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?

From analysis of independent categorical variable with the dependant variable, the following was observed:

- Summer and fall season have more bike users.
- More bikes are used on non holidays.
- more users on clear weather
- june to september have more users.
- The users have increased in 2019 compared to 2018
- Weekdays doesn't play much importance with the target variable as count is almost similar for all days.

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

On creation of dummy variable, a sparse matrix is created for all the categories present in the particu lar column, say there are 4 categories in a feature A,B,C,D if sparse matrix is created for all 4 the m ulticollinearity increases, which can lead to overfitting of the model. And by dropping the first it aut omatically will take that dropped category also into consideration for eg: A,B,C,D are 4 categories a nd sparse matrix is created for all B,C,D as sparse matrix consist of only 1 and 0 if all the three cate gories B,C and D are 0 then automatically it says the category is A.

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

Temp and atemp has the highest correlation with the target variable. Since temp and Atemp have hig h positive correlation to remove multicollinearity only temp was considered for modeling.

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

The assumptions were validated by the following methods:

- Normality of errors: Error terms are normally distributed
- Linear relationship between independent and dependent variables
- Homoscedasticity: similar variance
- Multicollinearity: very less multicollinearity among variables

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

The top 3 features are:

- Temperature
- Season
- Windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical algorithm that aims to find the linear relationship between a dependent variable (y) and one or more independent variables (x) by fitting a straight line through the data. The equation for a simple linear regression model with one independent variable is:

$y = \beta0 + \beta1x + \varepsilon$

where y is the dependent variable, x is the independent variable, β0 and β1 are the coefficients of the regression line, and ε is the error term that accounts for the randomness in the data that cannot be explained by the model.

The goal of linear regression is to estimate the values of β0 and β1 that minimize the sum of squared errors between the actual and predicted values of the dependent variable. This is done by finding the values of β0 and β1 that minimize the following objective function:

$SSE = \Sigma(yi - \beta0 - \beta1xi)2$

where SSE is the sum of squared errors, yi is the actual value of the dependent variable for the ith observation, and xi is the value of the independent variable for the ith observation.

To estimate the values of β0 and β1, we can use the least squares method, which consists in finding the values of β0 and β1 that minimize the SSE. This can be done using the infinitesimal calculus, and the resulting formulas for the estimates of β0 and β1 are:

$\beta1 = \Sigma(xi - \bar{x})(yi - \bar{y}) / \Sigma(xi - \bar{x})2$

$\beta0 = \bar{y} - \beta1\bar{x}$

where $\bar{x}$ and $\bar{y}$ are the mean values of the independent and dependent variables, respectively.

Once we have estimated the values of β0 and β1, we can use them to predict the value of the dependent variable for new values of the independent variables. For example, if we want to predict the value of y for a new value of x, we can use the following formula:

$\hat{y} = \beta 0 + \beta 1x$

where $\hat{y}$ is the predicted value of the dependent variable.

Linear regression can be extended to multiple independent variables using multiple linear regression. The equation for multiple linear regression with p independent variables is:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp + \varepsilon$

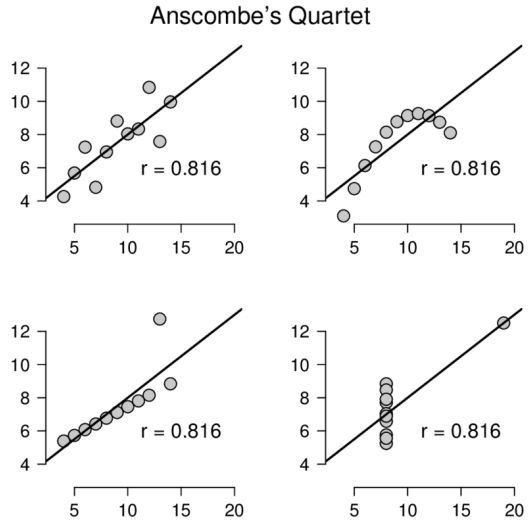
where x1, x2, ..., xp are the p independent variables, and β0, β1, β2, ..., βp are the coefficients of the regression equation. The least squares method can be used to estimate the values of β0, β1, β2, ..., βp, and the resulting regression equation can be used to make predictions for new values of the independent variables.

