Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

We have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of May, June, Jul, Aug and Sept. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious. Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

To avoid multicollinearity, we use drop_first=True during dummy variable creation.

When dealing with categorical variables in machine learning models, we often convert them into numerical representations using a technique called one-hot encoding. This involves creating a new binary column for each category, where 1 indicates the presence of that category and 0 indicates its absence.

This occurs when one variable is a perfect linear combination of other variables. In the context of dummy variables, this means that the sum of all dummy variables for a particular categorical variable always equals 1.
By dropping the first dummy variable, we implicitly assume that the omitted category is the baseline or reference category. This means that the coefficients of the remaining dummy variables represent the difference in the outcome variable compared to the baseline category.

Example:

Consider a categorical variable "Region" with three categories: "East", "West", and "South".

Without drop_first=True:

- Region_East , Region_West, Region_South

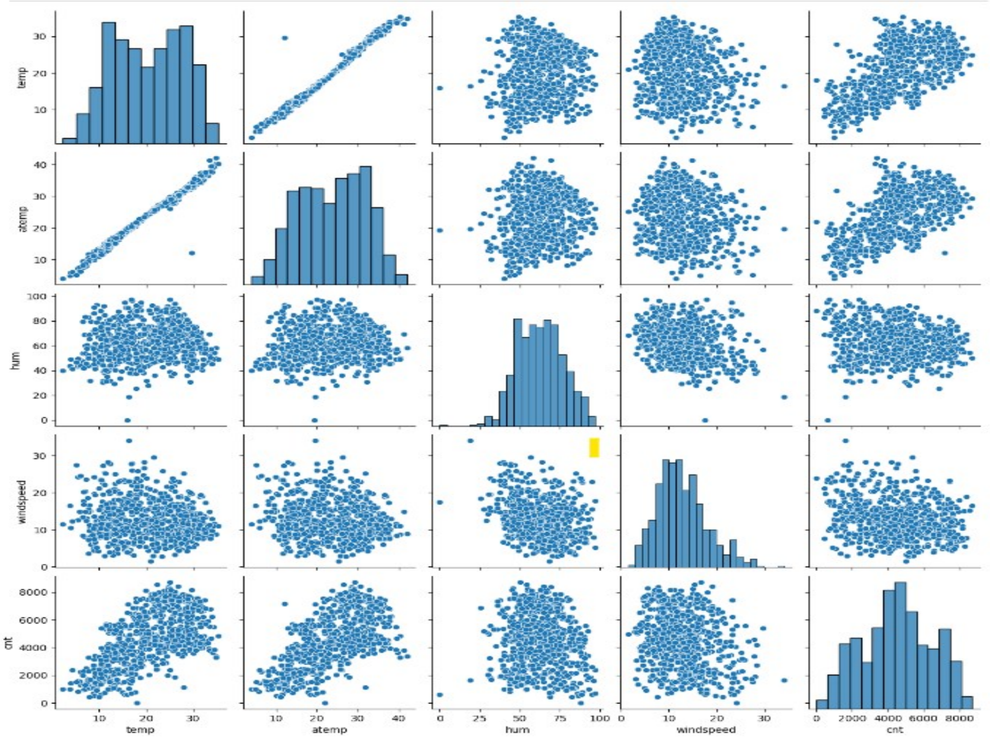
With drop_first=True:

- Region_West, Region_South

Here, "East" becomes the baseline category. The coefficients of "Region_West" and "Region_South" will represent the difference in the outcome variable compared to the "East" region.

By using **drop_first=True** helps to avoid perfect multicollinearity, improves model stability, and makes the interpretation of model coefficients more straightforward.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)



“temp” has the highest correlation with the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. **Linearity –**
    - Draw a scatter plot and check the relationship between X and y. It should displays some linear relationship Region_East , Region_West, Region_South
    - Residual plots: against the predicted values and it should be randomly scattered around zero
2. Error terms are independent of each other: The error terms should not be dependent on one another.
3. Normal Distribution of Error terms: Error terms are normally distributed with mean zero

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

- Temp

- yr_2019
- weathersit_snow

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<**Linear regression** is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning the change in the dependent variable is proportional to the change in the independent variables.

Types of Linear Regression

Simple Linear Regression: A single independent variable is used to predict the value of a numerical dependent variable.

Multiple Linear Regression: When more than one independent variable is used to predict the value of a numerical dependent variable.

