Assignment-based Subjective Questions

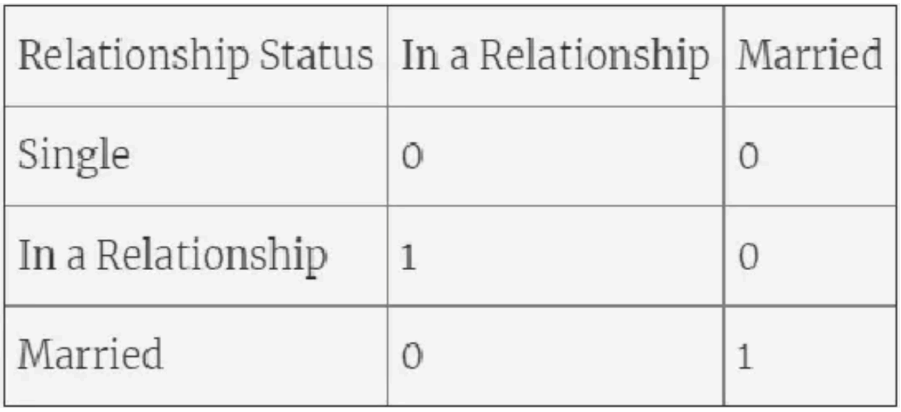
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

I used box plot to study the categorical variables impact on dependent variables and these are the observation.

- The fall season had highest booking with a median of near about 5000 and spring season had lowest booking. The summer and winter had count value in middle . This infer fall is the optimal weather condition for bike riding.
- Bike rental getting popular and increasing year on year and in 2019 median is much higher as compare to 2018.
- September had maximum rental and this co-relate relation with fall season as well.
- More rental on non-holidays day infer people spent more time with family on holidays.
- Median is near about same for all weekday but less booking on Sunday.
- Median is near about same for working or non-working day and had little impact on the higher booking side.
- The highest count is observed on clear weather and snow indicating unfavourable weather with less booking count.

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

A variable with n levels can be represented by n-1 dummy variables. If we can remove the first column then also we can represent the data . In case of Relationship status example (Single, In A relationship, Married) can be represented with In a relation ship and Mrraied field as shown below.



Relationship Status	In a Relationship	Married
Single	0	0
In a Relationship	1	0
Married	0	1

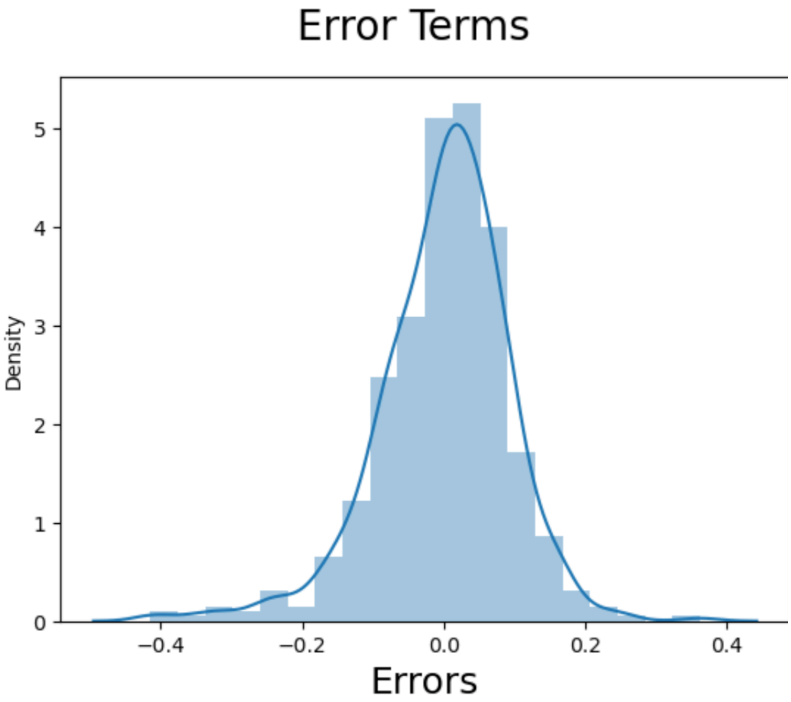
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

The temp and atemp field shows highest co-relation with a value of .63 with target variable.

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

We validate the assumption by plotting a distribution plot of the residuals and to check its normally distributed or not and mean is zero or not .

I received below distribution.



rrors are normally distributed here with mean 0. So everything seems to be fine

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

The 3 contributing features are .

