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?

Ans:- After performing bivariate analysis on the categorical variables, the following inferences were made about their effects on the dependent variable, 'cnt':

- **Season**: fall has the highest number of bookings with a median of over 5000, followed by summer, winter and spring. This indicates that season can be a good predictor for the target variable.
- **yr** - According to the dictionary, the year 2018 is represented by '0' and 2019 by '1'. We saw that the year 2019 recorded more bookings. The variable' yr' is a good predictor for the dependent variable.
- **mnth**: We saw that April, May, June, July, August, September and October record bookings with a median of over 4000. Also, we saw that the median of bookings rose from the month of Feb, reached its peak in July, and then gradually fell as the months went by. June and September received the highest bookings, whereas Jan received the least.
- **holiday**: According to the dictionary, 1 indicates it is a holiday, and 0 indicates it is not. According to the analysis, most bike bookings occur when it is not a holiday.
- **weekday**: There is not much drastic variance in bookings regarding the days of the week. However, Sunday records a slightly lower number of bookings—all the days of the week record bookings in the median range of 4000 to 5000.
- **workingday**: The median bookings are slightly higher on working days, i.e. Working days received slightly more bookings than non-working days.
- **weathersit**: Majority of the bookings occur in fair_weather situations, i.e. when it is Clear, Few clouds, Partly cloudy, Partly cloudy. Moderate weather recorded the second-highest bookings, followed by unfavorable weather. Weather situations show clear booking trends and can be a good predictor variable.

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

Ans:- When creating dummy variables, drop_first=True is essential to avoid creating an unnecessary column and prevent correlation issues among the dummy variables. This option ensures that out of n categorical levels, only n-1 dummies are generated, removing the first level. For instance, knowing A and B automatically implies C if we have categories A, B, and C, making the third variable redundant.

Including all the dummy variables without dropping the first one can lead to multicollinearity, where highly correlated variables can cause issues in the model. Multicollinearity can make it challenging to interpret the coefficients of the variables and can also lead to unstable and unreliable model predictions. Therefore, drop_first=True streamlines the dummy variable creation process, enhancing efficiency and avoiding redundancy.

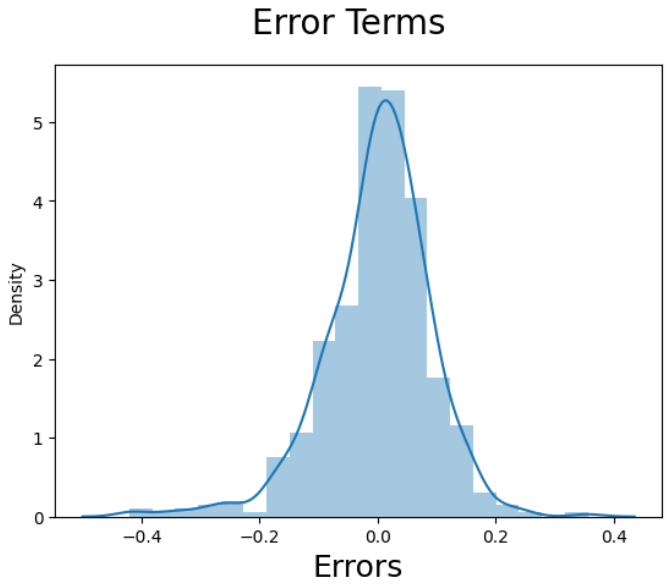
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans:- From the pair-plot of the numerical variables, '**temp**' and '**atemp**' have the highest correlation with the target variable 'cnt'. They show a linear relationship with the target variable.

