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 plotted below box plot and following inferences.

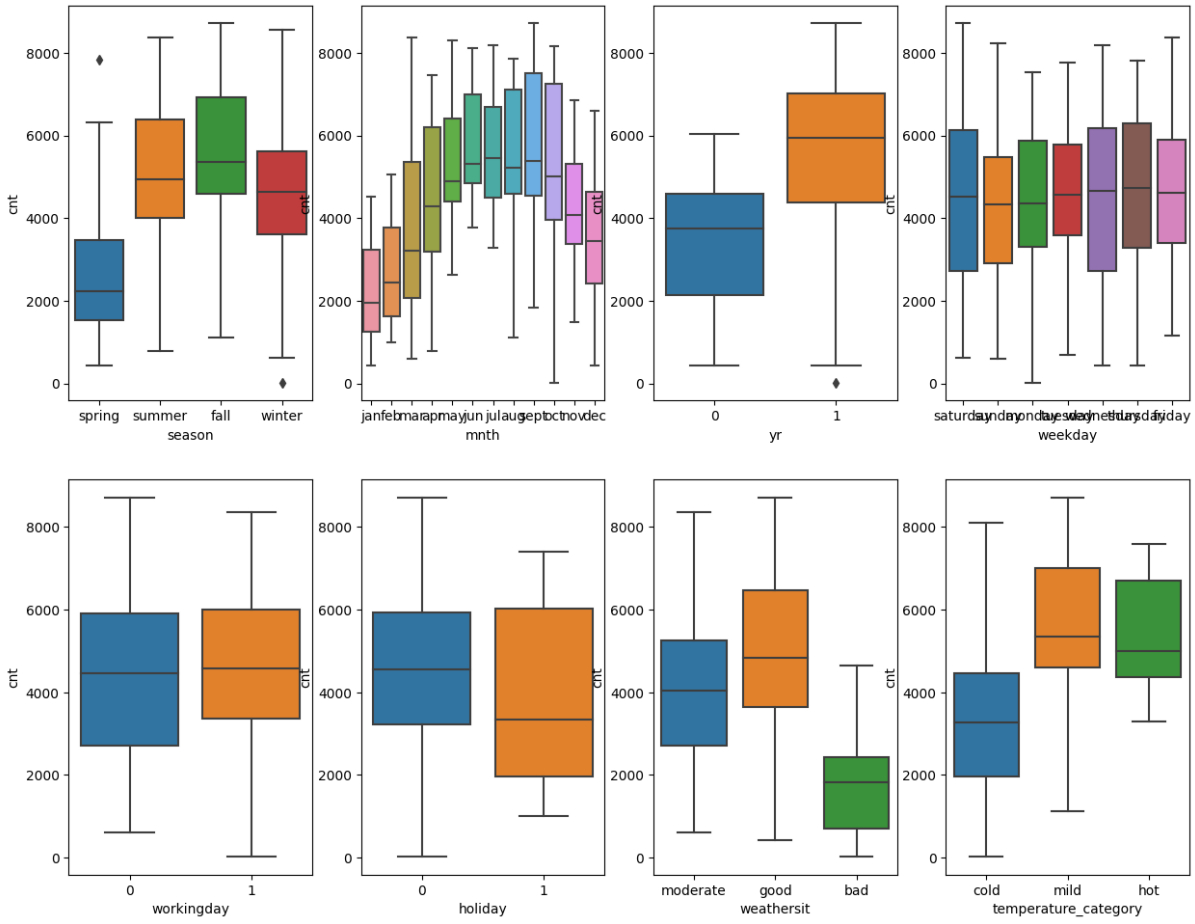
a) Spring season have lowest count

b) Lowest count of bike ride happens in month of January, best months for business lies between March to October

c)Bike counts increased considerably in 2019 as compared to 2018

d)As expected count of back rides are too much impacted with bad weather.

e) From the new temperature category that we created we can deduce that cold weather have lowest bike ride



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

Using drop_first=True during dummy variable creation avoids multicollinearity and improves interpretability in regression models.

