Assignment Questions

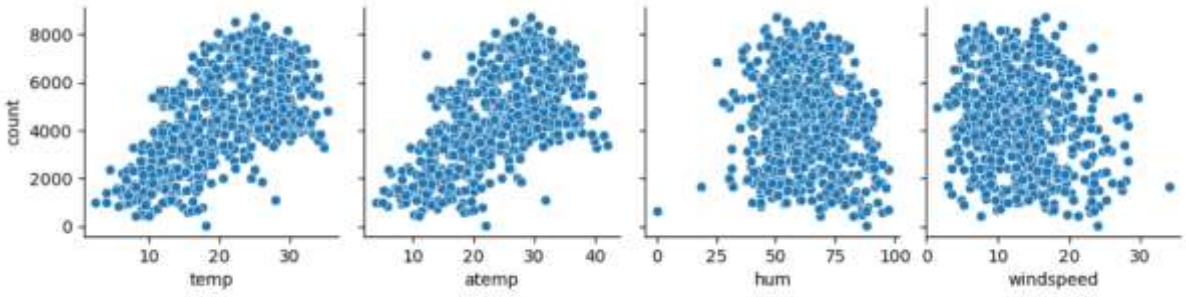
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3-marks)**
   a. The categorical variable in the dataset were 'season', 'yr_2019', 'holiday', 'workingday', 'weathersit', and 'month'.
   b. Season - Fall season had maximum value of count and Spring, the least.
   c. weathersit - Highest count was seen when weathersit was 'Clear', 'Partly cloudy'.
   d. yr_2019 - BoomBikes attracted a greater number of bookings in 2019 compared to 2018.
   e. Holiday - Rentals are less on holidays
   f. month - Trend increased from start of the year till September at peak, and started decreasing till the end of the year.
   g. workingday -> the number of bookings is almost equal on working and nonworking days
   h. weekday -> Thu, Fri, sat, sun have a slightly higher number of bookings compared to start of the week.

2. **Why is it important to use drop_first=True during dummy variable creation? (2-marks)**
   While creating dummy variables for categorical columns, drop_first = True is used to exclude the first category from the set of created dummy variables. This is important for various reasons:

   To avoid multicollinearity: Features including the dummy variables can lead to multicollinearity which can question the model reliability, by dropping one category we create a reference category and this helps in avoiding the perfect multicollinearity and improves model stability

   Model Efficiency: Having more variables in our model may increase the execution time and the model might become very complex which may result in an inefficient model. Hence dropping one category can be very useful.

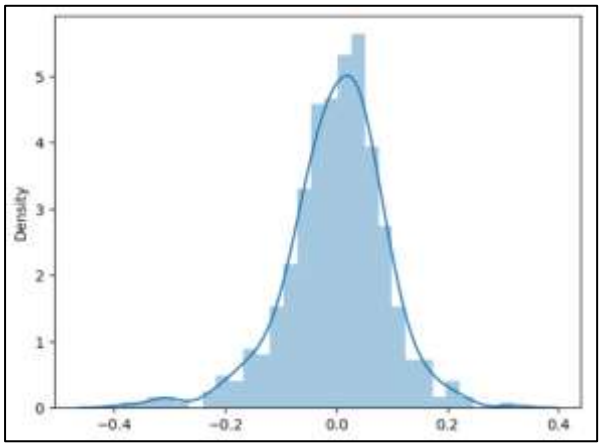
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1-mark)**
   Pair-plot indicates that temp and atemp variables show highest correlation with the target variables

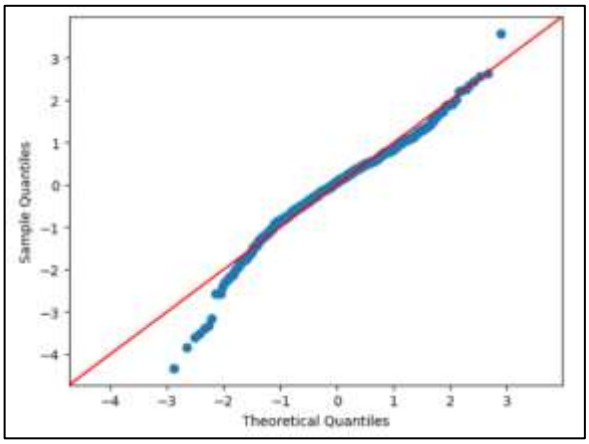


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3-marks)**

After training the model on the train dataset, the following assumptions were validated:

a. **<u>Normality:</u>** the predicted target values were found using the predict function and with that, the residuals were calculated. These residuals when plotted using distplot function, The graph was found to be normally distributed.





