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: The models indicates the values of R-squared (82.5%) and adjusted R-squared (82.1%) has the selected 10 features which are significant for prediction. P(F-statistic) of the model is close to 0 which indicates that this model is a good fit. Coefficients of the features indicate how they are related with count. Top 3 significant features that contribute towards explaining towards the demand of the shared bikes.

   Temperature with coefficient 4725.82 indicates that as temperature increases, the demand for bikes increase. Year with coefficient 1989.56 indicates that the bike demand has increased from last year and is expected to increase in future. Season_4 with coefficient 1132.48, indicates that the Organization should take more measures in this season to create a very high demand for bikes.

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

   Ans: A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups.

   We need to drop first dummy variable because one level of your categorical feature become the reference group during dummy encoding for regression and is redundant. Alongside we are trying to avoid multicollinearity. If you don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

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

   Ans: Observation from Pair Plot:

   Looking at the correlation between Count and Temperature, shows higher temperatures lead to an increasing number of bike rents and lower temperatures decrease the number of rents. For Humidity and Windspeed, Count seems to be distributed uniformly

   Based on the Correlation Heat map

   We can see high correlations between different seasons for different Months. For example, Season_4 is correlated with Months 10,11,12. Temperature is correlated with Count of Bikes being rented. Temperature is also correlated negatively for few months indicating lower temperature for those months and positively correlated for few months indicating higher temperatures Working Day is negatively correlated to Weekday_6 which is Saturday Different Weather Situations are correlated to Humidity. Year is highly correlated with Count of bikes rented indicating 2019 has mode rented bikes than 2018.

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

   Ans. The validation and the assumptions of Linear Regression are as follows:

   - Linear relationship between X and Y: This was proved with the help of the pair plot
   - Error terms are normally distributed :Residual Analysis was performed
   - Error terms are independent of each other
   - Error terms have constant variance (homoscedasticity)

   We have calculated the RFE (Recursive Feature Elimination) and compared the p-values, VIFs and R-squared scores, dropping/eliminating redundant features one by one, in order to achieve the best fitted line graph and reducing multicollinearity

   Strategies we utilised:

   - When a feature has high p-value and high VIF, then it can be dropped first.
   - When a feature has high p-value and low VIF, then we will drop it as it is not significant.
   - When a feature has low p-value and high VIF, we try to drop other features which have higher p-value and if there is none, we will remove this feature and build the model and check the R-squared.
   - When a feature has low p-value and low VIF, then it is significant and not explained by other features and hence we keep these features in the model.

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

   Ans: Based on the final model:
   For the market, Data suggests the following remarks for the demand of rental bikes:

   - Year on year, the demand for rental bikes is increasing which is a good sign for the Organization.
   - People prefer using the rental bikes during higher temperatures. So, measures should be taken to attract customers during days with lower temperatures.
   - Summer and Winter seasons have been preferred by customers to use rental bikes
   - People don't really prefer using Rental bikes on Holidays, So discounts or offers has to be given to people to use rental bikes on Holidays to attract more customers.
   - As wind speed increases, customers tends to ignore the rental bikes.

   Therefore, top three features contributing towards the demand of shared bikes are:-

   a. Weather Situation
   b. Working Vs Holiday statistics
   c. Season

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans. Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression is a method of modelling a target value based on independent predictors. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

Regression analysis is used for three types of applications:

- Finding out the effect of Input variables on Target variable.
- Finding out the change in Target variable with respect to one or more input variable.
- To find out upcoming trends.

Types of Linear Regression:

- Linear Regression
- Multiple Linear Regression
- Logistic Regression
- Polynomial Regression

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

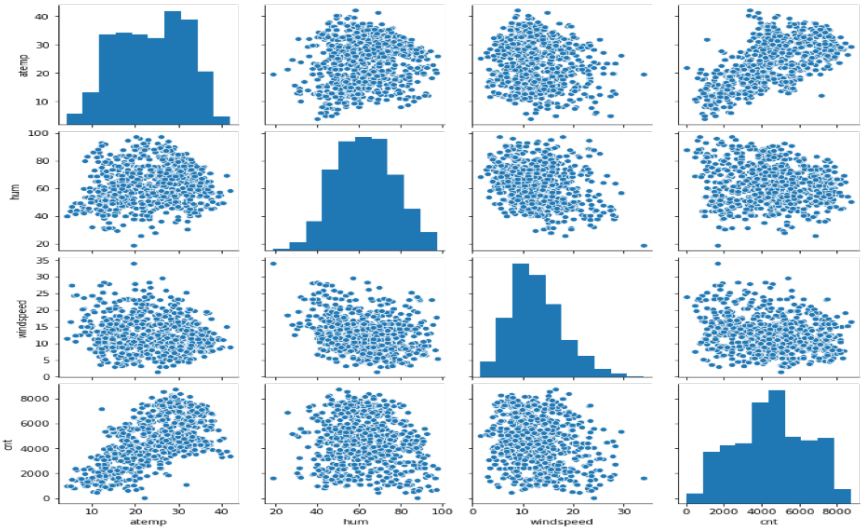
$$a\,(intercept) = \frac{n\sum y - b\,(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

