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)

Ans. From analysing the categorical variables of the dataset, the following inferences can be made about their effect on the dependent variable:

1. **Seasonal Influence:** Demand for bike rentals peaks in the **Fall Season**, suggesting that seasonality plays a significant role in bike rental behaviour. Fall's favourable conditions likely contribute to this surge in demand.

2. **Year-on-Year Growth:** There is a noticeable **Year-on-Year Increase** in demand between 2018 and 2019. This indicates a growing trend in bike rental popularity, possibly due to increased awareness, improved bike infrastructure, or changes in lifestyle preferences.

3. **Monthly Demand Patterns:** Demand is highest during the **Dry months of June-October**, reflecting that weather conditions significantly impact bike rental demand. Customers tend to rent more bikes when the weather is dry and predictable.

4. **Holiday Effect:** Although demand is lower on holidays, the distribution of demand shows high variability. This suggests that while fewer people rent bikes overall on holidays, those who do rent tend to exhibit diverse rental behaviours, possibly due to varying holiday plans or activities.

5. **Weather's Role: Good Weather Situations** are associated with higher bike rental demands. Favourable weather conditions, such as clear skies and mild temperatures, encourage outdoor activities, including bike rentals.

6. **Working Days vs. Non-Working Days:** Demand is higher on working days, which also exhibit lower variability in bike rentals compared to non-working days. This suggests that people are more consistent in renting bikes during their daily commutes or routine activities on workdays.

7. **Day of the Week Effect: Wednesday, Thursday, and Saturday** show higher demand compared to the other days of the week. This could reflect mid-week commutes and weekend recreational activities, making these days particularly favourable for bike rentals.

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

Ans. When a categorical variable with `k` levels is converted into dummy variables, `k` new variables are generated. To reduce this to `(k-1)` variables, we use `drop_first=True`.

This approach addresses the issue of multicollinearity. Multicollinearity occurs when several independent variables in a model are highly correlated. It results in unreliable statistical inferences and is therefore undesirable.

For example, if a category 'Grade' has three levels (A, B, and C), and you create dummies for 'Grade', you would have three dummy variables: one for each level. However, it is redundant to keep all three dummies because the value of 'Grade' is known to be one of the three levels. If a data point does not belong to the dummy for A or B, it must belong to C. This redundancy can cause problems in regression analysis, so one of the dummies must be removed. Using `drop_first=True` is a common method to achieve this.

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

Ans. `temperature` has the highest correlation (0.63) with the dependent variable `count`.

Since `apparent_temperature` and `temperature` have a perfect correlation (0.99) with each other, `apparent_temperature` also has the highest correlation (0.63) with the dependent variable `count`. `year` has the second highest correlation (0.58) with `count`.

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

Ans. After building the Linear Regression model on the training set, the assumptions of the model were validated through a series of diagnostic checks:

1. **Residual Analysis:** The distribution of residuals was analysed by plotting a **histogram** of the residuals fitted with a **kernel density estimate (KDE)**. This allowed for a visual inspection of whether the residuals were normally distributed, a key assumption of linear regression.
2. **Linearity:** The linearity assumption was validated by plotting the **residuals against the fitted values**. If the residuals displayed a random scatter without any clear pattern, it confirmed that the relationship between the predictors and the outcome variable was linear.
3. **Homoscedasticity:** Along with linearity, the plot of **residuals versus fitted values** was used to check for homoscedasticity, ensuring that the variance of the residuals was constant across all levels of the fitted values. A random scatter of residuals with no discernible pattern would confirm this assumption.
4. **Independence of Errors:** The **Durbin-Watson test** was employed to check for autocorrelation in the residuals, confirming the independence of errors. A Durbin-Watson statistic close to 2 would indicate that the residuals are independent.
5. **Normality of Errors:** A **Q-Q plot** was generated to check the normality of the residuals. If the residuals closely followed the diagonal line in the Q-Q plot, it indicated that the errors were normally distributed.
6. **No Multicollinearity:** The **Variance Inflation Factor (VIF)** was calculated for each predictor to assess multicollinearity. A VIF value below 10 was considered acceptable, confirming that multicollinearity was not a significant issue 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? (2 marks)

Ans. Temperature and Year are the most positively significant features for explaining the demand of the shared bikes. Spring Season is the next significant feature but it has a deterring (negative) impact on the demand.

