Assignment-based Subjective Questions

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

Solution:

To analyze the effect of categorical variables on the dependent variable, we can use statistical techniques such as chi-square tests, contingency tables, and logistic regression. The specific method we choose will depend on the nature of the data :

Here are some general steps we follow to analyze the effect of categorical variables:

1. Identified the categorical variables in our dataset and determined the categories for each variable.

2. Calculated the chi-square test statistic to determine if there is a statistically significant relationship between the categorical variables and the dependent variable. The chi-square test measures the difference between the observed frequencies in the contingency table and the expected frequencies if the variables were independent. If the chi-square test statistic is significant, it indicates that the variables are not independent and there is an association between them.

3. Calculated the effect size to determine the strength of the relationship between the categorical variables and the dependent variable. Cohen's d or Cramer's V are commonly used effect size measures for categorical variables.

4. Used logistic regression to model the relationship between the categorical variables and the dependent variable. Logistic regression is a statistical technique that models the probability of an event occurring given the values of the independent variables. In this case, the dependent variable is binary (e.g., yes/no), and the independent variables are categorical. The logistic regression model can provide information on the strength and direction of the relationship between the variables, as well as the significance of individual categories within each variable.

The inference about the effect of the categorical variables on the dependent variable will depend on the results of the analysis. If the chi-square test is significant and the effect size is large, it indicates a strong association between the categorical variables and the dependent variable. If the logistic regression model shows a significant effect for one or more categories within a variable, it indicates that those categories have a significant effect on the probability of the dependent variable occurring. The direction of the effect (positive or negative) will depend on the coding of the categorical variable and the specific research question.

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

Solution:

When creating dummy variables from categorical variables, it is important to use drop_first=True to avoid multicollinearity in the model.

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other. This can cause problems in the regression analysis, such as unstable coefficients and inflated standard errors, which can lead to inaccurate predictions and incorrect conclusions.

Using drop_first=True when creating dummy variables helps to avoid multicollinearity by removing one of the dummy variables for each categorical variable. This means that each category is compared to a reference category, rather than being directly compared to each other. For example, if you have a categorical variable with three categories (A, B, and C), using drop_first=True will create two dummy variables (B and C) and use A as the reference category. This ensures that the model does not compare category B directly to category C, which could cause multicollinearity.

In addition, dropping the first category can also simplify the interpretation of the coefficients in the model. The coefficient for each dummy variable represents the difference in the mean of the dependent variable between that category and the reference category. By dropping the reference category, the intercept represents the mean of the dependent variable for the reference category, and the coefficients for the dummy variables represent the differences in means between the other categories and the reference category.

Overall, using drop_first=True when creating dummy variables is important for avoiding multicollinearity and simplifying the interpretation of the coefficients in the model.

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

Solution:

By analysing the data from the histogram we can conclude that 'registered' column has highest correlation with target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Solution:

After building a linear regression model on the training set, it is important to validate the assumptions of linear regression to ensure that the model is reliable and accurate. Here are the steps we have followed to validate the assumptions of linear regression:

1. Linearity: Checked for linearity by plotting the actual values of the dependent variable against the predicted values from the model. The plot should shows a linear relationship between the two variables.

2. Normality: Checked for normality by plotting a histogram of the residuals (the differences between the actual values and predicted values). The histogram is roughly symmetric and bell-shaped. we also has a normal probability plot to check for normality, which shows the points following a straight line.

3. Homoscedasticity: Checked for homoscedasticity by plotting the residuals against the predicted values. The plot shows no pattern or trend, and the spread of the residuals is roughly constant across all values of the predicted variable.

4. Independence: Checked for independence by plotting the residuals against the order of the observations in the dataset. There is no pattern or trend in the plot, indicating that the residuals are not dependent on the order of the observations.

5. Outliers: Checked for outliers by examining the residuals plot and identifying any points that are far away from the rest of the data. we can also use diagnostic plots such as Cook's distance or leverage plot to identify influential data points.

If any of these assumptions are violated, we may need to adjust the model or the data to ensure that the assumptions are met.

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

Solution:

Following are the columns that are contributing significantly towards explaining the demand of the shared bikes:

- registered: count of registered users
- casual: count of casual users
- instant: record index
- atemp: feeling temperature in Celsius

General Subjective Questions

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

Linear regression is a statistical modeling technique that aims to establish a relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as the predictor or explanatory variables). The goal of the linear regression algorithm is to estimate the parameters of a linear equation that best describes the relationship between these variables.

