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?

- Summer and fall are the best times for people to rent bikes.
- September and October have the highest bike rental rates.
- More bicycles are leased on Saturdays, Wednesdays, and Thursdays.
- The majority of bike rentals occur when the weather is clear.
- 2019 saw an increase in bike rentals.
- Whether it is a working day or not, there is no significant impact on the cost of renting a bike.
- On holidays, bike rental rates are higher.

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

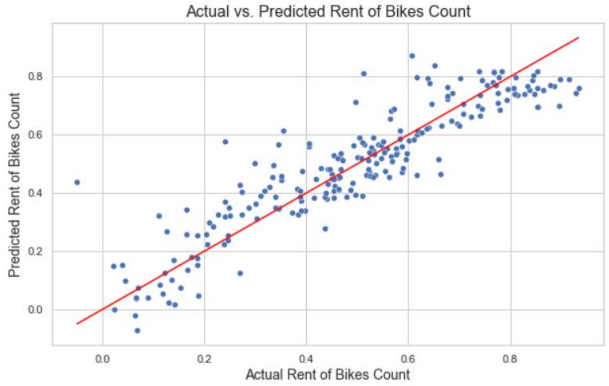
- When constructing dummy variables, drop_first = True is used to remove the base/reference category. This is done in order to prevent the model from becoming multi-collinear if all dummy variables are used. When all of the other dummy variables in a given category are equal to 0, it is simple to determine which category is the reference category.

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

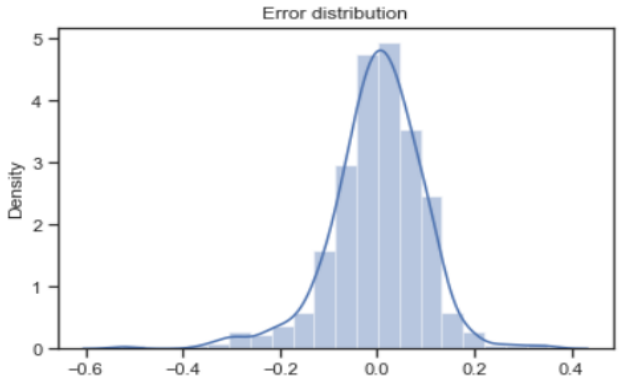
- The variable "temp" has a 0.63 correlation with the target variable, making it the most correlated variable.
- Since "atemp" is produced from temperature, humidity, and windspeed during model development, it is not taken into account.
- Since the values of these columns add up to the target variable, the casual and registered variables are essentially a component of that variable, hence their correlation is ignored.

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

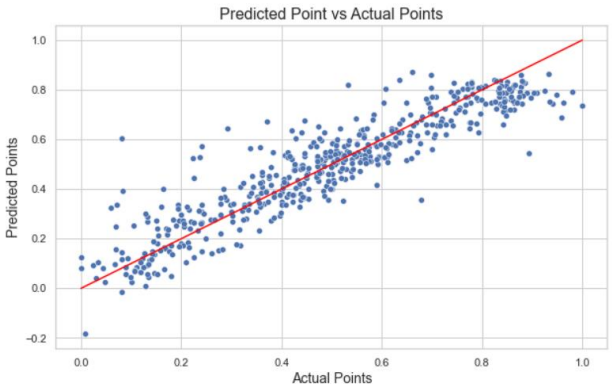
- Linear relationship between independent and dependent variables



- Error terms are independent of each other
- Error terms are normally distributed



- Error terms have constant variance (homoscedasticity)



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

- Temperature
- Year
- Holiday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a Machine Learning algorithm used for supervised learning. It assists in forecasting a dependent variable (goal) using the provided independent variable(s). Regression techniques often establish a linear relationship between a dependent variable and the other independent variables. Simple linear regression and multiple linear regression are the two different types of linear regression. When a single independent variable is used to forecast the value of the target variable, simple linear regression is performed.

When several independent factors are used to forecast the numerical value of the target variable, this is known as multiple linear regression. Regression lines are linear graphs that depict the connection between dependent and independent variables. When both the dependent and independent variables are on the X-axis, there is a positive linear connection. It is a negative linear relationship, though, if the value of the dependent variable falls as the value of the independent variable rises on the X-axis.