Following is the list of the top three predictors for reference:

- `temperature` (coefficient=0.4254, p-value=0.000)
- `year_2019` (coefficient=0.2496, p-value=0.000)
- `season_spring` (coefficient=-0.1551, p-value=0.000)

General Subjective Questions

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

Ans. **Linear Regression** is a type of supervised machine learning algorithm that models the linear relationship between a dependent variable (i.e., the target) and one or more independent variables. It is used to predict the value of the dependent variable for unknown data by fitting a linear equation to a related and known dataset. The general form of the linear regression model is:

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

where,

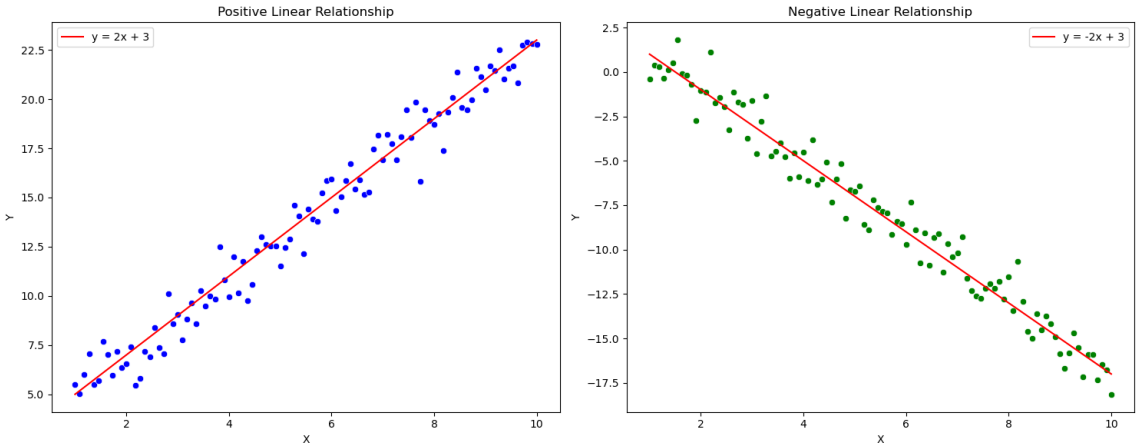
- Y is the dependent variable (**target**)
- $X_1, X_2, \ldots, X_p$ are the independent variables (**features**).
- β_0 is the **intercept** of the regression line.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the **coefficients** of the independent variables.
- p is the **number of predictors** (independent variables).
- ϵ represents the error term or **residuals**.

Types of Linear Regression:

- **Simple Linear Regression:** When there is only one independent variable.
- **Multiple Linear Regression:** When there are two or more independent variables.

Types of Linear Relationships:

- **Positive Linear Relationship:** In a positive linear relationship, as one variable increases, the other variable also increases proportionally. The slope of the line in a scatterplot of the data will be upward, indicating a direct relationship between the two variables.



- **Negative Linear Relationship:** In a negative linear relationship, as one variable increases, the other variable decreases proportionally. The slope of the line in a scatterplot of the data will be downward, indicating an inverse relationship between the two variables.

Objective: To find the **best-fit linear equation** by minimizing the **cost function**, such as the Mean Squared Error (MSE) or the Residual Sum of Squares (RSS), between the observed values and the predicted values.

Feature Engineering:

- Feature Encoding: Feature encoding is the process of converting categorical variables into a numerical format that can be provided as input to machine learning algorithms.
- Feature Scaling: Feature scaling is vital in linear regression, particularly with gradient descent, as models can be sensitive to the scale of input features. Scaling ensures that all features are treated equally, leading to more accurate and efficient training.

Estimation Methods: To find the best-fit line, the algorithm typically uses the Ordinary Least Squares (OLS) method or the Gradient Descent method.

- **Ordinary Least Squares (OLS)**: This method finds the parameters that minimize the sum of the squared residuals (the differences between observed and predicted values).

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  where, Y_i & $\hat{Y}_i$ are the actual and predicted values respectively for the observation i, and n is the number of observations.
- **Gradient Descent:** This is an iterative optimization algorithm that updates the parameters in the direction of the negative gradient of the cost function.

Interpretation of Intercept and Coefficients:

- **Intercept:** The intercept represents the expected value of the dependent variable when all predictor variables are equal to zero. It serves as the baseline level of the dependent variable.
- **Coefficients:**
  - **Positive Coefficients:** A positive coefficient indicates a positive relationship between the predictor and the dependent variable.
  - **Negative Coefficients:** A negative coefficient indicates a negative relationship between the predictor and the dependent variable.

