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 in the provided dataset are: (a) Season, (b) Year, (c) Month, (d) Holiday, (e) Weekday , (f) Workingday and (g) Weathersit. The individual levels on each of these had a significant impact on the final model created. Here are the insights from the final model:

(a) Fall season has the highest number of demand for rentals

(b) Yearly demand has grown.

(c) Months of March, August, September and October have a positive correlation coefficient which signifies the demand for bikes in these months.

(d) Among weekdays, Saturday seems to be the day when the bike rentals are higher

(e) Weather situation of mist and light rain, the demand is less as we can see they are negatively correlated.

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

During the dummy variable creation, one column is created for each of n levels of the categorical variables. Since one of the columns could be easily derived based on the non-existent values of other variables, that could be dropped.

For example, in the current dataset, we had a categorical variable called season which took 4 values - spring (1) , summer (2), fall (3) and winter(4). Here n = 4. While creating dummy variables, the presence of the column value is denoted by 1 and 0, otherwise. So, the column Spring will have value of if season variable has 1 and 0, if it has 2, 3 or 4. After creating the 4 dummy variables, we end up with 4 combinations: 1000, 0100, 0010 and 0001.

Since one of the variables could be easily determined based on the remaining combinations, it is not required while doing the modeling. In this case, if we drop Spring column, the values are going to be 0 for other 3 variables - summer, fall and winter, which is very obvious. Hence the **drop_first=true** is required to remove this variable/ column from the dataset for further analysis so as to be non-redundant and to ensure we keep only the required variables for analysis.

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

Looking at the pair-plot among numerical variables, **temp and atemp** have the highest correlation with target variable **cnt** at **0.64 and .65** respectively.

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

Validation the assumptions on the Linear regression after building the model on the training set was done as follows:

A. Residual Analysis

- The errors are normally distributed with a mean of 0
- Actual and Predicted values show the same pattern

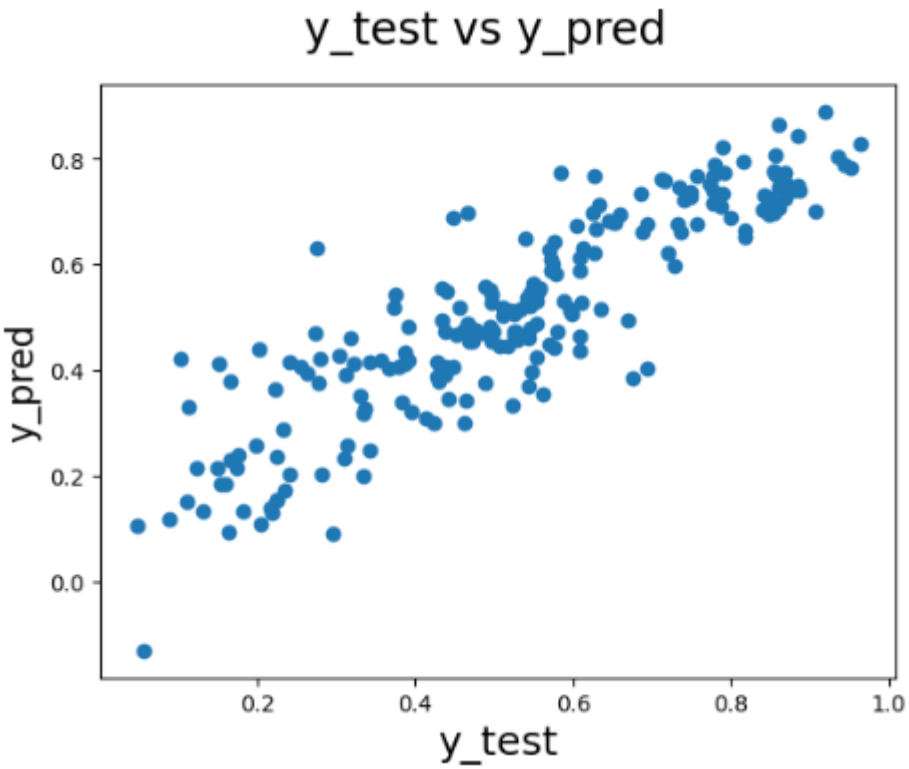
- Error terms are independent of each other

B. R2 validation on the test dataset

- R2 value on the test dataset (0.773) is very close to the R2 value on the train data set (0.782). So we can safely conclude that the model is performing well on the unseen test dataset as well.

C. Plot Test vs Predicted value set

- Scatter plot on the Test and Predicted values were very close and we can visually observe a linear pattern. Hence, we can conclude that the model is performing well.



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 in the final model that contribute significantly towards explain the shared bikes demand are:

1. Year (positively correlated)
2. Fall season (positively correlated)
3. Months of Sep and October (Positively correlated)

Among the factors that affect the model negatively are: the windspeed and weather_mist and weather_light as they negatively correlated.

