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: From my analysis of the effect of Categorical variable on the dependent variable is as follows,

- yr (year variable): Count of bike rentals increased and become popular in 2019 as compared to 2018.
- weathersit (weather variable): The count of bike rental is more during clear weather. The demand is low during light snowy weather.
- Season variable: There is a peak demand in fall and summer season, and they are more favourable for bike rentals. The demand is very low in spring season.
- Weekday: The demand is less on Sunday.
- Mnth (Month variable): The demand is high in June, July, August, and demand will reduce after that.

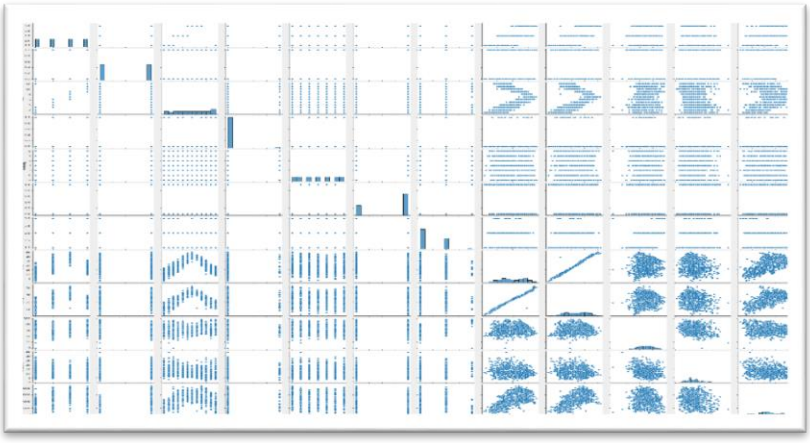
2. **Why is it important to use drop_first=True during dummy variable creation?**

Ans: we need to drop first column during dummy variable creation because:

- If we don't drop, the dummy variables will be correlated and effects the model adversely and the effect is stronger.
- We need to drop first column to avoid redundant features.

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

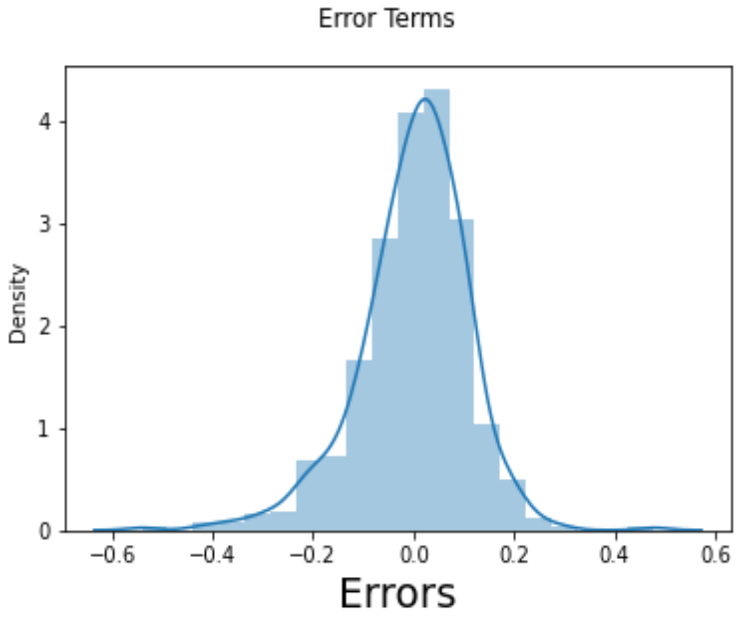
Ans: By observing at the pair-plot among the numerical variables cnt(Target variable) and temp(Temperature) has significantly high correlation as compared to the other variables i.e., 0.63.



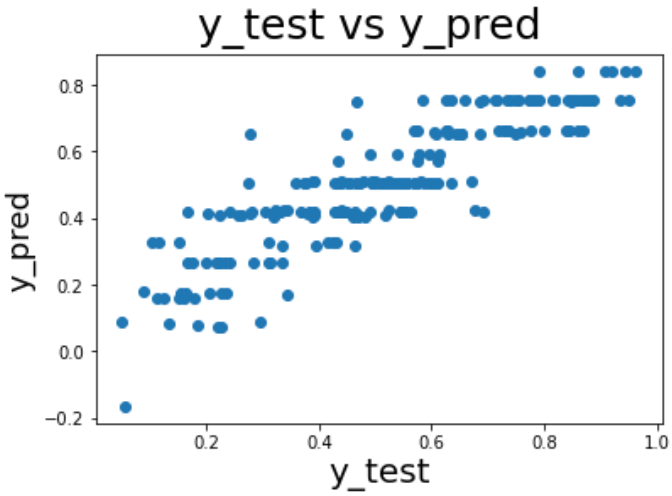
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: The assumptions of Linear Regression after building the model on the training set are as follows:

- Residual Errors follows Normal distribution, and this was validated after plot distribution.



