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> In the bike sharing dataset while performing EDA, I visualized the relationship between the categorical variables and the target variable. the demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018. Also, during model building on inclusion of categorical features such as yr,season etc. we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

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

Ans> During dummy value creation (dummy encoding) it is advisable to use drop_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding. Hence, drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.

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 graphs we can say that temp and atemp 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?

Ans> We can validate the assumptions of Linear Regression after building the model on the following training set by below method:

- Fitted regression line is linear.
- Identified the relationship between dependent and independent variable.
- Checked for multicollinearity between the variable via VIF.

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

Ans> Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- Demands increases in the “ yr ”
- Demand decreases if it is snow, thunderstorm or working day

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans> 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.

Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages – (1) analyzing the correlation and directionality of the data,

(2) estimating the model, i.e., fitting the line, and

(3) evaluating the validity and usefulness of the model.

First, a scatter plot should be used to analyze the data and check for directionality and correlation of data. When the scatter plot indicates a positive relationship between the two variables. The data is fit to run a regression analysis.

The first step enables the researcher to formulate the model, i.e. that variable X has a causal influence on variable Y and that their relationship is linear.

The second step of regression analysis is to fit the regression line. Mathematically least square estimation is used to minimize the unexplained residual.

When we fit a line through the scatter plot, the regression line represents the estimated job satisfaction for a given age. However the real observation might not fall exactly on the regression line. We try to explain the scatter plot with a linear equation of y = b0 + b1x. The distance between the regression line and the data point represents the unexplained variation, which is also called the residual ie- The method of least squares is used to minimize the residual.

Now, we evaluate the validity and usefulness of the equation. The key measure to the validity of the estimated linear line is R^2. R^2 = total variance / explained variance.

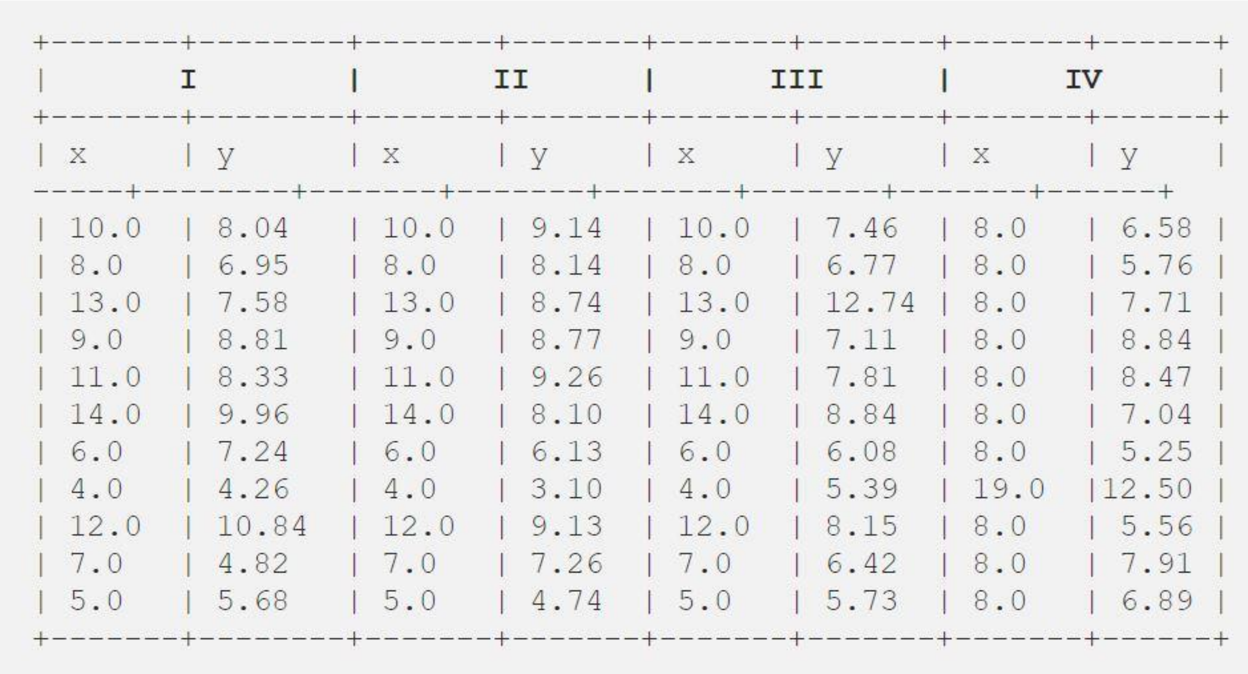
The number of observations and of course the number of independent variables increases the R^2. However, over-fitting occurs when the model is not efficient anymore. To identify whether the model is fitted efficiently a corrected R^2 is calculated. the larger the sample size the smaller the effect of an additional independent variable in the model.

The last step for the linear regression analysis is the test of significance. Linear regression uses two tests to test whether the found model and the estimated coefficients can be found in the general population the sample was drawn from. Firstly, the F-test tests the overall model. The null hypothesis is that the independent variables have no influence on the dependent variable.

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 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?

Ans> In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

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

Ans> 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 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.

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.

MinMax Scaling : 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).

Standardization : x = x – mean(x) / sd(x)

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?

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).

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

Ans> Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

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