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 effects of categorical variable on the dependent variable are:

   - season: This variable represents seasonal variations (like spring, summer, fall, winter). Warmer seasons like spring and summer increase bike demand, while colder seasons like winter decrease bike demand due to less favourable biking conditions.
   - weathersit: It reflects weather conditions such as clear, mist, and light rain/snow. Clear weather boosts bike usage, while poor weather conditions (e.g., snow, rain) reduces demand as people prefer not to bike in adverse conditions.
   - yr: This is a binary variable (0 for 2018 and 1 for 2019). It indicates how demand of bike sharing has increased over the time. The positive coefficient for yr suggests higher bike demand in 2019, likely due to increased popularity of the service.
   - mnth: The mnth variable accounts for the specific month of the year, with demand commonly peaking during warmer months like May through September and decreasing in colder months like December and January.
   - holiday: This binary variable indicates whether the day is a holiday. Demand patterns differ on holidays, as people are more likely to use bikes for recreational purposes.
   - workingday: This variable differentiates between working days and weekends or holidays. Working days tend to have more structured demand related to commuting, while weekends/holidays might see more casual bike rentals.
   - weekday: The weekday variable identifies the specific day of the week. Some days may see higher commuting activity (e.g., Monday through Friday), while weekends (Saturday and Sunday) could show more recreational usage.

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

   drop_first=True is a parameter used in get_dummines function of panda library in python. As many times, we have non-numeric variables in our dataset which is also known as categorical variable. We cannot use these variables directly into the model as they are non-numeric. So dummy variable is to be created for each discrete categorical variable for a feature.

If we don't use drop_first= True then if there are n levels then n dummy variables will be created and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap. The multicollinearity arises because the sum of all dummy variables equals 1, which leads to perfect correlation.

So, we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level. This prevents the dummy variable trap, where perfect multicollinearity occurs. Furthermore, the dropped category becomes the reference point. The remaining dummy variables are interpreted relative to this reference. This makes model easier to interpret and avoid redundancy.

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

   According to the pair-plot temp variable has the highest correlation with the target variable cnt among all the numeric variable (as registered and causal are not feature variables). This strong correlation reflects that registered variable contribute significantly to the total bike demand.

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

   After developing the model on the training data, I performed the following steps to confirm the linear regression model assumptions:

   - Residual Plot: I plotted the residuals to check for random distribution, which helps verify the assumptions of linearity and homoscedasticity. In my model, the residuals were roughly randomly distributed, indicating that the model appropriately captures the relationship between the independent and dependent variables. Furthermore, the spread of the residuals was consistent across the range of fitted values, supporting the assumption of homoscedasticity.
   - Distribution of Residual: To assess the distribution of residuals, I created a histogram. This test evaluates whether the residuals follow a normal distribution, a key assumption of linear regression. While there was some variation, the histogram displayed a generally bell-shaped distribution, suggesting that the residuals are approximately normally distributed with no significant outliers.
   - Q-Q Plot: I also generated a Q-Q plot to further investigate the normality of the residuals. This plot compares the distribution of residuals to a theoretical normal

distribution. In my model, the Q-Q plot showed some deviations at the tails, which indicate potential issues with normality. However, the majority of points fell close to the reference line, suggesting that the residuals are reasonably normally distributed and acceptable for the linear regression model.

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

   According to my final model output, the top three factors that strongly explain the demand for shared bikes are:

   - temp: With a coefficient of 0.4412 and a t-value of 14.680, this variable is significant. The positive relationship shows that when temperatures rise, so does the demand for bike rentals, indicating a strong preference for biking in warmer weather.
   - yr: This variable has a positive coefficient of 0.2321 and a t-value of 27.138, implying that bike demand tends to increase over the years. This suggests that the popularity of bike-sharing services is growing over time.
   - hum: This variable has a negative coefficient of -0.1285 and a t-value of -3.270, indicating a significant negative impact on bike demand. Higher humidity levels tend to deter individuals from renting bikes, demonstrating that adverse weather conditions significantly influence bike usage.

General Subjective Questions

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

   Linear regression is a supervised method that computes a linear relationship between the dependent variable and one or more independent features by fitting a linear equation to the observed data. When there is only one independent feature, it is known as Simple Linear Regression, and when there is more than one feature, it is known as Multiple Linear Regression.