General Subjective Questions

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

The Linear Regression basically helps understand how one or more independent variables are related to a dependent variable. The relationship is based on coefficient that are multiplied to these independent variables. The product of independent variables and their coefficients are summed up along with a constant value called intercept that eventually derives the dependent variable.

There are two types of Liner Regression model:

a. Simple Linear Regression
b. Multiple Linear Regression

a. **Simple Linear Regression:** In this model, we estimate the relationship between two quantitative variables. There is one independent variable and one dependent variable and a constant. The formula for simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- y is the predicted variable based on different values of X
- β_0 is the intercept, the predicted value of y when the X is 0.
- β_1 is the coefficient – a value that determines how much y changes when x increases
- X is the independent variable - the one that influences the changes to the value of y
- ε is the error of estimate, how much error there is in our estimate of regression coefficient.

Linear regression finds the best fit line through the data by plotting the value of Y for different values of X. While doing this, it finds the best coefficient value (β_1), which minimizes the error (ε).

b. Multiple Linear Regression: Multiple Linear Regression is an extension of Simple Linear Regression model in that it has more than one independent variable. The formula for calculating multiple linear regression is given as follows:

Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$

- Y is the dependent variable or predicted variable
- X_j is the j^{th} independent variable
- β_0 is the intercept
- β_j is the coefficient of the j^{th} independent variable, which is practically the effect on Y based on one unit increase of X, while holding all other independent variables constant.
- ε is the error term in the model

while plotting the regression in a graph, it gives a sloped straight line that describes the relationship between independent and dependent variables as shown below:

The red line in the graph indicates the best fit line.

One good example for a regression is the salary of an employee in an organization based on the years of experience. Here, the salary will be the dependent variable and number of years of experience is the independent variable. The salary is dependent on the number of years of experience. This information can be used to predict the future salary as the person's experience increases over the years.

On a final note, Linear Regression can be positive or negatively related.

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

Anscombe's quartet is used to illustrate the need for visualizing data and not just rely on summary statistics alone for interpretation. It also emphasizes the importance of using exploratory data analysis and visualization techniques by spotting trends, outliers, and other details that might not be obvious with the numerical statistical observations.

Anscombe quarter comprises of a set of four datasets, that have an identical statistical property such as mean, variance, R2 and correlations but having different representations while spotting on a graph.

This method was demonstrated by statistician Francis Anscombe in 1973 to outline the importance of visualizing data and not just rely on the summary statistical observations, which could be misleading the interpretation.

Consider the following 4 datasets:

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

(source: https://www.geeksforgeeks.org/)

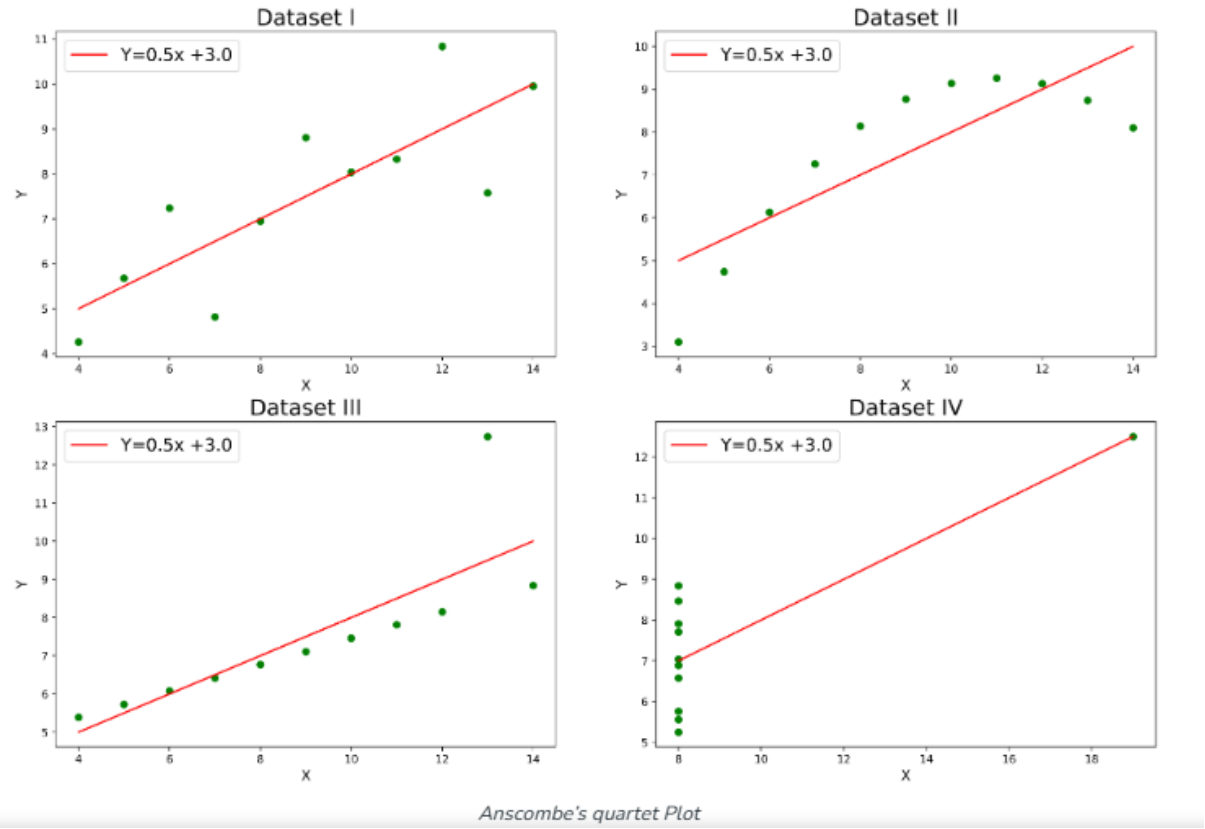
The summary statistics of these datasets look like below:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

