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?

Some of the categorical variables have below effects

Season - Summer and Winter are positively correlated to cnt.
Season - Spring is negatively correlated to cnt

Yr - 2019 is correlated to cnt

Month - Dec, Jan, July , Nove are negatively correlated
WeatherSit - Mist, Snow, Sun are negatively correlated

```
----------------------
const          0.2552
holiday       -0.0997
temp           0.4353
windspeed     -0.1590
spring        -0.0702
summer         0.0339
winter         0.0920
2019           0.2342
Dec           -0.0469
Jan           -0.0517
July          -0.0474
Nove          -0.0432
Sept           0.0670
Sun           -0.0498
Mist          -0.0835
Snow          -0.2986
```

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

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.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished

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

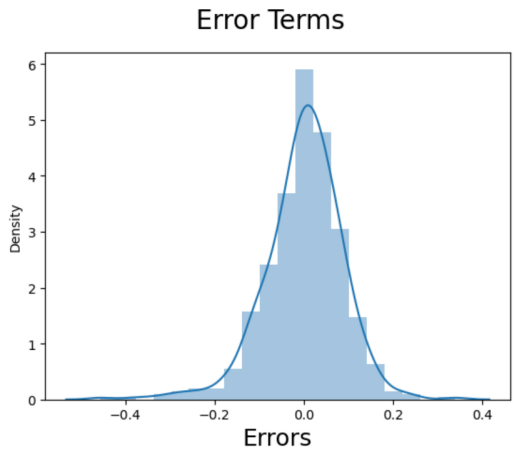
Temp variable has the highest correlation

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

No Multicolinearlity in the data

Each observation is unique

Error terms are normally distributed



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

temp , 2019 and snow are the top 3 features

const	0.2552
holiday	-0.0997
temp	0.4353
windspeed	-0.1590
spring	-0.0702
summer	0.0339
winter	0.0920
2019	0.2342
Dec	-0.0469
Jan	-0.0517
July	-0.0474
Nove	-0.0432
Sept	0.0670
Sun	-0.0498
Mist	-0.0835
Snow	-0.2986

General Subjective Questions

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

Linear regression is a statistical algorithm that is used to establish the relationship between two continuous variables. In simpler terms, it predicts the outcome of one variable based on the values of another.

Linear regression works by fitting a straight line through a set of data points to obtain a mathematical formula that predicts the value of one variable based on the value of the other. The formula for the line, known as the regression equation, is expressed by the equation: $Y = mX + b$

Where:

- Y is the dependent variable (the variable that we want to predict or analyze)
- X is the independent variable (the variable that is used to predict the dependent variable)
- m is the slope of the line that is fitted to the data
- b is the y-intercept of the line

The process of developing a linear regression model involves several steps:

1. Data collection: Collect data on the variables of interest. The data should be in a continuous format and should be representative of the population of interest.
2. Data preprocessing: Clean up the data to remove any outliers, missing values, or errors in the data.
3. Data exploration: Examine the relationship between the variables using graphical techniques like scatter plots, correlation plots, and histograms.
4. Model selection: Choose the best model that fits the data. This can be done by comparing different models like simple linear regression, multiple linear regression, and polynomial regression.
5. Model training: Train the model using the available data to find the best regression equation that accurately predicts the value of the dependent variable.
6. Model evaluation: Evaluate the performance of the model using statistical measures like the R-squared value, root mean squared error, and mean absolute error.
7. Model deployment: Use the trained model to make predictions on new data.

In summary, linear regression is a powerful algorithm for predicting the value of a dependent variable based on the value of an independent variable. It is widely used in various domains like finance, marketing, healthcare, and education to make data-driven decisions.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, including the same mean, standard deviation, correlation coefficient, and regression line. This means that if you were to analyze the data based purely on these statistical measures, you would come to the same conclusions for each dataset. However, a graphical analysis of each dataset reveals that they are vastly different and paint a completely different picture when it comes to the relationship between the variables.

The quartet was created by statistician Francis Anscombe in 1973 to underscore the importance of visualizing data and not relying solely on summary statistics. The four datasets consist of 11 (x,y) data points each and are as follows:

1. Dataset 1: Every x-coordinate is the same, and the y-coordinates increase linearly. This dataset has a strong positive linear relationship.
2. Dataset 2: Same as dataset 1, with one outlier. The relationship is still primarily linear.
3. Dataset 3: All the data points have the same x-coordinate, but the y-coordinate varies wildly. In this case, the relationship is nonlinear.
4. Dataset 4: This data set contains an obvious outlier, which influences the slope and correlation coefficient of the regression line. The relationship appears to be strong linear, but this conclusion is misleading due to the outlier.

The takeaway from Anscombe's quartet is that descriptive statistics alone are insufficient to fully understand the relationship that exists between two variables. Visual examination of the data is essential, as it allows researchers to identify outliers, unusual patterns, and other details that may be overlooked by summary statistics alone.

Additionally, Anscombe's quartet highlights the importance of careful data analysis and the potential for data to be misleading. It is crucial to approach data with skepticism, examine it from multiple angles, and recheck assumptions to ensure that the conclusions we draw are valid.

3. What is Pearson’s R?

Pearson’s R, also known as Pearson’s correlation coefficient or the Pearson product-moment correlation coefficient, is a statistical measure that describes the

strength and direction of the linear relationship between two continuous variables. The coefficient is denoted by the symbol r and ranges from -1 to 1.