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

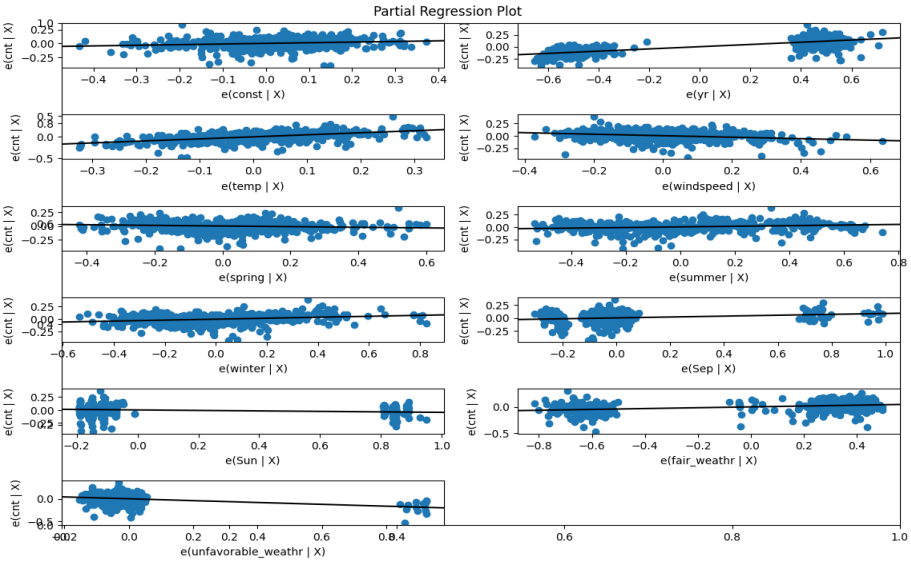
Ans:- The final model was validated by making sure that the assumptions of Linear Regression were met:

1. **Residual analysis**: The residuals were normally distributed



From the histogram above, we can see that the residuals are normally distributed; hence, this assumption of error term normally distributed in this linear regression is met.

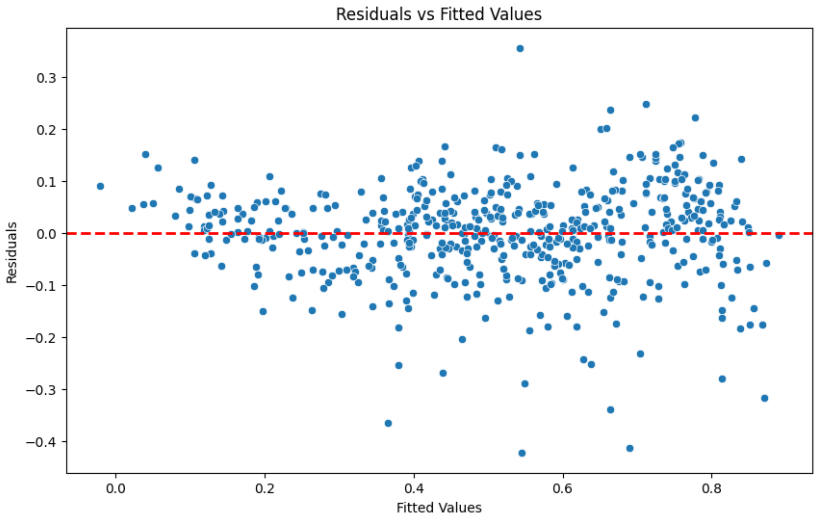
2. **Partial regression plots**: Checked if some of the predictor variables displayed a linear relationship with the target variable.



The partial regression plots above show that the predictor variables display a linear relationship with the target variable, 'cnt'. Hence satisfying one of the assumptions of linear regression as part of model validation.

3. **Multicollinearity**: The predictor variables should not be associated with each other. We calculated the VIF (Variance Inflation Factor) for each predictor to detect multicollinearity. We assumed that VIF values above 5 may indicate an issue. The VIF of each features are as follows: temp = 4.722, wind speed = 4.628, fair_weathr = 2.840, yr = 2.073, spring = 2.024, summer = 1.845, winter = 1.616, Sep = 1.217, Sun = 1.179 and unfavorable_weathr = 1.11
All the features in the final model have VIF below 5 and hence have very low or negligible multicollinearity.

4. **Homoscedasticity**: The residuals should be independent.



The residual values are homoscedastic according to the plot of Residuals vs fitted values above. The error terms do not display any visible patterns but appear to be evenly distributed noise around zero, which is ideal.

5. **R-squared and Adjusted R-squared**: The R-squared and Adjusted R-squared were calculated as 0.834 and 0.830, respectively, which are excellent scores indicating that the model is reliable.

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

Ans:- According to our final model, lr6, the top three predictor variables that contributing significantly towards explaining the demand of shared bikes are:

1. **temp** (Temperature): A unit increase in 'temp', increases the bookings by 0.4799 units.
2. **yr** (year): A unit increase in 'yr', increases the bookings by 0.2343 units.
3. **unfavorable_weathr**: A unit increase in 'unfavorable_weathr', decreases the bookings by 0.2062 units.