1. Temp- A co-efficient of .4794 indicates a positive increase in demand as temperature increases.
2. Yr (Year): A co-efficient of .2344 indicates a positive increase in demand as year increases.
3. Weather Situation (Light_snowrain) : A negative co-efficient of .2818 indicates a bike demand decrease by a factor of .2818.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a statistical method that is used to predict the value of a dependent variable (y) based on the value of one or more independent variables (x). It is a widely used statistical method that is applicable to a wide variety of problems. The linear regression algorithm finds the best-fitting line that minimizes the sum of the squared residuals between the actual values of y and the predicted values of y.

The equation for a linear regression line is:

$$y = mx + b$$

where:

- y is the dependent variable
- x is the independent variable
- m is the slope of the line
- b is the y-intercept

The slope of the line (m) indicates the direction and the strength of the relationship between x and y. A positive slope indicates that y increases as x increases, while a negative slope indicates that y decreases as x increases. The y-intercept (b) indicates the value of y when x is zero.

There two type of linear regression algorithm

- Simple linear regression: Only one independent variable.
- Multiple linear regression: Multiple independent variables. In this case equation changed to

  $$y = b0 + b1x1 + b2x2 + ... + bnxn$$

  where:

  y is the dependent variable

  x1, x2, ..., xn are the independent variables

  b0 is the y-intercept

  b1, b2, ..., bn are the regression coefficients

Steps for linear regression algorithm:

1. Gather the data: The first step is to gather the data that you need to train the model. This data should include the dependent variable and the independent variables.
2. Split the data into training and test sets: The data should be split into two sets: a training set and a test set. The training set will be used to train the model, and the test set will be used to evaluate the model's performance.

3. Fit the model: The next step is to fit the model to the training set. This is done by finding the values of the slope (m) and the y-intercept (b) that minimize the sum of the squared residuals.
4. Evaluate the model: The model's performance should be evaluated using the test set. This can be done by calculating the R-squared value or the adjusted R-squared value.
5. Use the model to make predictions: Once the model has been evaluated and found to be satisfactory, it can be used to make predictions about future values of the dependent variable.

The linear regression algorithm can be used for a variety of tasks:

- Predicting the sales of a product based on its price and marketing campaign.
- Predicting the risk of a loan default based on the borrower's credit score and income.
- Predicting the growth of a population based on its birth rate and death rate.

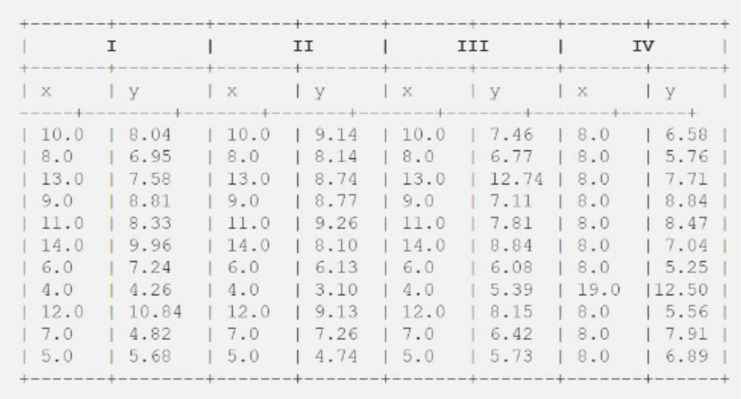
2. **Explain the Anscombe's quartet in detail**.

Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before performing statistical analysis. Anscombe's quartet is a set of four data sets that have nearly identical summary statistics, but have very different distributions and appear very different when graphed. Each dataset consists of 11 (x,y) points.

The four datasets are:

- Dataset 1: A perfect linear relationship.
- Dataset 2: A quadratic relationship.
- Dataset 3: A strong curvilinear relationship.
- Dataset 4: A random scatterplot.

