and summer season, and then a decrease in the bike rental count in fall and winter season.
4. There is no significant change in bike demand with working days and non-working days.
5. During clear, partly cloudy weather, the bike rental count is the highest, second-highest during misty cloudy weather, and followed by 3rd highest, during light snow, light rain weather and Scattered clouds. No bike rentals on Heavy rain and Ice pallets weather condition.
6. Significant categorical variables are: Year, Month, Holiday, Season, Weather condition.

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

Answer:

If we have a categorical variable with n categories and create n dummy variables, the last dummy variable becomes redundant. This redundancy can lead to multicollinearity in our model, which can make interpretation difficult and can impact the model's performance. To address this redundancy, we can use the drop_first parameter when creating 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' and 'atemp' has the highest correlation.

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

Answer:

Residual analysis will be carried by plotting histogram and scatter plot for the error terms to validate the assumptions

- Residues distribution is normal distribution with zero mean value.

- Residues or error terms of variables are independent.

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:

1. Temp coefficient is (0.5682) high and +ve shows registrations are more in high temperatures due to clear weather.
2. Bike sharing total count is increased from 2018 to 2019 (year coefficient 0.2334) after Covid-19. People are preferred to use bikes than public transport.
3. Next significant variable is weather condition 3 (coefficient -0.2535): Light Snow, Light Rain + Thunderstorm + Scattered clouds causes for low registrations.
4. Next significant variable is wind speed (coefficient -0.1455): -ve shows bike riding is difficult hence registrations are reduced.

General Subjective Questions

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

Answer:

Linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Assumptions of Linear regression Model

1. Independent variables (input variables) have linear relationship with dependent or output variable.

2. Residue terms or error terms are normally distributed with zero Mean.

3. Residue terms or error terms are will have same variance and independent to each other implies the data homogeneity or homoscedasticity.

4. The independent variables are measured without error.

5. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

6 Equation: $Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$

Y= Dependent or output variable, ϵ = error

$\beta0$ = Y intercept or constant,

X1, X2...Xp = Independent variables.

$\beta1,...\beta p$ = Co-efficient of independent variables.

Coefficients are obtained by minimizing the sum of squared errors, the least squares criteria

RSS = e1^2 + e2^2+.....en^2

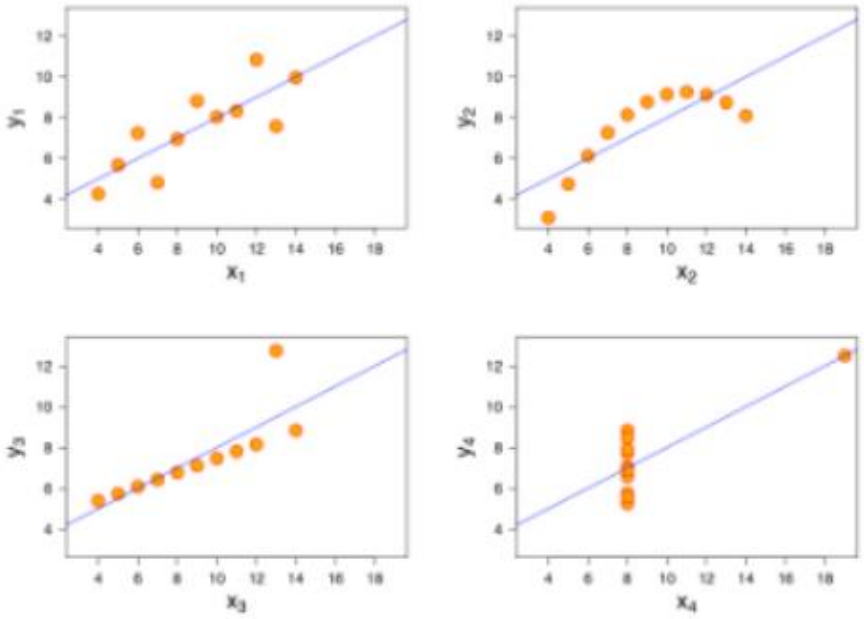
e = Y – Ypred

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

Answer:

The quartet 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. The datasets are as follows.

These phenomena can be best explained by the Anscombe's Quartet, shown below:



As we can see, all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So we should never ever run a regression without having a good look at our data.

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

Answer:

Pearson's R correlation coefficient which is the correlation coefficient in the linear regression model. This correlation coefficient is designed for linear relationships.

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

Answer:

Scaling is the factor which is used to represent the object size. The size of the object can be shown by increasing or decreasing its original size.

Having features with varying degrees of magnitude and range will cause different step sizes for each feature. Therefore, to ensure that gradient descent converges more smoothly and quickly, we need to scale our features so that they share a similar scale.

StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.

Normalized scale or Min-Max scale shrinks the data within the given range, usually of 0 to 1.

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

Answer:

An infinite VIF (Variance Inflation Factor) in linear regression indicates a **perfect multicollinearity** between the independent variables. This means that one or more independent variables can be perfectly predicted by a linear combination of the others.

This happens because of following reasons:

1. **Linear Dependence:** When two or more independent variables are linearly dependent, it means that one variable can be expressed as a linear combination of the others.
2. **Matrix Invertibility:** In the VIF calculation, a matrix is inverted. If the matrix is singular (i.e., it has a determinant of zero), it cannot be inverted. This happens when there is perfect multicollinearity.
3. **Infinite Variance:** The VIF measures the variance inflation factor, which is a measure of how much the variance of the estimated coefficient of a variable is inflated due to

multicollinearity. When there is perfect multicollinearity, the variance becomes infinite, hence the "inf" value.

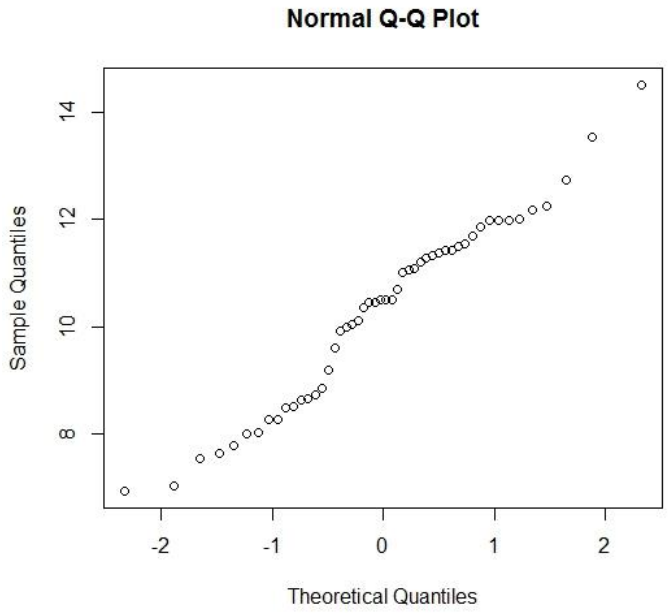
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Answer:

The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A QQ 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 QQ plot when both sets of quantiles truly come from normal distributions.



Now what are "quantiles"? These are often referred to as "percentiles." These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64. The following R code generates the quantiles for a standard normal distribution from 0.01 to 0.99 by increments of 0.01.