A value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases in value, the other also increases proportionally. On the other hand, a value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases in value, the other decreases proportionally. A value of zero indicates no linear relationship between the two variables.

The Pearson correlation coefficient is widely used in different fields such as finance, social sciences, and health. It can be used to determine the correlation between any two continuous variables such as height and weight, age and income, interest rates and stock prices, or any other two variables that have a linear relationship.

In statistical analysis, Pearson's R is used to analyze and interpret data in a variety of ways, such as:

1. To determine the strength and direction of the relationship between two variables.
2. To test hypotheses regarding the correlation between two variables.
3. To test the reliability and validity of instruments and questionnaires used in research.
4. To control for the effects of one variable on another when conducting multiple regression analysis.

In summary, Pearson's R is a statistical measure that enables researchers to quantify the extent to which two variables are related, providing a useful tool for analyzing and interpreting data in many different fields.

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

Scaling is the process of transforming data to bring them onto a similar scale or range. It is performed to standardize the range and variance of different variables so that they can be more easily compared.

The main reasons for performing scaling are as follows:

1. To prevent features with large values from disproportionately influencing the model.

2. To make the computations faster and to reduce the chance of numerical errors that can occur with very large or very small values.
3. To help different features converge at similar rates in the optimization process, thereby improving the accuracy of the model.

There are two common scaling techniques:

1. Normalized Scaling: Normalization scales the values between 0 and 1. It is used when the scaling is required for optimizing the variables in machine learning algorithms, such as K-nearest neighbors, k-means clustering, and logistic regression. It is also known as Min-Max scaling.
2. Standardized Scaling: Standardization scales the data to have zero mean and unit variance. It is used when data is normally distributed and has outliers. Standardized scaling enables algorithms to perform better as it eliminates the outliers and produces a more consistent prediction.

The main difference between normalized scaling and standardized scaling is that normalized scaling transforms the data to a specific range, whereas standardized scaling transforms the data to have zero mean and unit variance. In normalized scaling, the minimum and maximum values are used to stretch or shrink the data range, whereas in standardized scaling, the mean and standard deviation are used to center and regulate the data. Standardized scaling is generally more appropriate when there are different scales to be adjusted for and when the distribution of data is known to be normal. On the other hand, normalized scaling is more appropriate when there is a need to adjust the range and prevent non-linear activation functions from saturating.

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

VIF is short for Variance Inflation Factor, a measure used in statistical analysis to detect multicollinearity where two or more predictor variables in a regression model are highly correlated with each other. It measures the degree of multicollinearity between the independent variables in a regression analysis.

A high VIF value indicates a high level of correlation between the independent variables and can negatively impact the accuracy of the regression model, leading to overfitting and reduced predictive power.

Theoretically, VIF values can range from 1 to infinity. A VIF of 1 indicates no correlation between the predictor variables, while a value of infinity means that the variables are

perfectly correlated. In practice, VIF values above 10 are considered problematic, with values above 5 indicating moderate correlation.

An infinite VIF can occur due to two main reasons:

1. Perfect multicollinearity: If there is perfect multicollinearity between variables, the VIF value will be infinite. Perfect multicollinearity occurs when two or more variables in the regression model are perfectly linearly related. For example, if we have two variables where one is the exact linear function of another variable, this leads to the perfect multicollinearity problem. In such cases, one of the variables should be removed from the model, or a different estimation technique should be used.
2. Zero variation in one of the variables: If one of the variables has zero variation, the VIF value for that variable will be undefined. In this case, the variable should be removed from the model.

In summary, an infinite VIF occurs when there is perfect multicollinearity between variables or when one variable has zero variation. To prevent and overcome these issues, it is important to carefully select variables to be included in the regression model, transforming variables when necessary, and to identify problems related to multicollinearity.

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

Q-Q plot is short for Quantile-Quantile plot, which is a graphical tool used to compare the distribution of a sample against a theoretical distribution to check whether they have the same distributional shapes. The Q-Q plot displays the quantiles of the sample data against the quantiles of the theoretical distribution, creating a scatter plot of data points.

The usefulness of the Q-Q plot in linear regression arises from the need to check the assumptions made regarding the distribution of the residuals, which is a crucial aspect of linear regression. The residuals are the differences between the observed values and the predicted values of the dependent variable. A Q-Q plot can be generated to visually examine whether the distribution of the residuals follow a normal distribution or not.

The use and importance of a Q-Q plot in linear regression can be explained as follows:

1. Checking for normal distribution: The Q-Q plot is used to check whether the distribution of the residuals in the regression model is normal or not. A normal distribution of residuals indicates that the model assumptions are valid, and the model is a good fit for the data. On the other hand, a non-normal distribution of residuals indicates that the model assumptions are not valid, and the model may require further modifications.
2. Identifying potential outliers: The Q-Q plot also helps to identify potential outliers in the data. Outliers are data points that are significantly different from the other data points, and they can significantly impact the regression analysis. Q-Q plots can help to identify any potential outliers in the residuals, which can be further investigated to identify the cause of their occurrence.
3. Assessing model fit: Finally, the Q-Q plots can be used to assess the goodness of fit of the model to the data. The plot helps to provide a visual indication of whether the data points follow a known distribution, such as the normal distribution. A better fit of the model to the data means that the data points fall closer to the straight line.

In summary, the Q-Q plot is a useful and important tool in linear regression since it helps to validate the normality assumption, detect outliers, and visually assess the goodness of fit of the regression model to the data. It provides a simple and effective way of checking the accuracy and reliability of the regression analysis.