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? (3 marks)

I have done analysis on categorical columns using the boxplot.

Below are the few points we can infer from the visualization –

1. Fall season have attracted more booking, whereas spring season attracted very less bookings compared to other seasons.
2. Booking amount increased from April till mid-year and show a decreasing trend towards year end, signifying booking rates being high mostly in mid-year.
3. Apparently, we can see a very steep decrease in booking as weather condition worsens.
4. Mid-weekdays seem to get slightly higher bookings, but not many variations are observed in weekdays.
5. Booking seemed to be almost equal either on working day or non-working day.
6. We see linear increase in bike rental in from year 2018 to 2019 which shows good yearly growth in bike rental.

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

- When creating dummy variables, we represent categorical variables with binary (0/1) variables.
- If we include all the created dummy variables in our model, it can lead to multicollinearity.
- By setting drop_first=True, we drop the first created dummy variable for each categorical feature, which effectively eliminates the multicollinearity issue.
- Dropping the first dummy variable ensures linear independence among the variables, preventing redundancy in the model.
- By using `drop_first=True`, we make sure that the first category (the one dropped) becomes the reference category. This category is not explicitly included as a dummy variable, and its absence or presence is inferred from the presence or absence of the other categories. This makes it easier to understand the effects of the different categories because we can compare them to the reference category. In simpler words, `drop_first=True` helps us understand how each category is different from the category that is not included explicitly.

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

Feature `atemp` and `temp` has highest correlation of around 62-65% with target variable.

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

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

- Normal distribution of error. Residuals mean almost comes around 0.
- There should be insignificant multicollinearity among variables – For that made sure that all the independent variable in model has VIF score less than 5. As we know that high VIF score indicates multicollinearity.
- Linear relationship validation Linear should be present among predictors and independent variable. To check this, Actual vs Predicted plot was checked to make sure it follows a rough diagonal line indicating linear relationship.
- Homoscedasticity - There should be no visible pattern in residual values. Residual vs Fitted plot (Error vs Actual) to check is error variance is constant throughout the plot.
- Independence of residuals - No autocorrelation between residual, i.e., Durbin – Watson test was performed. And the score came very close to 2, indicating no autocorrelation.

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

Top 3 feature contributing in explaining the target variable were, year, feeling temperature and season – spring.

General Subjective Questions

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

Linear regression is a widely used statistical algorithm for modeling the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting straight line (or hyperplane in multiple dimensions) that minimizes the overall distance between the observed data points and the predicted values.

Here's a step-by-step explanation of the linear regression algorithm:

Data preparation:

Gather the dataset consisting of observations of the dependent variable (also known as the target or response variable) and one or more independent variables (also called predictors or features).

Ensure the data is in numerical format and handle any missing values or outliers as necessary.

Model representation:

Assume a linear relationship between the dependent variable and the independent variables.

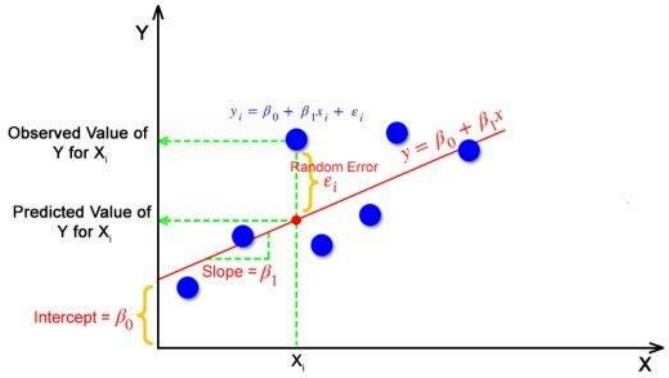
For a single-variable linear regression, the model can be represented as: $Y = \beta_o + \beta_1 X + \varepsilon$, where Y is the dependent variable, X is the independent variable, β_o is the y-intercept, β_1 is the slope, and ε is the error term.

For multiple-variable linear regression, the model extends to include additional independent variables: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$, where p is the number of predictors.