Assumptions: If the assumptions of linear regression (linearity, independence, homoscedasticity, and normality) are violated, the model's predictions and inferences may be unreliable:

- **Linearity:** If the relationship between the variables is not linear, consider applying transformations (e.g., log, square root) to the dependent or independent variables.
- **Independence:** If residuals are not independent, time series models or autoregressive methods might be needed.
- **Homoscedasticity:** If the variance of residuals is not constant, consider weighted least squares or transforming the dependent variable.
- **Normality:** If residuals are not normally distributed, non-parametric methods or bootstrapping may be necessary.
- **No-Multicollinearity:** If there is high multicollinearity (i.e., independent variables are highly correlated with each other), it can lead to unstable estimates and inflate the standard errors of the coefficients. To address multicollinearity, we must drop one of the correlated variables.

Hypothesis Testing: This testing is done to determine whether the coefficients of the predictors are statistically significant, which is measured by the p-Value. This helps in understanding whether each predictor has a meaningful impact on the target variable.

- **Null Hypothesis (H_0):** $\beta_i = 0$ (The predictor has no effect on the target variable.)
- **Alternative Hypothesis (H_1):** $\beta_i \neq 0$ (The predictor does have an effect on the target variable.)
- **p-Value:** The p-value is obtained from the t-distribution and indicates the probability of observing the computed t-statistic under the null hypothesis. A low p-value (typically less than 0.05) suggests that the coefficient is significantly different from zero, leading to the rejection of the null hypothesis.

Multicollinearity Testing: This testing is done to identify and assess the extent of multicollinearity among predictor variables in a regression model. Multicollinearity occurs when independent variables are highly correlated, which can lead to unreliable estimates of regression coefficients and affect the stability of the model.

- **Variance Inflation Factor (VIF):** The Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A high VIF (typically greater than 5) indicates that the predictor is highly correlated with other predictors, suggesting potential multicollinearity.

Model Evaluation: To assess the performance of a linear regression model, several evaluation metrics, such as the R-squared, can be used.

- **R-squared (Coefficient of Determination):** Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. R-squared ranges from 0 to 1, with 1 indicating a perfect fit.

  $$R^2 = 1 - \frac{RSS}{TSS}$$

  where, RSS is the residual sum of squares and TSS is the Total sum of squares.

Shortcomings: Despite its wide usage, linear regression has some limitations:

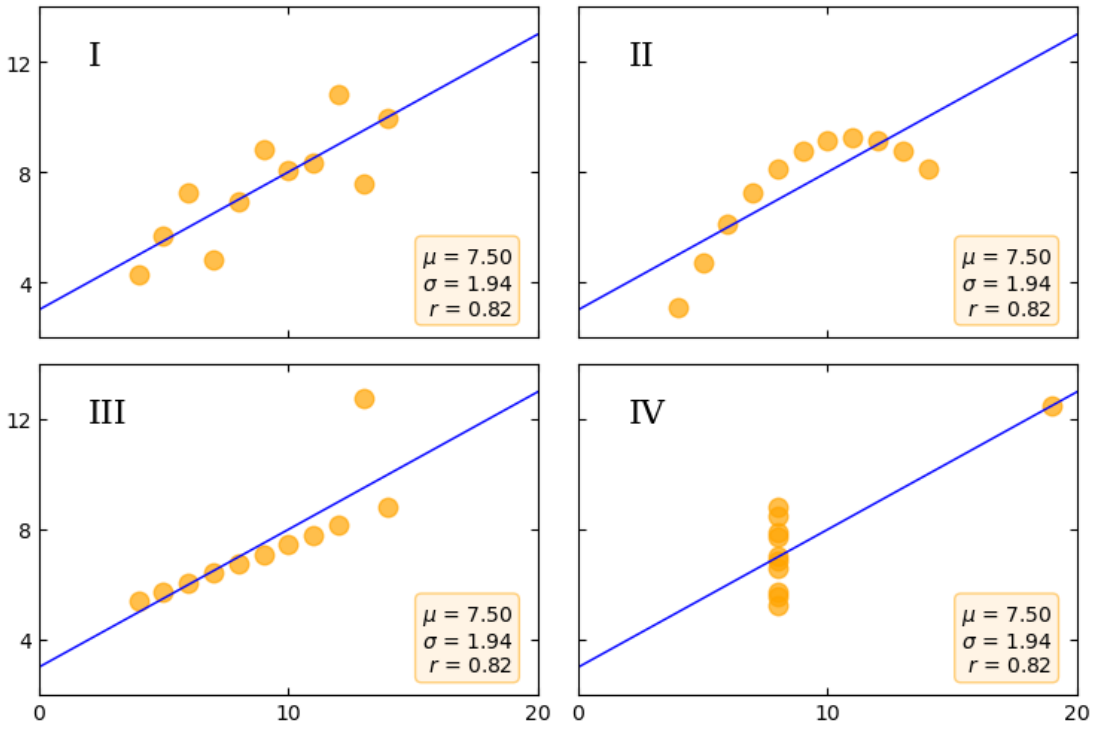
- **Non-Linearity:** If the relationship between variables is non-linear, a linear regression model will not capture this relationship accurately.
- **Outliers:** Linear regression is sensitive to outliers, which can skew the model and lead to misleading results.
- **Assumptions:** Violations of the model's assumptions (e.g., multicollinearity, homoscedasticity, independence of errors) can lead to unreliable results.
- **Overfitting:** Overfitting occurs when the model is too complex and captures noise in the data along with the underlying pattern. This results in excellent performance on the training data but poor performance on unseen data, as the model fails to generalize.
- **Underfitting:** Underfitting occurs when the model is too simple to capture the underlying pattern in the data, leading to poor performance on both the training and testing data. The model may be too rigid and unable to capture the true relationships between variables.