B = Slope of the line.
A = y-intercept of the line.
X = Independent variable from dataset
y = Dependent variable from dataset

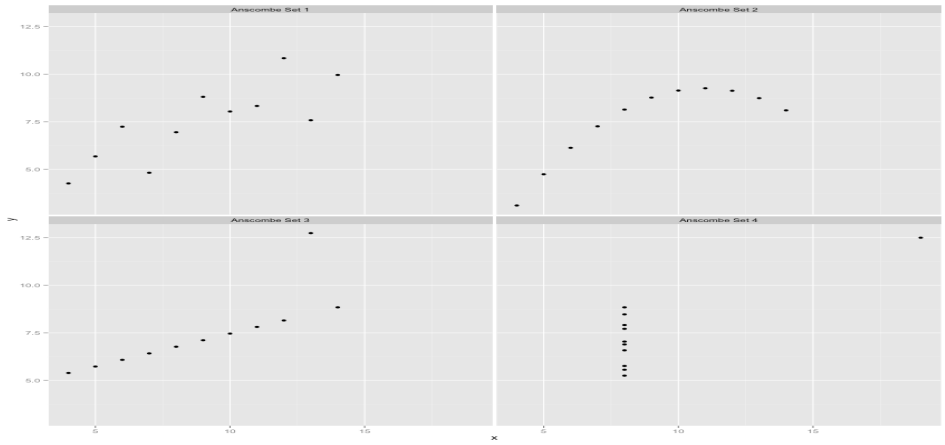
2. Explain the Anscombe's quartet in detail.

Ans. Anscombe created the dataset to demonstrate the importance of visualizing data and also to highlight the effect that outliers can have on a statistical findings of a dataset. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed.



For Example,

	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



When we plot these four datasets on an x & y coordinate plane, we can observe that they show.

3. **What is Pearson's R?**

Ans. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used. Pearson's correlation coefficient (r) for continuous interval level, data ranges from -1 to +1. Problem with Pearson's correlation is not being able to tell the difference between dependant and independent variables.

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

Ans. Scaling/ Feature Scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model.

Most of the times, our dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, there is a problem. To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

There are two types of scaling techniques:

- Normalised scaling
- Standardised scaling

Difference between Standardized and Normalised Scaling:

Standardisation:

The result of standardization (Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively. The equation is shown below :-

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

This technique is to re-scale features value with the distribution value between 0 and 1 is useful for the optimization algorithms, such as gradient descent, that are used within machine learning algorithms that weight inputs (e.g., regression and neural networks). Rescaling is also used for algorithms that use distance measurements.

CODE:-

```
#Import library
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
sc_X = sc_X.fit_transform(df)
```

Max-Min Normalization:

Another common approach is the so-called Max-Min Normalization **(**Min-Max scaling). This technique is to re-scales features with a distribution value between 0 and 1. For every feature, the minimum value of that feature gets transformed into 0, and the maximum value gets transformed into 1. The general equation is shown below:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

CODE:-

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaler.fit(df)
```

In contrast to standardisation, we will obtain smaller standard deviations through the process of Max-Min Normalisation.

Max-Min Normalisation typically allows us to transform the data with varying scales so that no specific dimension will dominate the statistics, and it does not require making a very strong assumption about the distribution of the data.

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

Ans. When there is perfect correlation, the VIF tends to increase towards infinity. The variance model also tends to become inflated, if there is a presence of multi-collinearity i.e. VIF tends to 10 or higher. The Regression model also gets affected and it hinders to conclude a best fit model.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique used for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile we mean fraction or percentage. Quantiles are breakpoints that divided our numerically ordered data into equally proportioned buckets

The advantages of Q-Q plot are:

- The sample size doesn't need to be equal.
- Many distribution aspects simultaneously tested.