2. Explain Anscombe's quartet in detail.

Four data sets make up Anscombe's quartet, which have essentially similar simple descriptive statistics but radically diverse distributions and visual appearances. There are eleven points in every dataset. The main goal of Anscombe's quartet is to emphasize the significance of visualizing a group of data before starting an analysis process because statistics simply cannot accurately describe two datasets being compared.

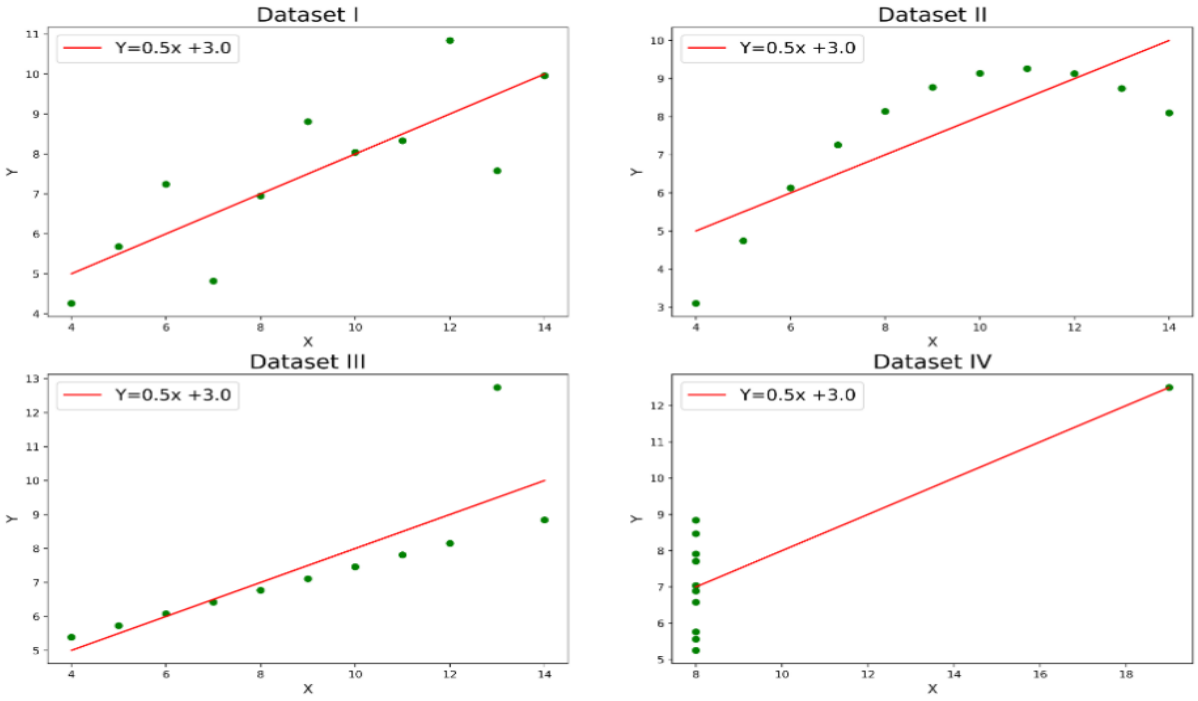
The four datasets of Anscombe's quartet.

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

Following is the statistical summary for the above dataset:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

And below is the plot for all 4 datasets:



Explanation of the above graph:

- In the first one(top left) if we look at the scatter plot we will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if we look at this figure we can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) we can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
- Finally, the fourth one(bottom right) shows an example of when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

A linear link between two quantities is established using Pearson's correlation coefficient. 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.

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

- Scaling is a pre-processing method used to standardize the independent feature variables in the dataset within a predetermined range.
- The dataset may contain a number of features that range widely in high magnitudes and units. There will be some discrepancy in the units of all the characteristics included in the model if scaling is not done on this data, which results in erroneous modeling.
- Standardization replaces the values with their Z scores, whereas Normalisation places all the data points in a range between 0 and 1.

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

When the two independent variables are perfectly correlated, VIF has an infinite value. In this instance, the R-squared value is 1. Given that VIF is equal to 1/(1-R2), this results in VIF infinity. According to this idea, multi-collinearity is an issue, and one of these variables must be eliminated in order to create a useful regression model.

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

To assess whether a dataset in question follows a certain distribution, such as a normal, uniform, or exponential distribution, the quantile-quantile (Q-Q) plot is used to plot quantiles of a sample distribution with a theoretical distribution. It enables us to determine whether the distribution of two datasets is the same. It is also useful to determine whether or not the errors in the dataset are typical.

***** END *****