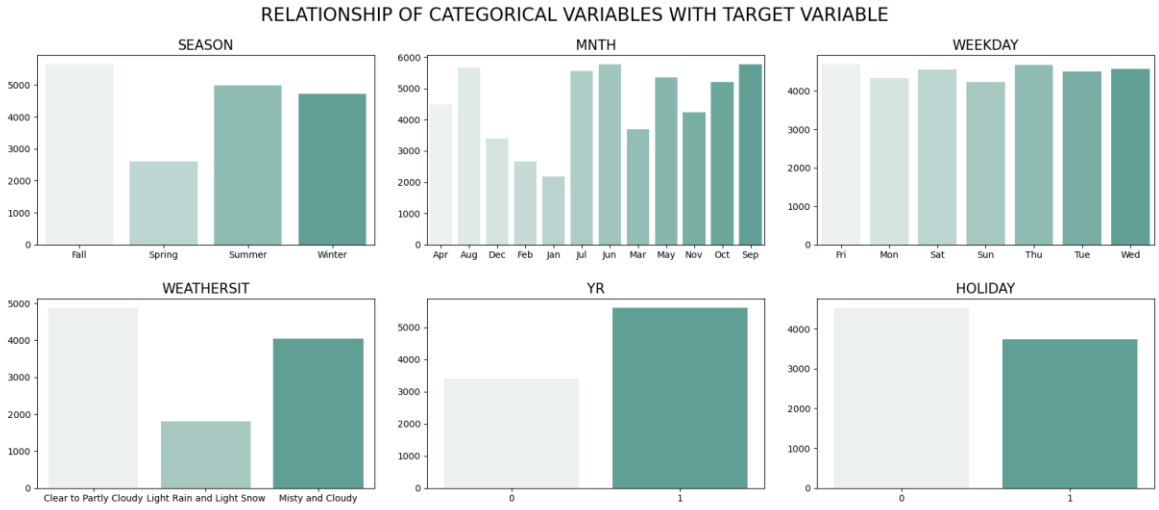
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?** (3 marks)



a. **Day of the Week and Bike Demand:**
   - The highest bike demand is observed on Friday, Saturday, Sunday, and Thursday.
   - This pattern suggests that bikes are used for a mix of purposes—both office commutes and leisure travel.

b. **Weather and Demand:**
   - Bike demand is exceptionally high on clear days.
   - Weather conditions play a significant role in influencing bike usage.

c. **Business Growth:**
   - The company has experienced significant business growth from 2018 to 2019.
   - This positive trend indicates the company's success.

d. **Holiday vs. Weekdays:**
   - Bike usage on holidays is slightly lower than on weekdays.
   - People seem to prefer bikes for daily commuting during weekdays.

e. **Working Days vs. Non-Working Days:**
   - Bike usage is similar regardless of whether it's a working day or not.
   - This suggests consistent demand throughout the week.

f. **Month:**
   - Bike usage is significantly less during Dec, Jan, Feb. And demand is more or less similar throughout other part of the year even though highest demand is during Jun, Jul, Aug.

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

`drop_first=True` is important to use, as it helps in reducing the redundant columns created during dummy variable creation.

Explanation with examples:

When you have a categorical variable with k categories, creating one dummy variable for each category leads to k new columns. However, the information they encode isn't fully independent. If all other dummy variables for a row are 0, then the remaining category must be the one associated with the missing dummy variable. This inherent redundancy creates multicollinearity, where variables are highly correlated.

By setting `drop_first=True`, you only create k-1 dummy variables, eliminating the redundant one and avoiding multicollinearity issues.

Example:

Consider a categorical variable representing fruit types: ["Apple", "Banana", "Orange"]. Creating dummies without `drop_first` yields:

Sample	Apple	Banana	Orange
Apple	1	0	0
Banana	0	1	0
Orange	0	0	1

But, if any row has Apple and Banana as 0, we already know it must be Orange (1). So, using `drop_first=True` removes the Apple column, keeping only Banana and Orange:

Sample	Banana	Orange
Apple	0	0
Banana	1	0
Orange	0	1

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

The numerical variable ‘registered’ showed the highest correlation (0.95) with the target variable ‘cnt’ when considering all features. However, after data preparation and dropping ‘registered’ and ‘casual’ due to multicollinearity, the numerical variable ‘atemp’ or ‘temp’ now exhibits the highest correlation (0.63) with ‘cnt’.

