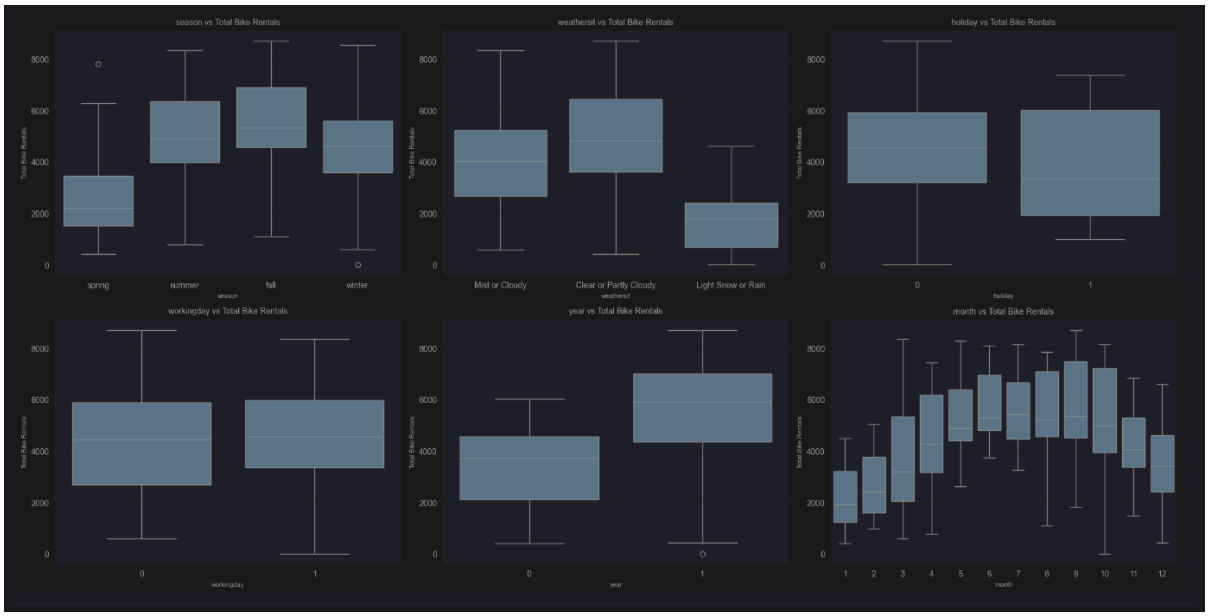
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)



Insight from categorial plot variable on dependent variable(total_count)
weathersit: During the weather of mist or clody and clear or partly cloudy, rent bike count is higher
holiday: People are renting bike when its holiday to enjoy family time and travel
year: People rented more bike in 2019 compare to 2018, initial year for new company, the growth increase, so count in 2019 is more compare to 2018
month: People rented more bike during the month from April to November
workingday: On non-working day, count of rent bike is more compare to working day
season: Majority of user prefer to rent bike during summer, fall and winter season and less during summer, it could be because of moderate temperature

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True ensures that we avoid the **dummy variable trap**, which leads to multicollinearity. It drops one category from each categorical variable, preventing redundancy and ensuring that the independent variables remain linearly independent keeping only n-1 dummy columns.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

From the pair plot and correlation heatmap, **temperature (temp)** has the highest correlation with total bike rentals (total_count). This makes sense as warmer temperatures generally lead to higher demand for bike rentals.

We also observed that 'temp' and 'atemp' are nearly identical

---

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of Linear Regression checks include:

- **Linearity**: Verified using scatter plots of independent variables against the dependent variable.
- **Homoscedasticity**: Checked using residual plots to ensure variance is constant.
- **Normality of residuals**: Verified using a **Q-Q plot** and a histogram of residuals.
- **Normality of Error terms:** Error term are normally distributed in graph plot
- **Multicollinearity**: Correlation between independent feature is very small or there is no multicollinearity between them. VIF is below 5, considered as good VIF

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The most significant features in the final regression model were:

- **Temperature (temp)** - positively correlated with demand
- **Year (year)** - shows the increasing trend in rentals over time
- **Humidity (humidity)** - negatively correlated with demand

