Assignment- based Subjective Questions

1. **From your analysis of the categorical variables of the dataset, what could you infer about their effect on dependent variable ? (3 marks)**

Answer: The analysis of the categorical variables is done using boxplot and barplots. The analysis inferences are as below,

- **Season -** 'Fall' season has more demand for bikes as compared to others. The demand for the year 2019 is increased than in 2018.
- **Month** - for the 'sept' month, the demand is highest. The rise in demand for bikes continued till 'sept' month and then it started decreasing till end of year.
- **Weekday** - 'Thursday', 'Friday', 'Saturday' and 'Sunday' attracted more bookings than other days.
- **Weathersit -** 'Clear' weather has more bookings.
- **Holiday -** When holiday is there, bookings seem to get increased and hence demand is more for holiday.
- **Working Day -** Demand seems to be equal on working days and non-working days as well.
- **Year -** Demand for bikes increased from year 2018 t0 2019.

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

Answer:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If there are 'n' number of variables, then dummy variables that need to be created are 'n-1'.

Syntax : drop_first = bool ; default value is False, which implies whether to get 'n-1' dummy variables or 'n' variables.

Example :
If we have 3 variables as Republican, Democrat, or Independent. Political affiliation can be represented with two dummy variables:

- X1 = 1, if Republican; X1 = 0, otherwise.
- X2 = 1, if Democrat; X2 = 0, otherwise.

In this example,we don't have to create a dummy variable to represent the "Independent" category. If X1 equals 0 and X2 equals zero, we know the voter is neither Republican nor Democrat. Therefore, voter must be Independent.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: I have validated the assumptions of Linear Regression based on following assumptions:

- **Linear Relationship Validation :** There should be visible linear relationship between target variable and predictors.
- **Multicollinearity:** There should not be any Multicollinearity between predictors.
- **Normality of Error terms:** Error terms should be normally distributed.
- **Homoscedasticity:** There should not be any visible pattern in residual variables.
- **Independence of Residual terms:** There should not be any auto-correlation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
1. temp - coefficient = 0.4782

2. Winter - coefficient = 0.0959
3. Sept - coefficient = 0.0909

General Subjective Questions

1. Explain the Linear Regression algorithm in detail. (4 marks)

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

The mathematical equation to find out the best fit line is,

Y = mX + c

Here, x and y are two variables on the regression line.

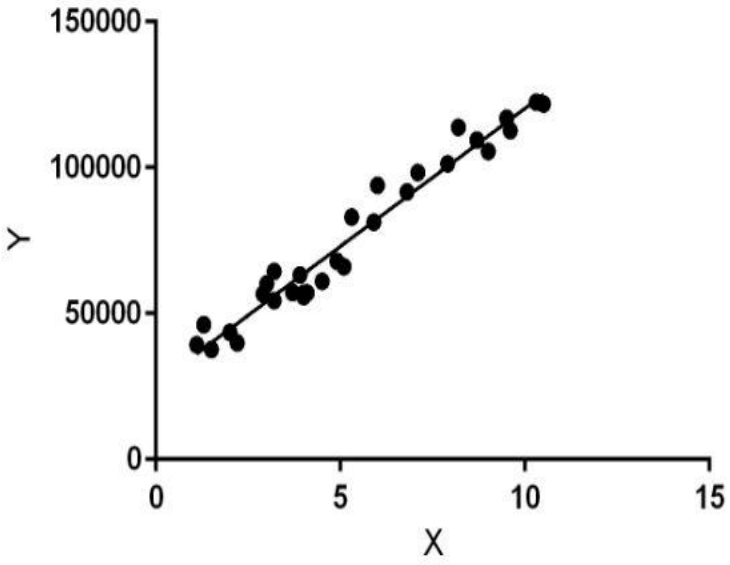
m = Slope of the line

c = y-intercept of the line

X = Independent variable from dataset

Y = Dependent variable from dataset.