   Simple Linear Regression:
   This is the simplest form of linear regression, and it contains only one independent variable and one dependent variable. The equation for simple linear regression is:

   $$Y=\beta_0 + \beta_1 X$$

   were
   - Y is the dependent variable,
   - X is the independent variable,
   - β0 is the intercept and β1 is the slope.

   Key concepts are:
   - A Linear relationship exists between X and Y.
   - Error terms are normally distributed.
   - Error terms are independent of each other.
   - The error terms exhibit constant variance (homoscedasticity).

   Multiple Linear Regression:
   Multiple Linear Regression extends simple linear regression to include multiple independent variables. It is essential when a single variable does not adequately explain the variability in the dependent variable. The equation for multiple linear regression is:

   $$Y=\beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_n Xn$$

   were
   - Y is the dependent variable
   - X1, X2, …, Xn are the independent variables
   - β0 is the intercept
   - β1, β2, …, βn are the slopes

   A lot of concepts are same or simply extension of single linear regression:
   - The model fits a hyperplane instead of a simple line, representing the relationship in a higher-dimensional space.

- Coefficients are obtained by minimizing the sum of squared errors, typically using the Ordinary Least Squares (OLS) method, which finds the best-fitting line by minimizing the squared differences between observed and predicted values.
- The assumptions from simple linear regression still hold, including zero mean, independent normally distributed error terms, and constant variance.

New considerations for MLR are:

- Overfitting: Adding more features can lead to a model that becomes overly complex, capturing noise rather than the underlying relationship, resulting in poor performance on unseen data.
- Multicollinearity: This occurs when predictor variables are highly correlated with each other, which can inflate the variance of coefficient estimates and make the model unstable.
- Feature Selection: Selecting an optimal set form a pool of a given features many of which might be redundant becomes an important task.

To assess the performance of linear regression models, several metrics can be used:

- R-squared: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables, providing insight into the model's explanatory power.
- Adjusted R-squared: Adjusted for the number of predictors in the model, useful for comparing models with different numbers of features while accounting for the potential inflation of R-squared due to additional variables.
- Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction, giving a clear picture of the model's performance.
- Root Mean Squared Error (RMSE): The square root of the average of squared differences between predicted and observed values, providing insight into the model's predictive accuracy and emphasizing larger errors.

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

   Anscombe's quartet was created in 1973 by statistician Francis Anscombe to demonstrate the significance of charting data before analysing it and developing a model.

Anscombe's quartet is a collection of four datasets with equal descriptive statistical qualities in terms of means, variance, R-squared, correlations, and linear regression lines, but different representations when plotted on a graph.

Anscombe's quartet has four datasets, each of which contains 11 x-y pairings of data. When plotted, each dataset appears to have its own connection between x and y, complete with distinct variability patterns and correlation strengths. Regardless of these differences, each dataset has the same summary statistics, such as the x and

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

y mean and variance, x and y correlation coefficient, and linear regression line.