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

After training the model, I conducted the following analysis based on Linear Regression assumptions:

a. Ensuring a Linear Relationship between X and Y:
   - ✦ We verify that there exists a linear relationship between the independent variable (X) and the dependent variable (Y).

b. We verify that there exists a linear relationship between the independent variable (X) and the dependent variable (Y).

- ✦ We verify that there exists a linear relationship between the independent variable (X) and the dependent variable (Y).
- ✦ If the scatter plot shows consistent variance (no funnel shape), the assumption of homoscedasticity is met.

c. Managing Multicollinearity:
- ✦ We select or exclude independent variables based on two criteria: VIF and p-values

d. Verifying Normality of Error Terms:
- ✦ We check whether the error terms (residuals) follow a normal distribution with a mean of zero.
- ✦ Normality is crucial for valid statistical inference.

e. Assessing Independence:
- ✦ By plotting residuals from the y-train dataset, we examine whether there is any pattern or correlation.
- ✦ Independence assumption holds if residuals show no discernible structure.

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

a. Temperature Impact:
- ✦ The most critical factor influencing demand is temperature. With a coefficient of 0.600201, for every 1-degree change in temperature, demand increases by a factor of 0.600201 (temperature × 0.600201). Impact is mostly during hotter months (especially in June, July, and August)

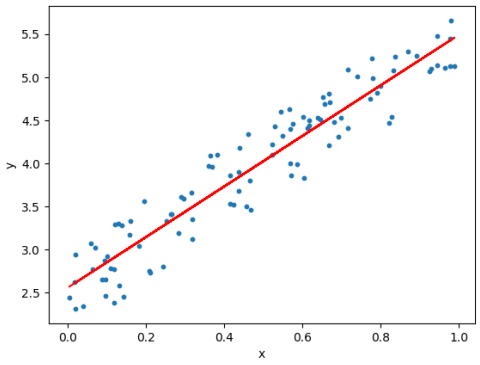
b. Weather Conditions:
- ✦ The second most important factor is Light Rain or Light Snow, with a coefficient of 0.289400. On days with light rain, demand is expected to decrease by 28.9%.

c. Annual Growth:
- ✦ The third significant factor is the year, with a coefficient value of 0.238668.

General Subjective Questions

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

Linear regression is a fundamental algorithm in machine learning and statistics. It aims to <u>discover the linear relationship</u> between a <u>dependent variable</u> (predicted variable) and one or more <u>independent variables</u> (the features influencing the prediction). By finding the line that best fits this relationship, we can make predictions about the dependent variable based on the values of the independent variables.



Key Concepts:

1. **Data Representation:**
Imagine you have data points scattered on a graph, with each point representing a pair of values: one for the independent variable (x) and the other for the dependent variable (y). Your goal is to find a straight line that minimizes the distance between itself and as many data points as possible.

2. **The Equation:**
Linear regression works by fitting a linear equation of the form:

```
y = mx + c
```

where:
y: the dependent variable (predicted variable)
x: the independent variable (the feature influencing the prediction)
m: the slope of the line, representing the change in y for every unit change in x
c: the y-intercept, representing the value of y when x is 0

3. **Types of Linear Regression:**
- Simple Linear Regression: Involves a single independent variable
- Multiple Linear Regression: Includes multiple independent variables

4. **The Algorithm:**
There are different algorithms to find the "best-fit" line, but the most common one is least squares regression. This method minimizes the sum of squared residuals, which are the vertical distances between each data point and the line.

5. **Visualizing the Process:**
Imagine you throw darts at a board, aiming for a bullseye. Each dart represents a data point, and the bullseye represents the ideal "best-fit" line. Least squares regression is like adjusting the position of the bullseye to minimize the total squared distance between each dart and its closest point on the board.

