# 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?**

   The Bivariate analysis of the categorical variable was done using the barplot and the following is the inference I could made by looking into the barplot of all the categorical variable with the dependent variable "cnt"

   - Number of bookings in fall(autumn) seems to be the most whereas in spring there was the least booking of the bikes
   - Count of booking seems to be more in 2019
   - Jan seems to have least number of booking whereas Jul, Sept seems to have the highest number of bookings
   - On holiday there seems to be less count of booking (since in a year number of holidays are going to be less than non-holiday)
   - Sunday has the least number of bookings. Maybe on Sunday many people don't need to go to work or school. And Saturday has the highest, maybe because after working/studying hard people seems to roam around on Saturday.
   - Number of vehicles was booked in working days seems to be more than that on non-working days
   - Rainy season have the least number of bookings.

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

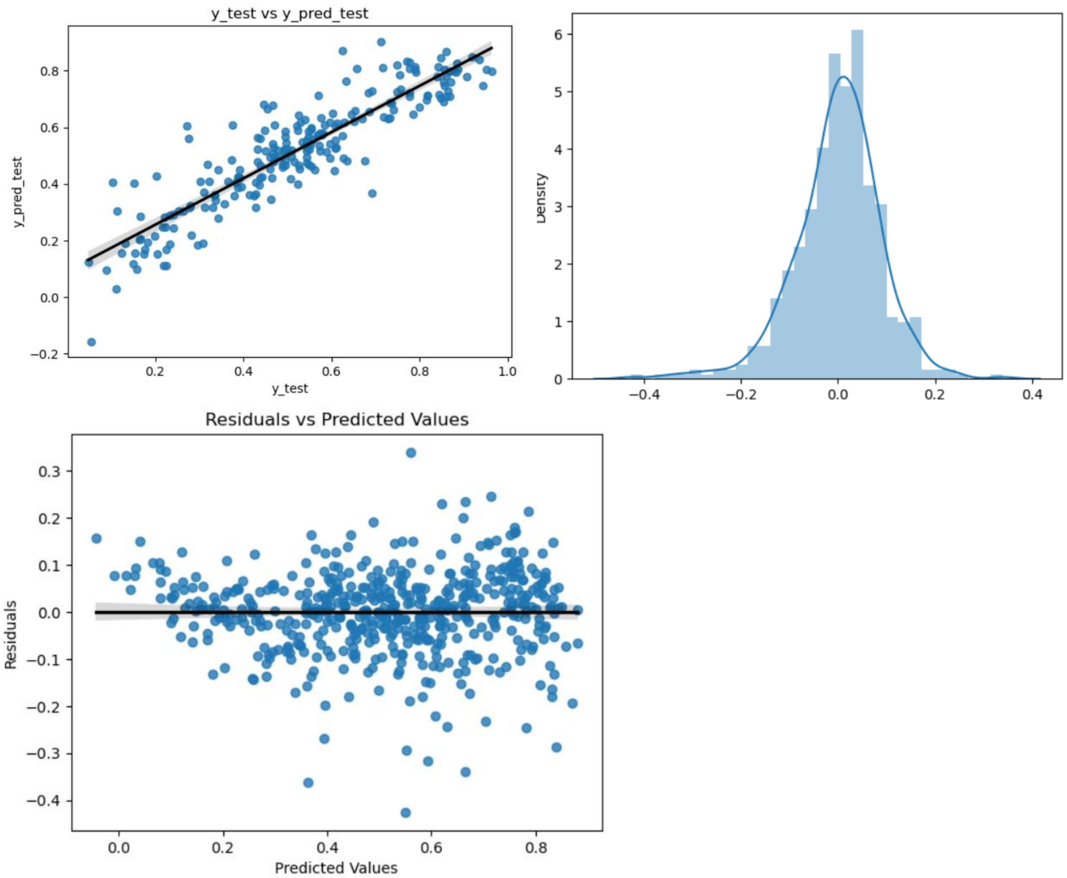
   Simple answer is, ***drop_first=True*** is used to avoid the issue of multicollinearity in linear regression analysis. Multicollinearity arises when two or more predictor variables in a regression model are highly correlated, which can lead to unstable estimates of the regression coefficients. So, when categorical variables are converted to dummy variables, one of the dummy variables can be explained by other dummy variables. Hence, that causes the correlation between variables.

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

   After looking at the pair-plot of numerical variables, ***"temp"*** has the highest correlation with the target variable *"cnt"*

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

   - There was linear relationship between X and Y
   - Error terms are normally distributed with mean zero
   - Error terms are independent of each other
   - Error terms have constant variance (homoscedasticity)



Above plot is the screenshot of the residual analysis to validate the assumptions of linear regression after building the model