Linear Regression model is represented by the equation

$$\mathbf{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon}$$

The Learning Process:

1. Data Preparation: - Collect and clean the data. Split the data into training and test sets
2. Model Training: Train the model on training data using OLS method and find squared residuals. The algorithm iteratively adjusts the coefficients to minimize the error
3. Model Evaluation: Evaluate the trained model on testing set. Common evaluation metrics are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2)

Limitations and Considerations:

1. **Liner Assumption:** The relationship between variables must be linear.
2. **M**ulticollinearity: High correlation between independent variables can affect the model's stability and interpretation.
3. Outliers: Outliers can significantly impact the model's performance.
4. Overfitting and Underfitting: The model may overfit the training data or underfit the data, leading to poor generalization.

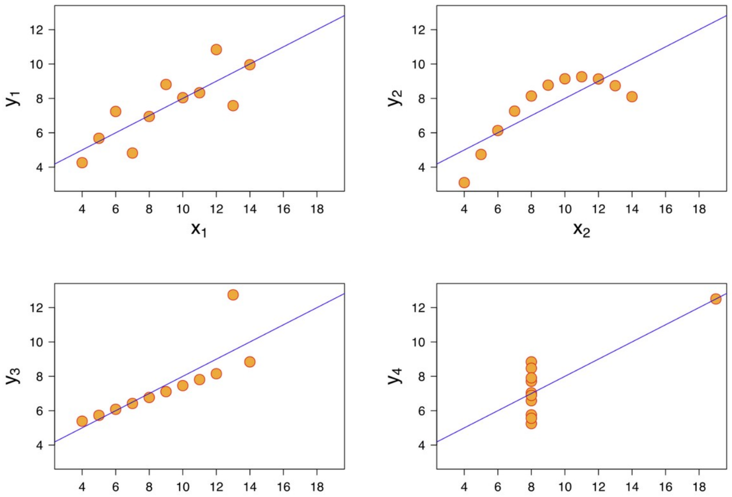
To check linear regression is suitable for any given data, a scatter plot is used. If the relationship is linear, we can go for linear models. But if it is not linear, we must apply some transformations to make the relationship linear. In case of univariate linear regression plotting a scatter, plot is easy. In multi variate analysis, two dimensional pairwise scatter plots can be plotted. >

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Anscombe's Quartet is a famous set of four datasets, to demonstrate the importance of data visualization. Each dataset consists of 11 (x, y) pairs of data points. It is crucial to visualize data before performing any statistical analysis.

The surprising fact is that these four datasets have nearly identical statistical properties, including **Mean,**

Variance, Correlation and Linear Regression.
when you visualize these datasets, they look dramatically different**:**



Dataset 1: This dataset shows a clear linear relationship between x and y.
Dataset 2: This dataset shows a perfect quadratic relationship between x and y, but the linear regression line still fits reasonably well due to the limited range of x values.
Dataset 3: This dataset is almost linear, except for one outlier point that significantly influences the regression line.
Dataset 4: This dataset has a constant x value for all but one point, making it difficult to fit a meaningful regression line. >

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Pearson's correlation coefficient, often denoted by the letter "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two variables.
The value of "r**" ranges from -1 to +1**

- +1: Perfect positive correlation (as one variable increases, the other also increases proportionally).
- -1: Perfect negative correlation (as one variable increases, the other decreases proportionally).
- 0: No correlation (no linear relationship between the variables) >

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and

consider smaller values as the lower values, regardless of the unit of the values.

Scaling performed to **Improve Model Performance, Faster Convergence, Better Interpretability**

Types of Scaling: Normalization (Min-Max Scaling) and Standardization (Z-score Normalization)

Difference between Normalized scaling and Standardized scaling

Normalization	**Standardization**
This method scales the model using minimum and maximum values	This method scales the model using the mean and standard deviation.
Values on the scale fall between [0, 1] and [- 1, 1].	Values on a scale are not constrained to a particular range
When features are on various scales, it is functional.	When a variable's Mean and Standard deviation are both set to 0, it is beneficial
Additionally known as scaling Normalization	This process is called as Z-score Normalization
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful

>

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< **A VIF (Variance Inflation Factor) value of infinity indicates perfect multicollinearity between two or more independent variables in your regression model.** This means that one or more of your variables can be perfectly predicted by a linear combination of the others>

It happens due to

- **Duplicate Variables:** You might have accidentally included the same variable multiple times in your model, either directly or in a transformed form.
- **Linear Combinations:** One variable might be a linear combination of others. For instance, if you have variables "Height in Inches" and "Height in Centimeters," they are perfectly correlated.
- **Dummy Variable Trap:** If you create dummy variables for a categorical variable with n categories, you should only include n-1 of them in your model to avoid perfect multicollinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< **A Q-Q plot, or Quantile-Quantile plot**, is a graphical tool used to compare two probability distributions. In the context of linear regression, it's primarily used to assess the normality of residuals>

Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data falls below and 70% falls above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2 sample tests.