Graphical Representation of Linear Relationship between variables,



A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

The goal of the linear regression algorithm is to get the best values for m and c to find the best fit line and the best fit line should have the least error. In Linear Regression, **Mean Squared Error (MSE)** or **RSS** cost function is used, which helps to figure out the best possible values for m and c, for the best fit line. Using the cost function, the values of m and c change such that the MSE value settles at the minima. Gradient descent is a method of updating m and c to minimize the cost function.

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

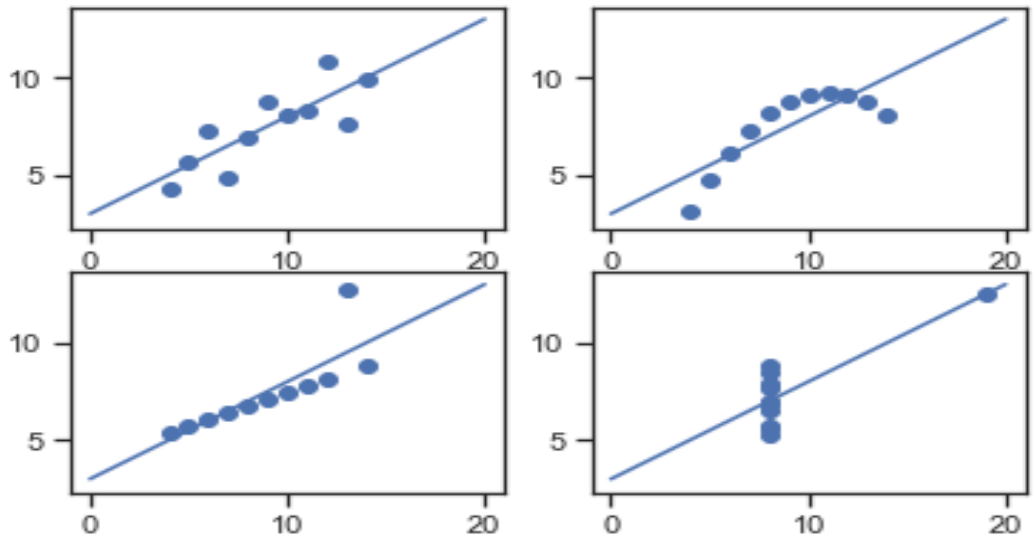
Answer: Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but difference or change is seen in the regression model once you plot each data set.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before analyzing it and building the model. These four data sets have nearly the same statistical observations for each x and y point in all four data sets. However, when plotted these data sets, they look very different from one another.

Datasets with statistical summary is as follow,

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm.



Data-set 1 - represents a linear relationship with some variance.

Data-set 2 - shows a curve shape but not a linear relationship.

Data-set 3 - a tight linear relationship between x and y and one outlier.

Data-set 4 - looks like the value of x remains constant, except for one outlier as well.

It emphasizes the importance of Data Visualization by plotting graphs in Data Analysis while building the model.

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

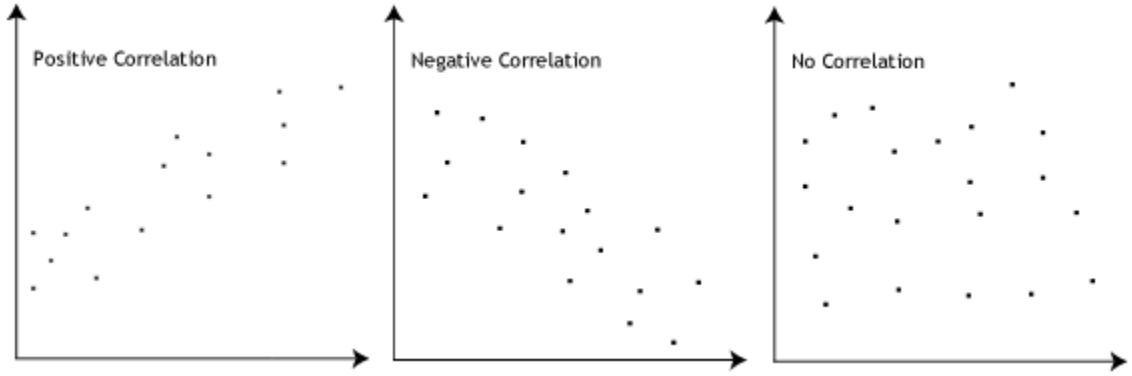
Answer : Pearson correlation coefficient is a measure of the strength of a linear association between two variables denoted by r.