b. **<u>Homoscedasticity:</u>** When plotted a scatterplot for the residuals, It was found that the error terms indicate no visible patterns which concludes that the error terms are homoscedastic

c. **<u>Multicollinearity</u>:** Multicollinearity occurs when two or more predictor variables are highly correlated. From the VIF table for the final model we can observe that all the features have the VIF less than 5 which concludes that the independent variables are strongly correlated with only the target variable

	Features	VIF
0	const	50.82
3	temp	2.25
4	hum	1.90
9	month_Jan	1.63
7	season_winter	1.55
13	weathersit_Mist & Cloudy	1.55
6	season_summer	1.43
10	month_Jul	1.43
8	month_Dec	1.26
12	weathersit_Light Snow & Rain	1.25
5	windspeed	1.20
11	month_Sep	1.20
1	yr_2019	1.03
2	holiday	1.02

d. **<u>Linear Relation between X and Y:</u>** Using pair plot, it was evident that there is a linear relationship between numeric variables temp, atemp and count

e. **<u>Error terms must be independent:</u>** From Durbin-Watson score of 2.065, it concludes that there is 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 based on the final model:
   1. Temp with co-eff = 0.572
   2. Yr_2019 with co-eff = 0.229
   3. Season_winter with co-eff = 0.126

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4-marks)**
   Linear Regression is a method for predictive analysis which shows linear relationship between independent variables *(X-axis)* and dependent variable *(y-axis)*. The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted and actual values of the dependent variable. A regression between a single independent variable and dependent variable is called **Simple Linear Regression** whereas a regression between multiple independent variables and dependent variable is called **Multiple Linear Regression.**

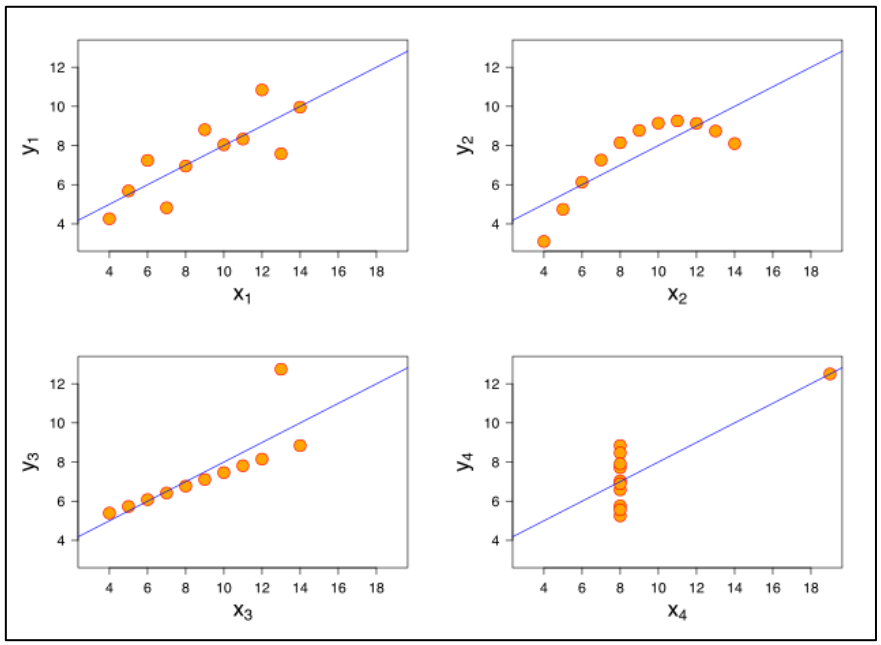
   Simple Linear Regression: $Y = \beta_0 + (\beta_1 * X)$
   Multiple Linear Regression: $Y = \beta_0 + (\beta_1 * X_1) + (\beta_2 * X_2) + (\beta_3 * X_3) + \ldots\ldots\ldots + (\beta_n * X_n)$

   Linear regression algorithm includes:
   a. **Data Cleaning:** Gather the data and identify the independent and dependent variables. The next step is to clean the data that involves handling missing values and outliers, converting the variables into right format etc.
   b. **Visualization:** Visualizing the data to get an initial understanding of how the variables are related to the target variables.
   c. **Data Preparation:** Create dummy variables for the categorical columns and concatenate the same with the main dataset. The next step is to scale the variables appropriately. This is done to avoid multicollinearity and increase the efficiency of the model
   d. **Model Building:** A model can be built using statsmodels or sklearn libraries and choose appropriate features based on p-values and significant values.
   e. **Residual Analysis and Prediction:** Once the model is built, it is time to validate the assumptions. They are
      - Independence of error terms
      - Multicollinearity
      - Linearity
      - Normality
      - Homoscedasticity

   Once the assumptions are validated, we can predict the y values for the test dataset and use metrics like R2, Adjusted R2, MSS values to conclude the model effectiveness.