(source: https://www.geeksforgeeks.org/)

By reading these observations, once can easily conclude that the data sets are identical in nature as they have the same mean variance, correlation coefficients and linear regression.

However, when these datasets are plotted in a graph, it looks like the following:



Anscombe's quartet Plot

(source: https://www.geeksforgeeks.org/)

Clearly, all these 4 graphs are different even though they have the same statistical properties. Interpretation of these graphs is as follows:

- **Dataset I**: It seems there is a linear relationship between independent variable (X) and dependent variable (Y)
- **Dataset II**: In this set, we can see that there is NO linear relationship between independent variable (X) and dependent variable (Y)
- **Dataset III**: In this set, the relationship between the two independent and dependent variables are linear but there is one outlier (which can be observed in the far top right in the graph)
- **Dataset IV**: In this set, one can observe the outliers in the data which cannot be fit with a linear regression model.

In summary, Anscombe quartet reveals the importance of using visualization techniques to spot the trends for interpretation and not just rely on the statistical observations of the dataset.

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

Pearson's R is the statistical measure that is used for assessing linear relationship between an independent and dependent variable. It quantifies both the strength and direction of the relationship of these variables. Here are some of the key points about this metric:

1. **Correlation Coefficient(r):** Pearson's correlation coefficient lies between -1 and 1. It indicates how closely the two variables are related.
    a. **Positive Correlation**: A value of 0 to 1 indicates a positive relationship between the two variables. If one increases, the other also tend to increase. Example: Temperature vs Ice Cream sales. When the temperature soars, the ice cream sales also increase, typically.
    b. **No Correlation (0)**: There is no relationship between the two variables. Example: Exercise vs Singing. There is absolutely no relationship between these two parameters.
    c. **Negative Correlation**: A value of -1 to 0 indicates a negative correlation. It means that when one variable increase, other one tends to decrease. Example: Cooking at home vs eating out. The more you cook at home, you tend to eat outside less.
2. **Interpretation:** The interpretation of these coefficient values is summarized in the table below:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

3. **Using Perason's R:** Pearson's correlation coefficient(r) can be used only when all of the following are true in a dataset
    a. Both variables are quantitative in nature.
    b. The variables are normally distributed. A histogram plot on each of the variables can reveal if the values are spread as normal.
    c. The data does not have any outliers. Scatterplot can be a good tool to check for outliers
    d. Relationship is linear between variables.
4. Formula: Pearson correlation coefficient is calculated using the formula given below:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

x, y are variables whose data is provided for coming with the correlation coefficient value

n – number of data points

r – is the resultant value of correlation coefficient.

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

Scaling is a concept of bringing data points of features into a common unit so as to make sure the measures / calculations are as accurate as possible and meaningful. Scaling avoids numerical issues, and improve performance of linear regression algorithm.

It has to be performed on data sets to ensure the output of linear regression models has a fair comparison against all features and they all contribute equally during the model building process. For example, let us assume one of the features in a data set has values ranging from 0 to 1 where as another feature has data range as 1000 to 2000. If the scaling is not done on these, the model would interpret large-scale feature as more important than the other one. This is not a fair comparison and would lead to biased results. Hence scaling on the datasets is very important and must be performed so that coefficients and prediction variables are properly computed.

Normalized Scaling: This method of scaling values on a feature (also known as min-max scaling) is to consider min and max values of the data points and computes scaled value as $(X_i - X_{min})/(X_{max} - X_{min})$. In this method all datapoints of the feature would be in the range of 0 to 1.

Standardized Scaling: In this method, we transform the feature's data such that mean is 0 and variance is 1. This is also called Z-score normalization.

