BOOMBIKES LR ASSIGNMENT

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?

From the dataset it is understood that the following are the categorical variables: Categorical - season, yr, mnth, holiday, weekday, workingday, weathersit.

- Season: The demand for bikes is low during spring and considerably high in fall (1: spring, 2: summer, 3: fall, 4: winter).
- Year: The bike selling demand has a very good incline in 2019 when compared to 2018.
- Holiday: The counter is a little higher on a regular day than on a holiday (0: holiday, 1: regular day).
- Weekday: Bikes are rented almost equally throughout the week.
- Weather situation: Of the 4 categories given in the dataset, category 1 has the highest demand and category 4 has no statistics to present.
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

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

This parameter is used when we must avoid retaining redundant features in the dataset. For example, in the training module, we learnt that the furnishing status can be split into furnished, semi-furnished and unfurnished. However, when we know the status/value of 2 variables, we can automatically determine the 3rd variable's value. Based on the binary values, if furnished and semi-furnished are 0, it is understood that unfurnished is 1 if that column is not retained in the dataset. This way, we can drop a few redundant columns and concise our dataset.

Therefore, if we have 'n' categorical variables in the dataset, 'n-1' columns should be used to represent the dummy variables.

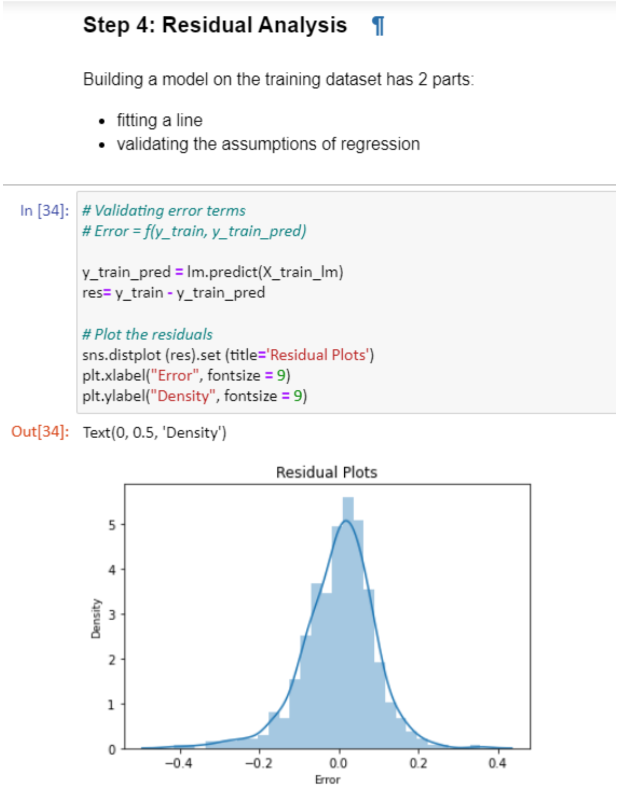
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

From the dataset it is understood that the following are numerical variables: Numerical - instant, temp, atemp, hum, windspeed, casual, registered and cnt. Out of these variables, we have dropped instant, casual, registered and atemp.

Comparing 'cnt' with temperature, humidity and windspeed, we can see a (almost) linear plot with a high correlation with **temperature**.

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

We plotted a residual analysis plot with ‘error’ on x-axis and ‘density’ on y-axis which helps us understand the distribution of the error terms (y_train and y_train_pred). The distribution should be such that it is centered around zero and it is approximately normally distributed. This is seen in the below histogram.



Step 4: Residual Analysis ¶

Building a model on the training dataset has 2 parts:

- fitting a line
- validating the assumptions of regression

In [34]:
```
# Validating error terms
# Error = f(y_train, y_train_pred)

y_train_pred = lm.predict(X_train_lm)
res= y_train - y_train_pred

# Plot the residuals
sns.distplot (res).set (title='Residual Plots')
plt.xlabel("Error", fontsize = 9)
plt.ylabel("Density", fontsize = 9)
```

Out[34]:
```
Text(0, 0.5, 'Density')
```

Also, we can check the R-squared value variations based on p-values or VIF values by adding or dropping variables. This is performed to avoid multicollinearity.