Note: According to the dictionary, unfavorable_weathr = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:- Linear regression is a statistical method used in machine learning for predictive analytics. It is a supervised algorithm that provides a linear relationship between a dependent variable (also known as the target variable) and one or more independent variables (aka features or predictor variables). There are two types of linear regression. When a model has only one independent variable, it is a simple linear regression, whereas if the model has more than one independent variable, it is a multiple linear regression. The primary goal is to discover the optimal linear relationship that characterizes the correlation between these variables. This relationship is expressed through an equation as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Where, y = *dependent variable*

$X_1, X_2, \ldots X_n$ = *independent variables*

β_0 = *intercept*

$\beta_1, \beta_2, \beta_n$ = *coefficients representing the impact of each independent variable on the dependent variable*

ϵ = *error terms*

Linear regression aims to determine the values of 0, 1, 2,... n, which minimizes the sum of squared differences between the observed and predicted values, commonly known as residuals. This approach is commonly denoted as "least squares" estimation.

The above equation helps us make an inference, but since we are still relying on inferences from samples, some uncertainty is introduced, and we cannot rely solely on this. To remedy this and take a more statistical approach, linear regression adopts some assumptions as follows:

1. There is a linear relationship between the dependent and independent variables.
2. The residuals or error terms are independent of each other.
3. The residuals are normally distributed.
4. The variance of residuals is constant (homoscedasticity).

The above assumptions apply to both simple linear regression (SLR) and multiple linear regression(MLR). One of the differences between the two is that in MLR, the model now fits a 'hyperplane' instead of a line.

Linear regression consists of the following steps:

- Data Collection: Gather data on the dependent and independent variables.
- Data Preprocessing: Handle missing values outliers, check for duplicate data, etc.

- Data Exploration: Perform exploratory analysis (EDA) on the dataset to understand the relationships between variables.
- Model Building: Fit the linear regression model to the training data.
- Model Evaluation: Assess the model's performance using metrics like R-squared, Mean Squared Error, etc.
- Prediction: Use the model to make predictions on the test data.

2. Explain the Anscombe's quartet in detail.

Ans:- Anscombe's quartet is a group of four datasets that provide a useful caution against applying individual statistical methods to the data without first graphing them. They share identical statistical properties, such as the same mean, standard deviation, and linear line regression parameters, but look quite different when they are graphed.

Note: The datasets in the table below were generated using pandas, and the plots below were plotted using Excel.

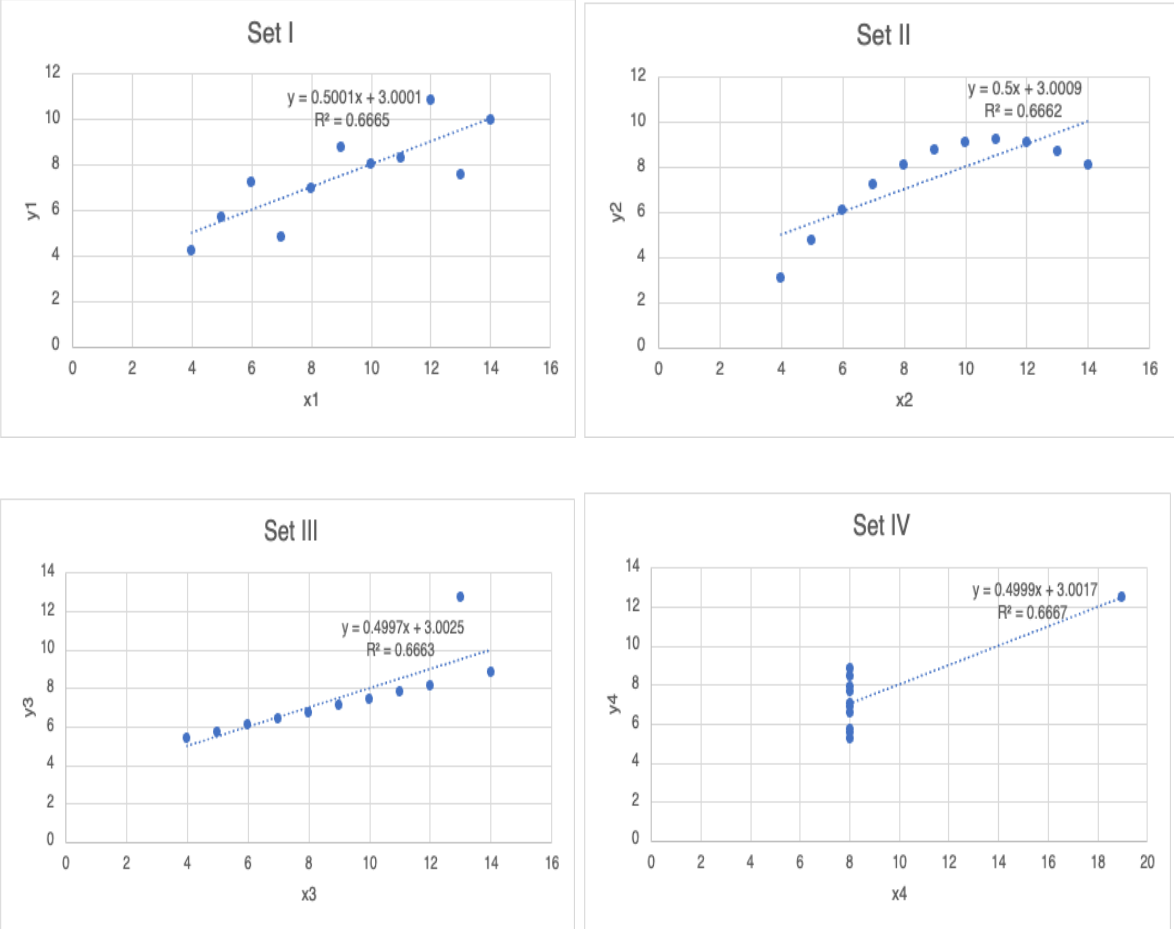
	Set I		Set II		Set III		Set IV	
	x1	y1	x2	y2	x3	y3	x4	y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
SUM	99	82.51	99	82.51	99	82.5	99	82.51
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Covariance	3.316	2.031	3.316	2.031	3.316	2.03	3.316	2.03

