# 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)

From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019. Additionally, we could discern that bike rental are higher on holidays.

---

**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)

drop_first = True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

---

**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)

The temp variable has the highest correlation with the target variable.

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

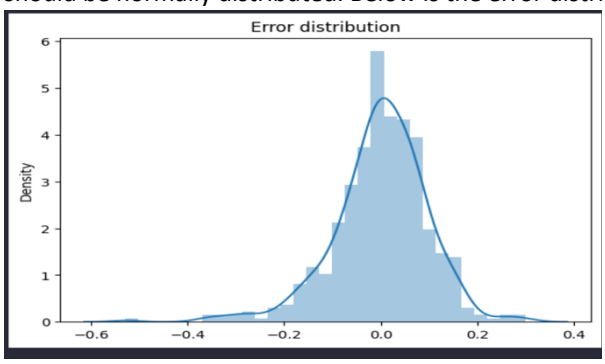
**Total Marks:** 3 marks (Do not edit)

**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

I have validated the assumption of Linear Regression Model based on below 5 assumptions -

**Normality of error terms**

- Error terms should be normally distributed. Below is the error distribution from residual analysis:




**Multicollinearity check**

- There should be insignificant multicollinearity among variables. I made sure the VIF for all the independent variables is not greater then 5 and eliminated them from the model after analysing the p-value as well. For ex: atemp, humidity was having the high VIF value hence were eliminated.

**Linear relationship validation**

- Linearity should be visible among the dependent variable and feature variable, after analysing the model it was clear that there was clear relationship between the feature variable and dependent variable.

**Homoscedasticity**

- There should be no visible pattern in residual values.

---

**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)
Below are the variables that were highly correlated among all the independent variable
Year (0.55),
temp (0.63),
month:

- Sep (0.23)
- Jan (-0.39)

---

# 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).

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−
**Positive Linear Relationship:**
 A linear relationship will be called positive if both independent and dependent variable increases.
**Negative Linear relationship:**
 A linear relationship will be called negative if independent increases and dependent variable decreases.
Linear regression is of the following two types −
 Simple Linear Regression

 Multiple Linear Regression

Assumptions -
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)

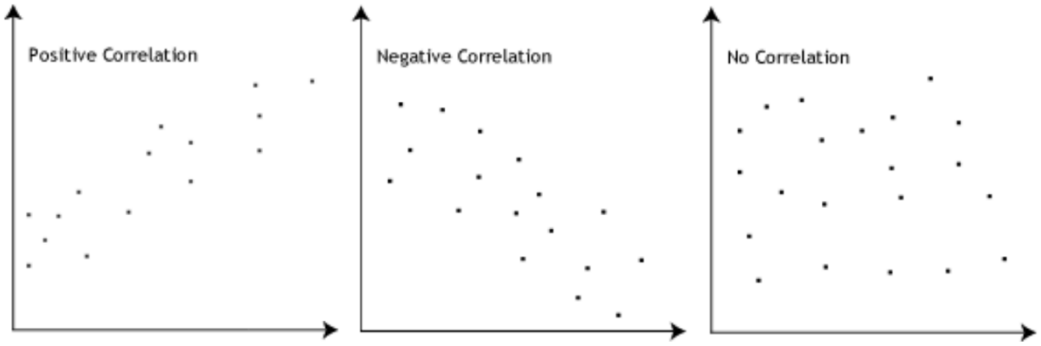
Anscombe's quartet is a group of four datasets that have the same mean, standard deviation, and regression line, but which are qualitatively different. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.
The datasets are as follows:
- Dataset I: A simple linear relationship between x and y.
- Dataset II: A non-linear relationship between x and y.
- Dataset III: A simple linear relationship between x and y, but with an outlier.
- Dataset IV: A simple linear relationship between x and y, but with a different outlier.

The datasets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when the data is plotted, it is clear that the datasets are qualitatively different.
Anscombe's quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

---

**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)

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It is a number between -1 and 1. A value of -1 indicates a perfect negative correlation, a value of 0 indicates no correlation, and a value of 1 indicates a perfect 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 process of changing a set of variables to fall into a specific range. It's performed to unify all values across a data set to certain range which will in turn produce realistic results in linear regression. Normalized Scaling scales the values to have a unit norm, while standardized scales the values between -1 and 1.

**Example:** If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

---

**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)

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite, it shows a perfect correlation between two independents variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

---

**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)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

**Use of Q-Q plot:**

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

**Importance of Q-Q plot:**

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.