---

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

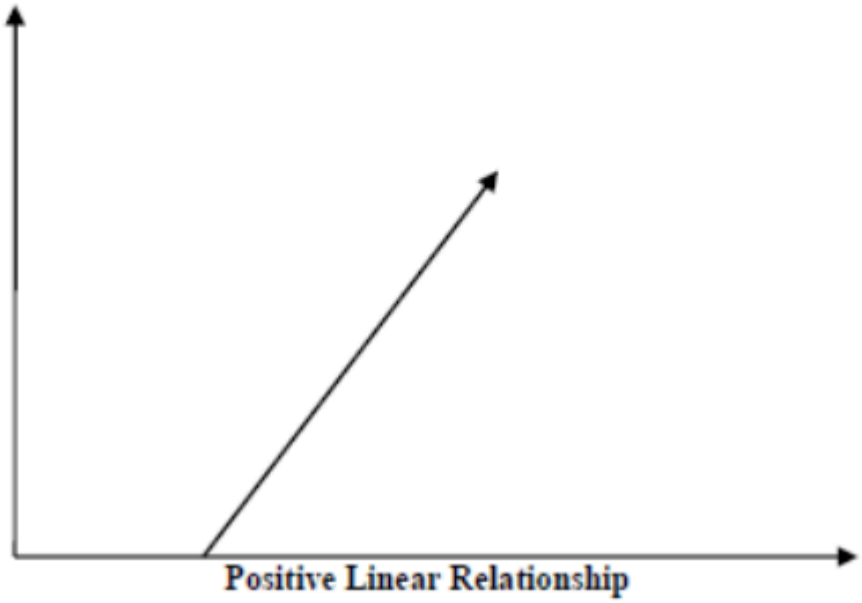
Linear regression may be defined as the **statistical** model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
It is ML algorithm used forSupervised Learning

Mathematically the relationship can be represented with the help of following equation –
Y = mX + c

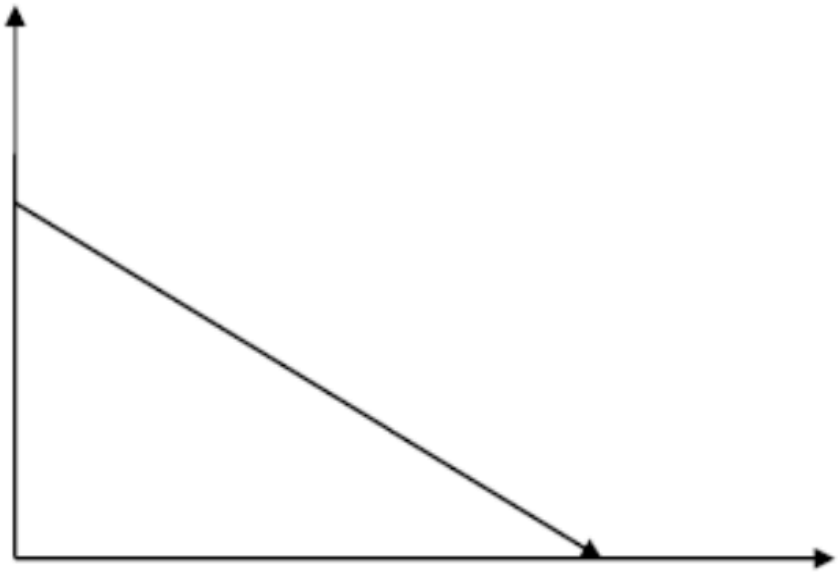
Here,
Y - is the dependent variable we are trying to predict.
X - is the independent variable we are using to make predictions.
m - is the slope of the regression line which represents the effect X has on Y
c - is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below–

1. Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.
   It can be understood with the help of following graph:



2. Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Linear regression is of the following two types **–**

- Simple Linear Regression
- Multiple Linear Regression

Assumptions in Linear Regression-

The following are some assumptions about dataset that is made by Linear Regression model:

Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms – Error terms should be normally distributed

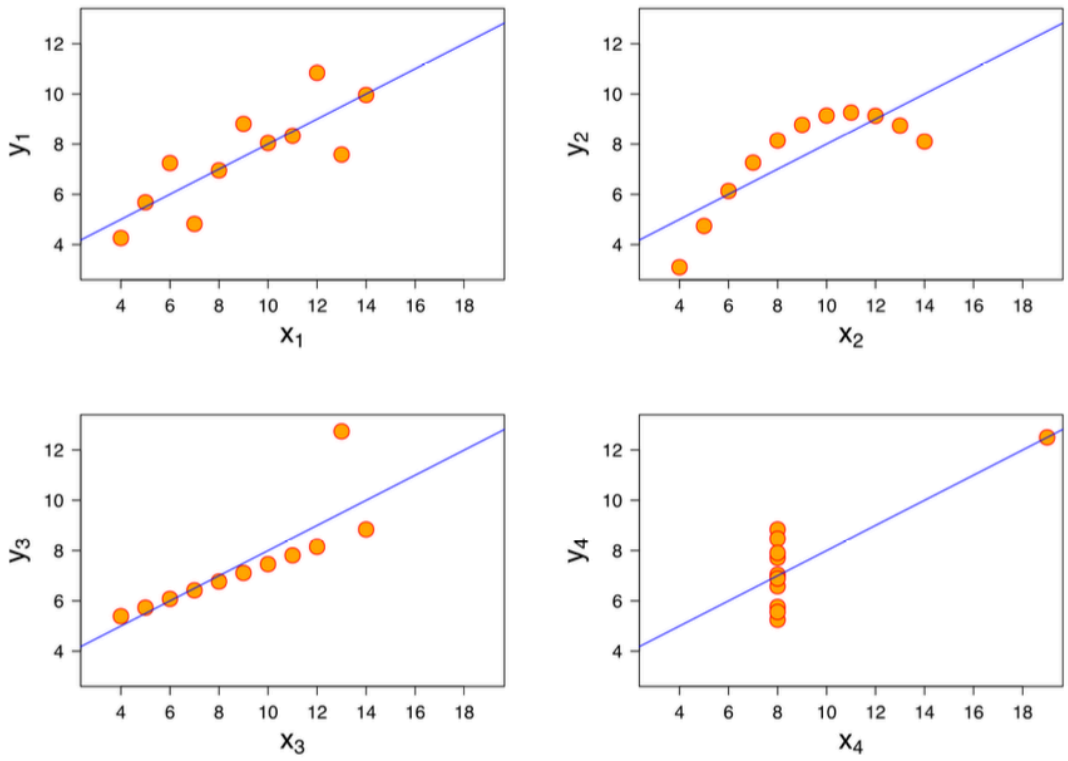
Homoscedasticity – There should be no visible pattern in residual values.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