Model training:

Determine the best-fitting line or hyperplane by estimating the values of the coefficients (β_0, β_1, β_2, ..., β_p).

The estimation is typically done using a technique called ordinary least squares (OLS), which minimizes the *sum of squared differences between the observed values and the predicted values*.



Model evaluation:

Assess the quality and performance of the linear regression model.

Calculate various evaluation metrics such as the coefficient of determination (R^2), which indicates the proportion of the variance in the dependent variable explained by the independent variables.

Additionally, examine other metrics like root mean squared error (RMSE) or mean absolute error (MAE) to assess the accuracy and precision of the model's predictions.

Model utilization:

Once the linear regression model is trained and evaluated, it can be used to make predictions on new, unseen data by plugging in the values of the independent variables into the equation.

The model can also be used for inference, such as determining the significance and impact of the independent variables on the dependent variable.

Linear regression is a powerful and interpretable algorithm commonly used in various fields, including economics, finance, social sciences, and machine learning. Its simplicity and intuitive nature make it a fundamental technique for understanding the relationship between variables and making predictions based on that relationship.

Linear regression relies on several assumptions to ensure the validity and accuracy of the model's results. These assumptions are as follows:

- *Linearity: The relationship between the dependent variable and the independent variables is assumed to be linear. This means that the change in the dependent variable is directly proportional to the change in the independent variables.*
- *Independence: The observations in the dataset are assumed to be independent of each other. There should be no systematic relationship or dependency between the residuals (errors) of the model.*

- *Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables. In other words, the spread of the residuals should be consistent throughout the range of the independent variables.*
- *Normality: The residuals are assumed to follow a normal distribution. This assumption is important for statistical inference, hypothesis testing, and constructing confidence intervals.*
- *No multicollinearity: The independent variables should not be highly correlated with each other. High multicollinearity can make it difficult to separate the individual effects of the independent variables on the dependent variable and can lead to unstable and unreliable coefficient estimates.*

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

Anscombe's quartet refers to a set of four datasets that have nearly identical statistical properties but exhibit very different patterns when plotted. These datasets were created by the statistician Francis Anscombe in 1973 to highlight the importance of visualizing data and the limitations of relying solely on summary statistics.

All the four datasets contains 11(x, y) data points.

The four datasets within Anscombe's quartet have the following characteristics:

Dataset I:

When plotted, it forms a relatively linear relationship, resembling a simple linear regression.

The summary statistics, such as the mean, variance, and correlation coefficient, are close to those of Dataset II.

Dataset II:

When plotted, it exhibits a non-linear relationship that follows a quadratic curve.

Despite the non-linearity, the summary statistics are like those of Dataset I.

Dataset III:

When plotted, it appears to have a linear relationship with an outlier, which significantly affects the regression line and summary statistics.

The outlier has a substantial impact on the correlation coefficient, making it similar to the previous datasets.

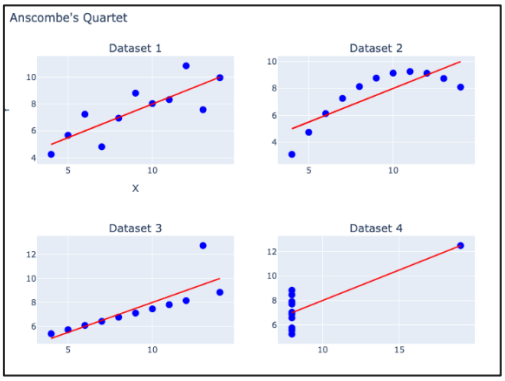
Dataset IV:

The last point in Dataset III, which is replaced with a completely different outlier. When plotted, it shows a strong linear relationship except for the outlier at the end.

The outlier drastically changes the regression line and correlation coefficient, highlighting the sensitivity of these measures.

The purpose of Anscombe's quartet is to emphasize the importance of data visualization in addition to summary statistics. Although these four datasets have nearly identical statistical properties, they

exhibit fundamentally different patterns when visually inspected. This demonstrates that relying solely on summary statistics can lead to misleading conclusions about the underlying relationships within the data.



