# 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?**
   As we cannot use categorical variables directly in our model, due to their non-numeric nature, so we used the concept of dummy variable creation to address this situation. For example- If a categorical variable has p-levels of categories in it, then (p-1) variables can explain the categorical variable.

   While building our model we found that
   - winter, spring from season,
   - year
   - July, September from months
   - Light Snow & Mist from weathersit

   were most significantly affecting predictions of the model.

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

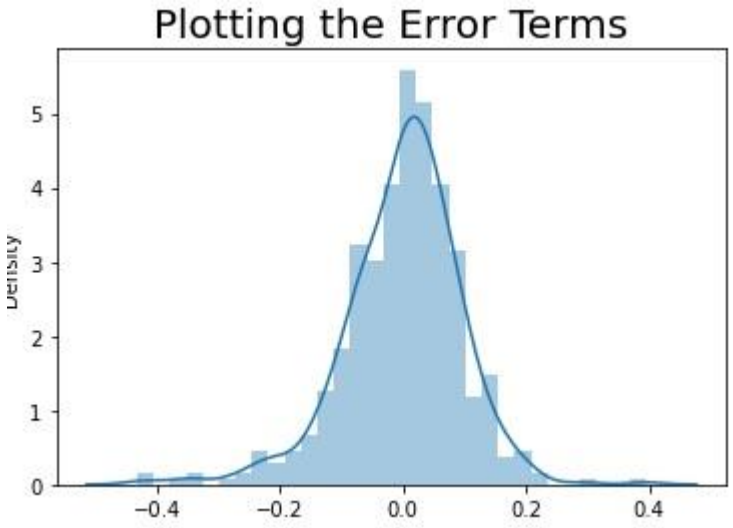
   While creating dummy variable using the code pd.get_dummies(), it actually creates p no of dummy variables for categorical variable having p-level of categories in it. But we actually need p-1 no of dummy variables to explain the categorical variable completely. Therefore, we use drop_first=True to delete one dummy variables.

   Example- Let's say we have 3 levels in Categorical column, namely- furnished, unfurnished & semi_furnished and we want to create dummy variables for that column. Logically if a variable is unfurnished and semi_furnished, then it's obvious that it's unfurnished. So, we do not need 3rd variable to identify the unfurnished.

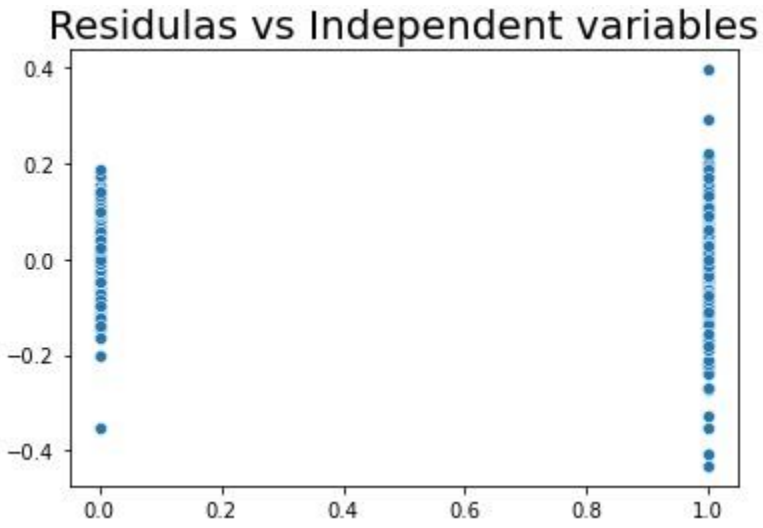
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Looking at the pair-plot, atemp(feel like temperature) variable has the highest correlation, which we later dropped and considered temp(temperature) variable as it had very high correlation with atemp variable, which suggested presence of multicollinearity.

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

   After building the final model, the difference in actual y & predicted y values using the model was computed which is nothing but the error. The error terms should be normally distributed, with mean centred at zero & that assumption was proved right when we plotted the result using distplot.




   We see no pattern between error terms and independent variables.




   Also, the p-value & VIF values suggests that the model built using the chosen variables are statistically significant and have no signs of multicollinearity as well.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   The top 3 features are-
   - Temp(temperature in Celcius): 0.4695 coefficient value
   - Yr(year): 0.2332 coefficient value
   - Light Snow(a sub-category of weathersit): -0.2993 coefficient value

## General Subjective Questions

6. **Explain the linear regression algorithm in detail.**
   Linear Regression is a machine learning algorithm based on supervised learning (i.e., they are trained on labelled data, which means the data contains information about input and output parameters).

   Regression basically means predicting a target variable using independent variables. So linear regression finds out the linear relation between them. There are two type of linear regression based on number of independent variable-
   - **Simple Linear Regression (SLR)**- One Independent variable
   - **Multiple Linear Regression (MLR)**- More than one Independent variable

   Formulae of Linear Regression-

   $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

   **where, for $i = n$ observations:**

   $y_i$ = dependent variable

   $x_i$ = explanatory variables

   $\beta_0$ = y-intercept (constant term)

   $\beta_p$ = slope coefficients for each explanatory variable

   $\epsilon$ = the model's error term (also known as the residuals)

   The motive of the linear regression algorithm is to find the best values for y-intercept and slope of each explanatory variable. In case of SLR we have one slope term and for MLR the number increases.

So, to find best values for y-intercept and slope, the model aims to predict target variable such that the error or difference between predicted and true value of target variable is minimum, which is indeed the cost function. And the best fit line is that line which is obtained when the cost function is minimum. This best fit line best explains the variance in the data.