We can see that the summary statistics for all the four datasets are the same:

- The mean is 9 across the x variables and 7.5 across the y variables of the four datasets.

- The variance is 3.316 across the x variables and 2.031 across the y variables of the four datasets.

However, when we plot these datasets, they all look very different (refer to the graphs below). The linear line regression parameter of all the four sets are very similar i.e. R square ≃ 0.6665 , intercept ≃ 3.0009 and the correlation coefficient ≃ 0.5.



From this demonstration of Anscombe's quartet and after visualizing them in graphs, we can conclude that statistics are just tools for analysis and should not be replaced with common sense and should be supported by anecdotal analysis and visualizations.

3. What is Pearson's R?

Ans:- Pearson's correlation coefficient, which is denoted as "r" or Pearson's "r," is a measure of the linear relationship or correlation between two continuous variables. It assesses the strength and direction of a linear association between the variables. The coefficient ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship, and
- 0 indicates no linear relationship.

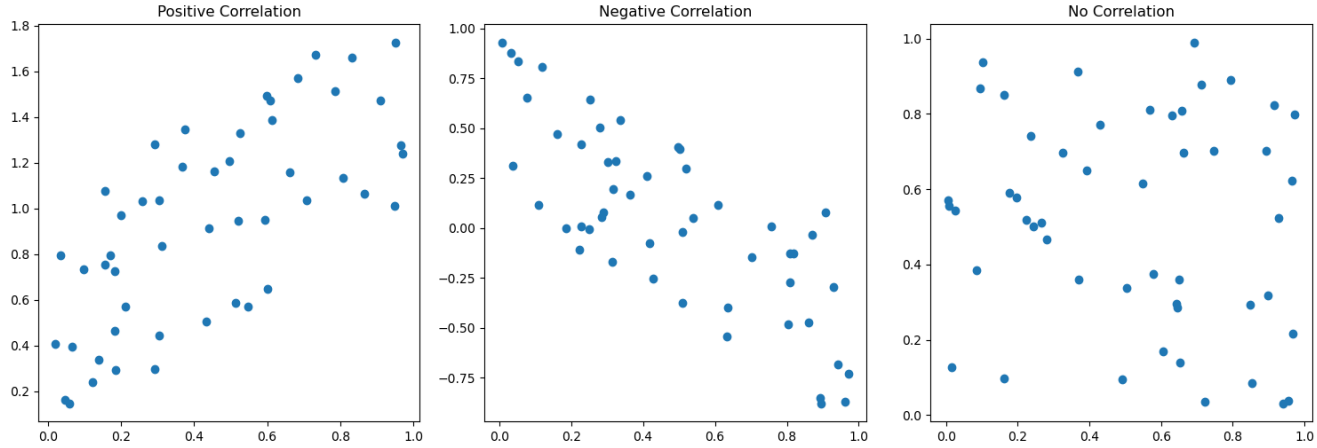
For two variables, X and Y, with n data points, the Pearson's coefficient is given by:

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 (Y_i - \overline{Y})^2}}$$

where:

- Σ denotes the sum of the values
- X and Y are the variables being compared
- $\overline{X}$ and $\overline{Y}$ are the means of X and Y, respectively

Pearson's R is particularly sensitive to outliers and assumes that the relationship between the variables is linear.



