Assignment-based Subjective Questions

Document by – Vishwanath G Patil

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

Analysis on categorical columns was performed using the box plot and bar plot. Below are few inferences from the visualization –

- Fall seems to have attracted more bookings. Furthermore, with each season there has been a significant increase in the number of bookings from 2018 to 2019.
- Higher number of bookings is recorded during the month of May, june, july, aug, sep and oct. Increasing trend was observed at the beginning of the year till mid of the year and then it is seen decreasing as we approached the end of year.
- Clear weather attracted more bookings, which seems obvious.
- In a particular week Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
- On a holiday, a smaller number of bookings were observed, which seems reasonable as on holidays, people may want to spend time at home and enjoy themselves with family.
- Booking seemed to be almost equal either on working day or non-working day.
- Comparatively the year 2019 attracted a greater number of bookings than the previous year, which shows good progress in terms of business.

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

Answer:

When creating dummy variables, **`drop_first=True`** is often used to prevent multicollinearity, a situation in which two or more predictor variables are highly correlated with each other.

Multicollinearity can cause problems in statistical models, such as overfitting, unstable parameter estimates, and difficulty in interpreting the importance of individual predictor variables.

By setting **`drop_first=True`**, one level of each categorical variable is omitted, and the remaining levels are represented by the binary dummy variables. This means that the dropped level becomes the reference category, and the interpretation of the coefficients of the dummy variables becomes more straightforward.

For example, suppose we have a categorical variable "color" with three levels: red, green, and blue. If we create dummy variables without dropping a level, we would have three binary variables, with "red" being the reference category. This would result in the model having perfect multicollinearity, where the sum of the three dummy variables always equals one.

However, if we drop the first level (red), we would only have two binary variables (green and blue), which would avoid perfect multicollinearity.

Therefore, using **`drop_first=True`** during dummy variable creation can help to prevent multicollinearity and improve the accuracy and interpretability of the statistical models.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

'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? (3 marks)

Answer:

I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- ✓ Normality of error terms
  - o Error terms should be normally distributed
- ✓ Multicollinearity check
  - o There should be insignificant multicollinearity among variables.
- ✓ Linear relationship validation
  - o Linearity should be visible among variables
- ✓ Homoscedasticity
  - o There should be no visible pattern in residual values.
- ✓ Independence of residuals
  - o No autocorrelation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks) Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- ➢ temp
- ➢ winter
- ➢ sep

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find a linear equation that best predicts the value of the dependent variable based on the values of the independent variables.

The equation for a simple linear regression model with one independent variable is:

$y = \beta0 + \beta1*x + \varepsilon$

where y is the dependent variable, x is the independent variable, β0 is the intercept, β1 is the slope, and ε is the error term.

The equation for a multiple linear regression model with n independent variables is:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta n*xn + \varepsilon$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, β0 is the intercept, β1, β2, ..., βn are the slopes, and ε is the error term.

The linear regression algorithm estimates the values of β0, β1, β2, ..., βn that minimize the sum of squared errors (SSE) between the predicted values and the actual values of the dependent variable in the training data. The SSE is calculated as:

$SSE = \Sigma(yi - \hat{y}i)^2$

where yi is the actual value of the dependent variable, $\hat{y}i$ is the predicted value of the dependent variable, and Σ is the sum across all observations in the training data.

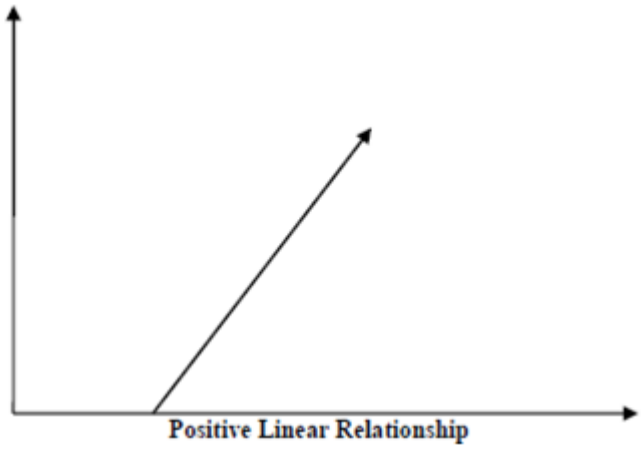
To estimate the values of β0, β1, β2, ..., βn that minimize the SSE, the algorithm uses a technique called ordinary least squares (OLS) regression. OLS regression finds the values of β0, β1, β2, ..., βn that minimize the SSE by taking the partial derivatives of SSE with respect to each coefficient and setting them equal to zero.

