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?**

Answer:

Based on the analysis of the categorical variables from the dataset, the following inferences can be made about their effect on the dependent variable:

- The **year** appears to have a significant impact on bike demand. There was a noticeable increase in bookings from 208 to 209, indicating a positive trend in business growth.
- The **season** also influences bike demand. The fall season, in particular, seems to attract more bookings. Furthermore, there is a significant increase in bookings from the start of the year until mid-year (May to October), after which the trend starts to decrease.
- **Weather conditions** play a crucial role in bike demand. Clear weather conditions tend to attract more bookings, which is an expected observation.
- The **day of the week** has an effect on bike demand. Thursdays, Fridays, Saturdays, and Sundays have a higher number of bookings compared to the start of the week.
- **Holidays** also influence bike demand. There are fewer bookings on holidays, which could be attributed to people preferring to spend time at home with family.
- The demand for bikes does not seem to differ significantly between working days and non-working days.
- Other factors such as **temperature**, **wind speed**, and specific **months** (December, January, November, September) and **weather types** (Snowy, Misty) also appear to influence bike demand.

Conclusion: These observations provide valuable insights into the factors that influence bike demand and can be used to make informed business decisions and strategies.

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

Answer:

The drop_first=True parameter in dummy variable creation is used to avoid the issue of multicollinearity, which can occur when one variable can be perfectly predicted by another. In the context of dummy variables, if we have n categories for a feature and we create n dummy variables for them, we actually introduce perfect multicollinearity. This is because the value of one dummy variable can be perfectly predicted from the others. For example, if we have a feature 'color' with categories 'red', 'blue', and 'green', and we create three dummy variables for them, knowing the values of two will automatically tell us the value of the third. This perfect multicollinearity is problematic because it can:

- Make the coefficients/weights unstable and their interpretation misleading.

- Lead to redundancy as we are essentially adding an extra variable that doesn't provide new information.

By using drop_first=True, we drop the first dummy variable, reducing the n dummy variables to n-, which eliminates this issue of multicollinearity. The dropped category's information is not lost but is rather used as the reference category against which the others are compared. For instance, in the 'color' example, if we drop 'red', a '0' in both 'blue' and 'green' would indicate 'red'.

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

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

Answer:

After building the linear regression model on the training set, I validated the assumptions of the model in the following ways:

1. **Normality of Residuals**: I checked whether the residuals (error terms) are normally distributed. This is a fundamental assumption of linear regression and can be validated using methods like Q-Q plots or statistical tests like the Shapiro-Wilk test.
2. **Absence of Multicollinearity**: I verified that there is no significant multicollinearity among the predictor variables. Multicollinearity can inflate the variance of the regression coefficients, making them unstable. Tools like the Variance Inflation Factor (VIF) can help detect multicollinearity.
3. **Linearity**: I confirmed that there is a linear relationship between the independent and dependent variables. This can be visually checked using scatter plots.
4. **Homoscedasticity**: I ensured that the residuals have constant variance at every level of the predictor variables. This is known as homoscedasticity. If the plot of residuals versus predicted values shows a funnel shape, it indicates heteroscedasticity, violating this assumption.
5. **Independence of Residuals**: I checked for autocorrelation in the residuals, which means that the residuals are not independent of each other. The Durbin-Watson test is commonly used to detect autocorrelation.

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

According to the final model, the top 3 features that contribute significantly towards explaining the demand of shared bikes are:

1)temp 2)weathersit_Light Snow/Rain 3)year

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical model that examines the linear relationship between a dependent variable and a set of independent variables. This relationship is linear, meaning that changes in the value of one or more independent variables will correspondingly change the value of the dependent variable.

The mathematical representation of this relationship is given by the equation:

$Y = mX + c$

Here:

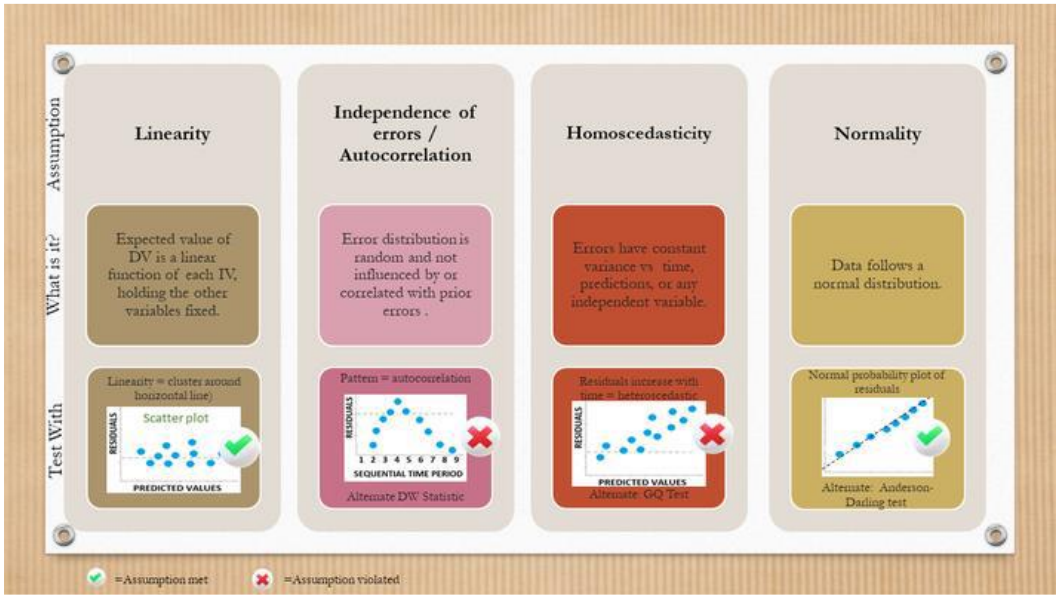
- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is the slope of the regression line, representing the effect X has on Y.
- c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

The linear relationship can be either positive or negative:

- Positive Linear Relationship: Both the independent and dependent variables increase.
- Negative Linear Relationship: The independent variable increases while the dependent variable decreases.

Linear regression can be categorized into two types:

1. Simple Linear Regression
2. Multiple Linear Regression

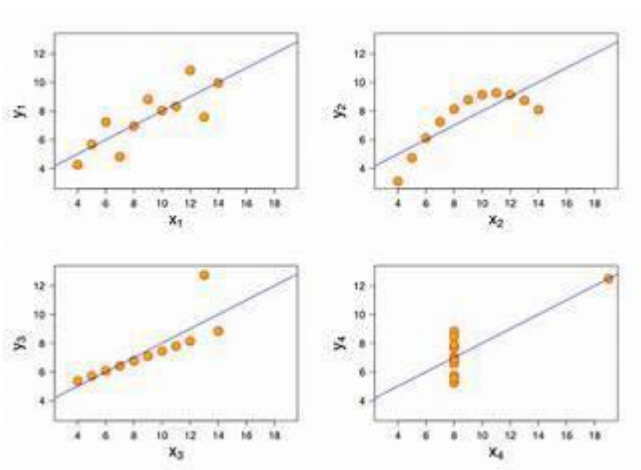


The Linear Regression model makes the following assumptions about the dataset:

- Multi-collinearity: There is little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables or features have dependencies among them.
- Auto-correlation: There is little or no auto-correlation in the data. Auto-correlation occurs when there is a dependency between residual errors.
- Relationship between variables: The relationship between response and feature variables is linear.
- Normality of error terms: Error terms should be normally distributed.
- Homoscedasticity: There should be no visible pattern in residual values.

These assumptions help ensure the validity and reliability of the linear regression model.

2. Explain the Anscombe's quartet in detail.
Answer:
Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties . The quartet is still often 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 following table summarizes the statistical properties of all four datasets:

Property	Value
Mean of x	9
Sample variance of x	
Mean of y	7.50
Sample variance of y	4.25
Correlation between x and y	0.86
Linear regression line y = a + bx	y = 3.00 + 0.500x
Coefficient of determination of the linear regression	R^2 = 0.67

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modeled as Gaussian with mean linearly dependent on x. For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate. In the third graph (bottom left), the modeled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from to 0.86. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables .

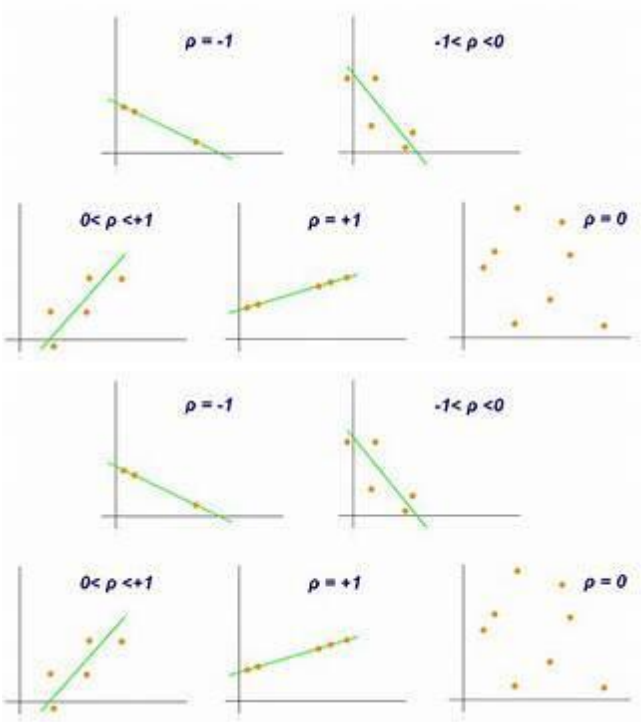
3. What is Pearson's R?

