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?**

Ans. Observations from the boxplots drawn for categorical variables:

- The year box plots indicates that more bike rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during September month.
- The weekday box plots indicates that more bikes are rent during Saturday.
- The weathersit box plots indicates that more bikes are rent during clear, Few clouds, partly cloudy weather.

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

Ans. We use get_dummies () method having parameter drop_first which allows us to keep or remove first level to get n or n-1 dummies out of n categorical levels.

It explains multi-collinearity in case of Multiple Linear Regression as having n dummies for n levels of a categorical variable can create redundancy of one level. The data available in n-1 of the combination can uniquely represent this redundant column. Hence, it is better to drop one of the columns and have n-1 dummies(columns) to represent n levels.

For example, assuming we have 4 types of values in Categorical column season and we want to create dummy variable for that column. If variables are summer, winter, spring and rainy then it is obvious if season is not winter, spring, and rainy then it is summer. So, we do not need four variables to identify the summer.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. This approach reduces multi-collinearity in the dataset, which is one of the important Assumption of Multiple Linear Regression

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

Ans. By looking at the pair plot **temp** variable has the highest **(0.63)** correlation with the target variable 'cnt' before dropping the column.

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

Ans Given below are the assumptions of Linear Regression:

1) **The Dependent variable and independent variable must have a linear relationship**
A pair plot of the data frame can help us see if the independent variables exhibit linear relationship with the Dependent Variable.

R-squared: 0.796 and Durbin-Watson: 1.909

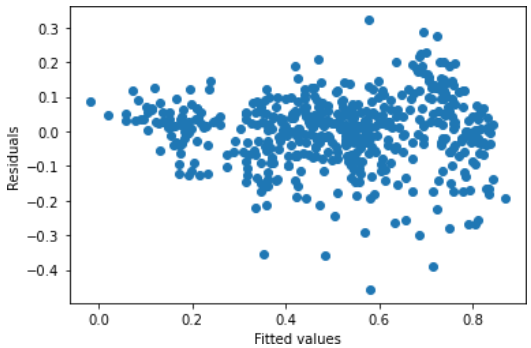
2) **No Autocorrelation in residuals means Independence of error terms**
Durbin-Watson: 1.909 for our final model

According to Durbin-Watson Test.
DW = 2 would be the ideal case here (no autocorrelation)
0 < DW < 2 -> positive autocorrelation
2 < DW < 4 -> negative autocorrelation

As we can see, Durbin-Watson:~ 2 which seems to be very close to the ideal case.

3) **No Heteroskedasticity means Constant variance of error terms**

Residual vs Fitted values plot can tell if Heteroskedasticity is present or not.
If the plot shows a funnel shape pattern, then we say that Heteroskedasticity is present.



We don't see a funnel like pattern in the final modal plot, so no heteroskedastic.

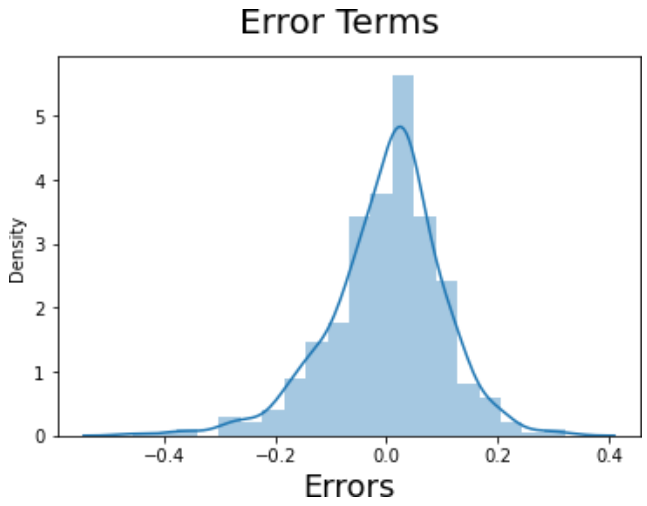
4) **No Perfect Multicollinearity.**
Heatmap and by calculating VIF (Variance Inflation Factor) values, we can check **Multicollinearity**

If VIF=1, Very Less Multicollinearity
VIF<5, Moderate Multicollinearity
VIF>5, Extreme Multicollinearity

The Variables with high Multicollinearity can be removed altogether, or if we can find out which 2 or more variables have high correlation with each other, we could simply merge these variables into one. We need to make sure that VIF < 5.

5) **Residuals must be normally distributed**

Use Distribution plot on the residuals and see if it is normally distributed or Another way how we can determine the same is using Q-Q Plot (Quantile-Quantile)



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

Ans: The Top 3 features contributing significantly towards the demands of share bikes are:

1) temp (Positive correlation).

2)Yr_2019 (Positive correlation)

3)Weathersit_Light (Negative correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: **linear regression** is a linear approach to modelling the relationship between a dependent variable and one or more explanatory variables or independent variables. When we use one explanatory variable is called **simple linear regression** and more than one explanatory variable, the process is called **multiple linear regression.** Linear regression model is used to predict the relationship between variables or factors. It is one of the very basic forms of supervised machine learning where we train a model to predict the behaviour of data based on some variables.

- In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.
- One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.
- Linear regression is used to predict a quantitative response Y from the predictor variable X.

If **y** denotes the dependent variable which we want to predict, and **x** denotes the independent variable which is used to predict **y,** then the mathematical relationship between them

$$y = mx + c$$

This equation is that of a straight line.