Differences between Normalized scaling and Standardized scaling are as follows:

Normalized Scaling	Standardized Scaling
Minimum and maximum values in a feature are used for scaling	Mean and Standard deviation is used for scaling
The values typically range from 0 to 1	Values does not fall within a specific range. It depends on the data that is being processed.
Affects by outliers	Outlier values have much lesser impact
Used when features are of different scales	Used when we need to ensure zero mean and 1 std deviation for the data
It is useful when we do not know about the how data is distributed	It is used when the datapoints of the feature are Normal

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

Variance Inflation Factor (VIF) measures the severity of multicollinearity in a regression analysis. It measures the amount of relationship two or more features have, which are part of the model. It is computed using the formula

$$VIF_i = 1 / (1 - R_i^2)$$

Where Ri2 represents the adjusted coefficient of determination for regressing the i^{th} independent variable over the remaining ones. General rule of thumb is that a VIF value of 5 or above indicates high multicollinearity might exist and further investigation is required.

VIF tend to infinity if the R_i^2 value is 1. This means that the given independent variable can be predicted perfectly (perfect correlation) by other variables in the model.

Once we find out that the multicollinearity exists in the dataset, they must be removed to ensure the model predicts correctly. Some measures to overcome this issue are:

a. removing redundant variables
b. Combine variables to form a dummy variable
c. Use ridge regression or lasso regression technique to penalise large coefficients
d. Applying Principal component analysis (PCA)

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

Q-Q plot or quantile-quantile plot is a graphical method for determining if a dataset follows a certain probability distribution, such as normal distribution.

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution. Some examples of quantiles are:

a. Median – It represents the 50^{th} percentile in a dataset which is the middle value when it is ordered from smallest to largest. It divides dataset into two halves.
b. Quartiles (25^{th}, 50^{th}, 75^{th} percentiles) – This method divides the dataset into 4 parts or quarters. The first quartile is all values below 25% in the dataset, the 2^{nd} quartile represents all values that falls below 50%. Similarly, the third quartile represent all values falling below 75% and the last quartile represent data between 75% and 100%.
c. Percentiles: Percentile is similar to quartiles but divide the dataset into 100 equal parts. For example, 95^{th} percentile indicates all values below 95% of the data

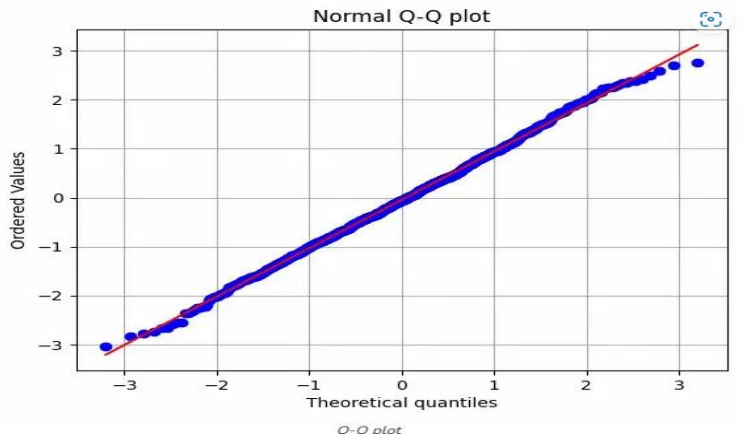
Q-Q plot is a plot of the quantiles of the first dataset against the quantiles of the second set to determine the relationship of the distribution.

The way Q-Q Plot is plotted is as follows:

1. Collect data for the analysis
2. Sort the data and identify quantiles
3. Choose the distribution that you want to compare the dataset against. Example: normal distribution
4. Compute the quantiles of this chosen distribution on the dataset
5. Plot the dataset in x axis
6. Plot the theoretical distribution values in y axis
7. Connect these two datapoints (x,y) and observe the trend

Interpretation: If the plotted values have a linear relationship, then it can be concluded that the theoretical distribution assumed on the dataset is correct. Otherwise, another distribution could be plotted and steps 3 to 7 described above can be followed to determine the fitness.

A sample Q-Q plot is given below:



Source: geeksforgeeks website

Other usages of Q-Q plot in linear regression model:

A Q-Q plot can be used to check for assumptions in regression models. For example, Q-Q plot can be used to check if the residuals of the model are normally distributed. This assumption is used for many parametric tests and confidence levels. It can be used to find out if their dataset has a constant variance, which is an assumption of homoscedasticity of the model.

References:

Pearson Correlation Coefficient (r) | Guide & Examples (scribbr.com)

https://www.geeksforgeeks.org/

Quantile Quantile plots - GeeksforGeeks

https://www.baeldung.com/cs/normalization-vs-standardization

Normalization vs Standardization - GeeksforGeeks

https://corporatefinanceinstitute.com/

Quantile Quantile plots - GeeksforGeeks