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)

Ans: There are many categorical variables in the dataset and their effect on the dependent variable is significant. After the performing the exploratory data analysis it is observed that

- Demand for sharing bikes increased in 2019 compared to 2018
- Demand for sharing bikes is high in fall and summer season and low in spring
- Demand for sharing bikes is low during holidays
- Demand for sharing bikes is low in light snow/ rain weather and high when weather is clear

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

Ans: The major idea in creating dummy variables is that for a categorical variable with '**n**' levels, you create '**n-1**' new columns each indicating whether that level exists or not using a zero or one. In python pandas library, we can do that with the help of get_dummies function with an option for **drop_first=True**. The **drop_first=True** argument helps in reducing the extra column created during the dummy variable creation thus maintain n-1 columns for n levels of a categorical variable and hence avoiding redundancy.

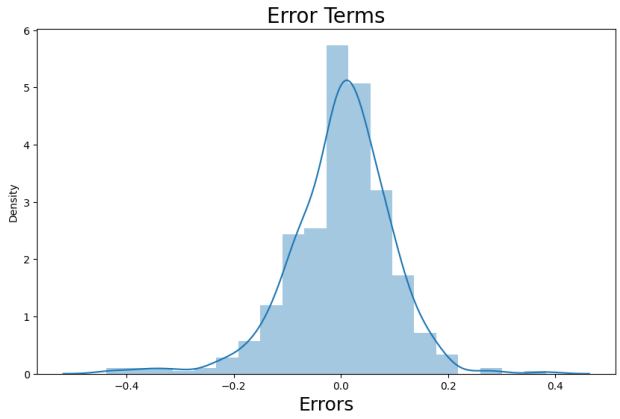
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: The numerical variable 'temp' has the highest correlation with the target variable 'cnt'

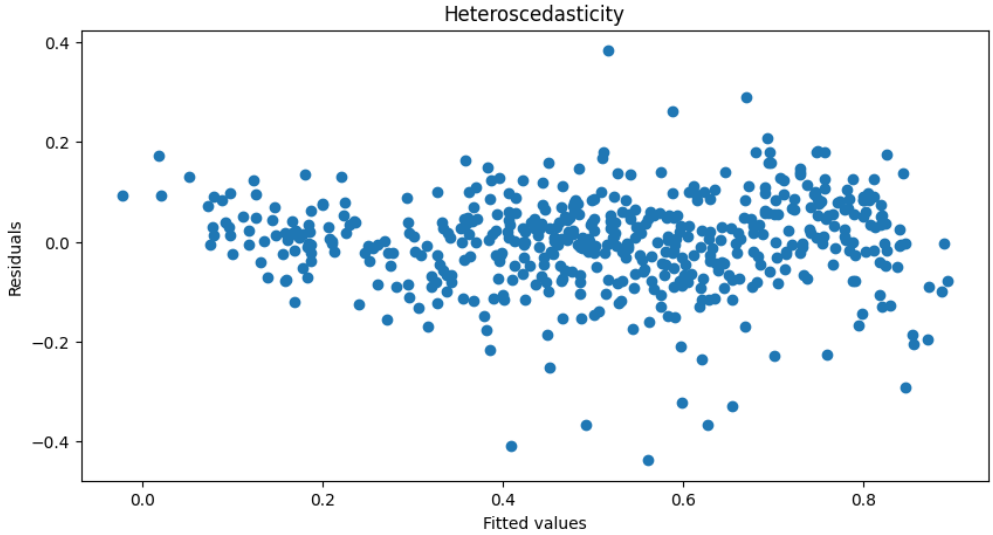
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Validating the assumptions of linear regression is a key process to ensure the reliability of the model. The assumptions of the linear Regression are validated by

