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)**
   Answer
   a. **Season:** Summer and fall seasons tend to correlate with higher bike rental counts compared to spring and winter.
   b. **Weather Situation:** Clear weather conditions generally lead to increased bike rentals, whereas mist, light rain, and heavy rain decrease rental counts.
   c. Month and Weekday: Months from May to October and weekends show higher bike rental activity, reflecting seasonal and weekly patterns in demand.
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer
   Using **drop_first=True** during dummy variable creation is important because it helps to avoid multicollinearity issues in regression models.
   Example:
   Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can lead to unreliable estimates of coefficients and decreased model interpretability

   Without drop_first=True:

Spring	Summer	Fall	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

   The presence of all four dummy variables would create a perfect multicollinearity because the sum of all four columns would always be equal to 1. This perfect multicollinearity can lead to issues such as unstable coefficients in regression models

   With drop_first=True:

Summer	Fall	Winter
0	0	0
1	0	0
0	1	0
0	0	1

   The interpretation of coefficients straightforward. Each coefficient represents the effect of being in that category relative to the baseline (dropped category). This improves the interpretability of the regression model.

   Using drop_first=True during dummy variable creation is crucial to avoid multicollinearity and enhance the interpretability of regression models. It ensures that the model remains stable and the coefficients are accurately interpreted in relation to the omitted category, thereby improving the overall reliability and performance of the regression analysis

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Answer
   Looking at the pair-plot and the correlation matrix among the numerical variables (cnt, temp, hum, windspeed), the variable temp (temperature) has the highest correlation with the target variable cnt (bike rental count).

	cnt	temp	hum	windspeed
cnt	1.000000	0.627044	-0.098543	-0.235132
temp	0.627044	1.000000	0.128565	-0.158186
hum	-0.098543	0.128565	1.000000	-0.248506
windspeed	-0.235132	-0.158186	-0.248506	1.000000

   From the correlation matrix, the highest correlation is between cnt and temp with a correlation coefficient of approximately 0.63. This indicates that temperature is the numerical variable most strongly correlated with the number of bike rentals
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Answer
   To validate the assumptions of Linear Regression after building the model on the training set, I employed several methods:
   Linearity: Checked scatter plots of residuals versus predicted values and reviewed pair plots and correlation matrices to confirm a linear relationship between predictors and the target variable.
   Homoscedasticity: Examined residual plots to ensure constant variance across the predicted values, indicating homoscedasticity
   No Multicollinearity: Calculated Variance Inflation Factors (VIF) for predictors to ensure no severe multicollinearity issues.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Answer
   model.coef_ will provide the coefficients for features.
   From the coefficients, you can identify the top 3 features with the largest coefficients (in absolute value), indicating their significant contribution to the demand for shared bikes.
   - Temperature (temp) - with a coefficient of 1123.62, indicating a strong positive impact on bike demand.
   - Year (yr_2019) - with a coefficient of 993.43, showing higher demand in 2019 compared to 2018
   - Season (season_winter) - with a coefficient of 527.01, suggesting higher bike rentals during the winter season.

```
# Print coefficients and intercept
print("Intercept:", model.intercept_)
print("Coefficients:", dict(zip(selected_features, model.coef_)))
```

```
Intercept: 4486.382352941177
Coefficients: {'workingday': 94.27522050927764, 'temp': 1123.6247369327864, 'hum': -220.455160551037
47, 'windspeed': -274.13910066566797, 'season_summer': 337.0690807948046, 'season_winter': 527.01057
77545944, 'weathersit_light_rain': -345.1087458916278, 'weathersit_mist': -211.5417114961825, 'yr_20
19': 993.4280859985616, 'mnth_Sep': 243.9489483592738}
```

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer
   Linear regression is also a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

   Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear equation that can predict the dependent variable based on the independent variables. Types of Linear Regression
   Simple Linear Regression - It involves only one independent variable and one dependent variable. The equation for single linear regression $y=\beta_0+\beta_1 X$
   Multiple Linear Regression - It involves more than one independent variable and one dependent variable. The equation for multiple linear regression is $y=\beta_0+\beta_1 X+\beta_2 X+\ldots\ldots\ldots\beta_n X$
   The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Answer
   Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph
   The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading

   The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

   Purpose of Anscombe's Quartet
   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's Quartet Dataset
The four datasets of Anscombe's quartet.

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

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

   Pearson Correlation Coefficient: Correlation coefficients are used to measure how strong a relationship is between two variables. There are different types of formulas to get a correlation coefficient, one of the most popular is Pearson's correlation (also known as Pearson's r) which is commonly used for linear regression. The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses

   Below is a formula for calculating the Pearson correlation coefficient (r)

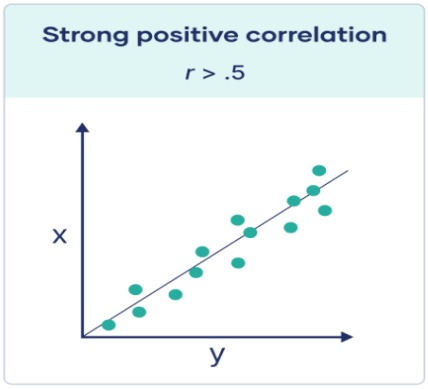
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