The linear regression algorithm can be represented mathematically as:

$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta n*Xn$

where Y is the dependent variable, β0 is the intercept or constant term, β1 to βn are the coefficients or parameters of the independent variables X1 to Xn respectively.

To obtain the best estimates of these coefficients, we use a technique called ordinary least squares (OLS). In OLS, we minimize the sum of the squared residuals, which is the difference between the predicted value of Y and the actual value of Y for each observation in the data set. The objective is to find the values of the coefficients that minimize the sum of the squared residuals.

The steps involved in the linear regression algorithm are as follows:

1. Data preparation: Gather the data for the dependent and independent variables. Ensure that the data is complete, accurate, and appropriate for the analysis.

2. Data exploration: Explore the data to identify any patterns, trends, or outliers. Identify the relationship between the dependent variable and the independent variables using scatterplots or correlation matrices.

3. Model development: Select the appropriate type of linear regression model (simple or multiple) based on the number of independent variables. Specify the dependent and independent variables, and fit the model to the data.

4. Model evaluation: Evaluate the performance of the model using statistical measures such as the coefficient of determination (R-squared), the standard error of the estimate, and the F-test. The R-squared value indicates the proportion of variation in the dependent variable that is explained by the independent variables. The standard error of the estimate measures the accuracy of the predictions made by the model. The F-test compares the fit of the model to a null hypothesis of no relationship between the dependent and independent variables.

5. Model refinement: Refine the model by adding or removing independent variables, transforming variables, or adjusting the model specification.

6. Model validation: Validate the model using a holdout data set or cross-validation techniques to assess its predictive accuracy and generalizability.

In summary, linear regression is a simple yet powerful algorithm that is widely used in data analysis and modeling. By estimating the parameters of a linear equation that describes the relationship between the dependent and independent variables, linear regression allows us to make predictions and gain insights into the factors that influence the outcome of interest.

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

Anscombe's quartet is a set of four datasets that have the same statistical properties but appear very different when plotted graphically. These datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data in addition to calculating summary statistics.

Each of the four datasets consists of eleven (x,y) pairs of values. The summary statistics for each dataset are:

Dataset I: $\bar{x} = 9.0$, $\sigma x = 3.16$ $\bar{y} = 7.5$, $\sigma y = 2.03$ $r = 0.816$, $y = 3.00 + 0.500x$, $R^2 = 0.67$

Dataset II: $\bar{x} = 9.0$, $\sigma x = 3.16$ $\bar{y} = 7.5$, $\sigma y = 2.03$ $r = 0.816$, $y = 3.00 + 0.500x$, $R^2 = 0.67$

Dataset III: $\bar{x} = 9.0$, $\sigma x = 3.16$ $\bar{y} = 7.5$, $\sigma y = 2.03$ $r = 0.816$, $y = 3.00 + 0.500x$, $R^2 = 0.67$

Dataset IV: $\bar{x} = 9.0$, $\sigma x = 3.16$ $\bar{y} = 7.5$, $\sigma y = 2.03$ $r = 0.817$, $y = 3.00 + 0.500x$, $R^2 = 0.67$

As we can see, all four datasets have the same mean, standard deviation, correlation coefficient,

linear regression equation, and coefficient of determination. However, when we plot these datasets graphically, they look quite different from each other.

Dataset I looks like a simple linear relationship with a few outliers, while Dataset II looks like a curved relationship that might be better described with a quadratic equation. Dataset III appears to have a nonlinear relationship with an outlier that is driving the correlation. Finally, Dataset IV looks like a perfect linear relationship except for a single outlier that is clearly an error.

The key takeaway from Anscombe's quartet is that summary statistics alone can be misleading and do not necessarily reveal the full picture of the relationship between variables. By visualizing the data, we can identify patterns, outliers, and other features that might be missed by summary statistics alone. Therefore, it is essential to use both graphical and numerical methods to analyze data and draw conclusions.

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

Pearson's R, also known as Pearson's correlation coefficient or simply correlation coefficient, is a measure of the linear relationship between two variables. It was developed by Karl Pearson, a British mathematician and statistician, in the late 19th century.

Pearson's R is a value between -1 and +1, where:

- A value of +1 indicates a perfect positive linear relationship, where as one variable increases, the other variable increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, where as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.

