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?**

Answer - Below are the categorical variables and their effect on dependent variable:

- **Season** – Bike rentals are more during the Fall season and then in summer. Demand is lowest during spring season.
- **Yr** - Bike rentals demand has increased in the year 2019 from 2018.
- **mnth** - The months from May to Sep has good amount of bookings which shows that the month has some role here and can act as a good predictor for the dependent variable.
- **Holiday** – This has a good influence on predictor as bike bookings happening on non-holiday is more.
- **Weekday** - Bike rentals are more Friday and Sunday. But not much influential on predictor still.
- **Workingday**- Demand is a bit more on working day.
- **weathersit** - Bike rentals are more during clear weather or in partly cloudy weather. Demand is less during snowy weather. This variables has good amount of impact on bike demand.

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

Answer – Although we can drop any one of the columns during dummy variable creation however drop_first = True is generally used and it is important because it reduces the redundant/extra dummy column which we can easily find out using the other dummy variables values.

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

Answer - Temp and atemp numerical variables have the highest correlation with the target variable (0.63 as per heat map)

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

With the help of pair plot found target variable is having linear relationship with few independent /predictor variables (e.g temp, atemp).
After plotting histogram of the error terms, it was concluded that error terms are following a normal distribution with mean 0. This was done
By plotting a distplot of residuals.
All the predictor variables in the final model have VIF value less than 5. So we can consider that there is insignificant multicollinearity among the predictor variables.

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

   Year ,Weather and atemp

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

Regression analysis is used for three types of applications:

- Finding out the effect of Input variables on Target variable.
- Finding out the change in Target variable with respect to one or more input variable.
- To find out upcoming trends.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

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

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.

Mathematically, we can write a linear regression equation as:

$Y = mx + c$
Where y is the target variable, m is the slope, x is the predictor variable and c is the intercept.

Y is also called **dependent** variable and x is the **independent** variable.

There are 2 types of linear regression models.
1. Simple Linear regression
2. Multiple Linear regression

In simple linear regression there is only one predictor variable wherein multiple regression there are more than one predictor variables.

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

Positive Linear Relationship:

If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

Finding the best fit line:

.
The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

Cost function-
The different values for coefficient of lines (b0, b1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
Cost function optimizes the regression coefficients. It measures how a linear regression model is performing.

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

R^2 or Coefficient of Determination –

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes.
Overall, the higher the R-squared, the better the model fits your data.
Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

RSS(Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

TSS(Total sumofsquares): Itis thesumoferrorsofthedatapoints from meanof response variable. Mathematically, TSS is:

$$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Assumptions of Linear Regression

- **Linear relationship between the features and target:**
  Linear regression assumes the linear relationship between the dependent and independent variables.

- **Small or no multicollinearity between the features:**
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

- **Homoscedasticity Assumption:**
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

- **Normal distribution of error terms:**
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

- **No autocorrelations:**
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

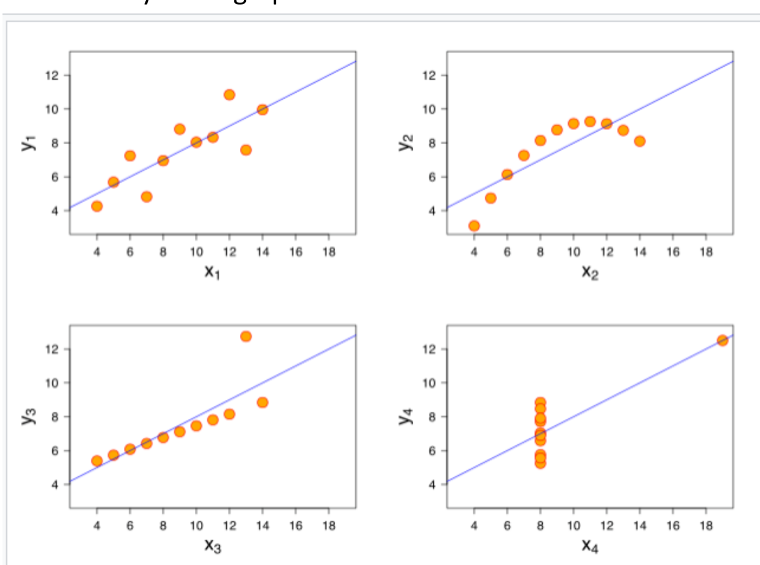
3. **Explain the Anscombe's quartet in detail.**

   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 analyzing it and the effect of outliers on statistical properties.

   Once Francis John "Frank" Anscombe who was a statistician of great repute 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.
All four sets are identical when examined using simple summary statistics ,but vary considerably when graphed.



4. What is Pearson's R?

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

Scaling 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 it leads to incorrect modelling. To solve this issue, we have to 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 none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Similarly, in many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.

There are 2 methods of scaling and below are their differences in the way they treat data:

- 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.

  Formula which is applied on data would be

  x= (x – min(x))/(max(x) – min(x))

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

  x= (x – mean(x))/sd(x)

  where sd is the standard deviation

sklearn.preprocessing.scale helps to implement standardization in python.

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

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

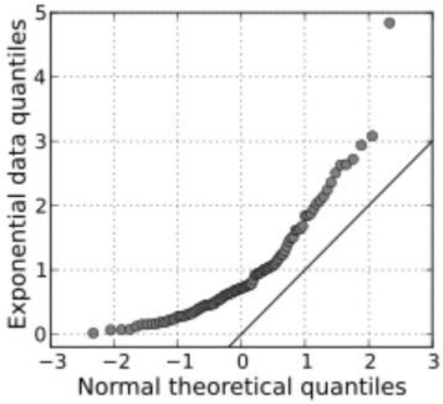
If there is perfect correlation which means R-squared is 1 , which lead to 1/(1-R2) infinity then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1.
To solve this problem we need to drop one of the variables from the dataset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

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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.

(Note-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.)



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.