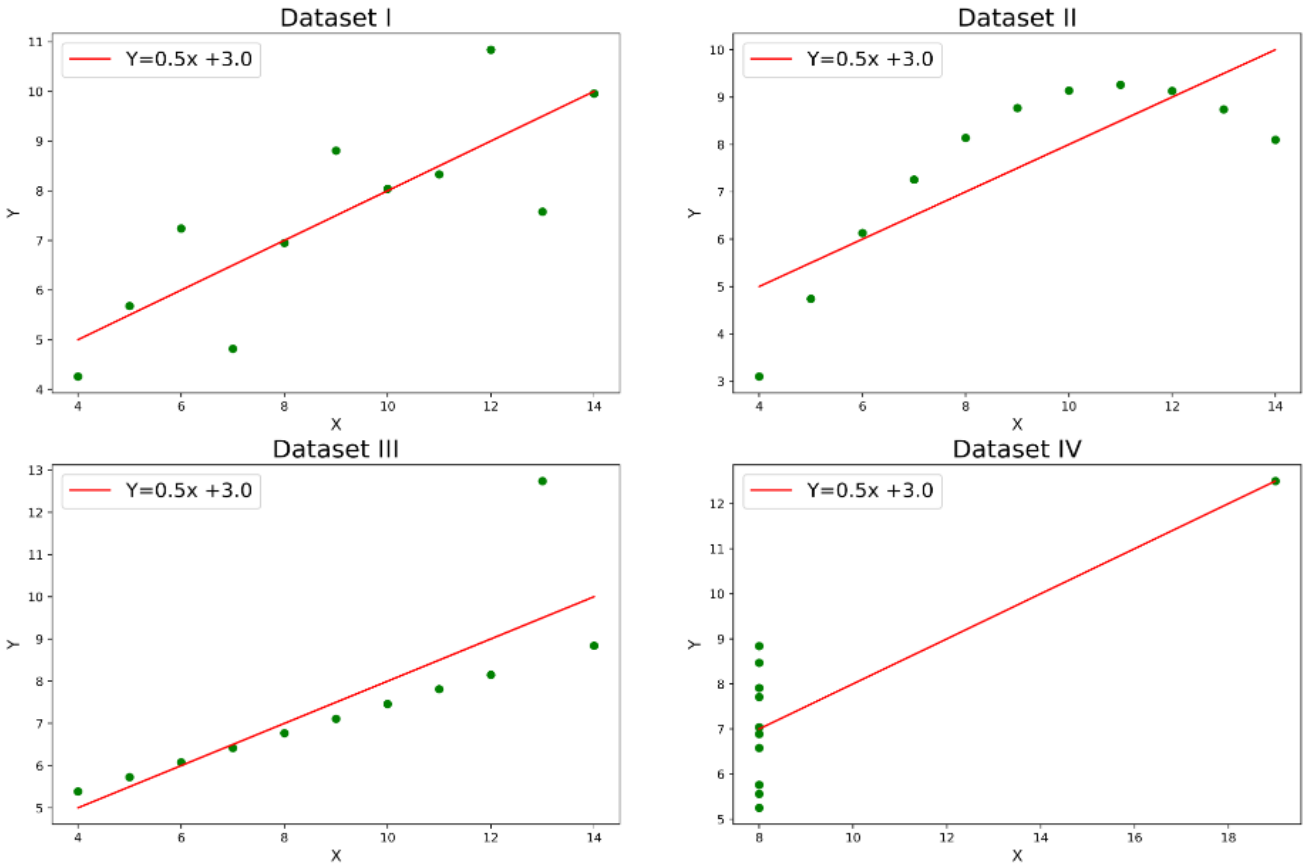
The four datasets are:

Statistics summary of all four dataset is:

	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

Clearly, we can observe the same descriptive statistics summary this uniformity in summary statistics may persuade one to conclude that the datasets are fundamentally the same.

However, the scatter plot shows the inherent difference:



- The scatter plot in the first one (top left) indicates a linear relationship between x and y.
- Looking at the second figure (top right), one may conclude that x and y have a non-linear connection.
- In the third (bottom left), one can state that there is a perfect linear relationship for all of the data points except one, which appears to be an outlier and is marked as being far away from that line.
- Finally, the fourth (bottom right) demonstrates how one high-leverage point is sufficient to achieve a high correlation coefficient.

Although Anscombe's Quartet's descriptive statistics appear homogeneous, the accompanying visualizations reveal unique patterns, demonstrating the importance of integrating statistical analysis with graphical exploration for reliable data interpretation.

3. What is Pearson’s R? (3 marks)

Pearson's R, often known as the Pearson Correlation Coefficient, is a method for measuring the linear correlation between two continuous variables.

Pearson Correlation Coefficient is a value ranging from -1 to 1 that indicates the degree and direction of the association between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Pearson's R is calculated as:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i are the individual sample points.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y.

There are some limitations of this method:

- Pearson's R measures linear relationships only. If the relationship is non-linear, R may not reflect the true strength of the relationship.
- Sensitive to outliers: One or a few extreme points can distort the value of R.

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

   Scaling or Feature scaling is one of the most important data pre-processing technique in machine learning. It transforms feature values to a similar scale,

ensuring all features contribute equally to the model. It also helps to increase the calculation speed in an algorithm.

The reason why scaling is essential is that, the majority of the time, the acquired data set includes features with wildly different magnitudes, units, and ranges. If scaling is not done, the method simply considers magnitude rather than units, resulting in erroneous modelling. To solve this problem, we need to scale all of the variables to the same magnitude.