Data Sets:

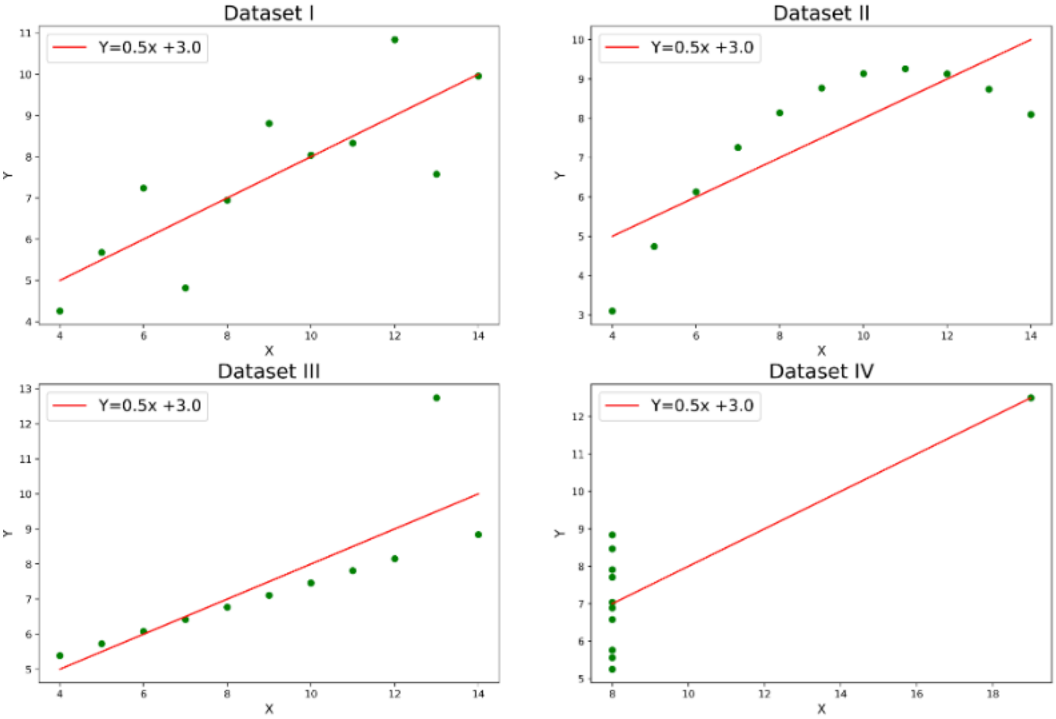


```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

Dataset are different but each dataset has the same summary statistics, such as the mean, variance, standard deviation, correlation coefficient, and linear regression line.

Table summarizing the summary statistics of the four datasets:

Dataset	Mean	Variance	Standard deviation	Correlation coefficient	Linear regression line
1	9.5	23.04	4.83	0.816	y = 0.5x + 3.0
2	9.5	81.00	9.00	0.816	y = 3.2x + 1.7
3	7.5	12.25	3.50	-0.533	y = -0.5x + 9.0
4	7.5	28.09	5.33	0.129	y = 2.8x - 2.5

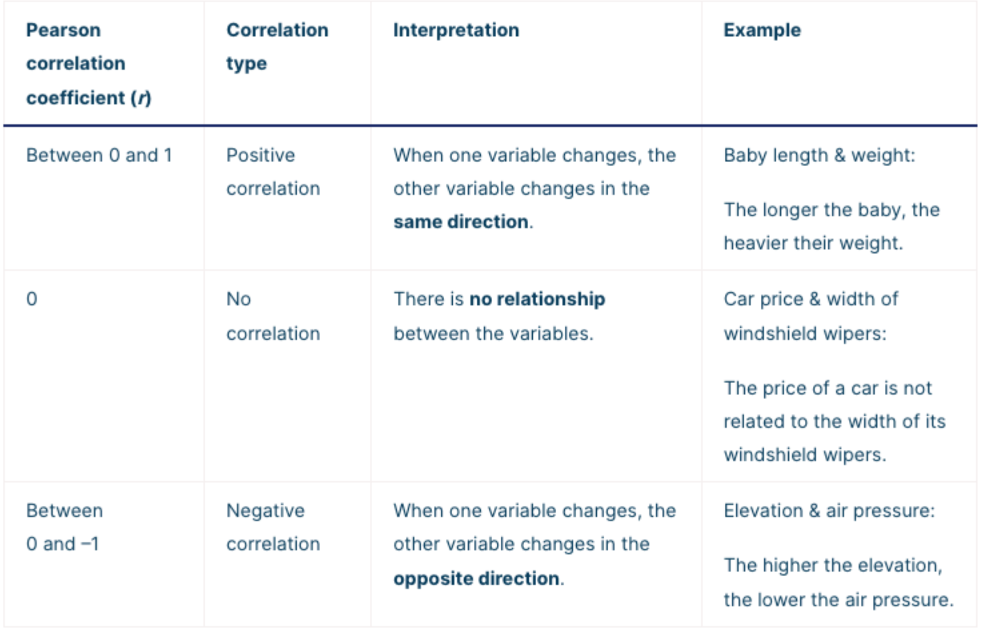


As you can see, the mean, variance, standard deviation, and correlation coefficient are all the same for all four datasets. However, the linear regression lines are very different, and the scatterplots look very different.

This shows that summary statistics can be misleading if the data is not visualized. We should always visualize the data to get a better understanding of its distribution and relationships.

3. **What is Pearson's R?**

Pearson's R is a statistical measure that is used to quantify the linear correlation between two variables. It is a dimensionless number that ranges from -1 to 1, where:



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 using the following formula for sample:

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

This formula can be rearranged as

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:

- x_i is the value of the first variable for the $i$th observation
- y_i is the value of the second variable for the $i$th observation
- $\bar{x}$ is the mean of the first variable
- $\bar{y}$ is the mean of the second variable

Pearson's R is a very popular statistical measure because it is easy to calculate and interpret. However, it is important to note that it is only applicable to linear relationships.

Limitations of Pearson's R:

- It is only applicable to linear relationships.

- It is sensitive to outliers.
- It cannot be used to compare correlations between different datasets with different scales.

If the relationship between the two variables is nonlinear, Pearson's R may not be a reliable measure of correlation. Despite its limitations, Pearson's R is a useful statistical measure that can be used to quantify the linear correlation between two variables. It is a simple and effective way to measure the strength of a relationship between two variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

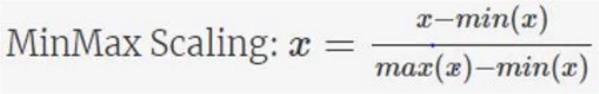
Scaling is the process of transforming the values of features in a dataset to a common scale. This is done to make the features comparable and to avoid features with larger ranges dominating the learning process. *It also helps in speeding up the calculations in an algorithm.*

Reasons why scaling is performed:

- To improve the performance of machine learning algorithms. Many machine learning algorithms are sensitive to the scale of the features. Scaling the features can help to improve the performance of these algorithms by making them less sensitive to the scale of the features.
