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)**

   The categorical variables for the given dataset are season, weathersit, workingday, holiday, mnth, yr and weekday. I used bar plot and box plot to analyse these variables. Below are the few points we can infer from the visualization –

   - **Yr:** There is a clear increase in the bookings for the year 2019 when compared to 2018. There is a 64.7% increase in bookings from the year 2018 to 2019.
   - **Mnth:** We can see that there are more bookings from May to October. With the month August having the maximum bookings.
   - **Season:** Here we can see that Fall and Summer are the two seasons during which there is maximum bookings.
   - **Holiday:** On holiday, we can see that there are less number of bookings. We can also say that since we have only 21 holidays to 709 working days in the dataset, the available data of holidays is less and therefore it is not conclusive evidence to say so. Nonetheless, the average of bookings on holidays is slightly lesser than working day's.
   - **Weekday:** Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
   - **Weathersit:** Clear weather attracted more booking which is understandable.
   - **Workingday:** Booking seemed to be almost equal either on working day or non-working day.

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

   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence, it reduces the correlations created among dummy variables.
   Syntax - drop_first: bool, default False
   Therefore by default we will get x dummies out of x categorical levels. To get x-1 dummies we have to say True, due to which we will just receive x-1 by removing the first level.
   Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

   temp and atemp are the two numerical variables that have high correlation with cnt which is the target variable. The correlation coefficient is 0.63 for both the variables with the 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:

- Normality of error terms:

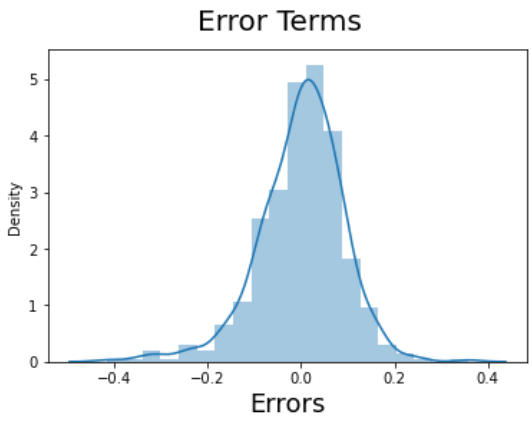
  Error terms should be normally distributed. We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0
- Multicollinearity check:

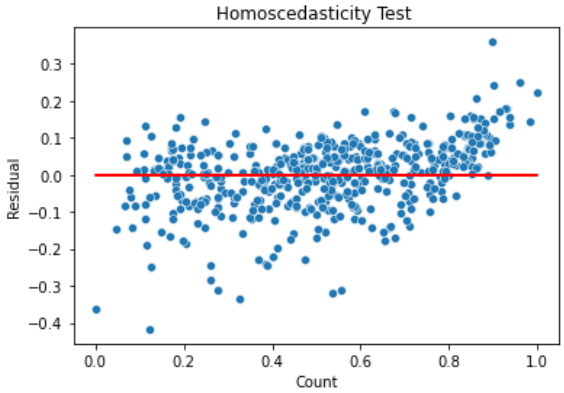
  There should be insignificant multicollinearity among variables.
- Linear relationship validation:

  Linearity should be visible among variables
- Homoscedasticity:

  There should be no visible pattern in residual values. This means they should have similar variance throughout the distribution and this can be validated by making a scatter plot of residuals and a horizontal line passing through 0. So we can observe that all the points are having almost similar variance throughout the distribution so we can say that Residuals have homoscedasticity.
- Independence of residuals:

  No auto-correlation





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

The top 3 features contributing significantly towards explaining the demand of the shared bikes along with their correlation coefficient are:

- temp: 0.4777
- weathersit_Light_snow/rain: -0.285(negative correlation)
- year: 0.2341

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input variables and the output variable, and the goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual values.

Linear regression is based on the popular equation "y = mx + c".
Regression is broadly divided into simple linear regression and multiple linear regression.

- Simple Linear Regression :
  SLR is used when the dependent variable is predicted using only one independent variable.
- Multiple Linear Regression :
  MLR is used when the dependent variable is predicted using multiple independent variables.