Once the values of β0, β1, β2, ..., βn are estimated, the linear regression model can be used to make predictions on new data by plugging in the values of the independent variables and solving for the dependent variable using the estimated coefficients.

Linear regression is a simple and powerful method for modeling the relationship between a dependent variable and one or more independent variables. However, it assumes that the relationship between the dependent variable and the independent variables is linear and that the error term is normally distributed and has constant variance. These assumptions should be checked and validated before using linear regression for prediction or inference.

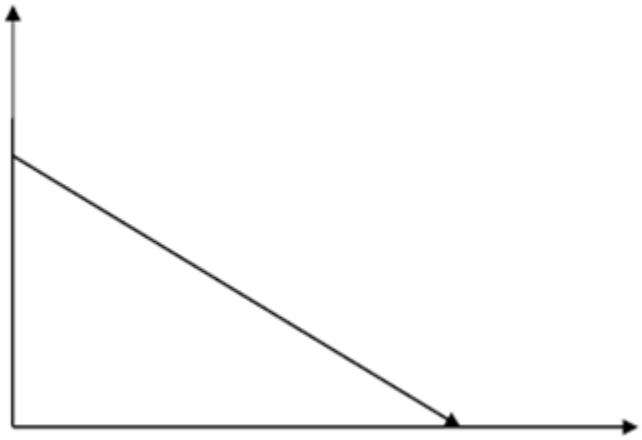
Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship:
  - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

- Negative Linear relationship:
  - A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

✓ multi-collinearity –

- Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

✓ Auto-correlation –

- Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

✓ Relationship between variables –

- Linear regression model assumes that the relationship between response and feature variables must be linear.

✓ Normality of error terms –

- Error terms should be normally distributed

✓ Homoscedasticity –

- There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties but exhibit vastly different patterns when plotted. The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in statistical analysis.

Each dataset in the quartet consists of 11 (x, y) pairs. The four datasets have the same mean, variance, correlation coefficient, and linear regression line, but they have different patterns of points when plotted.

Dataset I: This dataset has a simple linear relationship between x and y, with a clear positive correlation.

Dataset II: This dataset also has a linear relationship between x and y, but with a clear outlier that exerts a strong influence on the linear regression line.

Dataset III: This dataset has no apparent linear relationship between x and y, but has a strong quadratic relationship when x is squared.

Dataset IV: This dataset has a strong linear relationship between x and y but is heavily influenced by an outlier that is not visible on the plot.

The purpose of the quartet is to demonstrate the importance of visualizing data before performing statistical analysis. The datasets have identical statistical properties but have vastly different patterns when plotted. A statistical analysis that is performed without visualizing the data may overlook important patterns and relationships.

The quartet also illustrates the importance of understanding the limitations of summary statistics such as mean, variance, and correlation coefficient. These statistics can be useful in describing the properties of a dataset, but they may not provide a complete understanding of the underlying patterns and relationships.

In summary, Anscombe's quartet is a powerful example of the importance of data visualization and understanding the limitations of summary statistics in statistical analysis.

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure of the linear relationship between two variables. It is denoted by the symbol "r" and ranges from -1 to 1, where 1 indicates a perfect positive correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation.

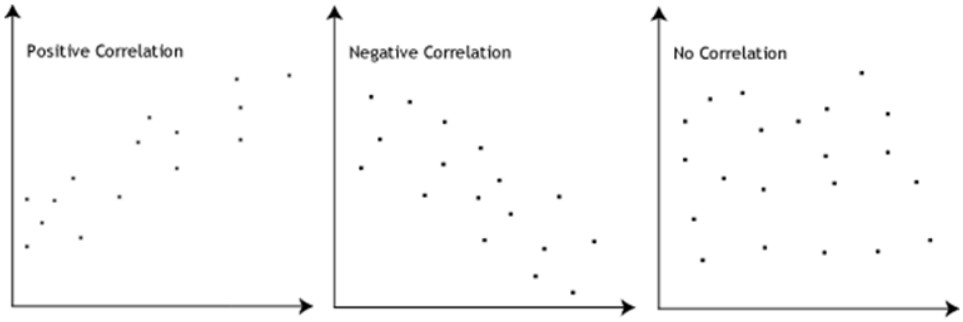
The formula for Pearson's R is:

r = (Σ((xi - x_mean)*(yi - y_mean))) / ((n-1)*s_x*s_y)

where xi and yi are the individual values of the two variables, x_mean and y_mean are their respective means, n is the number of observations, and s_x and s_y are the standard deviations of the two variables.

