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 from the dataset are season, yr (year), mnth (month), holiday, workingday, weekday, and weathersit (weather).

- In the spring season, the number of users is comparatively lesser than in the other seasons.
- In Fall, summer, and winter the number of users is nearly the same.
- The number of bike users was higher in 2019 than in the year 2018, so the business has improved
- A greater number of rental bike users are seen in the months of August, June, July, May, September, and October.
- Fewer numbers of rental bike users are seen in the months of December, January, February, and March.
- The number of bike users is almost the same in both working days and non-working days.
- The users avail more bikes in the months of May and June, especially on Sundays.
- In general (from Jan to Dec), the bike user's demands are high on Thursday, Friday, and Saturday.
- More bike users are seen in clear clouds or few clouds and during mist cloudy.
- The number of users starts decreasing when there is a little change in snow, thunderstorms, and rain.

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

When creating a dummy variable, it is important to use drop_first=True because

- It reduces the extra column after the dummy variable is created.
- Also, it reduces the correlations created among dummy variables.

Example: Consider there is a column named c1.

```
c1 = ['furnished', 'furnished', 'semi-furnished', 'furnished', 'furnished']
```

After creating dummies,

	furnished	Semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

Now, three columns are not needed. The furnished column can be dropped, as the type of furnishing can be identified with just the last two columns where —

- 00 will correspond to furnished
- 01 will correspond to unfurnished
- 10 will correspond to semi-furnished

So, if the first variable is dropped i.e., drop_first=True

	Semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0

Hence, it reduces a column and correlations created among the dummy variables.

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

   - The highest correlation with the target variable (count) is with “temp”, and ”atemp” which has a value of 0.63
   - Since “atemp” is a derived parameter from “temp”, this feature is not considered for modeling.

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

   The assumptions of Linear Regression after building the model of the training set are validated by

   1. The error terms should be normally distributed and centred at 0
   2. By checking the error patterns. There should not be a correlation between error (residual) terms (Autocorrelation)
   3. The error terms should have constant variance (Homoskedasticity)
   4. The independent variables should not be correlated. This is checked by the values of VIF, i.e., (Multicollinearity)
   5. Linear relationship between dependent and independent variables.

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

   The top 3 features that contribute significantly explaining the demand for shared bikes are:
   - Temp (Temperature)
   - Year (yr)
   - Season_4 (Winter)

General Subjective Questions

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

   - Linear regression is an algorithm in statistics and machine learning that provides a linear relationship between one or more independent variables(X) and a dependent variable (Y) to predict future outcomes.
   - Linear regression algorithm is used for predictive analysis.

   There are two types of linear regression.

   1. Simple linear regression
   2. Multiple linear regression

Simple Linear Regression:

With only one independent variable, the basic form of the regression model is given as

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

Where,

y – dependent variable

β_1 – slope of the line

X – independent variable

β_0 – intercept

Linear regression aims to find the best-fit line that minimizes the difference between the actual values of the data and the predicted values of the linear model. This is done by the Ordinary Least Square (OLS) method. This method minimizes the sum of squared differences between the actual and predicted values.

Multiple Linear Regression:

Linear regression can be extended to handle multiple independent variables which is called multiple linear regression.

The equation is given as,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_n X_n$$

Where,

$\beta_1, \beta_2, \ldots \ldots \beta_n$ – coefficients associated with each independent variable.

Algorithm of linear regression:

1. Data Collection
2. Pre-processing of the data – prepares the data for modeling. Checking missing values, scaling, mapping categorical variables
3. Splitting into the training set and testing set – Training set to train the model and testing set to evaluate the performance.
4. Model Training – Fits the linear regression model with the training data.
5. Model Evaluation – evaluates the performance of the model on the testing data. The parameters like R squared, MSE (mean squared error), and MAE (mean absolute error) are evaluated
6. Prediction – once the model is trained and evaluated, it can be used to make predictions on new data

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

- Anscombe's quartet was created by the statistician Francis Anscombe in 1973.
- Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics, but drastically different distributions and relationships when graphed.
- This quartet was created to emphasize the importance of visualizing data rather than solely on summary statistics.
- Each dataset in Anscombe's quartet consists of 11 paired observations of two variables.
- Despite having the same mean, variance, correlation coefficient, and linear regression

line, the datasets exhibit distinct characteristics when plotted. This shows how summary statistics alone may not fully capture the underlying patterns or relationships in the data.

The four datasets in Anscombe's quartet are typically described as follows:

1. Dataset I: A simple linear relationship with minimal noise.
2. Dataset II: A non-linear relationship that closely resembles a quadratic function.
3. Dataset III: A perfect linear relationship except for one outlier, which heavily influences the regression line
4. Dataset IV: No linear relationship between the variables, but the summary statistics are similar to the other datasets.