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

Ans. **Anscombe's quartet**, is a collection of four datasets that are nearly identical in their statistical properties (e.g., mean, variance, correlation, and linear regression line) but exhibit strikingly different distributions when graphed. Anscombe created these datasets to emphasize the importance of visualizing data before analysing it.

	I		II		III		IV	
	x1	**y1**	**x2**	**y2**	**x3**	**y3**	**x4**	**y4**
	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	82.51	99	82.51	99	82.5	99	82.51
AVERAGE	9	7.5	9	7.5	9	7.5	9	7.5
STD. DEV.	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
CORRELATION		0.82		0.82		0.82		0.82

The phenomena can be visually shown as:



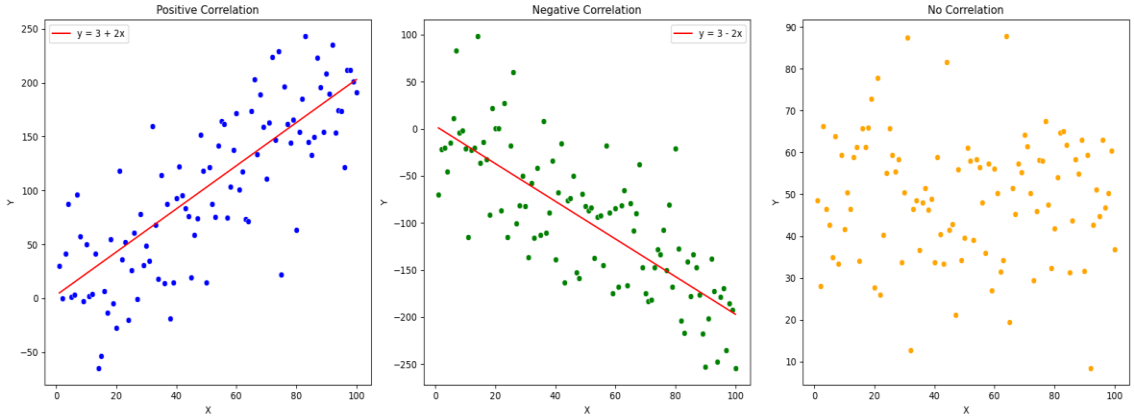
The quartet consists of four datasets, each with 11 (x, y) pairs. The key details are:

- **Dataset 1:** This is a **classic linear relationship** between x and y, with a strong positive correlation. The linear regression model fits the data well, and the residuals are evenly distributed, making this dataset ideal for linear regression analysis.
- **Dataset 2:** This dataset features a clear **non-linear relationship**. Even though the summary statistics (e.g., mean, variance) are similar to Dataset 1, a linear regression model would be inappropriate here. The relationship is curved, so using a linear model would result in poor predictions and misleading insights.
- **Dataset 3:** In this dataset, there is an **outlier** that significantly influences the regression line. While the summary statistics still resemble those of the other datasets, the presence of the outlier distorts the regression model, leading to a slope that doesn't represent the true relationship between the majority of the data points.
- **Dataset 4:** Here, most of the data points have the **same x-value**, resulting in a vertical line. The regression model fits the outlier perfectly, but this is misleading since the data does not exhibit a meaningful linear relationship. This dataset highlights the danger of blindly applying linear regression without examining the data visually.