The value of Pearson's R indicates the strength and direction of the linear relationship between the two variables. A positive value of r indicates a positive linear relationship, meaning that as one variable increases, the other variable also tends to increase. A negative value of r indicates

a negative linear relationship, meaning that as one variable increases, the other variable tends to decrease. A value of 0 indicates no linear relationship between the two variables.



Pearson's R is widely used in statistics and data analysis to measure the strength and direction of the relationship between two variables. However, it assumes that the relationship between the two variables is linear, and that the data are normally distributed. If these assumptions are not met, other correlation measures such as Spearman's rank correlation or Kendall's tau may be more appropriate.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is the process of transforming the values of variables to a common scale, so that they can be compared on equal grounds. It is an important step in data preprocessing that is often performed before applying machine learning algorithms.

Scaling is performed to bring all the variables to the same scale, which is important when variables have different units or ranges of values. If variables are not on the same scale, some variables may dominate over others, which can lead to bias in the model.

There are two commonly used types of scaling: normalized scaling and standardized scaling.

Normalized scaling, also known as min-max scaling, rescales the values of variables to a range between 0 and 1. The formula for normalized scaling is:

x_norm = (x - min(x)) / (max(x) - min(x))

where x is the original value of the variable, min(x) is the minimum value of the variable, and max(x) is the maximum value of the variable.

Standardized scaling, also known as z-score normalization, rescales the values of variables to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

x_std = (x - mean(x)) / std(x)

where x is the original value of the variable, mean(x) is the mean of the variable, and std(x) is the standard deviation of the variable.

The main difference between normalized scaling and standardized scaling is the scale to which the variables are transformed. Normalized scaling rescales variables to a range between 0 and 1, while standardized scaling rescales variables to have a mean of 0 and a standard deviation of 1. In normalized scaling, the minimum and maximum values of the variable are used in the transformation, while in standardized scaling, the mean and standard deviation of the variable are used.

In summary, scaling is an important step in data preprocessing that is performed to bring variables to a common scale. Normalized scaling and standardized scaling are two commonly used types of scaling that rescale variables to different scales. Normalized scaling rescales variables to a range between 0 and 1, while standardized scaling rescales variables to have 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?
(3 marks)

Answer:
The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables in a regression model. It quantifies how much the variance of the estimated regression coefficient is inflated due to collinearity among the predictor variables.

In some cases, the value of VIF can be infinite, which indicates perfect multicollinearity among the predictor variables. This happens when one or more of the predictor variables can be perfectly predicted by a linear combination of the other predictor variables. In other words, one or more of the predictor variables is a linear combination of the other predictor variables, so they provide no additional information to the model.

When there is perfect multicollinearity among the predictor variables, the regression coefficients cannot be estimated using standard regression techniques. This is because there is not a unique set of regression coefficients that can explain the relationship between the response variable and the predictor variables.

To identify the variables causing perfect multicollinearity and to resolve the issue, one can examine the correlation matrix among the predictor variables and remove the variables that are highly correlated with each other. Another approach is to use regularization techniques such as Ridge Regression or Lasso Regression that penalize the regression coefficients to reduce the effect of multicollinearity.

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

Answer:
A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to compare the distribution of a sample of data to a known probability distribution, such as a normal distribution. The Q-Q plot is a scatter plot of the quantiles of the sample data against the quantiles of the theoretical distribution. If the sample data follows the theoretical distribution, the points on the Q-Q plot will fall approximately along a straight line.

In linear regression, Q-Q plots are used to check the assumption of normality of residuals. The residuals are the differences between the observed values and the predicted values of the response variable. The assumption of normality of residuals is important because if the residuals are not normally distributed, it may indicate that the model is not capturing all the relevant factors that affect the response variable or that there is heteroscedasticity in the data.

To create a Q-Q plot of the residuals, we first standardize the residuals by subtracting their mean and dividing them by their standard deviation. We then plot the standardized residuals against the expected quantiles of the normal distribution. If the points on the Q-Q plot fall along a straight line, it indicates that the residuals are normally distributed. If the points deviate significantly from a straight line, it may indicate non-normality of residuals.

The Q-Q plot is a useful tool in linear regression because it provides a visual assessment of the assumption of normality of residuals. If the assumption is violated, we can take corrective measures such as transforming the data or using a different model. The Q-Q plot can also help us to identify outliers or other patterns in the residuals that may require further investigation.