- Maintains Linear Relation between the Dependant variable (test and predicted)
- The p-value of the selected variables were found to be lesser than 0.05 and this shows the variables are significant.



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 that are significant variables in identifying the Target variable:

- **Year**: 2019 year has more demand for bikes rentals 0.2459.
- **Season_spring:** -0.2435 demand of shared bikes more because of season climate condition spring**.**
- **Weather_light snow:** -0.3264 demand of the bike rentals is more because of weather condition.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Ans: Linear Regression is one of the algorithms in the Machine Learning world which comes under supervised learning. Regression models predict a dependent (target) value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Mathematically, we can write a linear regression equation as:

Y = ax+b

Here x and y are two variables on the regression line.

a=slope of the line.

b=y-intercept of the line

x=Independent variable from dataset

y=Dependent variable from dataset.

2. Explain the Anscombe's quartet in detail.

- 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.
- Each graph has different statistics when we look into it.

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

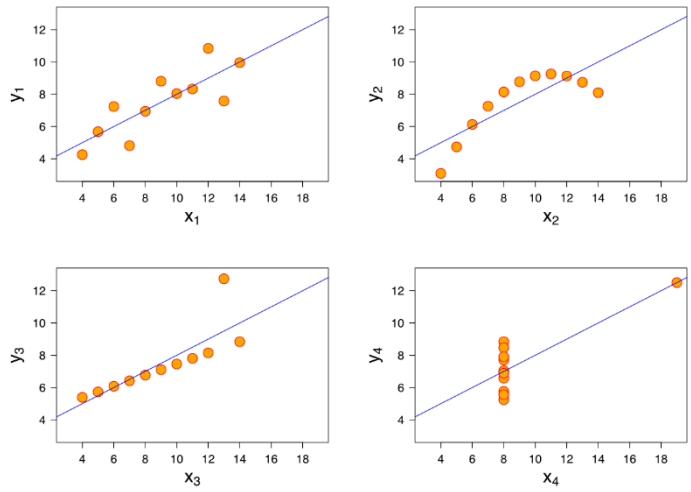
Quartet's Summary Stats:

Mean of x is 9 and mean of y is 7.50 for data set.

Sample variance of x is 11 Sample variance of y is 4.13 for each dataset.

Correlation between x and y is 0.816.

We plot these four datasets on x/y coordinate plane, Lets see

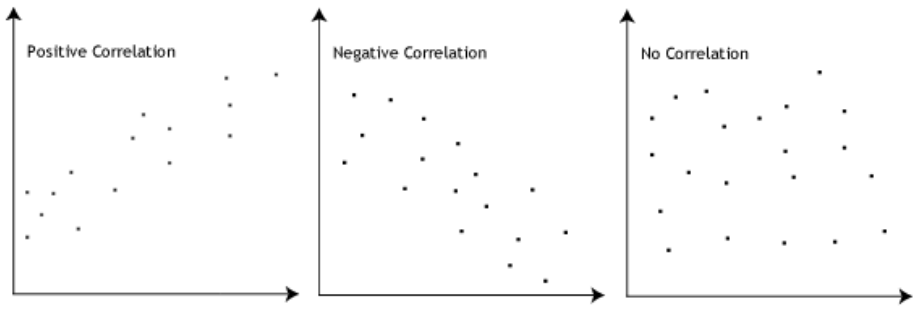


- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Ans: The Pearson correlation coefficient (PCC), also referred to as Pearson's r,

- the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data.
- It is the covariance of two variables, divided by 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.



The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction).
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions).
- r = 0 means there is no linear association.

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases and vie verse.

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. Scaling 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:

- Minimum and maximum value of features are used for scaling
- Scales values between 0 and1.
- Scikit-Learn provides a transformer called MinMaxscalar for Normalization.
- It is really affected by outliers,

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

Standardized 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**).

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

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

Ans: VIF = Variance Inflation Factors

Variance inflation factor measures how much the variance of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables.

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where 'i' refers to ith variable

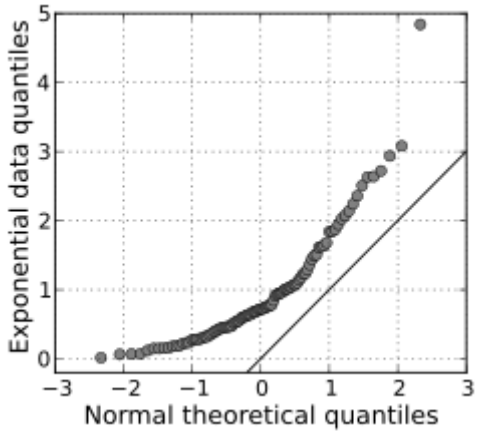
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. This shows a perfect correlation between two independent variables.

In the case of perfect correlation we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

A Q Q plot shown below:



- 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.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.