- To make features comparable. When features have different scales, it can be difficult to compare them. Scaling the features can help to make them comparable so that they can be analyzed more easily.
- To avoid features with larger ranges dominating the learning process. When features have different scales, features with larger ranges can dominate the learning process. Scaling the features can help to ensure that all features have an equal impact on the learning process.

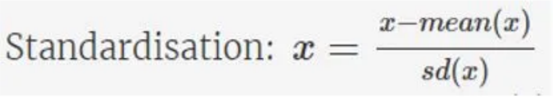
There are two main types of scaling:

- Normalization: This is the process of transforming the values of features so that they have a mean of 0 and a standard deviation of 1. *It brings all of the data in the range of 0 and 1.* ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Standardization: This is the process of transforming the values of features so that they have a mean of 0 and a variance of 1. *Standardization replaces the values by their Z scores.* ***klearn.preprocessing.scale*** *helps to implement standardization in python.*



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it ***loses*** *some information in the data, especially about* ***outliers****.*

Normalization is typically used when the features have different scales, while standardization is typically used when the features have similar scales.

Table summarizing the key differences between normalized scaling and standardized scaling:

Feature	Normalized scaling	Standardized scaling
Mean	0	0
Variance	1	1
Distribution	Good for Non Gaussian distribution	Gaussian (normal)
Use cases	Features with different scales	Features with similar scales

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

This happens when two or more independent variables are perfectly correlated. Perfect correlation means that the two variables are perfectly related to each other. In this case, the variance of one variable can be perfectly predicted by the other variable, and the VIF will be infinite.

```
VIF = 1 / (1 - R^2)
```

A VIF of 1 indicates that there is no multicollinearity, while a VIF of infinity indicates that there is perfect multicollinearity. A VIF of 5 or greater is often considered to be a sign of high multicollinearity.

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

A Q-Q plot is a graphical way to compare two distributions. In a Q-Q plot, the quantiles of the two distributions are plotted against each other. If the two distributions are the same, the points in the Q-Q plot will fall on a straight line. If the two distributions are different, the points in the Q-Q plot will deviate from the straight line. It is used to assess if a set of data follows a particular distribution, such as a normal distribution.

In linear regression, a Q-Q plot can be used to assess the normality of the residuals. The residuals are the differences between the actual values of the dependent variable and the predicted values of the dependent variable. If the residuals are normally distributed, the points in the Q-Q plot will fall on a straight line. If the residuals are not normally distributed, the points in the Q-Q plot will deviate from the straight line.