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

Anscombe's quartet is a set of four dataset that indicates identical statistical properties but display different patterns when graphically represented. It shows how important data visualization is and limitations of solely relying on the summary statistics



The datapoints for the same graphical representation is

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- The first plot indicates a simple linear relationship between two variables.
- The second graph while a relationship between the two variables is is not linear, and the Pearson coefficient is not relevant.
- In the third graph the modelled relationship is linear, but should have a different regression line.
- Finally, the fourth graph shows an example when one high leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

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

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula for Pearson's correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Where:

- x_i and y_i are the individual data points of the two variables.
- $\bar{x}$ and $\bar{y}$ are the means of the two variables.
- *n* is the number of data points.

The resulting value of Pearson's R ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases in a linear fashion. A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases in a linear fashion. 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 is a process of transforming numerical data to a standardized range or distribution. It involves adjusting the values of the variables in a dataset to ensure they are on a similar scale, which can be beneficial for various data analysis and modelling techniques.

Scaling is performed for several reasons:

a. **Comparison:** Scaling allows for a fair and meaningful comparison between variables with different units or scales. It ensures that no single variable dominates the analysis or model fitting process based solely on its larger magnitude.
b. **Algorithm requirements:** Many machine learning algorithms and statistical techniques rely on the assumption that the features are on a similar scale. Scaling helps meet these assumptions, leading to better performance and more reliable results.
c. **Convergence and efficiency:** Scaling can improve the convergence and efficiency of optimization algorithms, particularly when the variables have significantly different ranges. It helps algorithms reach an optimal solution faster and prevents them from being stuck in local minima.

There are two commonly used scaling techniques: normalized scaling (also known as min-max scaling) and standardized scaling (also known as z-score scaling).

a. **Normalized Scaling (Min-Max Scaling):** Normalized scaling transforms the data to a specific range, typically between 0 and 1. It follows the formula:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

where "min" is the minimum value in the dataset, and "max" is the maximum value in the dataset. Normalized scaling preserves the relative relationships and distribution of the data but constrains it to a specific range.

b. **Standardized Scaling (Z-Score Scaling):** Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1. It follows the formula:

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

where "mean" is the mean of the dataset, and "standard deviation" is the standard deviation of the dataset. Standardized scaling shifts the data distribution to have a mean of 0 and adjusts the spread of the data by scaling it with the standard deviation.

The key difference between normalized scaling and standardized scaling lies in the resulting distribution. Normalized scaling preserves the original distribution and range of the data, while standardized scaling transforms the data to have a standardized distribution with zero mean and unit variance. The choice between these techniques depends on the specific requirements of the analysis or modelling task at hand.

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

VIF value is infinite when perfect multicollinearity is present. Perfect multicollinearity means that one or more predictor variables can be perfectly predicted from a linear combination of other predictor variables. This situation typically arises when there are redundant or linearly dependent variables in the regression model

When perfect multicollinearity occurs, one or more variables can be expressed as a linear combination of the other variables. In this case, the regression coefficients cannot be estimated uniquely, and the VIF formula involves dividing by zero, resulting in an infinite value.

Perfect multicollinearity often arises due to data or model specification issues, such as including duplicate variables, including derived variables that are linear combinations of other variables, or when there are data errors.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3marks)**

   A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess the distributional similarity between two datasets. In the context of linear regression, a Q-Q plot is commonly employed to evaluate the assumption of normality for the residuals or the dependent variable.

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

   1. **Assessing Normality Assumption:** Linear regression assumes that the residuals (the differences between the observed values and the predicted values) follow a normal distribution. The Q-Q plot allows you to visually inspect whether this assumption holds. If the points in the plot align closely with the straight line, it indicates that the residuals are normally distributed, which is desirable for making reliable statistical inferences.
   2. **Detecting Departures from Normality:** A departure from the straight line in the Q-Q plot indicates a deviation from the normality assumption. If the points diverge from the line in a systematic pattern, it suggests that the residuals deviate from a normal distribution. This can have implications for the validity of the regression model and the reliability of statistical tests and confidence intervals.
   3. **Identifying Skewness or Kurtosis:** In addition to detecting departures from normality, a Q-Q plot can also reveal skewness (asymmetry) or kurtosis (heaviness of the tails) in the data. If the points deviate from the line at the extremes, it indicates heavy-tailed or light-tailed distributions compared to the assumed normal distribution.