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?**

Four variables 1. Seasons, 2. Month, 3. Years, 4. Weekdays, 5, weather where analysed by plotting boxplot and bar plots

In seasons vs total number of users
In Fall we have the highest number of users
And spring we have the least

In Months vs total number of users
The Month of September showed the highest number of users
The Month of Jan showed the least number

In Years vs total number of users
The year 2019 saw a rise in total number of users in comparison to 2018.

In Weekdays vs total number of users
Weekdays had more number of users as compared to weekends but weekdays we have 5 days in comparison to weekend of two days.

Weather vs Total number of users
The number of users where more on a clear weather day followed by mist and cloudy day

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

The default value of drop_first is False. If drop_first is set to False, that is One-Hot encoding. This will result in creation of more number of columns. To perform dummy encoding, we need to mention drop_first as True. So when categorical variables are of n levels, we get n-1 columns to represent the dummy variables. Hence it reduces the correlations created among dummy variables

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

The temp : temperature in Celsius and - atemp: feeling temperature in Celsius has the highest correlation with the target variable.

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

The Linear relationship between y_train and y_train_predict is established via scatter plot , Linearity Test Plot.
In, 'Plot Residuals' in residual analysis and model evaluation we can see that most of the errors are uniformly distributed around the zero.
Also We can see that in error plot, the error is normally distributed around zero.

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

Our equation of Multi-Linear regression line is:

Y= -0.3753 +(0.2313* workingday) +(0.4227 *temp)-(0.6877* spring) -(0.3539*July)- (0.2747*Sat)-(1.2958 *

Light Snow&Rain)-(0.3505*Mist&Cloudy)+(1.0415 * year2019)

The Top three features based on the coefficient that can explain the model are 1. Weather conditions as Light Snow and rain, then Misty and cloudy shows high influence. 2. The year 2019 which means that an year on increase could occur 3. The third is the season ,the spring which has as 0.6877 coefficient.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Simple Linear Regression

In this, we predict the outcome of a dependent variable based on the independent variables, the relationship between the variables is linear. Hence, the word linear regression.

Y is the predicted variable

X is the independent variable

Y = mX+ C

C is the constant

Positive Linear Relationship

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

Negative Linear Relationship

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

The straight line in the diagram is the best fit line. The main goal of the simple linear regression is to consider the given data points and plot the best fit line to fit the model in the best way possible.

In multiple Linear Regression we consider the effect of all independent variables on the dependent variable Y.

X1, X2 .. Xn are the independent variable

Y = C+ a1 X1+ a2 X2 + a3 X3…. an Xn

Steps involved in Linear regression Using Python

1. Import the relevant libraries

2. Define data set

3. Perform EDA on data set

4. Visualise Data

5. Split Data to test and train data

6. Scale numerical data

7. Encode Categorical data One hot encoding or dumb encoding

8. Perform Recursive feature elimination

9. Find the P values and VIF values and the eliminate multicollinearity and insignificant data and select relevant data.

10. Finalise Model using the selected relevant features that can explain the target variable.

11. Find the R square score of train data set, later this score can be compared with the test data set.

12. Evaluate model using residual analysis and multiple plots. Like plotting the errors to verify if the distribution is normal, the plot residual against all selected features to find the distribution is uniform across the mean.

13. Use the model on test data set to predict the values. Predict the target variable using the test data set.

14. Find the R square score of test data.

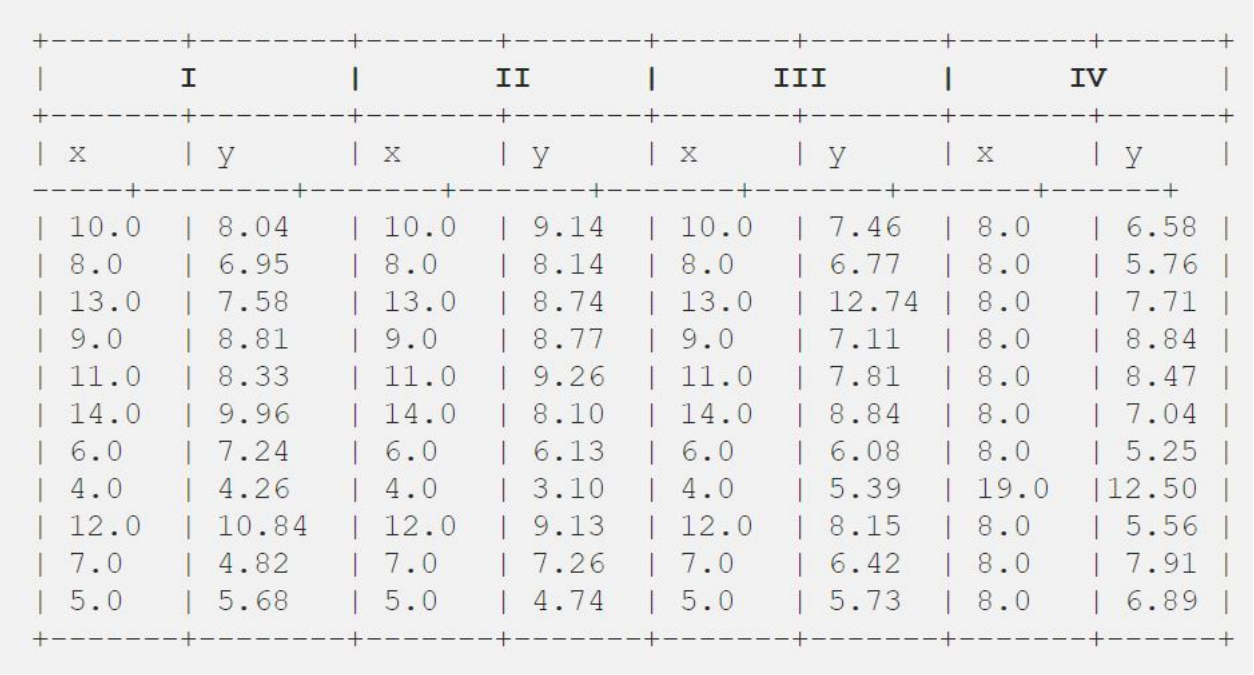
15. Compare the R square score with the test data set and train data set.

16. Plot the predicted target variable and the test data actual variable to find the model compatibility.

2. 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.

3. What is Pearson's R?

Pearson's correlation coefficient R is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship.

The Pearson coefficient is a measure of the strength of the association between two continuous variables.

To find the Pearson coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association.

3. What is scaling? Why is scaling performed?

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 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.

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

Normalization:

Typically means rescales the values into a range of [0,1].

Minimum and maximum value of features are used for scaling
It is used when features are of different scales.
It is really affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
It is useful when we don't know about the distribution

Standardization

typically means rescales data to have a mean of 0 and a standard deviation of 1

It is used when we want to ensure zero mean and unit standard deviation
It is not bounded to a certain range.
It is much less affected by outliers.
Scikit-Learn provides a transformer called StandardScaler for standardization.
It is useful when the feature distribution is Normal or Gaussian.

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

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. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

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 y=x 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.

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.