Anscombe's quartet serves as a reminder that exploring and visualizing data are crucial steps in understanding the nuances and complexities that may not be evident from summary statistics alone.

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

Pearson's correlation coefficient, often denoted as Pearson's R, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Pearson's R takes values between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The calculation of Pearson's R involves several steps:

- Standardize the variables.
- Calculate the covariance.
- Calculate the correlation coefficient.
- Mathematically, Pearson's R is expressed as:

$$R = Covariance(X, Y) \;/\; (Standard\ Deviation(X) * Standard\ Deviation(Y))$$

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

Scaling is the process of transforming numerical variables to a standardized range or distribution. It involves adjusting the values of the variables so that they fall within a specific scale or range. Scaling is performed to bring different variables to a common scale, making them directly comparable and avoiding biases in the analysis.

Here are the differences between two common types of scaling: normalized scaling and standardized scaling:

Normalized scaling (also known as min-max scaling):

Normalized scaling rescales the variables to a specific range, typically between 0 and 1. It calculates the normalized value of each data point using the minimum and maximum values of the variable. The formula for normalized scaling is:

$$normalized\_value = (value - min) / (max - min)$$

This type of scaling maintains the original distribution shape but compresses the variable's range to a specific interval.

Standardized scaling (also known as z-score scaling or standardization):

Standardized scaling transforms the variables to have a mean of 0 and a standard deviation of 1. It subtracts the mean value from each data point and divides by the standard deviation. The formula for standardized scaling is:

$$standardized\_value = (value - mean) / standard\_deviation$$

Standardized scaling results in a distribution centered around 0, with a spread of 1. It maintains the shape of the distribution but changes the scale.

The choice between normalized scaling and standardized scaling depends on the specific requirements and characteristics of the data and the analysis. Normalized scaling is often used when the actual minimum and maximum values of the variable are important and need to be preserved. Standardized scaling is commonly employed when the focus is on the relative position of each data point in the distribution, or when the algorithm or analysis requires variables to have a mean of 0 and standard deviation of 1.

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

(3 marks)

The value of VIF (Variance Inflation Factor) can sometimes be infinite. This occurs when there is perfect multicollinearity among the independent variables in the linear regression model. Perfect multicollinearity means that one or more of the independent variables can be perfectly predicted by a linear combination of the other independent variables.

When perfect multicollinearity exists, the VIF calculation breaks down because it involves dividing by zero. The formula for VIF includes the calculation of the variance of an independent variable as the model is fitted with that variable as the dependent variable, while all other independent variables are treated as predictors. However, if perfect multicollinearity is present, the model cannot be properly fitted because one of the variables is a linear combination of the others, leading to an infinite variance.

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 plot, short for quantile-quantile plot, is a graphical tool used to assess whether a given dataset follows a specific probability distribution. In the context of linear regression, a Q-Q plot is commonly used to examine the normality assumption of the residuals (errors) in the regression model.

Use of Q-Q plot ->

- A Q-Q plot compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically the normal distribution.
- The observed data is sorted in ascending order, and the corresponding quantiles are calculated.
- The expected quantiles from the theoretical distribution are determined based on the number of observations and their probabilities in the distribution.
- The observed quantiles are plotted on the y-axis, and the expected quantiles are plotted on the x-axis.
- If the observed data follows the theoretical distribution, the points on the plot should approximately fall along a straight line (45-degree reference line).

Importance of a Q-Q plot in linear regression:

Assessing normality assumption: One of the key assumptions of linear regression is that the errors (residuals) are normally distributed.

Visual identification of departures from normality: By plotting the residuals on a Q-Q plot, we can visually inspect whether the residuals conform to a normal distribution. If the points on the plot deviate from the 45-degree reference line in a systematic manner, it suggests a departure from normality.

By examining the Q-Q plot of the residuals, researchers and analysts can gain insights into the normality assumption, identify potential issues, and take appropriate steps to address them, leading to more reliable and robust linear regression results.