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: The effect on the dependent variable by analysing the categorical variables from the dataset, we can infer that:

- The Fall season has the highest demand of shared bikes with a median of about 5500, followed by the summer season
- The demand for the bike shares has increased with the upcoming year 2019 compared to 2018 as more people are getting aware of the shared bikes concept.
- The months June-Sept has comparatively high shared bikes demand compared to other months as this fall under the fall season.
- The demand for the bike rentals seems to be more during no-holiday days than during holidays.
- The demand for the bike shares seems to be comparatively very high when the weather situation is Clear or partly cloudy, whereas the demand is very low when there is light snow or light rain.

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

Ans: It is advised to drop one dummy variable as it helps in reducing the extra column that will be created during the dummy variable creation. Each dummy variable will be highly correlated with the rest, using all the dummy variables for creating the model leads to a dummy variable trap.

Example: As in our case we have created the dummy for seasons and dropped one season 'Spring' because if the season is not summer, fall and winter then it is definitely Spring. Hence, we drop one dummy variable.

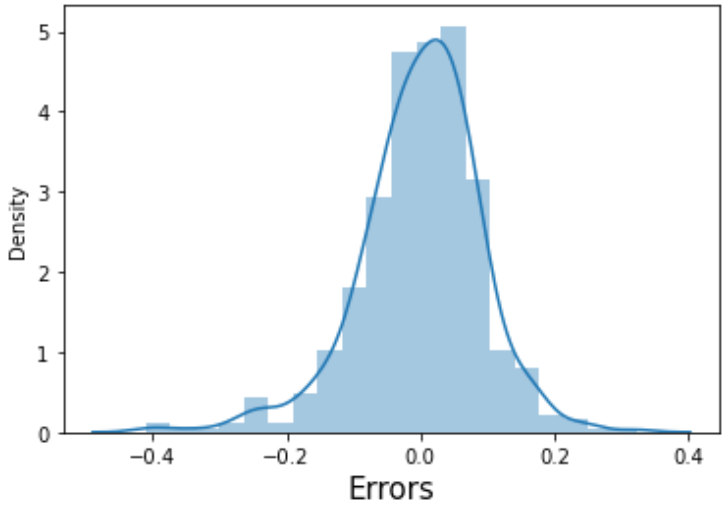
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: By looking at the pair-plot for the numerical columns we can conclude that 'temp' & 'atemp' seems to have highest correlation with the target variable 'cnt'.

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 by marking a distribution plot of the residual terms of the train dataset to check if it follows a normal distribution and their mean is centred to zero. We can even check that the error terms are distributed randomly and not following ani pattern by plotting a scatter plot.

Example:



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

Ans: The top predictors for our model which are affecting the bike sharing are

- **Temperature**: A Coeff. value of 0.5649, indicates a unit increase in temperature increases the bike shares by 0.5649 units
- **Weather Sit.** (Light Snow): A Coeff. value of (-0.2630), indicates a unit increase in weather situation decreases the bike shares by 0.2630 units
- **Year**: A Coeff. value of 0.2345, indicates that a unit increase in year variable increases the bike shares by 0.2345 units

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans: Linear regression is a supervised learning method that is used by the Train dataset which finds a linear equation that describes the correlation of the independent variables with the dependent variable. This is achieved by fitting a line to the data using ordinary least squares method. This line tries to minimize the sum of squares of the residuals. The residual is the distance between the line and the actual value of the independent variable. Finding the line of best fit is an iterative process.

The Linear regression equation looks like:

$$Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots\ldots.$$

Here, y is the dependent variable and x_1,x_2…are the independent variable, whereas β denotes the coefficients of the independent variables. The sign of the coefficients denotes whether it is positively or negatively correlated with the dependent variable.