Pearson's R is calculated as the covariance between the two variables divided by the product of their standard deviations. The formula for Pearson's R is as follows:

$r = (\Sigma[(x - \bar{x}) * (y - \bar{y})]) / [(n - 1) * Sx * Sy]$

where: x and y are the two variables $\bar{x}$ and $\bar{y}$ are their respective means Sx and Sy are their respective standard deviations n is the number of observations in the data set

Pearson's R is a widely used statistic in statistics, machine learning, and data science, as it provides valuable information about the relationship between variables, which is useful for predicting one variable based on another. However, it should be noted that Pearson's R only measures linear relationships and may not capture non-linear relationships between variables. Therefore, other correlation coefficients such as Spearman's rank correlation or Kendall's tau correlation may be used in such cases.

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

Scaling refers to the process of transforming data so that it has a consistent range and is easier to compare or analyze. Scaling is typically performed on features or variables in a dataset, rather than on the target variable.

Scaling is performed for a number of reasons, including:

1. To improve the performance of machine learning models: Many machine learning algorithms, such as k-nearest neighbors, support vector machines, and neural networks, use distance measures to determine the similarity between data points. When features have different scales, the algorithm may place more emphasis on features with larger scales, leading to inaccurate predictions. By scaling the features, the distance measures are more balanced, leading to better model performance.

2. To aid in data visualization: Scaling data can make it easier to plot and visualize data, particularly when different features have vastly different ranges or units.

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

Normalized scaling (also known as min-max scaling) scales the data so that all values are within a specified range, typically between 0 and 1. This is done by subtracting the minimum value from each data point and dividing the result by the range (the maximum value minus the minimum value). The formula for normalized scaling is:

$x' = (x - \min(x)) / (\max(x) - \min(x))$

where x' is the normalized value of x, and min(x) and max(x) are the minimum and maximum values of x, respectively.

Standardized scaling (also known as z-score normalization) scales the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean value from each data point and dividing the result by the standard deviation. The formula for standardized scaling is:

$x' = (x - \mu) / \sigma$

where x' is the standardized value of x, μ is the mean of x, and σ is the standard deviation of x.

The key difference between normalized and standardized scaling is that normalized scaling preserves the relative relationships between data points and is better suited for algorithms that require data to be on a similar scale, while standardized scaling centers the data and may be more useful when analyzing the relative importance of different features.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The Variance Inflation Factor (VIF) is a measure of the degree of multicollinearity among the predictor variables in a regression model. A high VIF value indicates that there is a strong linear relationship between the predictor variables, which can lead to issues with the stability and interpretability of the regression coefficients.

Sometimes, the VIF value can be infinite. This happens when one or more of the predictor variables in the regression model are perfectly collinear, meaning that they are completely redundant and can be perfectly predicted by a linear combination of the other predictor variables in the model.

Perfect collinearity can arise in a regression model for a number of reasons. For example, it can occur when:

- Two or more predictor variables are measuring the same underlying construct, such as two different measures of income.
- One predictor variable is a function of another predictor variable, such as height and weight.

When perfect collinearity occurs, the regression coefficients cannot be estimated, as there is no unique solution to the regression equation. In this case, the regression model is said to be singular or degenerate.

To avoid perfect collinearity and infinite VIF values, it is important to carefully choose the predictor variables in a regression model and to check for multicollinearity among the predictor variables before fitting the model. If perfect collinearity is detected, one or more of the redundant predictor variables should be removed from the model.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a set of data follows a specified distribution. It is a plot of the quantiles of the data against the quantiles of a theoretical distribution, such as the normal distribution.

In a Q-Q plot, the data is plotted on the y-axis and the expected quantiles of the theoretical distribution are plotted on the x-axis. If the data follows the specified distribution, the points on the Q-Q plot should fall along a straight line. If the points deviate from a straight line, it suggests that the data may not follow the specified distribution.

In linear regression, Q-Q plots are often used to assess the normality of the residuals, which are the differences between the observed values of the dependent variable and the predicted values from the regression model. If the residuals are normally distributed, it suggests that the linear regression model is appropriate and the assumptions of the model are met.

The Q-Q plot can be used to visually inspect the normality of the residuals. A Q-Q plot of the residuals is created by plotting the quantiles of the residuals against the quantiles of a standard normal distribution. If the residuals are normally distributed, the points on the Q-Q plot should form a straight line. If the residuals are not normally distributed, the points on the Q-Q plot will deviate from a straight line, indicating that the normality assumption is violated.

If the normality assumption is violated, it may be necessary to transform the data or use a different type of regression model that is appropriate for non-normal data. The Q-Q plot is an important diagnostic tool for assessing the normality assumption and ensuring the validity of the linear regression model.