Assignment based Subjective Questions with Answers

1. **From analysis of categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans.: cnt which is count of total rental bikes will be more in **winter season** and in **September month**. If **holidays** are less, then demand will be more for bikes. It will be less in **spring season**, and in **July month.** Also if weather is **cloudy** and **thunderstorm** is there, then demand will be less.

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

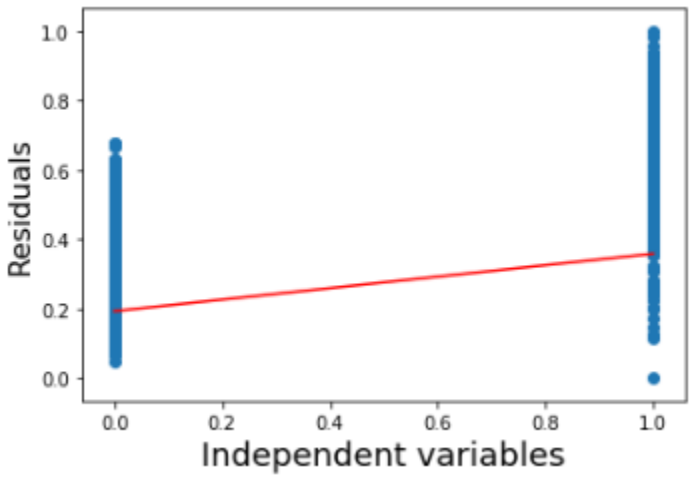
Ans.: Drop_first=True used to drop extra column created during dummy variable creation. If we don't drop the first column, then may affect some models adversely and the effect is stronger when the cardinality is smaller. Hence, **Drop_first=True** is important to use, as it helps in reducing the extra column which are created during dummy variable creation.

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

Ans.: A **temp** and **atemp** numerical variables are having highest correlation with the target variable**.**

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

Ans.: The given plot shows the Independent variables verses residuals:



There is a linear relation between variables to get the target variable. Thus the residuals are linearly dependent on independent variables.

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

Ans.: Top 3 features are: a)temp: Temperature in Celsius; b) yr: Year and
c) weathersit_Thunderstorm: weather situation

General Subjective Questions with Answers

1. **Explain the linear regression algorithm in detail.**

Ans.: Linear regression is a statistical method for modeling relationships between a dependent variable with a given set of independent variables. linear regression is used to determine the extent to which one or more variables can predict the value of the dependent variable.

There are two main types of linear regression:

1. **Simple linear regression:** This involves predicting a dependent variable based on a single independent variable.
2. **Multiple linear regression:** This involves predicting a dependent variable based on multiple independent variables.

Once a linear regression model has been trained, it can be used to make predictions for new data points. The scikit-learn LinearRegression class provides a method called predict() that can be used to make predictions.

Linear regression is implemented in scikit-learn using the LinearRegression class. This class provides methods to fit a linear regression model to a training dataset and predict the target value for new data points.

2. **Explain the Anscombe's quartet in detail**

Ans.: Anscombe's quartet consists of 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.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

Anscombe's Quartet exhibits diverse patterns in scatter plots, illustrating the importance of visualizing data for meaningful insights beyond numerical summaries.

3. **What is Pearson's R?**

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. Pearson's r, r, or Pearson's Correlation is a measure of the correlation between two variables.

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

Ans.: Scaling is a process to change the data in such a way that the model can process it without any problems. And Feature Scaling is one such process in which we transform the data into a better version.

Skewed data and outliers can negatively impact the performance of machine learning models. Scaling the features can help in handling such cases. By transforming the data to a standardized range, it reduces the impact of extreme values and makes the model more robust.

In normalized scaling and standardized scaling, we are transforming the values of numeric variables so that the transformed data points have specific helpful properties. The difference is that: in scaling, we are changing the range of your data, while. in normalization, we are changing the shape of the distribution of data.

5. **Sometimes the value of VIF is infinite. Why does this happen?**

Ans.: The greater the VIF, the higher the degree of multicollinearity. So, if there is perfect correlation, then VIF is infinity. A large value of VIF indicates that there is a greater correlation between the variables.

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

Ans.: Q-Q plots are also known as Quantile-Quantile plots. they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

The advantages of the q-q plot are: The sample sizes do not need to be equal. Many distributional aspects can be simultaneously tested.

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.