Here is a detailed explanation of the linear regression algorithm:

i. Problem Formulation:
- Identify the problem or question you want to answer.
- Determine the dependent variable (also known as the target variable or output variable) that you want to predict based on the independent variables (also known as features or input variables).

ii. Data Collection:
- Gather a dataset that contains observations of both the dependent and independent variables.
- Ensure the dataset is representative and has enough variability to capture the relationship between the variables.

iii. Data Preprocessing:
- Clean the dataset by handling missing values, outliers, and other data quality issues.
- Split the dataset into training and testing subsets.

iv. Model Representation:
- Linear regression assumes a linear relationship between the independent variables and the dependent variable, represented by the equation:
  y = b0 + b1*x1 + b2*x2 + ... + bn*xn

  - y represents the dependent variable.
  - x1, x2, ..., xn are the independent variables.
  - b0, b1, b2, ..., bn are the coefficients or weights that need to be estimated.

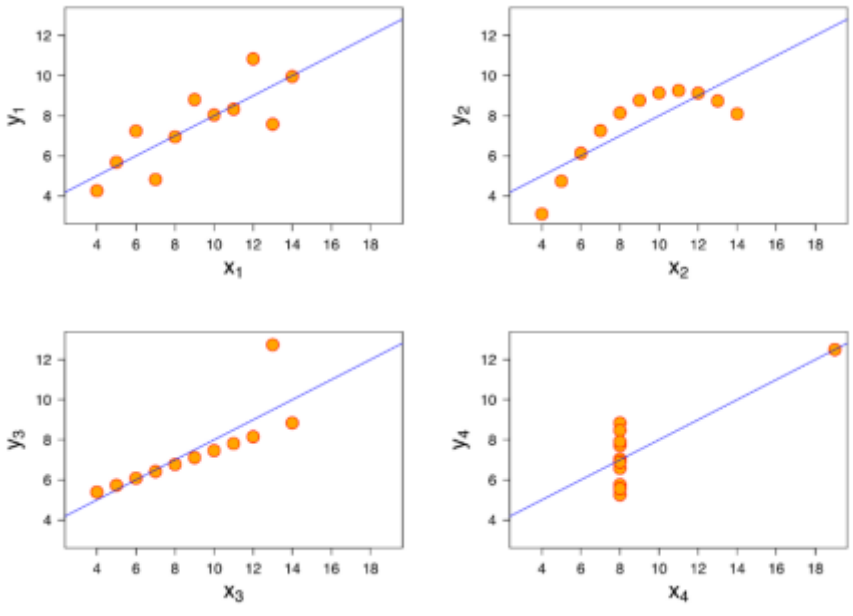
v. Training the Model:
   Train the model by utilizing the training dataset that is obtained after splitting the complete dataset.
vi. Model Evaluation:
   Checking if the model satisfies the assumptions of linear regression model.
vii. Making predictions utilizing the model.

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 vastly different patterns when plotted graphically. These datasets were created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not solely relying on summary statistics.

The four datasets in Anscombe's quartet consist of 11 data points each and share the same means, variances, and correlation coefficients. However, when examined individually, they reveal distinct patterns.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03



• Dataset I appears to have clean and well-fitting linear models.

• Dataset II is not distributed normally.

• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

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

   Pearson's correlation coefficient, often denoted as Pearson's R or simply as the correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the association between two continuous variables.

   Pearson's R is a value between -1 and 1, where:

   - A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
   - A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
   - A value of 0 indicates no linear relationship between the variables.

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

   Scaling, in the context of machine learning, refers to the process of transforming data to a specific scale or range. It involves adjusting the values of the variables so that they fall within a consistent and comparable range. Scaling is performed to address issues related to the different magnitudes or units of measurement across variables, which can affect the performance of certain machine learning algorithms.

   There are two common scaling techniques: normalized scaling and standardized scaling:

   - Normalized Scaling (Min-Max Scaling):
     Normalized scaling, also known as min-max scaling, transforms the variables to a specific range, typically between 0 and 1.
     The formula for normalizing a variable x is: x_normalized = (x - min(x)) / (max(x) - min(x)).

   - Standardized Scaling (Z-Score Scaling):
     Standardized scaling, also known as z-score scaling or standardization, transforms the variables to have a mean of 0 and a standard deviation of 1.

The formula for standardizing a variable x is: x_standardized = (x - mean(x)) / standard_deviation(x).
Standardization ensures that the variable has zero mean and equal variance, regardless of the original distribution.

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

VIF stands for variance inflation factor.
The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity. Where, R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity".

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 (quantile-quantile) plot is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data against the quantiles expected from the theoretical distribution.

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 fall below and 70% fall 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.

The use and importance of a Q-Q plot in linear regression are as follows:

- Checking Normality Assumption
- Detecting Skewness and Outliers
- Assessing Model Fit