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)

atemp – Feeling Temperature is the most important factor in renting a bike. When temperature is warm, bike renting will be high
Weather – If there is snow or rain, bike renting will be considerably lesser than the Misty day
Season – Winter time will attract more bike renters than in Summer time
Wind Speed – When windspeed is high bike renting will be less
Workingday – Bike renting will be slightly more during working days than in weekends or holidays
Year – 2019 is having higher number of bike renting than in 2018.

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)

In Pandas 'drop_first=True' helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

For example, if we have 3 types of values in Categorical column with Blue, Green, Red as its values and we want to create dummy variable for that column.

If one variable is not Blue and Red, then It is obvious it is Green. So, we do not need 3rd variable to identify the Green Value. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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 variable 'atemp' and 'temp' is having the highest correlation (0.63) with the target variable cnt

---

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)

1. **Linear relationship** – By plotting a scatter plot between y-test and y-predicted and observed that the points are **symmetrically distributed around a diagonal line.**
2. **No auto-correlation between the residuals -** Conduct a Durbin-Watson (DW) statistic test. It is available in the summary output of statistical models. The values should fall

between 0-4. If DW=2, no auto-correlation; if DW lies between 0 and 2, it means that there exists a positive correlation. If DW lies between 2 and 4, it means there is a negative correlation. **The Durbin-Watson (DW) value obtained was 2.01 which shows no auto-correlation between the residuals, hence validated this assumption**.

3. **No Multicollinearity** - The independent variables shouldn't be correlated. This can be

determined by determining the VIF (Variance Inflation Factor). VIF<=5 implies no multicollinearity. In the final model **all the independent variables are having the VIF < 5, Hence validated.**

4. Normal distribution of error terms – The error terms should form a normal distribution with mean value as zero. This is tested by plotting a distribution plot of error = (y_{actual} - $y_{predicted}$). The **error term distribution plot shows a normal distribution with mean value as zero. Hence validated this assumption.**
5. **Homoscedasticity -** Homoscedasticity means the residuals have constant variance at every level of x. Created a scatter plot that shows residual vs fitted value**. The data points are spread across equally without a prominent pattern, it means the residuals have constant variance (homoscedasticity)**.

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)

The top 3 features contributing significantly towards the demand of the shared bikes are:
1. **Atemp** - feeling temperature
2. **Light Snow, Light Rain** + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3. **Year** – Bike reting increases with year

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)

<Your answer for Question 6 goes here>
Linear regression is a supervised machine learning algorithm that predicts a numerical value by modeling the relationship between a dependent variable and one or more independent variables. It's a popular method in data analysis and machine learning. In Linear regression, the target variable will be numerical continuous.

It is a form of regression, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables.
The Equation of linear Regression is

$$y = m1x1 + m2x2 + m3x3 + \ldots\ldots\ldots + m(n)\,x(n) + c\ .$$

Where y is target variable and x1, x2 , x3 …… xn are predictor variables .
And we have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error.

We need to get the best fit line which is the line that has the minimum error.
In linear regression, when the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression. Error function is explained by 'e = - y', and error depends on the values of 'm' and 'c'.
Our aim is to build an algorithm which can minimize the error. And in order to do so we use cost function of Linear Regression, Which is: J (mi, c)= (1/2n)Σ(yi – yp)^2 Where yi and yp are expected

values and predicted values.
Our main aim is to minimize J by changing m and c and it can be done using Gradient Descent Algorithm. Cost function measures the performance of a Machine Learning model for given data.
There are two types of linear regression models:
Simple Linear Regression: In this, Linear Regression performs the task to predict a dependent variable(y) based on one independent variable.
The assumptions of simple linear regression were:
1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