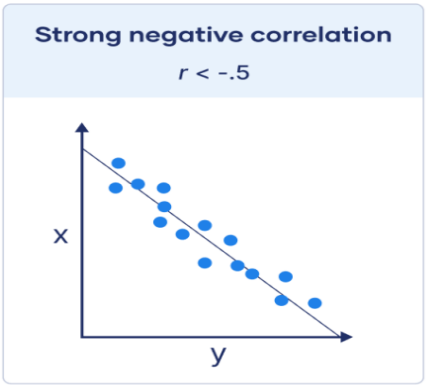
   The Pearson Correlation Coefficient, denoted as r, is a statistical measure that calculates the strength and direction of the linear relationship between two variables on a scatterplot. The value of r ranges between -1 and 1, where:

   1 indicates a perfect positive linear relationship,
   -1 indicates a perfect negative linear relationship, and
   0 indicates no linear relationship between the variables.

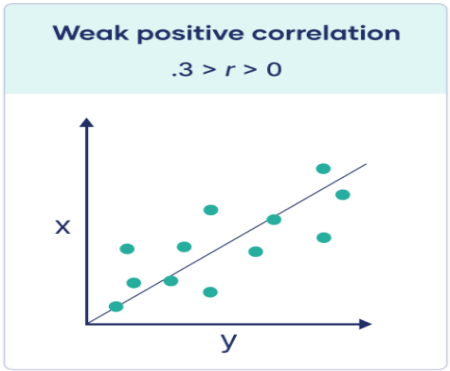
Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

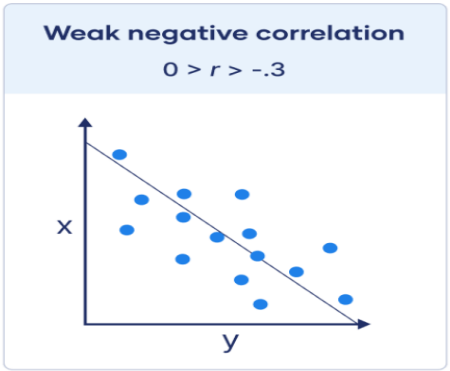
When r is 1 or –1, all the points fall exactly on the line of best fit:



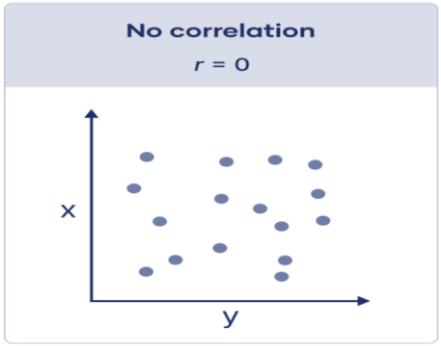


When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit





When r is 0, a line of best fit is not helpful in describing the relationship between the variables:



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Answer
   Scaling is the process of adjusting the range of features in your data so that they fit within a specific scale. This is especially important in machine learning because many algorithms are sensitive to the scale of input data, which can impact model performance and convergence speed.
   Scaling is performed to improve model performance and to optimize gradient descent to achieve faster and more stable convergence
   Type of scaling – Normalized Scaling (Min-Max Scaling), Standardized scaling (Z-Score Normalization)
   Normalized Scaling: Rescales the feature to a fixed range, usually 0 to 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   - **Use Case**: Useful when the distribution of data is not Gaussian (not normally distributed) or when you want features to have the same scale but different variances.
   - **Example**: If a feature ranges from 10 to 200, after normalization, it will range from 0 to 1.

   Standardized Scaling: Transforms the data to have a mean of 0 and a standard deviation of 1. where μ is the mean of the feature and σ is the standard deviation.

$$x' = \frac{x - \mu}{\sigma}$$

   - **Use Case**: Useful when the data follows a Gaussian distribution or when algorithms assume that data is centred around zero.
   - **Example**: If a feature has a mean of 50 and a standard deviation of 10, a value of 60 will be transformed to 1 after standardization.

Difference: **Normalization** scales data to a range of [0, 1] or [-1, 1], preserving the distribution shape. **Standardization** transforms data to have a mean of 0 and a standard deviation of 1, centring and scaling the distribution.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   Answer
   The value of the Variance Inflation Factor (VIF) can be infinite when there is perfect multicollinearity in the data, meaning one predictor variable is a perfect linear combination of one or more other predictor variables. This leads to division by zero in the VIF calculation
   Understanding VIF and Infinite Values

$$\text{VIF}_i = \frac{1}{1-R_i^2}$$

   An infinite VIF value indicates perfect multicollinearity, which occurs when $R_i^2 = 1$. This situation implies that one predictor variable is a perfect linear combination of one or more of the other predictor variables.

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

   A Q-Q (quantile-quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution. It plots the quantiles of the sample data against the quantiles of the specified theoretical distribution. If the sample data come from the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line.

   Key Features of a Q-Q Plot

   - **X-axis**: Theoretical quantiles from the specified distribution.
   - **Y-axis**: Sample quantiles from the data.
   - **Line of Equality**: A 45-degree reference line where points would lie if the sample data followed the theoretical distribution perfectly.

   Use and importance of Q-Q plot in Linear Regression

   - To validate residuals (errors) are normally distributed
   - To identify the residuals are skewed and if they have heavy/ light tails
   - The reference line indicates the outliers in the residuals

.