Assignment-based Subjective Questions

Q 1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans) Using Linear Regression model the analysis of categorical variables through plotting graph, coefficients and VIF values obtained have shown that they have both positive and negative effect on the dependent variable.

For example

- the variable holiday has got negative coefficient value and has got high VIF value (>5)
- one of the dummy variables summer created for the categorical variable season has got positive coefficient and it has got low VIF value (<5)
- dummy variables Monday to Sunday have got mixed sign of coefficients and except for Saturday others have got inf (infinity) VIF values
- dummy variables (Clear, Mist+Cloudy, Light_Snow_Rain) created for the categorical variable weathersit (renamed to weather_condition) have got both positive coefficients and high VIF values (>5)

Q2) Why is it important to use drop_first=True during dummy variable creation?

Ans) It helps in reducing the extra column created during the dummy variables creation. Due to this it reduces the correlations created among dummy variables

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans) The variable temp has got highest correlation with the target variable

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans) It is validated by the following steps

- Predicting the bike rental count values on the test data set.
- Comparing the r2 score of the test data set with the final value of the R-Squared from the summary of the statsmodels.
- The result of R-Squared on training set is 0.727 and the result of R2 score on the test set is 0.723 which is very close to the one found on the training set.
- Finding the residuals between the actual output value and the predicted value. Residual = actual output - predicted output.
- The residuals are plotted and found that the distribution curve was normal, and the mean value was at 0

Q5) 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

1. season (dummy variables fall with coefficient value 0.321 and VIF value 1.72, summer with coefficient value 0.254 and VIF value 1.76 and winter with coefficient value 0.225 and VIF value1.62
2. yr(renamed to year) with coefficient value 0.248 and VIF value 1.94
3. weekday (dummy variable Saturday with coefficient value 0.056 and VIF value 1.58

General Subjective Questions

Q1) Explain the linear regression algorithm in detail

Ans

) Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

There are two main types:

1. Simple Regression
2. Multivariable Regression

1. Simple Regression – It uses the tradition slope-intercept form and has got the formula

   Y = mx + b

   Where,

   m is called the co-efficient and it determines the slope of the linear regression line. This value can either positive or negative

   b is the constant and it determines the intercept of the line. It is also called as the bias

   x is the input data(value)

   y represents the prediction

2. Multivariable Regression – Multivariable Regression makes use of multiple input data (values) to predict the output value. It is represented by the below formula

   y = w1x1 + w2x2 + w3x3 + …… wnxn

   Where

   x1 to xn – represents different input variables in the data set

   w1 to wn – represents the coefficients for the variables x1 to xn

   y – represents predicted output value(data)

   One example of this type of regression would be the prediction of sales with respect to the amount spent on the different advertising channels

   Sales = w1Radio + w2TV + w3NewsPaper

Associated with this algorithm there few terminologies

Sum Squared Error – It is the difference between actual observation and the predicted values

Mean Square Error – It represents the average of the sum squared errors. It is also called as mean squared deviation (MSD)

Gradient Descent – It is an iterative optimization algorithm for finding the local minimum of a function. To find the local minimum of a function using gradient descent, we must take steps proportional to the negative of the gradient (move away) from the function at the current point.

Gradient Descent was originally proposed by CAUCHY in 1847. It is also known as the steepest descent

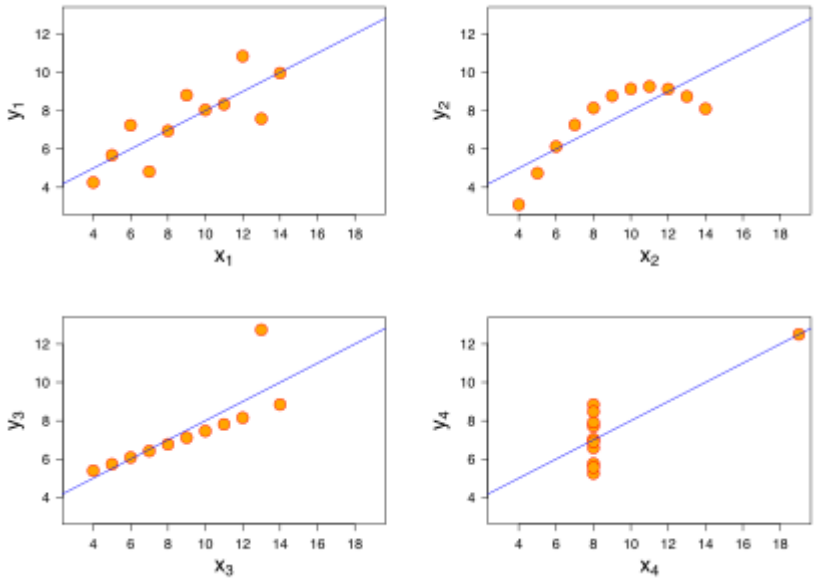
The linear regression algorithm model is evaluated based on the R-squared error, Adjusted – R squared error, p-values, F-function and VIF values

Q2) Explain the Anscombe's quartet in detail.

Ans) Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and 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 when analysing it, and the effect of outliers and other influential observations on statistical properties.

The plots are shown below



The first scatter plot (top left) shows that it is a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

The second graph (top right) shows that it is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.

The third graph (bottom left), where the distribution is linear, but should have a different regression line. Here the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

The last (fourth) graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Q3) What is Pearson's R?

Ans) It is s a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. Thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

It is named after Karl Pearson who formulated it during 1880s

It is the covariance of two variables divided by the product of their standard deviations.

The formula is

$pX,Y = COV(X,Y)/(STD(X) * (STD(Y))$

Where,

- COV – is the covariance
- STD(X) – is the standard deviation of X
- STD(Y) – is the standard deviation of Y

It is also called as Pearson product-movement correlation coefficient (PPMMC) or bivariate correlation.

An example would be, to find the correlation between the age of child with its height. The child's height increases with age.

There are certain requirements for this type of correlation

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should not be outliers in the data

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans) Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Difference between normalized scaling and standardized scaling is that in normalized scaling it rescales values between the range 0 to 1 whereas standardized scaling rescales the values in such a way that it will a mean of 0 and standard deviation of 1

The formula for normalized scaling is

(X-Xmin)/(Xmax-Xmin)

We can observe that when X becomes Xmax the equation results into a value of 1

The formula for standardized scaling is

(X – *u) /* standard deviation(X)

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans) This situation happens when here is a perfect correlation, then VIF = infinity. This tells a perfect correlation between two independent variables.

The value of R2 equal to 1 will cause this situation which leads 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

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Ans) It is called as Quantile-Quantile plot. It is a plot of two quantities that are opposite to each other. One example would be the mean which is a quantile where 50% of the data is below it and the other 50% is above it.

Quantile is explained as it is a fraction where certain values fall below that quantile. Its purpose is to find out whether two sets of data come from the same distribution or not.

To find the relation first a 45 degree is plotted on the Q_Q plot. When it is plotted the pointed will fall on that reference line, if the two data sets belong to a common distribution. The points in this type of plot lie on the line which is told as y = x when the two distributions to be compared are similar.

The Q-Q plot is used to compare the shapes of the distributions where it provides the graphical representation of how few properties such as skewness, location, scale are different or similar in two distributions.