Note: The plots above were generated using numpy and matplotlib.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:- Scaling is a technique used in the preprocessing step of machine learning to standardize or normalize the features of a dataset. It involves transforming the numerical values of variables so that they are on a similar scale.

Scaling is performed for the following reason but not limited to:

- Easy interpretation: Scaling makes comparing and interpreting the coefficients of different features in linear models easier.
- Convergence for gradient descent method becomes faster: Some machine learning algorithms are sensitive to the scale of features. For example, distance-based algorithms like k-nearest neighbours or gradient descent-based algorithms converge faster when features are on a similar scale.

Scaling is necessary because if we do not scale, the feature with the higher values will get the unnecessary benefit statistically. We should not give any feature importance for its scale. We should emphasize any feature because of its relationship with the target.

For example, when predicting the salary, let us say we have two features: age and blood pressure(bp). Age ranges from 23 to 65, and bp ranges from 90 to 140. Now, blood pressure is not more important than age because the value or magnitude is larger. We cannot give it more

statistical weightage just based on magnitude. In fact, age will likely be a better predictor for salary than BP. Therefore, scaling is essential so that all the features are transformed to scale neutral and none benefit from scale.

Difference between Normalised Scaling and Standardised Scaling:

- **Normalised Scaling:**
- Range: Scales the values to a specific range, typically between 0 and 1.
- Formula: $X_{normalized} = \frac{X - min(X)}{max(X) - min(X)}$
- Advantage: Useful when the distribution of the data is not normal or when the distribution of the data is skewed or has outliers

- **Standardised Scaling:**
- Range: Scales the values with a mean of 0 and a standard deviation of 1.
- Formula: $X_{standardized} = \frac{X - \overline{X}}{\sigma}$

  where : $\overline{X}$ = mean of X

  σ = standard deviation of X
- Advantage: Appropriate when the data follows a normal distribution and when features have similar ranges.

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

Ans:- When we observe that the value of Variance Inflation Factor (VIF) is infinite, there is an indication of perfect multicollinearity among the predictor variables in a multiple regression model. Multicollinearity refers to the case when two or more independent variables are highly correlated.

VIF can be calculated 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 the rest of the independent variables.

When perfect multicollinearity exists, it means that one or more variables are redundant or highly correlated with each other. As a result, the VIF calculation involves dividing by zero, leading to an infinite value.

If the VIF for a variable is infinite, it indicates that the variable is perfectly predictable from the other variables in the model. In such cases, it is necessary to identify and address the issue of multicollinearity by removing one or more variables or finding other ways to handle the high correlation among the variables.

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

Ans:- A Q-Q plot (Quantile-Quantile plot) is a graphical tool that is used to assess the distributional similarity between a sample of data and a theoretical distribution. It compares the quantiles of the observed data against the quantiles of a specified theoretical distribution, it can be normal distribution or uniform distribution or some other distribution.

Use in Linear Regression:

- In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed.
- The Q-Q plot is a visual tool to assess the normality of residuals. If the residuals are normally distributed, the points on the Q-Q plot should form a straight line.

For example, the 30% quantile is where 30% of the data falls below, and 70% falls above that point. The Q-Q plot includes a 45-degree reference line. Points should align with this line if both sets share the same distribution. Deviations from it suggest different distributions.

Importance in Linear Regression :

- Checking the normality of residuals is crucial because many statistical tests and interval estimates in linear regression are based on the assumption of normality.
- Deviations from normality might affect the validity of hypothesis tests and confidence intervals.

In practical terms, a Q-Q plot helps assess if two datasets have a common distribution. If they do, estimators for location and scale can combine both datasets. Conversely, differences between two samples can be better understood through Q-Q plots than with tests like chi-square and Kolmogorov-Smirnov. The plot visually highlights departures from the expected line, providing valuable insights into potential distribution differences.