6. **Evaluating the Model:**
Once you have a line, how good is it? We use metrics like:

- **R-squared:** tells you how much of the variance in y is explained by the model (0-1, higher is better)
- **Mean squared error:** measures the average squared distance between data points and the line
- **P-value:** tests if the relationship between x and y is statistically significant

7. **Assumptions and Limitations:**
Linear regression assumes a linear relationship between x and y. It also assumes normally distributed errors and homoscedasticity (constant variance of errors). Violating these assumptions can affect the model's reliability. Below is the list of assumptions:
- Linear Relationship
- Independence of Residuals.
- Homoscedasticity
- Normality of Residuals

8. **Applications:**

Linear regression finds use in various fields like:

Finance: predicting stock prices

Marketing: understanding customer behavior

Healthcare: analyzing medical data

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

It is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset contains of eleven (x, y) pairs as follows: -

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

Properties of Anscombe's Quartet:

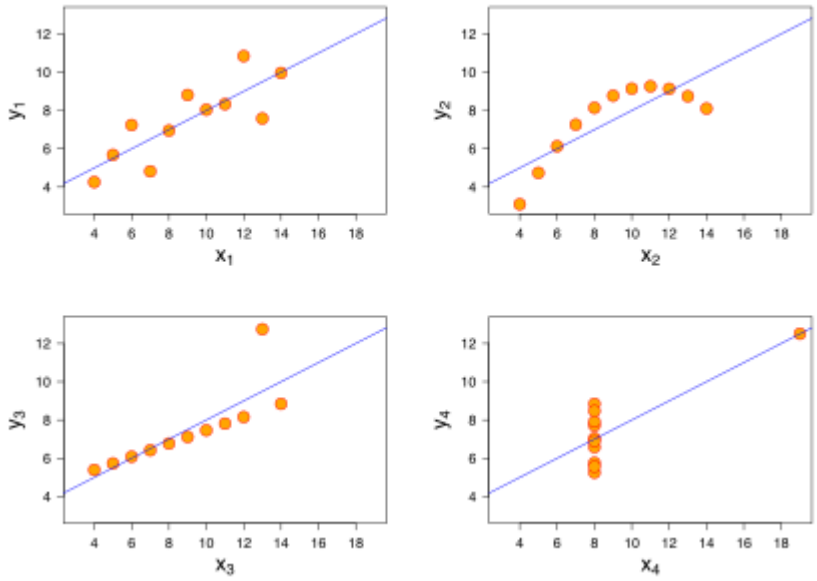
All the summary statistics for each dataset are identical

1. The average value of x is 9.
2. The average value of y is 7.5.
3. The variance for x is 11 and y is 4.12
4. The correlation between x and y is 0.816
5. The line of best for is $y = 0.5x + 3$.

Graphical Exploration:

But the plots tell a different and unique story for each dataset.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.



- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Pearson's R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}},$$