The Pearson's correlation coefficient varies between -1 and +1.

r = 1 - the data is perfectly linear with a positive slope. I.e. as the value of one variable increases, so does the other variable increases.

r = -1 - means the data is perfectly linear with a negative slope I.e. as the value of one variable increases, the value of the other variable decreases.

r = 0 means there is no linear association.



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

Answer:

What is scaling?

Scaling is a technique or step in data preprocessing for building a model which is applied to features to normalize the data within a particular range. It standardizes features having highly varying magnitudes to affixed range.

Why is scaling performed?

When the data is collected, it contains features having highly varying units or magnitudes and range. For the correct analysis and calculations while building the

model, it becomes necessary to scale up the features and bring all of them to the same level of magnitude and range. Hence, scaling is performed.

Example: If we have designed the model or algorithm in such a way that it would calculate or consider the values which are above 10 kg only. But in one of column we have a value mentioned in grams, let's say 5000 gm (5kg). Here, by algorithm this value will get selected but in actual it should not get selected as it is less than 10kg. Hence, here our predictions can get wrong and hence scaling is performed.

Difference between normalized scaling and standardized scaling?

Sr. No.	Normalization	Standardization
1	It is Min-Max Scaling. It uses Maximum and Minimum values of features for scaling.	It uses Mean and standard deviation of features for scaling.
2	It brings all the data in range of 0 and 1 or -1 to 1.	There is no boundation or range.
3	It is affected by outliers.	It is very less affected by outliers.
4	It is used for different scaled features.	It is used to bring the data to zero mean and unit standard deviation.
5	In Scikit-learn it is, sklearn.preprocessing.MinMaxScaler	In Scikit-learn it is, sklearn.preprocessing.scale

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

Answer: This happens when there is a perfect linear relationship between two independent variables. VIF is inverse of (1- R Squared). When there is perfect correlation between two independent variables, R squared value is 1, which yields VIF = infinity.

Example: Let's say we have to predict whether the person is more prone to heart attack or not. We have data containing variables like BP, Weight, Height, Stress, Age

and BMI. Here we can say that there is high Multicollinearity between BMI and Weight or BMI and Height as BMI is calculated from weight and height of the person only. Henec in this case VIF will be higher.
We need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of Q-Q plot in linear regression? (3 marks)

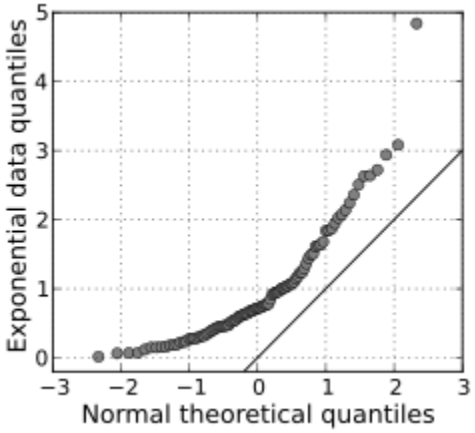
Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a dataset came from some theoretical distribution such as a Normal, exponential or Uniform distribution. This is helpful in determining whether the datasets are from the same populations with the same distributions when we receive the train and test dataset separately.

Use of Q-Q plot.

A Q-Q plot is a plot of one quantile of the first dataset plotted against quantiles of the second one. Here, quantile means the fraction of points below the given value. **Example**, if the fraction is 35%, the 35% of data fall below this point and 65% will be above this line.

If two datasets are from the same population, then the line with 45 degrees is plotted and the points are approximately along this line.

A Q-Q plot showing 45 degrees reference line,



Importance of Q-Q plot in linear regression:

1. With the use of Q-Q plots, some distributional aspects like shifts in location and scale, symmetry changes can be detected. Also we can check the presence of outliers with this plot.
2. It checks if two data sets came from populations having common distribution, common location and scale, similar shape in distributions, same tail behavior.