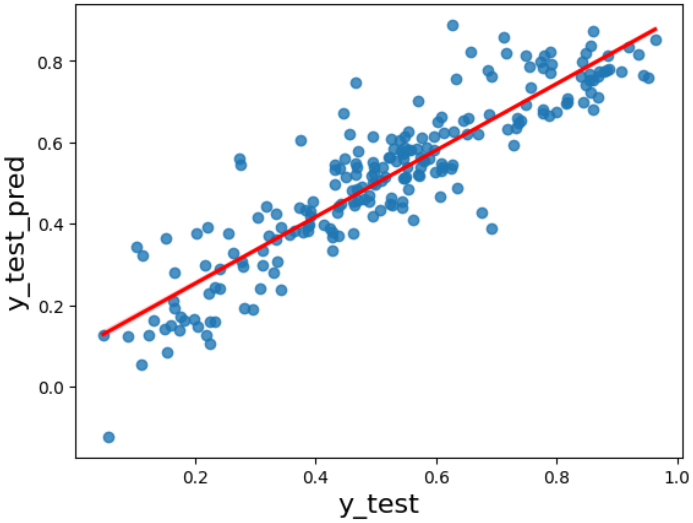
- Calculating the residuals/errors checking the whether the residuals follows normal distribution with approximately zero mean which indicates residuals are **independent** and **evenly distributed around zero**



- Checking for **homoscedasticity** of the residuals by plotting a scatterplot between predicted values and residuals



- There is a linear relationship between target variables and other independent variables



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

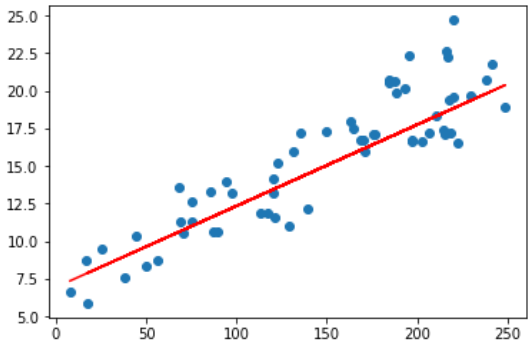
Ans: The top 3 features contributing significantly towards the explaining the demand of the shared bikes are

- Temperature (+0.470719)
- Year (+0.233312)
- Light snow + Rain (-0.304382)

General Subjective Questions

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

Ans: Linear regression explains the relationship between a dependent variable and one independent variable using a straight line. The goal is to find the best-fitting line that minimizes the sum of the squared differences between the observed values and the values predicted by the model.



The equation of the line regression is in the form of straight line i.e.,

$$y = \beta_0 + \beta_1 x$$

Where y is dependent variable which is predicted using independent variable x. β_o is the intercept and β_1 is the slope of the line.

If there are multiple independent variables, then it's a multiple linear regression problem, which explains the relationship between one dependent variable and several independent variables by fitting a linear hyperplane i.e.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

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

Ans: Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. The quartet was created by the statistician Francis Anscombe in 1973.

The quartet consists of four datasets, each containing 11 (x, y) pairs

Dataset I:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset II:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Dataset III:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset IV:
x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Each dataset has the same mean, variance, correlation, and regression line parameters when analyzed numerically, the graphical representation tells a different story. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

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

Ans: Pearson coefficient (R) is used to check a linear relationship (positive/negative) between two quantities x and y. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

$$R = \frac{\sum(X_i - \tilde{X})(Y_i - \tilde{Y})}{\sqrt{\sum(X_i - \tilde{X})^2 \sum(Y_i - \tilde{Y})^2}}$$

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

Ans:
Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

The dataset could have several features which have high magnitudes and different units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.
Standardization:

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

Min-max scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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

1. When there is perfect collinearity between two variables, i.e., when both variables are conveying same information will lead a VIF value of infinite
2. When R^2 value is equal to 1
3. Infinite VIF value can also come if the sample size is very less, which lead to overfitting and unstable model.

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

(3 marks)

Ans: The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution.

We can check whether the residuals after building the model follows normal distribution to validate the linear Regression assumptions.