One important thing to note here is that scaling just affect the coefficient and none of the other parameters, such as t-statistic, F-statistic, p-value and R-squared.

The two famous techniques for feature scaling are normalization and standardization.

Normalization:

It is used to transform features to be on a similar scale. The new point is calculated as:

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

This limits the range to [0, 1], or possibly [-1, 1]. Normalization is useful when there are no outliers, as it cannot deal with them. We usually utilize normalization to scale age rather than incomes because just a few people have high incomes while the age distribution is nearly uniform.

Standardization:

It is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

Outliers have no effect on standardized results because there is no fixed range of converted features.

Difference between Normalization and Standardization:

Normalization	**Standardization**
This approach scales the model by specifying minimum and maximum values.	This approach scales the model by using the mean and standard deviation.
When features are on different scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale range between [0, 1] and [-1, 1].	Values on a scale are not limited to a specific range.

Normalization is also known as scaling normalization.	Standardization is also known as Z-score normalization.
When the feature distribution is unknown, it is useful.	When the feature distribution is uniform, it is beneficial.

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

A variance inflation factor (VIF) measures the degree of multicollinearity in regression analysis. Multicollinearity occurs when there is a correlation between numerous independent variables in a multiple regression model. This can have an unfavourable effect on regression results. Thus, the variance inflation factor can estimate how much a regression coefficient's variance is inflated as a result of multicollinearity.

The VIF is given by:

$$VIF= 1/1-R^2$$

where R^2 is the coefficient of determination obtained by regressing the variable of interest on all other predictor variables.

The common heuristic we follow for VIF values is:

- >10: VIF is definitely high and variable must be eliminated.
- >5: can be okay, but it is worth inspecting.
- <5: Good VIF value, no need to eliminate this variable.

The value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between two or more predictor variables in a regression model. In other words, it occurs when one independent variable is a perfect linear combination of one or more other independent variables.

This happens because the VIF is calculated as:

$$VIF = \frac{1}{1 - R^2}$$

If R^2 is 1, it indicates perfect multicollinearity, meaning the variable can be perfectly predicted from the others. As a result, the denominator becomes zero:

$$VIF = \frac{1}{0} = \infty$$

Hence, the VIF value becomes infinite in such cases.

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 is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically a normal distribution. The plot displays the quantiles of the data against the quantiles of the reference distribution, helping to visually assess whether the data follows a specific distribution.

   For instance, if you're testing whether the distribution of exam scores in a class follows a normal distribution, a Q-Q plot would compare the quantiles of the actual exam scores against the quantiles from a normal distribution. If the exam scores are normally distributed, the points in the Q-Q plot will approximately fall along a straight 45-degree line. If they diverge significantly, it indicates a departure from normality.

   Structure of Q-Q plot:

   - The x-axis represents the quantiles of the theoretical distribution (e.g., normal distribution).
   - The y-axis represents the quantiles of the observed data.
   - If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a 45-degree line.

   Uses of Q-Q plot are as follows:

   - One of the assumptions of linear regression is that the residuals are normally distributed. A Q-Q plot helps you visually check if this assumption holds. If the residuals form a straight line in the Q-Q plot, it indicates that they are normally distributed, supporting the validity of the linear regression model.
   - If the points deviate from the 45-degree line (e.g., curve or S-shape), it suggests that the residuals are not normally distributed. This could indicate the presence of outliers, skewness, or heavy tails in the residuals, all of which violate the assumptions of linear regression and could affect the accuracy of the model.
   - In regression, it is essential to ensure that the model's assumptions are valid. The Q-Q plot is a critical diagnostic tool for detecting problems in the model's residuals, which can lead to biased or inefficient estimates if left unchecked.

   Importance of Q-Q plot in linear regression:

   - Ensuring Assumption Validity: In linear regression, checking for normally distributed residuals is vital for hypothesis testing, confidence intervals, and

prediction intervals. The Q-Q plot helps confirm whether the assumptions are met.

- Improving Model Accuracy: If the residuals are not normally distributed, it indicates that the linear regression model might be mis-specified. Addressing this issue can lead to better model performance and more reliable predictions.