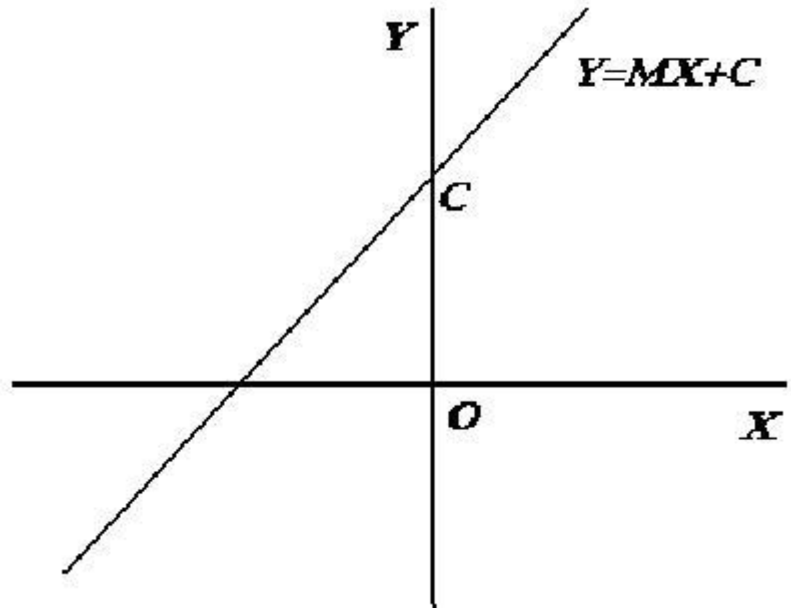
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

After checking the lm.summary(), top three features contributing significantly towards explaining demand of the shared bikes are:

- *temp*
- *sep*
- *winter*

# General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is one of the basic and most important algorithms in Machine Learning. It is a statistical method which shows the relationship between a dependent variable and one (or more) independent variables. Linear Regression can be explained by the Regression Line, which is nothing but the basic straight line having intercept and slope which we have studied in school.



The equation of the above linear line can be written as:

$$y = mx + c$$

So, based on this, we can infer it to the linear regression line where, X (independent variables), y (dependent/target variable), $\beta_0$(intercept), $\beta_1$(slope) as:

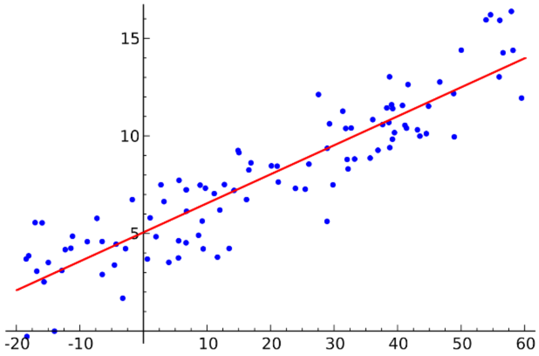
$$y = \beta_0 + \beta_1 X$$

So, now if we consider there are n number of X(independent variables) the equation becomes

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_n X_n + \in$$

Where, ∈ is error terms or residuals.

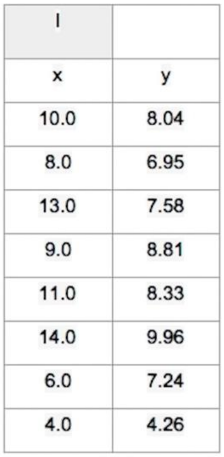
And the actual graph of the linear regression looks like:



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

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, variance and regression line, but which are qualitatively different.

Let us understand this with an example:



| I | |
|---|---|
| x | y |
| 10.0 | 8.04 |
| 8.0 | 6.95 |
| 13.0 | 7.58 |
| 9.0 | 8.81 |
| 11.0 | 8.33 |
| 14.0 | 9.96 |
| 6.0 | 7.24 |
| 4.0 | 4.26 |

This is the dataset where the mean, standard deviation, variance along with the plot and regression like are as follows:

$\bar{x} = 9$

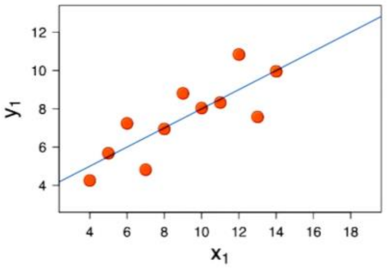
$\bar{y} = 7.5$

$\sigma_x^2 = 11$

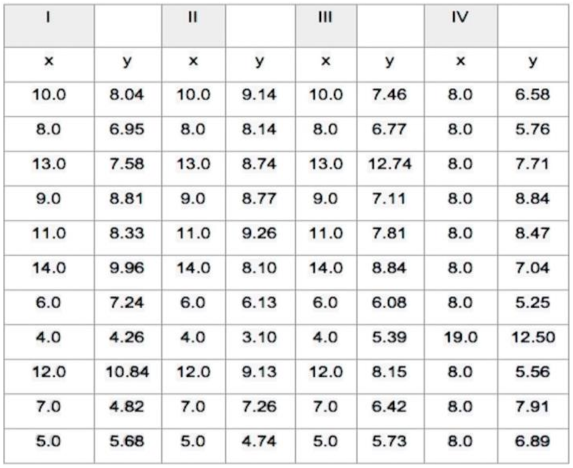
$\sigma_y^2 = 4.12$

$r = 0.816$

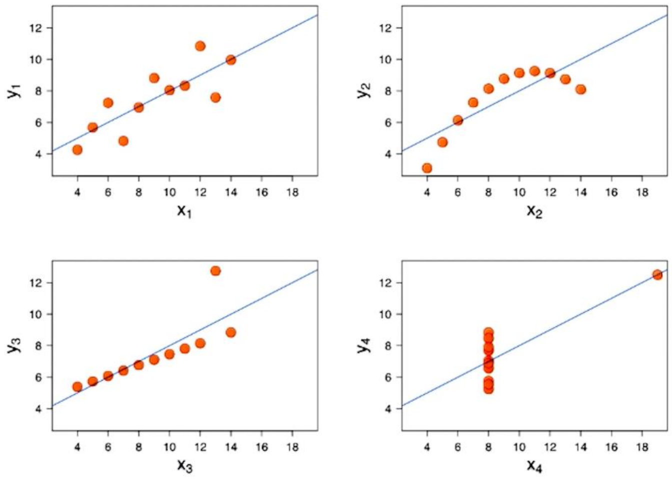
$y = 0.5+3$



Now, if we take other three dataset having same mean, standard deviation, variance and regression line, but the visualization is different



| 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 |



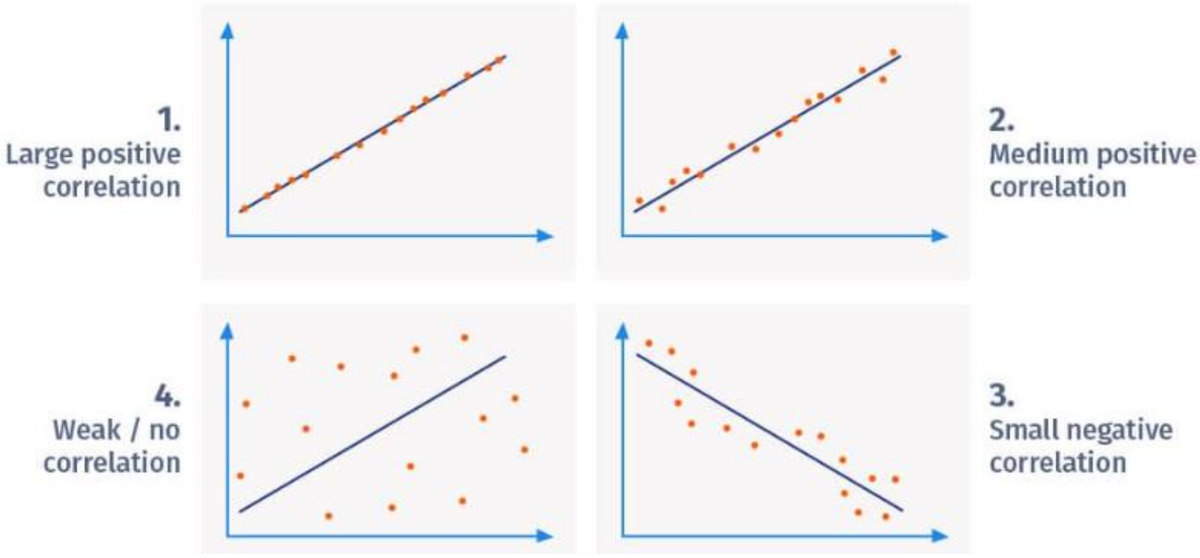
3. **What is Pearson's R?**

Pearson's R is also called as Pearson correlation coefficient. It is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

For example:

If variables have Pearson's R value between 0 and 1, it means they are positively correlated. In addition, if the value is higher then it means they are positively strongly correlated and vise versa. If the Pearson's R value is 0, it means they are not at all correlated. And similarly, if its between -1 and 0, it means they are negatively correlated.

The same can be visualized by the following graph



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

Scaling (also called as Feature Scaling) is a preprocessing step in machine learning that involves transforming the values of features (independent variables) to a common scale.

The scaling is performed to standardize or normalize the range of independent variables so that they contribute equally to the computation of distances and gradients in various machine learning algorithms.

| | Normalized Scaling | Standardized Scaling |
|---|---|---|
| Purpose | It transforms values to min=0, max=1 | It transforms values to mean=0, std=1 |
| Range of Transformed Values | [0,1] | No any specific range, but centered around 0 |
| Sensitive to outliers | More sensitive | Less sensitive |
| Algorithm Preference | Distances-based algorithms, specific range requirements | Gradient-based algorithms, when normal distribution is assumed |

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

As we know the formula of VIF is as follows:

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

Here, ${R_i}^2$ is nothing but, R_square value. If the R_square value is 1, the VIF becomes infinite.

The reasons behind this happening are:

- Perfect Multicollinearity: If two or more variables are perfectly correlated, then ${R_i}^2$ becomes exactly 1, leading to an infinite VIF
- Data Issues: Numerical instability or extreme values in the data can also lead to issues in VIF calculations.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution. It is particularly useful for checking the normality assumption of a dataset. The Q-Q plot compares the quantiles of the observed data against the quantiles of a specified theoretical distribution, often the normal distribution.

The Q-Q plot is used in linear regression to:

- Check Normality Assumption
- Detect Skewness and Outliers
- Validate Regression Assumptions
- Assure Residual Validity
- Identify Model Misfit