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

Based on final model, we can conclude that the variables contributing significantly towards explaining the demand of the shared bikes by looking at their coefficients are **year, temperature and weathersit_Light Snow.**

General Subjective Questions:

1. Explain the linear regression algorithm in detail

Linear Regression is a supervised machine learning technique with a continuous and constant slope as a predicted projected output. Rather than aiming to classify data into categories, it's used to predict values within a continuous range. Linear regression is mainly categorized as

a. **Simple Linear Regression:**

   This technique uses the most traditional concept where we have 'b' as the slope and 'c' as the intercept (x: input variable and y: prediction variable):

$$\mathbf{Y = mx + c}$$

   We will predict the model using just 1 independent variable

b. **Multiple Linear Regression:**

   This technique uses multiple independent variables against 1 predictor variable. By adding more variable from the dataset, we will be able to obtain more information regarding the variance in 'y'. By adding more variables, the variance will either remain the same or increase, but never decrease.

Either of the methods have the following steps to finally obtain the best fit model.

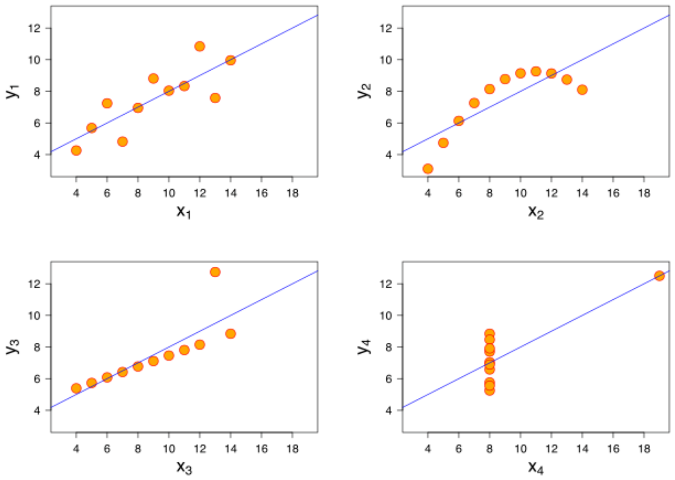
1. Reading and understanding the data
2. Preparing the data for modelling
3. Training the model
4. Residual Analysis
5. Predictions and evaluation on the test set

Assumption of Regression Model

- **Linearity**: The relationship between dependent and independent variables should be linear.
- **Homoscedasticity**: Constant variance of the errors should be maintained.
- **Multivariate normality**: Multiple Regression assumes that the residuals are normally distributed.
- **Lack of Multicollinearity**: It is assumed that there is little or no multicollinearity in the data.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
These provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.



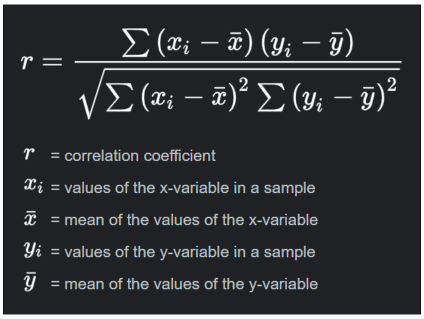
- The first scatter plot (top left) appears to be a simple linear relationship
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line. We can see 1 outlier which can have an effect on correlation.
- Finally, the 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.

3. What is Pearson's R?

There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R)

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

The formula reads as follows:



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.

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

Scaling is a data pre-processing technique that is applied on independent variables in order to normalise the data within a given range. It also aids in the acceleration of algorithmic calculations.

Most of the time, the acquired data set contains features with a wide range of magnitudes, units, and ranges. If scaling is not done, the algorithm will only consider magnitude rather than units, resulting in erroneous modelling. To solve this problem, we must scale all of the variables to the same magnitude level.

Scaling only changes the coefficients and not the other parameters such as the t-statistic, F-statistic, p-values, R-squared, and so on.

a. 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.

$$\boldsymbol{MinMax\ Scaling} : \boldsymbol{x} = \frac{\boldsymbol{x} - \boldsymbol{min}(\boldsymbol{x})}{\boldsymbol{max}(\boldsymbol{x}) - \boldsymbol{min}(\boldsymbol{x})}$$

b. 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).

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

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?

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 R^2 =1, which lead to 1/(1-R^2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

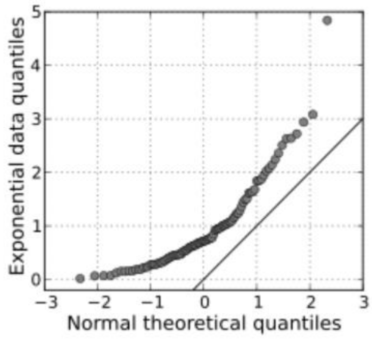
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity

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 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.

A Q-Q plot showing the 45-degree 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.