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

- Pearson's correlation coefficient denoted as $\mathfrak{r}$, is a measure of the linear relationship between two continuous variables.
- It quantifies the strength and direction of the relationship between two continuous variables.
- The values range from -1 to 1

$\mathfrak{r}$ = 1 indicates a perfect positive linear relationship
$\mathfrak{r}$ = -1 indicates a perfect negative linear relationship
$\mathfrak{r}$ = 0 indicates no linear relationship between the variables

The Pearson's coefficient is calculated by

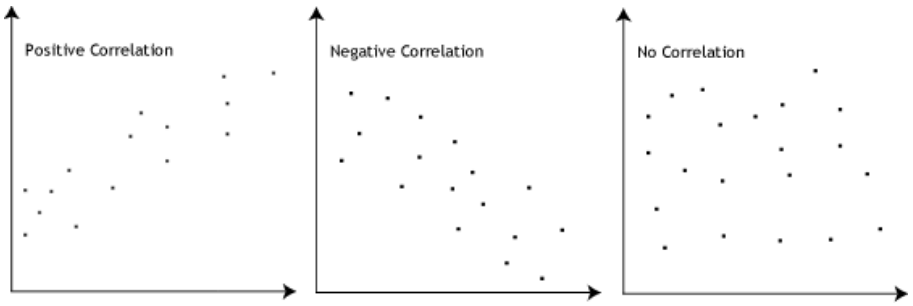
$$\mathfrak{r} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

$x_i \ and \ y_i$ – individual data points for the two variables.
$\bar{x} \ and \ \bar{\bar{y}}$ - means of the two variables respectively.

- Pearson's correlation coefficient measures only linear relationships and cannot capture non-linear associations between variables.

Below graph represents the Pearson's coefficient and its correlation.



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

- Scaling is a pre-processing step in machine learning used to normalize the range of independent variables or features of a dataset.
- The purpose of scaling is to bring all features to a similar scale or range, which can improve the performance and convergence speed of many machine learning algorithms

Scaling is performed for

1. Improving algorithm convergence
2. Improving model performance
3. Interpretability

There are two methods of scaling:

1. Normalized scaling
2. Standardized scaling

Difference between normalized scaling and standardized scaling

S. No.	Normalized scaling	Standardized scaling
1.	Minimum and maximum values of the features are used for scaling.	Mean and standard deviation are used for scaling.
2.	$normalization = \frac{(x - x_{min})}{(x_{max} - x_{min})}$	$standardization = \frac{x - mean(x)}{\sigma}$
3.	Values are between 0 to 1 and -1 to 1	The values are not bounded.
4.	It is affected by outliers.	It is less affected by outliers.

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

- The Variance Inflation Factor (VIF) is used to detect multicollinearity.
- Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other.
- When it is correlated to each other, the R_i^2 value equals to 1.

The formula to calculate is

$$VIF_i = \frac{1}{(1 - R_i^2)}$$

- If R_i^2value is 1, then VIF becomes infinite.
- When VIF values are infinite, it represents severe multicollinearity problem and needs attention.
- Hence, affected variables needed to be removed from the model or transformation or additional data cleaning are necessary to resolve the multilinearity issue.

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 plot, short for quantile-quantile plot, is a graphical tool used to assess whether a set of data follows a particular distribution, such as the normal distribution.
- It compares the quantiles of the dataset against the quantiles of a theoretical distribution (typically the normal distribution).
- The points on the plot represent the quantiles of the dataset against the quantiles of the theoretical distribution, and if the data points fall approximately along a straight line, it suggests that the data closely follows the specified distribution.

The use and importance of Q-Q plots in linear regression include:

1. Assumption checking:
    - Q-Q plots are used to assess whether the residuals from a linear regression model follow a normal distribution.
    - This is an essential assumption of linear regression, known as the normality of residuals assumption.
2. Detecting non-normality:
    - Deviations from a straight line in the Q-Q plot indicate departures from the normal distribution.
    - If the points exhibit systematic deviations, it suggests that the residuals may not be normally distributed.
    - Non-normality of residuals can affect the validity of statistical inferences and predictions made by the linear regression model.
3. Model diagnostics:
    - Q-Q plots are part of a set of diagnostic tools used to evaluate the assumptions and performance of linear regression models.
    - By visually inspecting the Q-Q plot, analysts can identify potential issues such as skewness, heavy tails, or outliers in the residuals.
4. Decision-making:
    - Q-Q plots help researchers and analysts make informed decisions about whether the assumptions of linear regression are met and whether the model is appropriate for the data at hand.
    - If substantial departures from normality are observed, alternative modeling techniques or data transformations may be considered.