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)

Used boxplot and bar plot for analysis of Categorical Variables.

Summary of findings listed below

1. Season

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019

2. Month

- Majority of the bookings have been done during the month of May, June, July, Aug, Sep and Oct compared to rest of months of the year. As you can see there is increasing trend starting of the year till mid of the year and then decreasing trend as we approached towards the end of year.

3. Weekday

- There is increasing trend visible as we approach from Sunday to Saturday, Saturday has highest number of bookings.

4. Weather Situation

- Clear weather conditions attract a greater number of bookings. (Do not edit)

5. Working Day / Holiday

- When it's holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Bookings seem to be almost irrespective of working day or non-working day

6. **Year (2018/2019)**

- 2019 attracted a greater number of bookings in comparison to 2018, which reflects 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)

In multiple linear regression (MLR), using **drop_first=True** when creating dummy variables is important to avoid a problem known as multicollinearity, specifically the dummy variable trap.

The dummy variable trap occurs when the inclusion of all categories of a categorical variable (converted into dummy variables) leads to perfect multicollinearity. This happens because one dummy variable can be perfectly predicted by the others, along with the intercept term, creating redundancy in the data. This redundancy can:

- Make it impossible for the regression algorithm to compute unique estimates for coefficients.
- Lead to numerical instability and unreliable results in your regression model.

How Does drop_first=True Solve This?

By setting **drop_first=True**, one category of the categorical variable (the "reference category") is excluded when creating dummy variables. This exclusion:

1. Prevents Multicollinearity: The excluded category serves as the baseline, and the coefficients of

the remaining dummy variables represent differences relative to this baseline.

2. Keeps Interpretability: Each coefficient shows the effect of being in a particular category compared to the reference category, making the model's output more interpretable.

3. Ensures Model Stability: It avoids redundant information, making the computation of the model more efficient and stable. (Do not edit)

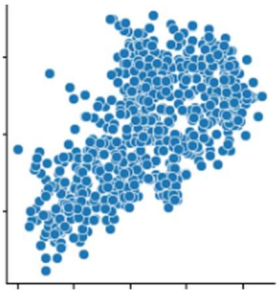
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

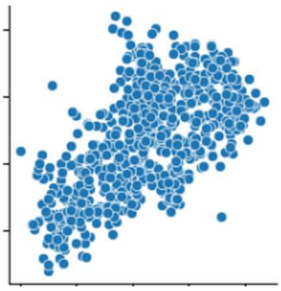
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Target variable is cnt and variables 'temp' and 'atem' have the highest correlation with the target variable (Do not edit)

temp



atemp



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)

Validating the assumptions of Linear Regression is a critical step to ensure the model's reliability and interpretability. I have validated the assumption of Linear Regression Model based on below listed assumptions:

1. Linearity

Assumption: The relationship between independent variables (features) and the dependent variable (target) should be linear.

2. Independence of Errors

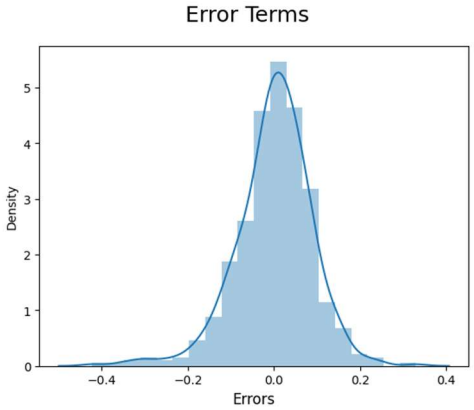
Assumption: The residuals are independent of each other (There shouldn't be any autocorrelation).

3. Homoscedasticity (Constant Variance of Errors)

Assumption: The residuals have constant variance across all levels of the independent variables.
No visible pattern should be observed from plot for residuals.
Durbin-Watson value of final model lr2 is 2.066, which signifies there is no autocorrelation.

4. Normality of Residuals

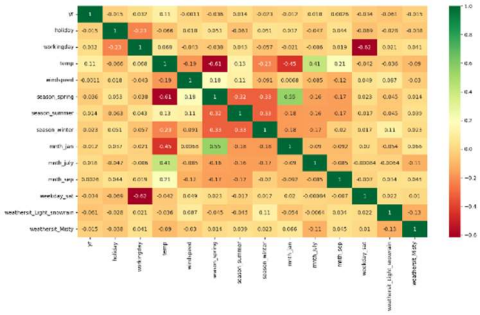
Assumption: The residuals are normally distributed.



Insights from the above histogram, we could see that the Residuals are normally distributed, which proves that assumption for Linear Regression is valid.

5. Multicollinearity

Assumption: The independent variables are not highly correlated with each other.



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)

Factors which help to spur bike rental demand.

- Temperature
- Year
- Month of Sep
- Winter Season
- Summer Season
- Working Day
- Saturday

Factors which impact negatively the bike rental demand

- Holiday
- Windspeed
- Spring Season

- Month of Jan and July
- When it rains or snow fall happens
- When misty condition is there (Do not edit)

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)
Please write your answer below this line. (Do not edit)
Linear regression is a supervised learning algorithm used to predict a continuous dependent variable (output) based on one or more independent variables (inputs). The goal is to find the linear relationship between the variables. Below is a detailed explanation of the algorithm:
Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Mathematically the relationship can be represented with the help of following equation −
Y = mX + c
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.
Furthermore, the linear relationship can be positive or negative in nature as explained below−
o **Positive Linear Relationship:**
  - A linear relationship will be called positive if both independent and dependent variable increases.

o **Negative Linear relationship:**
  - A linear relationship will be called positive if independent increases and dependent variable decreases.

- **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.