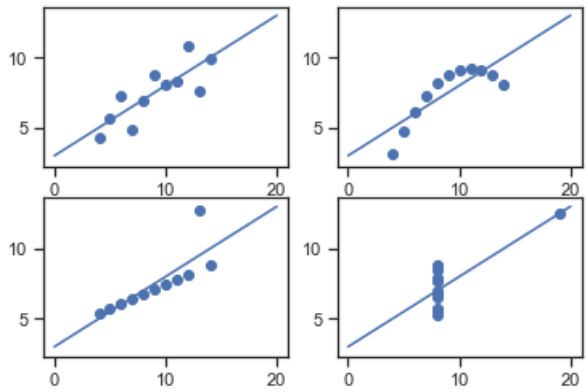
For example: The linear regression equation for the boom shared bikes looks like:

cnt = 0.108067 + (yr * 0.234515) - (holiday * 0.071614) + (temp * 0.564971) - (windspeed * 0.150211) + (summer * 0.084870) + (winter * 0.134806) + (sep * 0.082420) - (light snow * 0.263087) - (mist * 0.076336)

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet demonstrates the importance of data visualization which was developed by Francis Anscombe to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each dataset consists of eleven(x,y) points. The basic thing to analyse about these datasets is that they all share the same descriptive statistics but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

Example:



- First image consists of a set of (x,y) points that represent a linear relationship with some variance
- Second image shows a curve shape but doesn't show a linear relationship
- Third image looks like a tight linear relationship between x and y, except for one large outlier.
- Fourth pic looks like the value of x remains constant, except for one outlier.

3.What is Pearson's R?

Ans: Pearson's R or which is also referred as Pearson's correlation coefficient is a statistic that measures the linear correlation between two variables. Like all the correlations, it also has a numerical value that lies between -1.0 and +1.0.

Pearson's R cannot understand the nonlinear relationships between two variables and cannot differentiate between dependent and independent variable. Pearsons'R is the covariance of the two variables divided by the product of their standard deviations. Pearsons'R is a numerical summary of the strength of the linear association of the variable. If the variable tends to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Ans: Scaling is the process of scaling the values of features in a dataset so that they proportionally contribute to the distance calculation i.e., to normalize the data within a particular range.

The collected data set contains features highly varying in magnitudes, unit and range. If the scaling is not done then algorithm only takes magnitude into account and not the units and possesses a threat to the model. Hence, we use scaling to bring all the variables to the same

level of magnitude. The two most commonly used scaling techniques are Standardization and MinMax Scaling.

Differences between normalized scaling and standardized scaling:

- In **Normalization scaling** it brings all of the data in the range of 0 to 1. Sklearn.preprocessing.MinMaxSaler helps to implement the normalization in python.

  The formula for MinMaxScaling: $x = \frac{x - \min(x)}{\max(x) - \min(x)}$

- In Standardization scaling it replaces the values by their z-scores. It brings all of the data into a standard normal distribution which has a mean – '0' and standard deviation – '1'. Sklearn.preprocessing. scale helps to implement standardization in python. One biggest disadvantage of normalization over standardization is that it loses some information in the data , especially about outliers.

  Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

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

Ans: The variance inflation factor(VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for checking the multicollinearity.

Sometimes the value of VIF is infinite this means there is a perfect correlation between the two independent variables which is displaying VIF as infinite. In the case of perfect correlation we get the $R^2 = 1$, which means that $1/(1-R^2)$ as infinity. In order to solve this issue, we can drop one of the variables from the dataset which is showing the vif as infinite.

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

Ans: A Quantile-Quantile or (Q-Q) plot are plots of two quantiles against each other. A quantile is a fraction where certain values fall below a particular quantile. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing the model against. Also, it helps to determine if two data sets come from populations with a common distribution.

It used to check:

- if the two datasets come from populations with a common distribution.
- If the two datasets have common scale
- If the two datasets have similar distributional shapes

In a Q-Q plot a 45-degree reference line is plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the distance from this reference line, the greater the

evidence for the conclusion that the two data sets have come from populations with different distributions.

Advantages of Q-Q plot: The sample sizes need not be equal to plot a Q-Q plot & many distributional aspects can be simultaneously tested.

By,

Gubbala Venkata Hemanth

DSC-48.