Linear regression is widely used in various fields such as economics, finance, social science and engineering to model and analyze the relationship between variables and make predictions based on the model.

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

The Anscombe quartet is a set of four data sets that have nearly identical statistical properties but appear graphically very different.

Each of the four data sets contains 11 points (x, y) and has nearly identical statistical properties in terms of means, variances, correlation coefficients, and linear regression lines. However, when graphed, they show very different patterns and trends, ranging from a linear relationship to more complex patterns



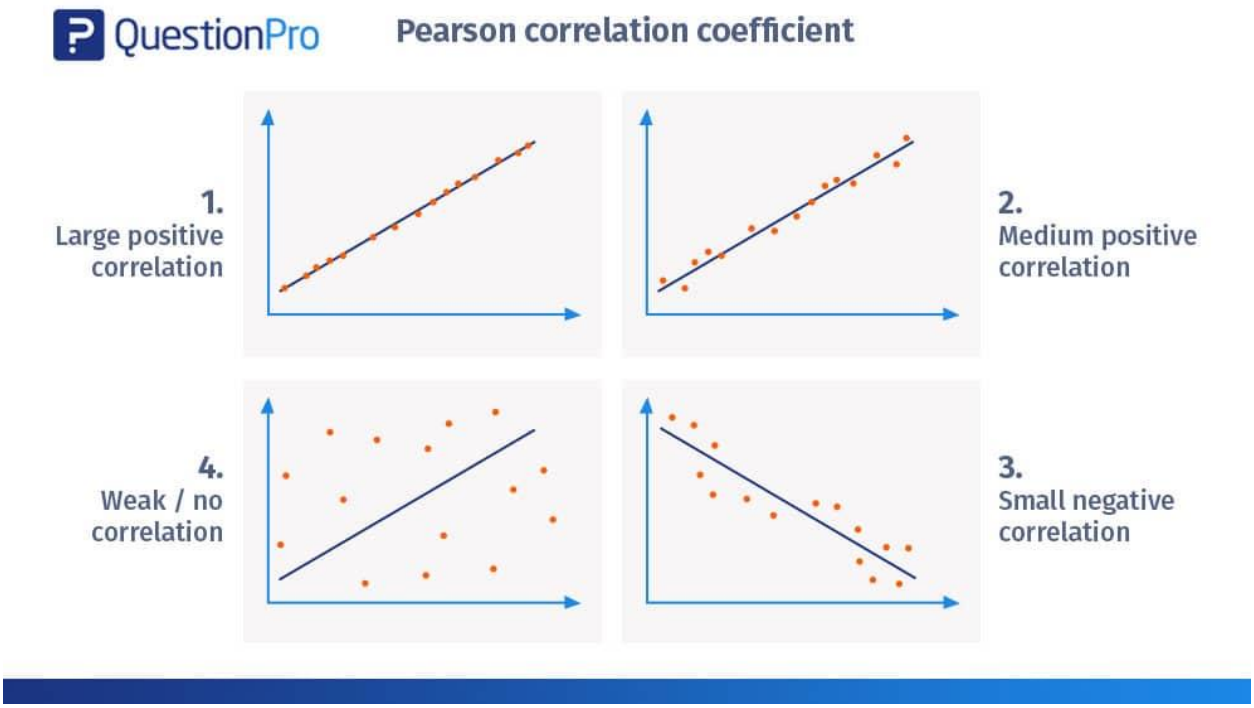
with outliers. The Anscombe Quartet is intended to illustrate that numerical summaries of data are sometimes insufficient to understand the underlying relationships and patterns in the data. By representing the data graphically, it is easier to see the nature of the relationships and identify any outliers or unusual patterns.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient or simply correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two variables. It is denoted by the symbol "r" and can take values between -1 and 1.