Answer:

The Pearson correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. It is a number between - and that measures the degree to which two variables are linearly related . A value of indicates a perfect positive correlation, while a value of - indicates a perfect negative correlation. A value of 0 indicates no correlation between the two variables .

The formula for calculating the Pearson correlation coefficient is:

r = (nΣxy - ΣxΣy) / sqrt((nΣx^2 - (Σx)2)(nΣy2 - (Σy)^2))



where x and y are the two variables, n is the number of observations, Σxy is the sum of the product of the deviations of x and y from their respective means, Σx and Σy are the sums of the deviations of x and y from their respective means, and Σx^2 and Σy^2 are the sums of the squares of the deviations of x and y from their respective means .

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

Answer:

Scaling is the process of transforming data so that it falls within a specific range. Scaling is performed to ensure that all features contribute equally to the analysis and to prevent features with larger values from dominating the analysis .

Normalization and standardization are two common scaling techniques. The main difference between normalized scaling and standardized scaling is that the values of a normalized dataset will always fall between 0 and 1, while a standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds .

The following table summarizes the differences between normalized scaling and standardized scaling:

Property	Normalized Scaling	Standardized Scaling
Range of values	0 to 1 or -1 to 1	Not bounded to a certain range
Suitable for	Algorithms that don't make any assumptions about the distribution of the data	Algorithms that create predictions about the data distribution
Formula	(x - min) / (max - min)	(x - mean) / standard deviation
Affected by outliers	Highly affected	Much less affected
Scikit-Learn transformer	MinMaxScaler	StandardScaler

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

Answer:

Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity among predictor variables.

When the VIF value becomes infinite, it usually indicates an extreme case of perfect multicollinearity. Perfect multicollinearity occurs when one or more predictor variables in a regression model are perfectly linearly related to other variables. This perfect correlation causes issues with the estimation of the regression coefficients.

Mathematically, an infinite VIF occurs when the correlation between one or more predictor variables is perfect or extremely close to perfect (e.g., if one variable can be expressed as an exact linear combination of others). As a result, the regression model cannot distinguish the individual effects of these perfectly correlated variables, leading to an inability to compute accurate coefficient estimates and causing the VIF to become infinite.

In practice, encountering infinite VIF values is a signal to investigate and potentially address multicollinearity issues in the dataset by considering variable selection techniques or combining correlated variables to mitigate the problem before performing regression analysis.

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

Answer:

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a certain theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset's empirical distribution against the quantiles of a specified theoretical distribution. Theoretical quantiles (from the chosen distribution) are plotted on the x-axis, and the corresponding quantiles from the actual data are plotted on the y-axis.

In linear regression, Q-Q plots serve several important purposes:

1. Normality Assumption Checking: Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot of the residuals is used to visually inspect if the residuals follow a normal distribution. If the points on the plot roughly align along a straight line, it suggests that the residuals are approximately normally distributed.

2. Identifying Outliers and Skewed Data: Deviations from the straight diagonal line in a Q-Q plot can indicate outliers or departures from the assumed distribution. Skewed data or outliers might appear as deviations from the expected linearity in the Q-Q plot, pointing to potential issues in the regression analysis that might need attention.
3. Model Validity Check: A Q-Q plot aids in validating the assumptions of the regression model. If the residuals violate the normality assumption, it might suggest problems with the model's validity, and corrective actions like transformation of variables or using robust regression techniques might be necessary.
4. Comparing Different Distributions: Besides testing for normality, Q-Q plots can also compare the empirical distribution against other theoretical distributions, helping to select an appropriate distribution for the data.
The Q-Q plot is a crucial diagnostic tool in linear regression analysis, enabling researchers and analysts to visually assess the goodness-of-fit of the model by examining the distributional assumptions of the residuals. Identifying departures from expected patterns assists in improving the model's accuracy and reliability by addressing potential issues related to normality, outliers, and the overall validity of the regression analysis.