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

temp

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

By Evaluating the model on test set data.

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

Year, Temperature & humidity(Negative impact)

General Subjective Questions

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

Linear regression is a widely used algorithm for modeling the relationship between a dependent variable (also known as the target variable or outcome) and one or more independent variables (also known as predictors or features).

The algorithm works by fitting a straight line to the data points that represents the best approximation of the relationship between the variables. This line is determined by finding the coefficients that minimize the sum of the squared differences between the predicted and actual values of the dependent variable.

To find these coefficients, linear regression uses a method called ordinary least squares (OLS), which calculates the slope and intercept of the line. The slope represents how much the dependent variable changes for a unit change in the independent variable, while the intercept represents the expected value of the dependent variable when all the independent variables are zero.

Linear regression assumes that there is a linear relationship between the independent variables and the dependent variable. However, it can handle multiple independent variables, allowing for more complex relationships. It also assumes that the errors (residuals) between the predicted and actual values follow a normal distribution and have constant variance.

Once the model is trained, it can be used to make predictions on new data by plugging in the values of the independent variables. The quality of the model can be assessed using various metrics, such as the coefficient of determination (R-squared) and root mean squared error (RMSE).

Linear regression is widely used in various fields, including economics, finance, social sciences, and machine learning. It provides insights into the relationship between variables, helps in predicting

outcomes, and can be extended to more advanced techniques like polynomial regression and regularization methods.

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

Anscombe's quartet is a set of four datasets that have identical statistical properties, including mean, variance, correlation, and regression line, but exhibit vastly different visual patterns when plotted. The quartet was introduced by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization and to caution against relying solely on summary statistics.

Each dataset in Anscombe's quartet consists of 11 paired x and y values. When these values are plotted, they form four distinct patterns:

Dataset I: This dataset forms a linear relationship with a positive slope. It is well-suited for linear regression analysis.

Dataset II: This dataset also has a linear relationship but with a slight curvature. It demonstrates that even when a linear regression model fits well, the underlying relationship might not be strictly linear.

Dataset III: This dataset has an apparent outlier that significantly influences the regression line. It shows the importance of identifying and handling outliers in data analysis.

Dataset IV: This dataset has a non-linear relationship that is completely masked by the presence of an outlier. It emphasizes the need for careful analysis and not relying solely on summary statistics.

The key message of Anscombe's quartet is that summary statistics alone can be deceptive. Visualizing the data provides insights into patterns, relationships, and potential issues that summary statistics may overlook. It highlights the importance of exploratory data analysis and the role of data visualization in understanding data.

Anscombe's quartet serves as a reminder that relying solely on numerical summaries can lead to misinterpretation and erroneous conclusions. It emphasizes the value of data visualization in gaining a deeper understanding of data and informing decision-making processes.

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

Pearson's R, or the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. It is widely used to assess the association between variables in statistics and data analysis.

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 variable values to a specific range or distribution. It is performed to ensure comparable scales and has three main benefits: improving model performance,

facilitating interpretation, and speeding up computation. Normalized scaling (min-max scaling) transforms variables to a range of 0 to 1, while standardized scaling (z-score normalization) transforms variables to have a mean of 0 and a standard deviation of 1. Both methods have their uses based on the specific needs of the analysis.

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

VIF values become infinite when there is perfect multicollinearity in the regression model. Perfect multicollinearity occurs when one or more independent variables can be perfectly predicted from the others. This situation leads to division by zero in the VIF calculation, resulting in an infinite value. Perfect multicollinearity can arise from including redundant variables or linearly dependent variables in the model. It is important to address multicollinearity to ensure accurate interpretation of regression coefficients and model stability.

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 is a graphical tool used to assess the distributional similarity between a sample dataset and a theoretical distribution. It is important in linear regression for checking the assumption of normality of residuals. By plotting the quantiles of the residuals against the quantiles of a normal distribution, it helps to evaluate if the residuals follow a normal distribution and detect outliers or deviations from normality. It validates the assumptions of the linear regression model and ensures the accuracy of the analysis.