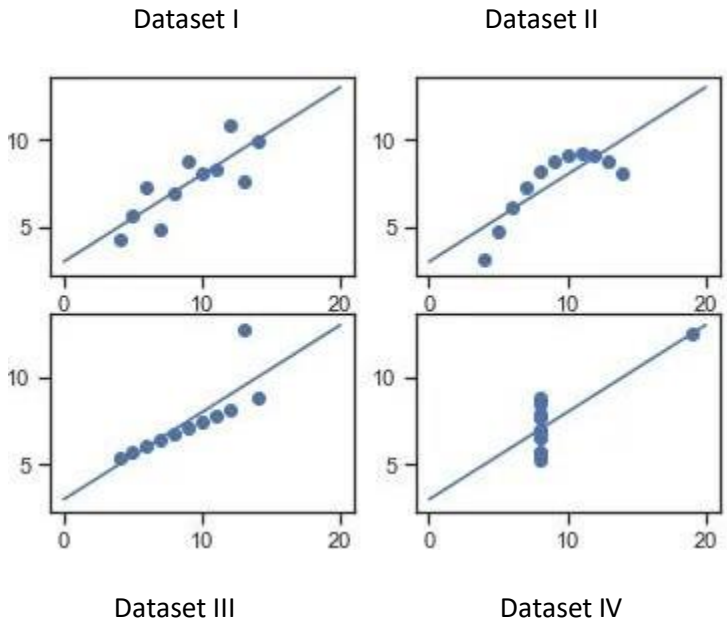
Multiple Linear Regression: Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables).
The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
The assumptions are steps are same as that of simple linear regression. In the assumptions there is one more assumption of No Multicollinearity among the independent variables.

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)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's Quartet underscores that numerical summaries alone can be misleading, emphasizing the crucial role of data visualization in uncovering patterns and the various anomalies present in the data like outliers, diversity of the data, linear separability etc. before applying various algorithms out there to build models out of these data.



Graphical Representation of Anscombe's Quartet

Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance
Data-set II — shows a curve shape but doesn't show a linear relationship
Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier
Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

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)

<Your answer for Question 8 goes here>

Pearson's R (correlation coefficient) is a statistical measure that evaluates the strength and direction of the relationship between two continuous variables by comparing their attributes and calculating a score ranging from -1 to +1. It is considered the most effective method for assessing associations due to its reliance on covariance. This coefficient not only reveals the magnitude of the correlation but also its direction. A high score indicates high similarity, while a score near zero indicates no correlation. This method is parametric and relies on the mean parameter of the objects, making it more valid for normally distributed data.

The Pearson correlation for two objects, with paired attributes, sums the product of their differences from their object means, and divides the sum by the product of the squared differences from the object means.

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

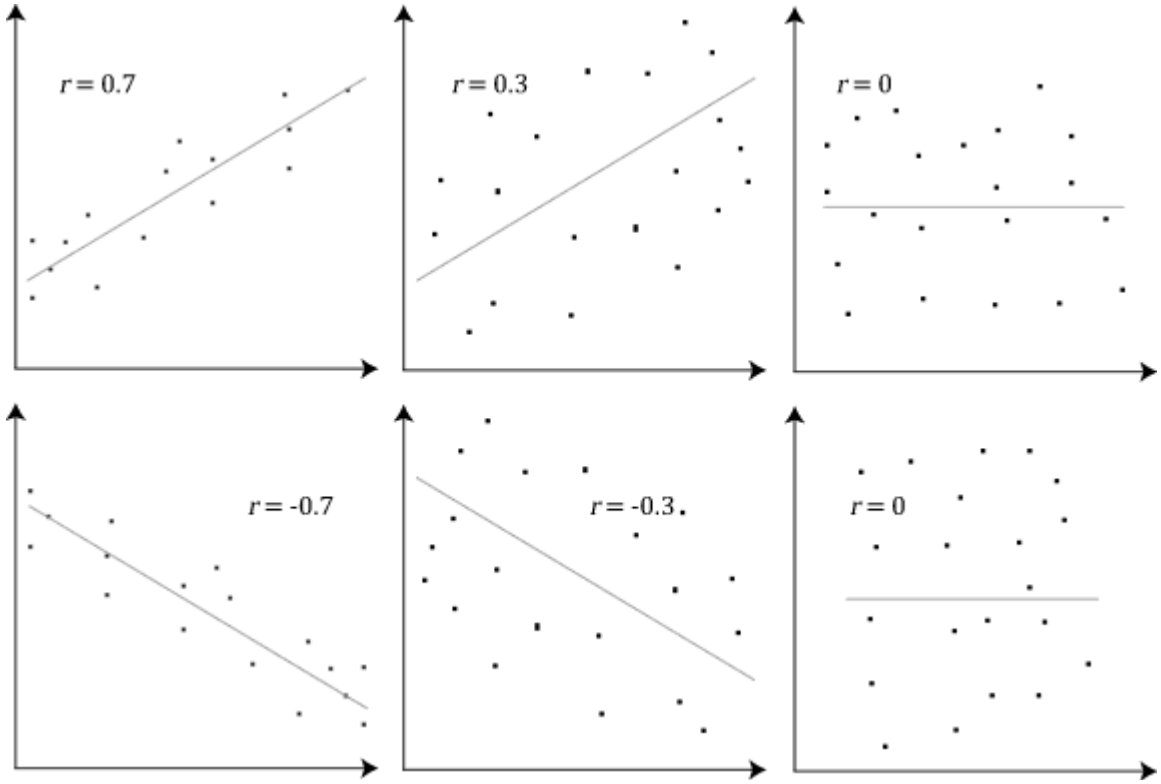
The Pearson correlation coefficient is a good choice when all the following conditions are true:

Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.

The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.

The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.

The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.



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)

<Your answer for Question 9 goes here>

Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. **MinMax Scaling**: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

Key Difference Between Normalization and Standardization

1. Standardization transforms data to have a mean of 0 and a standard deviation of 1, whereas normalization scales the data to a specific user-defined range between 0-1 or -1-1.

2. Normalization makes no assumption about the underlying data distribution, while standardization is often used when the data is assumed to be normally distributed.

3. Standardization is preferred for algorithms that are sensitive to feature scale or assume normality, such as Logistic Regression and Support Vector Machines, while normalization is better suited for distance-based algorithms like k-nearest neighbours (KNN).

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)

<Your answer for Question 10 goes here>

If there is a perfect correlation between the independent variables, the value of R^2 becomes 1 and then the value of VIF becomes infinite.

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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)

<Your answer for Question 11 goes here>

Q-Q Plots are plots of two quantiles against each other. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view toward the given data. A 45 ^{o} angle is plotted on the Q-Q plot, if the sample data is normally distributed, it will fit on the Q-Q plot line. If not, then the data is skewed.

QQ plots is very useful to determine

If two populations are of the same distribution

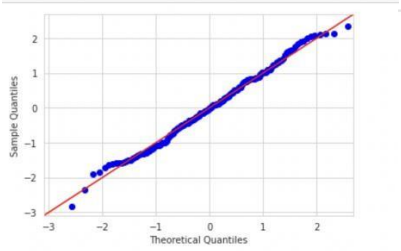
If residuals follow a normal distribution.

Having a normal error term is an assumption in regression and we can verify if it's met using this.

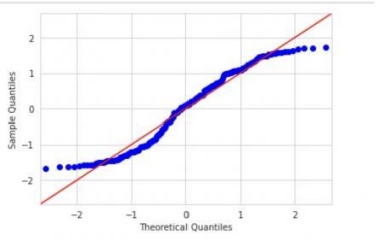
Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

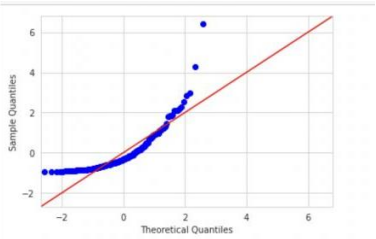
If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of Q-Q plots for different distributions.



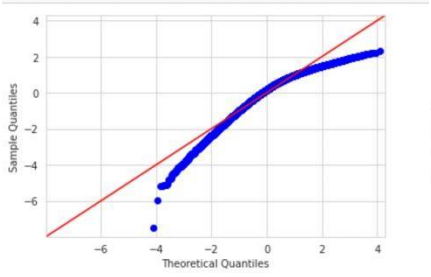
Normal Distribution



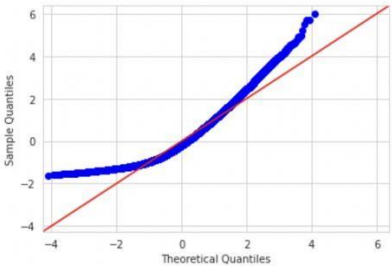
Uniform Distribution



Exponential distribution



Left Skewed Q-Q Plot



Right Skewed Q-Q Plot