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?

Answer:

Categories in the dataset are Season, Year, Month, Holiday, Weekday, Working day, Weather, Part month, Quarterly.

- The number of count are less in spring season.
- The number of count increases with year.
- The number of count is very less on January month. Most counts are in the month June to Sep.
- The number of bike users are less on holidays.
- Among all the weekdays, the count on Sundays are less.
- The number of count is more on working days.
- When there is light snow, light rain, thunderstorm the number of bike users decreases. From the data it is evident that, when there is heavy rain, ice pallets, snow and fog, no one uses the bike.
- The count of bike users in the third half of the month is less.
- In the middle of the year the count is increasing. And in the first quarter of the year the count is very less.

Q.2 Why is it important to use drop first=True during dummy variable creation?

Answer:

As it helps in reducing the redundant variable created while creating dummy variables. It will reduces the correlation created among the dummy variables.

For example:

If we have a categorical variable named “TIME” with three levels namely “Morning”, “Afternoon”, and “Evening”. Then if we are using drop first = True while creating dummy variable for the TIME variable, we will get two dummy variables TIME_Afternoon and TIME_Evening. So, if it is neither afternoon not evening obviously it will be TIME_Morning.

For n levels of a category, we can create n-1 dummy variables.

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

Answer:

From the pair plot it is very clear that the count (target variable) is highly positively correlated with the temperature variable with correlation value of 0.63.

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

Answer:

We can validate the assumptions of linear regression by conducting a residual analysis. We can check it by using regression plot.

1. There should be a linear and additive relationship between dependent variable (target) and independent variable (predictor). A linear relationship suggest that a change in dependent variable due to one unit change in independent variable is constant regardless of the value of independent variable. This can be validated by using a regression plot or scatter plot.
2. The residuals must be normally distributed, this can be validated using q-q plot or plotting a distribution plot of errors.
3. There should be no correlation between the residual terms. If correlation is there between the residuals then it is called Autocorrelation. We can validate the autocorrelation by plotting a scatter plot of error terms against the index. If the plot has any pattern we can conclude that there is auto correlation.
4. The error terms must have constant variance, and is called homoscedasticity. And the absence of constant variance is called Heteroscedasticity. This can be validated by plotting a scatter plot of error with respect to index.
5. The independent variables should not be correlated. If it is correlated then it is called multicollinearity. We can validate the multicollinearity by calculating the variance inflation factor for all the independent variables and in most cases it is good to have VIF value less than 05.

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

Answer:

As per our final model. The top three features for predicting the dependent variable i.e. factor influencing the demand of the booking are:

1. Wind chill (derived variable from temperature and wind speed):

Have coefficient value of '0.5028', which indicates that a unit increase in wind chill variable increases the bike hire numbers by 0.5028 units.

2. Weather_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds):

Have coefficient value of '-0.2810', which indicates that a unit increase in weather_3 decreases the bike hire numbers by 0.2810 units.

3. Year_2019:

Have coefficient value of '0.2425', which indicates that a unit increase in year variable increases the bike hire number by 0.2425 units.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail?

Answer:

Linear regression model is one of the very basic form of machine learning under supervised learning where we train a model to predict the behavior of the dependent variable based on one or more independent variable. In simple terms linear regression is the method of fitting a best fit line to the given data and the best fit line tells us the relationship of the dependent variable and the independent variable.

It is mostly done by using the Sum of Squared Residual Method.

There are two types of linear regression, and they are:

- Simple linear regression
- Multiple linear regression

Simple linear regression:

In simple linear regression there will be only one independent variable.

$Y = \beta o + \beta 1x + error$

Where βo is the y-intercept and β1 is slope coefficient
X = independent variable
Y = dependent variable

Multiple linear regression:

On the other hand in multiple linear regression there will be more than one independent variable.

$Y = \beta o + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots\ldots\ldots + \beta nXn + error$

Where βo is the y-intercept and β1, β2, upto βn are slope coefficient of X1, X2 upto Xn respectively
X1, X2….Xn = independent variables
Y = dependent variable

Q.2 Explain the Anscombe's quartet in detail?

Answer:

- Anscombe's quartet comprises of four different datasets that have nearly same identical simple descriptive statistics yet have very different distributions and appears very different when graphed.
- They were constructed in 1973 by Francis Anscombe to demonstrate both the importance of graphing data(data visualization) before analyzing it and the effect of outliers on statistical properties
- These four datasets are intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence all the important features in the datasets must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

Q.3 What is Pearson's R?

Answer:

- Correlation measures the strength of association of two variables as well as the direction. There are mainly three types of correlations are measured. One significant type is Pearson's correlation coefficient.
- In statistics Pearson's correlation coefficient also referred as Pearson's R, Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It cannot capture the nonlinear relationship between the two variables and cannot differentiate between dependent and independent variables.
- In technical terms, the Pearson's R is the covariance of two variables divided by the product of their standard deviation. The range of the Pearson's R is -1 to +1. -1 means two variables are negatively correlated and +1 means two variables are positively correlated.
- There are certain requirement for Pearson's correlation coefficient, they are:
  1. Scale of measurement should be in interval or ratio.
  2. Variables should be approximately normally distributed
  3. The association should be linear
  4. There should be no outliers in the data
- The formula for calculating the Pearson's correlation coefficient is,

$$r = \frac{\sum(x_I - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum(x_I - \overline{x})^2 \quad \sum(y_i - \overline{y})^2}}$$

Where, r is correlation coefficient, x_i is values of x variable in a sample, and $\overline{x}$ is the mean of the values of x variable, y_i is the values of y variable in a sample and $\overline{y}$ is the mean of the values of y variable.

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

Answer:

What is Scaling?

Scaling is the step of preprocessing that is applied to independent variables or the features of data. It basically helps to normalize the data within a particular range. It also helps in speeding up the calculations in algorithm.

Why is scaling performed?

Most of the time the data we collect will have different magnitudes, with different units and with different ranges. If scaling is not done then the algorithm only takes magnitudes in account and not the unit hence will leads to incorrect modelling. So to solve this issue, we have to do scaling to bring all the variables to the same magnitude level.

Normalized vs Standardized scaling?

1. **Normalization:**
   Normalization is a scaling technique in which the values are shifted and rescaled so that they end up ranging between 0 and 1. It is called Min Max Scaling.
   The formula is

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. **Standardization**:
   Standardization is another scaling technique where all the values are centered around the mean with a unit standard deviation. This means the mean of the attribute becomes zero and the resultant distribution will has a unit standard deviation.
   The formula is

$$X' = \frac{X-\mu}{\sigma}$$

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

Answer:

- If there is perfect correlation, then VIF will become infinity. This shows a perfect correlation between two independent variables
- In the case of perfect correlation, we get R2 = 1, which leads to 1 / (1-R2) equals to infinity. To solve this problem we need to remove one of the variable from the dataset which is causing the multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Answer:

Q – Q plot:

Q-Q plot (Quantile - Quantile plot) are plots of two quantiles against each other.

Use and importance:

- Q-Q plot is used to find out if two sets of data comes from the same distribution, A 45 degree is plotted on the Q-Q plot. If the two datasets comes from the common distribution, the point will fall on the reference line.
- This helps in scenario of linear regression where we have training set and test data set received separately and then we can confirm using Q-Q plot that both the datasets are from population with same distributions
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry and presence of outliers can all be detected from this plot