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)

The quartet was created by statistician Francis Anscombe in 1973 to demonstrate the importance of plotting data before analyzing it. It shows how summary metrics can be misleading and how outliers and other influential observations can affect statistical properties.

Data

Each dataset in the quartet consists of 11 (x, y) pairs. The summary statistics for all four datasets are nearly identical, including:

The average x value is 9

The average y value is 7.50

The variance for x is 11

The variance for y is 4.12

The correlation between x and y is 0.816

The linear regression (line of best fit) is y = 0.5x + 3

Appearance

When graphed, the datasets look very different from one another.

Key Lessons from Anscombe's Quartet

1. Always Visualize Data: Before drawing conclusions, use scatterplots or other visual tools to examine relationships and patterns.
2. Be Cautious with Outliers: Outliers can disproportionately influence summary statistics and regression lines.
3. Check for Model Assumptions: Assumptions such as linearity should be validated visually or statistically.
4. Avoid Blind Trust in Statistics: Identical summary statistics do not imply identical data distributions or relationships.
5. Highlighting Data Context: Statistical analyses must be accompanied by context, visuals, and domain expertise to ensure accurate conclusions.

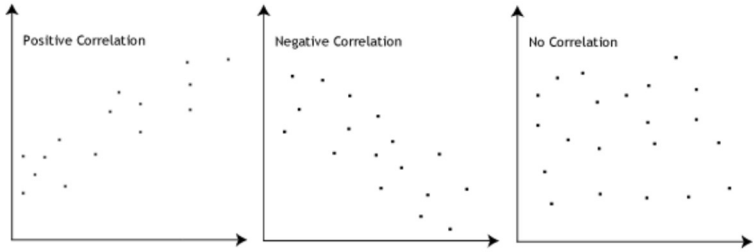
Purpose

Anscombe's quartet is often used in statistical education and data analysis training to stress the importance of exploratory data analysis (EDA). With the help of Python libraries matplotlib, seaborn and R make it easy to plot data and detect patterns visually.

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)

The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.
To find the Pearson correlation coefficient, two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association.
Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.



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.
Why we need scaling
1. Ease of interpretation
2. Faster convergence of gradient descent methods
How to achieve the Scaling
1. Standardization
2. MinMax Scaling
There are two major methods to scale the variables, i.e. standardization and MinMax scaling.
Standardization basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.
MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

Some important facts about scaling

1. Model Accuracy won't change as prediction won't change if you do the changes in fit or scale of variable.

Scaling does change the Coefficient.

2. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

3. Scaling should always be done after the test-train split since we don't want the test dataset to learn anything from the train data. So, if you're performing the test-train split earlier, the test data will then have information regarding the data like the minimum and maximum values, etc.

4. Standardized scaling will affect the values of dummy variables but MinMax scaling will not.

5. For Training dataset we use scaler.fit_transform() while for Test dataset we use scaler.transform()

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)

The **Variance Inflation Factor (VIF)** quantifies the degree of multicollinearity in a set of independent variables in a regression model. A VIF value can approach infinity under certain circumstances, which signals a severe problem with multicollinearity. It's one of way to deal with multicollinearity

Why **Variance Inflation Factor** (VIF)

- Sometimes pairwise correlations aren't enough
- Instead of just one variable, the independent variable might depend upon a combination of other variables
- VIF calculates how well one independent variable is explained by all the other independent variables combined

The VIF is given by:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

where *'i'* refers to the i-th variable which is being represented as a linear combination of rest of the independent variables. You'll see VIF in action during the Python demonstration on multiple linear regression.

The common heuristic we follow for the VIF values is:

> 10: Definitely high VIF value and the variable should be eliminated.

> 5: Can be okay, but it is worth inspecting.

< 5: Good VIF value. No need to eliminate this variable.

Consequences of Infinite VIF

When the value of VIF is infinite, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. •

Regression coefficients for the affected variable(s) cannot be estimated.To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

How to Resolve Infinite VIF

1. **Identify the Collinearity**:
    - Use the correlation matrix to spot highly correlated variables.
    - Check for relationships among dummy variables if categorical data is involved.
2. **Remove or Combine Variables**:
    - Drop redundant or highly correlated variables.
    - Combine correlated variables into a single feature (e.g., using Principal Component Analysis).
3. **Avoid Dummy Variable Trap**:
    - When encoding categorical variables, always drop one dummy variable (the reference category).
4. **Increase Sample Size**:
    - Collect more data to reduce the effect of multicollinearity in cases of insufficient observations.

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 (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specified theoretical distribution, most commonly the normal distribution. It compares the quantiles of the data against the quantiles of the theoretical distribution.

Purpose of a Q-Q Plot

- To visually check if a dataset follows a particular distribution (e.g., normality).
- To detect deviations from normality, such as skewness, kurtosis, or the presence of outliers.
- To validate assumptions in statistical models that require normality or other specific distributional assumptions.

Importance of a Q-Q plot in linear regression

It plays a critical role in linear regression by helping to validate one of its key assumptions: that the residuals (errors) are normally distributed. Properly assessing the distribution of residuals ensures the validity of statistical inference in regression models, such as hypothesis testing and confidence intervals.

Ensures Valid Statistical Inference:

Many statistical tests, such as ttt-tests for regression coefficients and FFF-tests for model significance, assume that residuals are normally distributed. A Q-Q plot helps confirm this assumption.

Model Evaluation:

A Q-Q plot helps detect potential violations of assumptions, enabling corrective actions (e.g., transforming variables, using a different regression approach).

Outlier Detection:

A Q-Q plot highlights residuals that deviate significantly from the expected distribution, pointing to potential outliers that may influence the model.

Improves Predictive Accuracy:

Identifying and addressing non-normality in residuals can lead to a better-fitting model, enhancing predictions and interpretations.