Cost Function (J): Root Mean Squared Error (RMSE)

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

In order to reduce cost function and achieve the best fit line the model uses Gradient Descent. The idea of which is to start with some random value converge to the point where values of y-intercept and slope comes out to be minimum. Finally, when we have the best values for y-intercept and slope, target variable can be predicted using independent variable.

So, this in is how linear regression works.

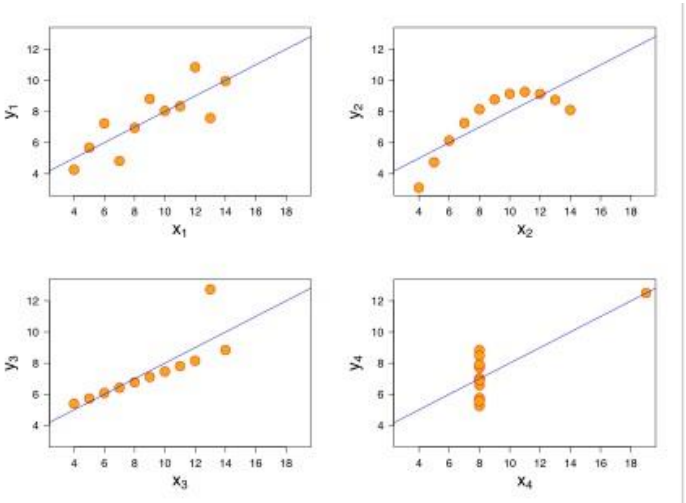
7. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet tells us about the importance of visualising the data before applying ML techniques to build models from them. Before proceeding with making models and predictions, one should understand the distribution of various anomalies present in the data like outliers, diversity of the data etc by plotting the data.

   An experiment carried out by statistician **Francis Anscombe** in 1973**,** where he took 4 sets of similar looking data with 11 data points each found that even though the data looked similar but when plotted their distribution had great difference.

**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 |



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

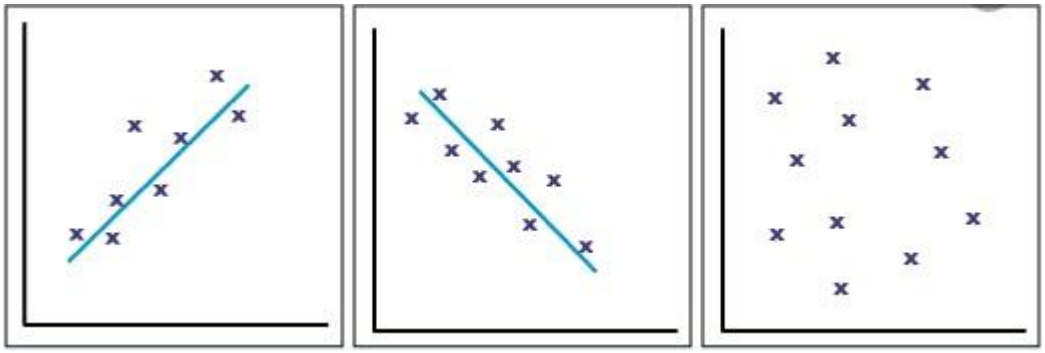
Pearson's R measures the strength of linear relationship between two variables. The value of the Pearson's R or Pearson correlation coefficient is always between -1 to +1. When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.

Formulae-

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

Where,
$r$ = Pearson correlation coefficient
$x$ = Values in the first set of data
$y$ = Values in the second set of data
$n$ = Total number of values.

Below is an example showing positive, negative and zero correlation respectively.



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

Scaling is a technique to standardize the independent features present in the data in a fixed range. Scaling gives the advantage of better interpretability of data and faster convergence of gradient descent function.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

There are two major types of scaling techniques-

- **Normalized scaling**- It scales all data points between the range 0-1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

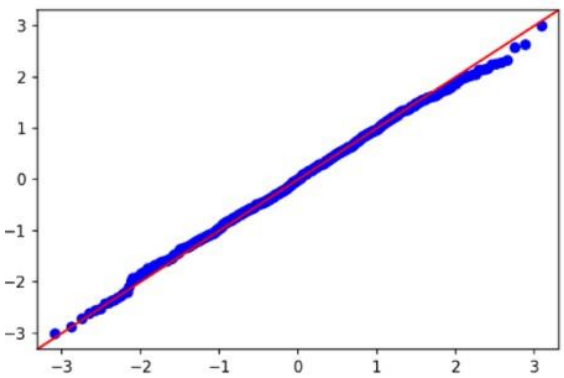
- **Standardisation Scaling**- It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. It follows **Standard Normal Distribution (SND)**

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

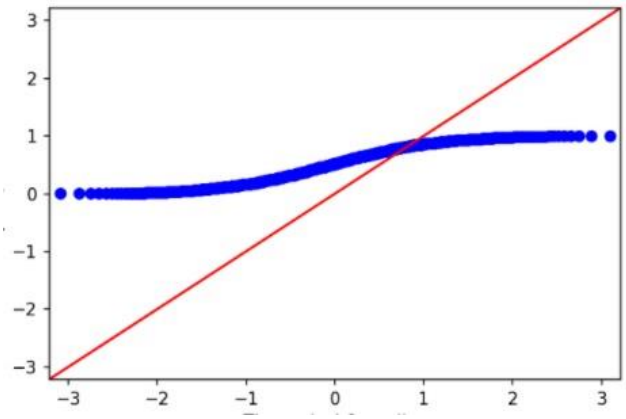
10. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). In other words, infinite VIF suggests variable has severe collinearity with other variables.

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

    Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. But it's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.




Q-Q plot showing normal distribution.




Q-Q plot showing not normal distribution.

Use and importance of a Q-Q plot in linear regression-

In a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. Or if we want to see that our assumption of error terms being normally distributed can also be verified using q-q plot.