Anscombe's quartet is mainly a group of four data sets which are nearly identical in simple descriptive statistics, but there are some strange behaviours in the dataset. They have very different distributions and appear differently when plotted on scatter plots.

Each dataset consists of eleven (x,y) points.



Observations from above data sets:
Data Set 1: fits the linear regression model well
Data Set 2: cannot fit the linear regression model because the data is non-linear
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model
We observed that Anscombe's quartet helps us to understand the importance of data visualizations to build a well-fit model.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Pearson's Correlation Coefficient (r) is a statistical measure of the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between

-1.0 and +1.0. Pearson's R cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
r = 1 means strong positive relationship. If one increases, the other also increases
r = -1 means strong negative relationship. If one increases, the other decreases
r = 0 means there is no linear relationship

Formula for Pearson r is:

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

Explanation of Terms:

- r = Pearson's correlation coefficient
- X_i = Individual values of the first variable (X)
- Y_i = Individual values of the second variable (Y)
- $\bar{X}$ = Mean of X
- $\bar{Y}$ = Mean of Y
- $\sum$ = Summation notation

In this project, **temperature and total_count** have a strong positive correlation.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is nothing but putting the numerical feature values in the same range. Scaling is very important because few variables may have values on a different scale (smaller/higher) compared to other variables. Scaling also helps in speeding up the calculation in an algorithm.

Scaling is performed to bring all the numerical features in the same range. If some feature having value in range of thousands or lakhs while other in the range on tens or hundreds, for example, if salary column has values ranging from 25k to 10L and no of years experienced is in the range 1-25, so the model will take magnitude in account and not the units resulting in wrong or incorrect modelling. Hence, it is very important to scale all numerical feature in the same range before building the model.

There are many types of scaling, but normalized and standardized scaling are popular and widely used.

Normalized scaling: Scaling which makes all numerical feature lie in the range of 0 and 1. One disadvantage of normalization is that it losses some information in the date such as outliers.

- **Min-Max Scaling**: Scales values between 0 and 1, maintaining distribution shape.

Standardized scaling (Z-score Scaling): Standardized scaling replaces the values by their Z scores. It brings all
the data into a standard normal distribution which has mean (μ) 0 and standard deviation (σ) 1.

Min-Max Scaling was used in this project for normalizing temperature, humidity, and windspeed.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Variance Inflation Factor (VIF) measures multicollinearity. An infinite VIF suggests **perfect collinearity**, meaning one independent variable is a perfect linear combination of others. In such cases, that variable should be removed.

VIF tells how much an independent variable is correlated with other independent variables. It detects the multicollinearity in the OLS regression analysis.
VIF below 5 is a good VIF. And VIF above 10 shows high correlation and should be removed.
If there is a perfect correlation between two independent features, then the VIF is infinite.
VIF is defined as:VIF = 1 / (1-R^2)
So, if r2 is 1 then the denominator becomes 0 and hence the VIF i.e., 1/0 is infinite. This means that variable is fully explained by some other variable in the model and hence does not make any sense to keep this feature in the model.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Quantile-Quantile or Q-Q plot is a plot of the quantiles of the theoretical set against the quantiles of the sample set. It helps us understand if a sample comes from a known distribution such as normal distribution. In Regression, we use Q-Q plot to check if the data in the sample is normally distributed. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed. If two distributions being compared are similar, the points on Q-Q plot will lie approx. on the line y=x

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions

Q-Q plot helps us to determine if two population are of the same distribution, if residuals follow normal distribution and if there is any skewness in the distribution.
If date sets, we are comparing are of the same type of distribution type, then plot would be a straight line.