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.
b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

When we have many independent variables, the equation can be written as

$$y = c + m_1x_1 + m_2x_2 + \ldots + m_nx_n$$

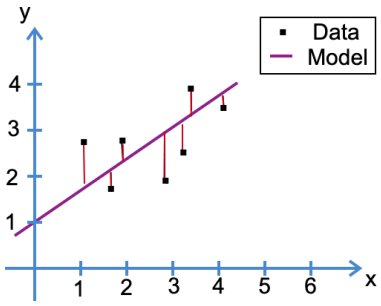
here **c** denotes the y-intercept (the point where line cuts the y-axis) and **m** denotes the slope of the independent variable **x**. Therefore, if we have an equation having dependent variable and independent variables, we can predict dependent variable by substituting values for independent variables.

Our goal is to find values of the ***m*** and **c** that minimize the difference between $\mathbf{y}_a$ (actual) and $\mathbf{y}_i$ (predicted).

Once we get the best values of these two parameters, we will have the **line of best fit** that we can use to predict the values of y, given the value of x.

To minimize the difference between $\mathbf{y}_a$ and $\mathbf{y}_i$, we use the method of **Least Square Method.**

Least Square Method



Graph showing distance between actual data and model line

Least Square method helps us in finding the line of best fit. The values of **m** (slope) and **c** (intercept) are found by keeping sum of the squared difference between $\mathbf{y}_a$ (actual)

Step 1: Calculate the mean of the *x* -values and the mean of the *y-values*

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Mean of x and y values

Step 2: The following formula gives the **m** (slope) of the line of best fit.

$$m = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

Formula to find **m** (slope)

Step 3: Compute the value ***c*** *(y*-intercept) of the line by using the formula:

$$c = \overline{Y} - m\overline{X}$$

Formula to find **c**
(y-intercept)

Step 4: Use the slope ***m*** and the *y*-intercept **c** to form the equation of the line.

An example for running a sales promotion and expecting a certain number of count of customers to be increased now what we can do is we can look the previous promotions and plot if over on the chart when we run it and then try to see whether there is an increment into the number of customers whenever we rate the promotions and with the help of the previous historical data we try to figure it out or we try to estimate what will be the count or what will be the estimated count for my current promotion .this will give us an idea to do the planning in a much better way about how many numbers of stalls maybe we need or how many increase number of employees we need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

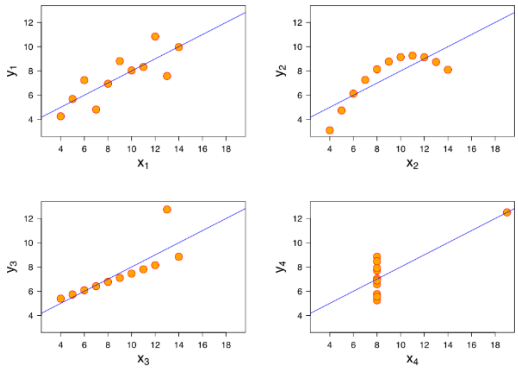
Some Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Anscombe's quartet datasets have the same mean, standard deviation, and regression line, but are qualitatively different. It illustrated the importance of analysing at a set of data graphically and not only relying on basic statistic properties.



Simple understanding:

Francis John “Frank” Anscombe who was a statistician found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson’s R?

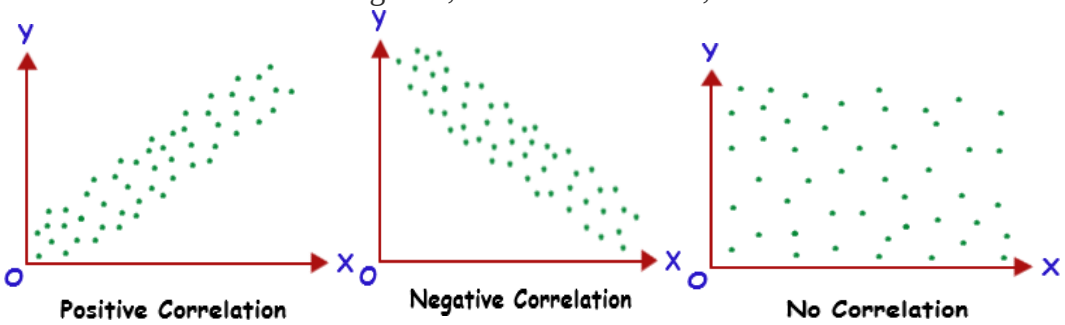
Ans: The Pearson correlation method is the most common method to use for numerical variables and measures the linear correlation between two variables. it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

There are below requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.



The Pearson correlation coefficient is denoted by the letter "r". The formula for Pearson correlation coefficient r is given by:

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

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.

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

Ans: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the

variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and no other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and

  1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

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

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

the choice of using normalization or standardization will depend on the problem and the machine learning algorithm we are using. There is no fix rule to tell when to normalize or standardize data. we can always start by fitting model to raw, normalized, and standardized data and compare the performance for best results.

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

Ans**:** If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

Below measures can be taken to reduce VIF (multicollinearity)

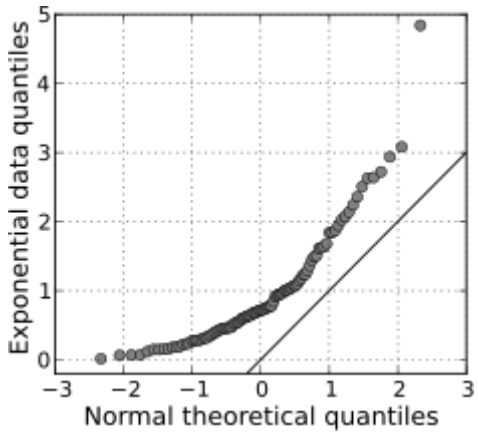
- Review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model.

- We can use principal component analysis and determine the optimal set of principal components that best describe your independent variables.
- We can increase the sample size by adding more data points to our model so that the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- We can transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- We can use a different type of model call ridge regression that better handles multicollinearity.

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

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.