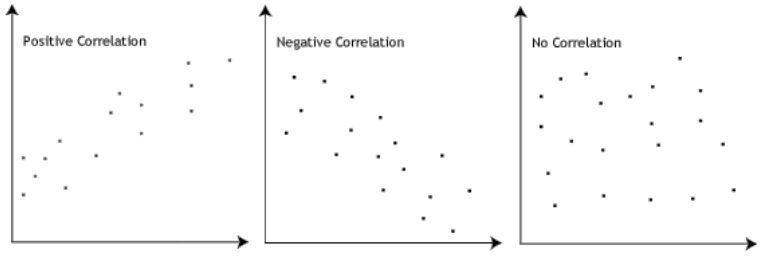
Where

- n is sample size
- x_i, y are the individual sample points indexed with i

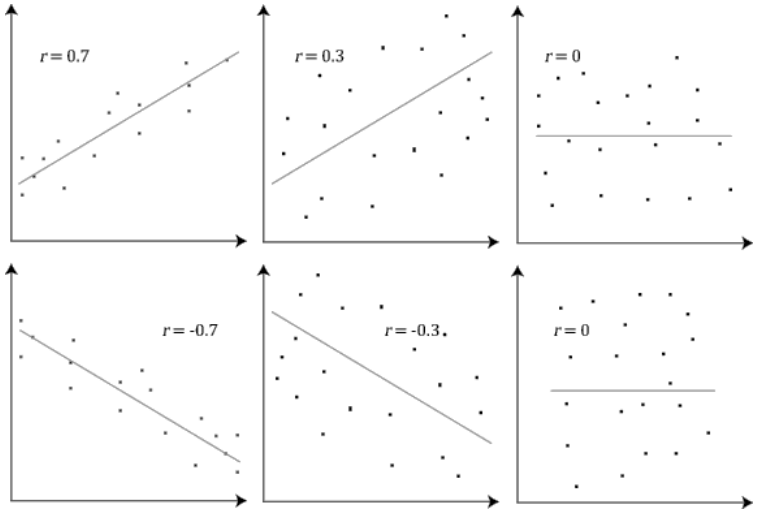
The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) It measures how closely two variables are related in a linear way. Think of it as drawing a straight line through data points. The value of (r) can range from -1 to +1.

- (r = 0): No association between the variables.
- (r > 0): Positive association (both variables increase together).
- (r < 0): Negative association (one variable increases, the other decreases).

This is shown in the diagram below:



When two variables are strongly associated, the value of (r) tends to be closer to either +1 or -1. A value of +1 or -1 means that all data points lie exactly on the best-fit line. Values of (r) between +1 and -1 (e.g., 0.8 or -0.4) indicate some variation around the best-fit line. The closer (r) is to 0, the greater the variation around the best-fit line. In summary, (r) quantifies the strength of association, and extreme values indicate a tight linear relationship, while intermediate values allow for some variability. Their correlation coefficients are shown in the diagram below:



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

1. What is Feature Scaling?
   - ✦ Feature scaling standardizes the independent features in a dataset to a consistent range.
   - ✦ It's crucial during data preprocessing to handle varying magnitudes or units.
2. Why Use Feature Scaling?
   - ✦ Ensures all features are on a comparable scale.
   - ✦ Prevents large-scale features from dominating the learning process.
   - ✦ Improves algorithm performance and stability.
3. Types of Scaling:
   - ✦ Normalized Scaling:
     - o Scales a variable to have values between 0 and 1.
     - o sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

     $$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$
   - ✦ Standardized Scaling:
     - o Transforms data to have a mean of zero and a standard deviation of 1.

       $$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$
     - o sklearn.preprocessing.scale helps to implement standardization in python.
     - o One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

4. Benefits of Scaling:
   - ✦ Uniform Treatment: All features contribute equally to the model.
   - ✦ Numerical Stability: Avoids issues like overflow or underflow.
   - ✦ Correct Modeling: Ensures both magnitude and units are considered.
5. Note:
   - ✦ Scaling affects coefficients but not other parameters (t-statistic, F-statistic, p-values, R-squared).

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

The Variance Inflation Factor (VIF) assesses the collinearity among predictor variables in a multiple regression model. It quantifies how much the variance of a given model's coefficients (betas) increases due to correlations between the variables. Specifically:

- ✦ VIF is calculated by dividing the variance of all the model's betas by the variance of a single beta if it were fitted independently.
- ✦ When there is **perfect correlation**, the VIF becomes **infinity**.
- ✦ A **high VIF** value indicates significant correlation between variables.
- ✦ An **infinite VIF** suggests that the corresponding variable can be expressed exactly as a linear combination of other variables (which also exhibit infinite VIF values).

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 assess whether a dataset plausibly follows a theoretical distribution (such as Normal, exponential, or Uniform). Additionally, it helps determine if two datasets originate from populations with a common distribution.

Here are the key points:

a. Purpose: In linear regression scenarios, where we have separate training and test datasets, the Q-Q plot confirms whether both datasets share the same underlying distribution.

b. Advantages:
- ✦ It works effectively with sample sizes.
- ✦ Detects various distributional aspects, including shifts in location, scale, symmetry changes, and the presence of outliers.

c. Scenarios Checked:
- ✦ If two datasets:
  - o Come from populations with a common distribution.
  - o Share a common location and scale.
  - o Exhibit similar distributional shapes.
  - o Display similar tail behaviour.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight i.e.