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

Ans. **Pearson's correlation coefficient**, commonly referred to as Pearson's r, is a statistic that measures the strength and direction of the linear relationship between two continuous variables. It quantifies how strongly two variables are related and is represented by a value ranging from -1 to 1 and helps you understand the strength and direction of this linear relationship.

1. $r = 1$: Perfect positive linear relationship (as one variable increases, the other increases proportionally).
2. $r = -1$: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).
3. $r = 0$: No linear relationship between the variables.
4. $0.7 \leq r < 1$: Strong linear relationship.
5. $0.3 \leq r < 0.7$: Moderate linear relationship.
6. $0 \leq r < 0.3$: Weak linear relationship.



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

Ans. **Scaling** is the process of adjusting the range of features in a dataset so that they fall within a specific range or distribution. This is typically done to ensure that different features contribute equally to machine learning models, especially when the features have different units or magnitudes. Scaling helps algorithms to converge faster and perform better.

Scaling is performed for several reasons:

1. **Improves Model Performance:** Many machine learning algorithms are sensitive to the scale of input data. If features are not on the same scale, the model might give more importance to features with larger ranges, leading to biased results.
2. **Speeds up Convergence:** In optimization-based algorithms (like gradient descent), scaling can help the model converge faster by ensuring that all features contribute equally to the calculation of the gradients.
3. **Prevents Dominance by Large Values:** Features with large ranges might dominate the learning process, leading to poor model performance. Scaling helps prevent this by making sure no single feature dominates due to its magnitude.

Difference Between Normalized Scaling and Standardized Scaling:

	Normalized Scaling	**Standardized Scaling**
Definition	Normalization typically scales the data to a specific range, usually between 0 and 1. Each feature is scaled based on the minimum and maximum values of that feature. It is also called as the Min-Max Scaling.	Standardization scales the data so that it has a mean of 0 and a standard deviation of 1. This involves centering the data (by subtracting the mean) and then scaling it by the standard deviation.
Formula	$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$	$X_{std} = \frac{X - \mu}{\sigma}$
When to Use	Normalization is useful when you know that the data has a bounded range and you want all features to contribute equally within this bounded range.	Standardization is useful when the data follows a normal distribution or when you want to standardize features to have zero mean and unit variance.

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

(3 marks)

Ans. A **Variance Inflation Factor (VIF)** value can become infinite when perfect multicollinearity exists between two or more predictor variables in a regression model. This means that one predictor variable is an exact linear combination of one or more other predictors.

VIF for a predictor variable is calculated as:

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

where, R^2 is the coefficient of determination when regressing that predictor variable against all other predictor variables in the model.

When perfect multicollinearity is present, the determination coefficient (i.e. R) between the variables is equal to 1 (or -1), and this leads to the denominator in the VIF calculation becoming zero, which results in an infinite VIF value.

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

(3 marks)

Ans. The Q-Q plot is a graphical tool that compares the quantiles of a dataset (Y-Axis) against the quantiles of a theoretical distribution (X-Axis), such as the normal distribution. It reveals how similar the distribution of the sample data is to the theoretical distribution. Deviations from the straight line indicate departures from the expected distribution.

Uses and Importance of a Q-Q plot in linear regression:

1. **Assessing Normality:** The Q-Q plot is used to assess whether the residuals follow a normal distribution. This is crucial for validating the assumptions of linear regression, which assume that residuals are normally distributed for hypothesis tests to be valid.
    a. **Points on the Line:** If the residuals are normally distributed, the points should approximately lie on the 45-degree reference line, where theoretical quantiles equal observed quantiles.
    b. **S-shaped Curve:** If the points deviate in an S-shaped curve or the tails bend away from the line, this suggests heavy-tailed or skewed data, indicating that residuals may not be normally distributed.
    c. **Systematic Deviations:** Any systematic deviations from the line, such as curves or patterns, suggest non-normality in residuals, which could indicate issues like model inadequacy or incorrect functional form.
2. **Model Diagnostics**: Identifying deviations from normality helps diagnose potential model issues. Non-normal residuals may indicate problems such as outliers, incorrect model specification, or violations of other assumptions like homoscedasticity.
3. **Enhancing Model Accuracy:** By addressing issues highlighted by the Q-Q plot, you can refine your model. For instance, if non-normality is detected, you might transform variables, add new predictors, or use robust regression techniques.