A value of -1 means a perfect negative linear relationship, 0 means no linear relationship, and 1 means a perfect positive linear relationship between the variables. Pearson's R is calculated by dividing the covariance between two variables by the product of their standard deviations.

Pearson's R is commonly used in fields such as psychology, economics, and biology to measure the strength of the relationship between two variables. It is an important tool for researchers to assess whether there is a relationship between variables and how strong that relationship is. However, it should be noted that Pearson's R only measures the strength of linear relationships and may not be appropriate for non-linear relationships.



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

Scaling is the process of changing data to conform to a specific format or distribution. The scaling process depends on the specific requirements of the data analysis and the characteristics of the data.

Scaling is done for many reasons, for example:

- Standardization: scaling can be used to transform data so that it falls within a specific range, such as 0 to 1 or -1 to 1. Data analysis techniques that require parallel data, such as distance-based algorithms in machine learning.
- Normalize distribution: Scaling can be used to transform a data distribution so that it follows a normal distribution.This is useful for some statistical analyzes where the data is assumed to be normally distributed.
- Reducing the effects of outliers: Scaling can be used to reduce the impact of very important outliers that can affect the results of some analyzes. By scaling the data, the effects of outliers can be minimized.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized scaling, also known as min-max scaling, scales data so that it falls within a certain range, usually between 0 and 1. The formula for normalized scaling is:

scaled_value = (value - min) / (max - min)

where min and max are the minimum and maximum values in the data set, respectively.

In standardized scaling, also known as z-score scaling, the data is scaled to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

scaled_value = (value - mean) / standard deviation

where mean is the mean value of the data set and standard_deviation is the standard deviation of the data set.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the original distribution of the data, while standardized scaling transforms the data into a normal distribution with a mean of 0 and a standard deviation of 1.

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

Sometimes the VIF value can be infinite. This is the case when one or more variables in the regression analysis are perfect, that is, they can be expressed as a combination of other variables in the model. In this case, the VIF values of the linear variables are infinite.When perfect collinearity occurs in a regression analysis, the coefficient estimates for the colinear variables are not exact because each linear combination of variables is the same. This can lead to problems in interpreting the results and makes it difficult to draw conclusions from the analysis.

To avoid missing VIF values, it is important to check for multiple variables and remove significant variables before performing the regression analysis. This can be done using techniques such as correlation analysis or calculating VIFs for each variable in the sample.

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

The Q-Q plot (quantile-quantile plot) is a graphical technique used to measure whether a data set follows a distribution, such as a normal distribution. It is a practical function that compares the data sets with the sets of the theoretical distribution.

For Q-Q plots, the observed value of the data set is plotted on the x-axis and the expected theoretical distribution amount is plotted on the y-axis. If the data set follows a theoretical distribution, the points on the Q-Q plot will be distributed along a straight line. A deviation from the straight line means a deviation from the theoretical distribution.

In linear regression, Q-Q plots can be used to test the normality and covariance assumptions. The normality assumption assumes that the residuals (the difference between the estimated and actual values) are normally distributed. The covariance assumption assumes that the variance of the residuals is constant across all values of the predictor variable.

A non-normal distribution of the residuals is an indication that the linear regression model is not appropriate for the data and a different modeling approach may be needed. Similarly, if the residuals have unequal variances

For example, this may indicate that a linear regression model is not appropriate because the model assumes that the variance of the residuals is constant across all values of the estimator.Q-Q plot can be used to test whether the residuals follow a normal distribution and whether the variance of the residuals is constant. If the residuals fall along a straight line on the Q-Q plot, the residuals are normally distributed. The nonlinearity of the points on the Q-Q plot indicates that the variance of the residuals is not constant (heteroskedasticity).

In summary, Q-Q plots are important tools for testing normality and covariance assumptions in linear regression. They can help identify differences between these hypotheses and suggest changes to the linear regression model to improve fit to the data.