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?**

As per our final Model, 'Weather Situation 3 (weathersit_3)' is the categorical variable having most influence on bikes booking with a coefficient value of -0.28. According to Data Dictionary, weathersit_3 refers to Light Rain/Snow. Since the coefficient value is having negative sign, it means that more the amount of rainfall or snow, less will be the sales of bikes. Apart from this, other categorical variables like month_9 and season_4 has coefficient values of 0.11 and 0.13 respectively. Since this are positive values so for the month of September and Winter season there can be a rise in sales.

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

We always use n-1 variables for 'n' number of levels during dummy variable creation. Let us take an example of Month column, while dummifying this variable there will be 12 new variables each denoting a single month (Month_1 is Jan, Moth_2 is Feb and so on) having 1 as True and 0 as False. Thus, instead of having 12 variables for this, we can have n-1 i.e. 11 variables. In such case, A row having all 0 values will denote the 1 for the dropped variable, hence minimizing the use of an extra column, so we use drop_first=True.

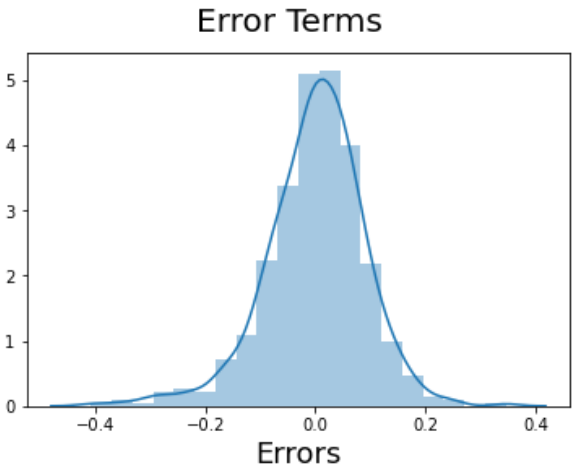
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

According to the pair-plot, 'temp' numerical variable has the highest correlation with 'cnt' target variable with correlation value as 0.63. Looking at the pair plot, in comparison with other it is more linear and positive.

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

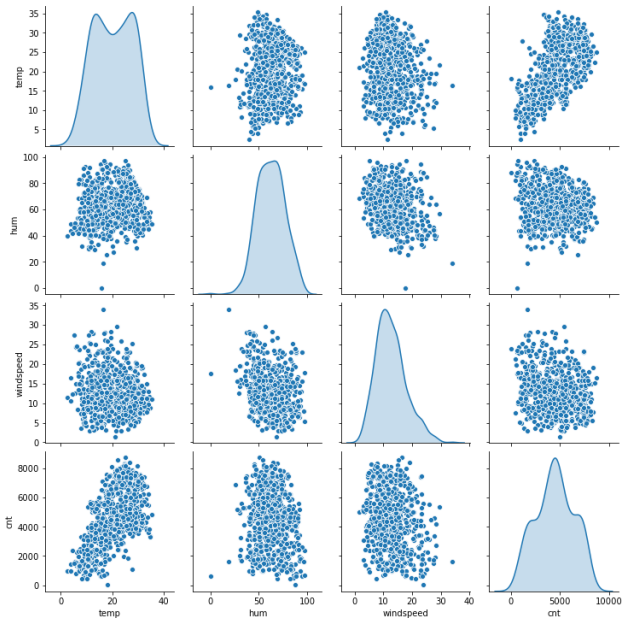
Validated the four assumptions of Linear Regression Models:

1. Normality: I have plotted the errors which is the difference of y_train_actual to y_train_predicted. For this error, I have done a distplot which gives me a histogram.



From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

2. Linearity: I have done a pair plotting between X predictor variables and y target variable:



As it is seen that, there is a linear relationship between 'temp' predictor variable and 'cnt' target variable. Hence our assumption for Linear Regression is valid.

3. Multicollinearity: From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5. Hence our assumption for Linear Regression is valid.

4. Homoscedasticity: From our observations, I can see that the variance of residual is the same for any value of X. Hence our assumption for Linear Regression is valid.

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

Based on the final model, the predicted top 3 features which contributes significantly towards explaining the demands on the shared bikes are **'temp', 'weathersit_3', 'yr'**.
1. **'temp'** indicates the temperature in Celsius and it has a positive coefficient value of **0.51** which implies that with each unit of increase in temperature can increase the demand of bikes by 0.51 units.
2. **'weathersit_3'** indicates Light Rainfall/Snow and it has a coefficient value of **-0.28**. As we can see that it has a negative sign, thus it implies that with increase in weathersit_3 by single unit there will be a 0.28 decrease in demand.
3. **'yr'** has values 0 and 1 where 0 denoted 2018 and 1 denoted 2019. It has a coefficient of 0.23. This implies that with every upcoming year, the demand for bikes rises to 0.23 units.
For all 3 parameters when we say, there is an increase/decrease in its unit, we assume that other parameters are kept constant.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables, it is refereed as Multiple Linear Regression.**

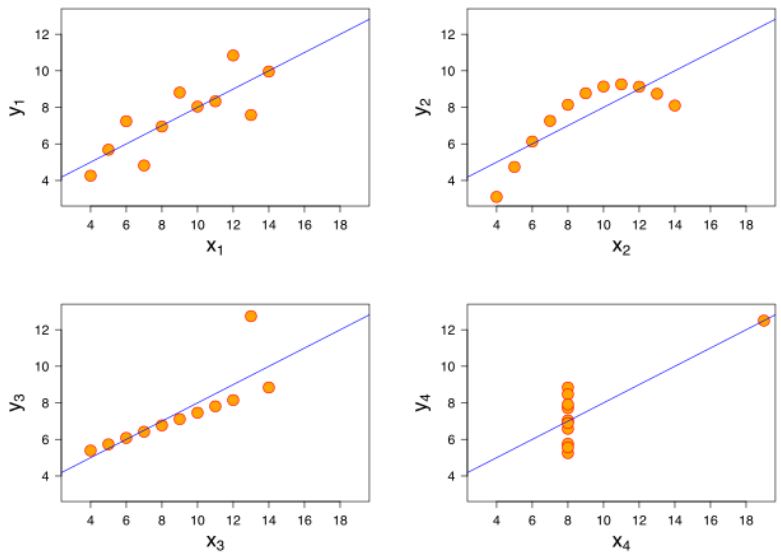
The equation for Linear Regression is:
y = B0 + B1*x + B2*x2 + …+ Bn*xn
where y is the target variable and X are the input parameters.
Here B1, B2,.. Bn are the coefficients and B0 is the intercept, which means the value of y when coefficients are 0.

It is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.

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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. To break the assumptions that numerical calculations are exact, but graphs are rough, he created this concept.

Below mentioned is the graphical representation of all 4 data sets:



As we can see that all 4 graphs follow a different trend, which means that even if by assumptions and numerical calculations if the data seems to be similar, it is better approach to visualize it.

3. **What is Pearson's R?**

In Statistics, Pearson's *R* also known as Pearson correlation coefficient or the bivariate correlation is a statistic that measures linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It was developed by Karl Pearson. By definition, it is the covariance of the two variables divided by the product of their standard deviations.

Formula for a Pearson coefficient is:

P X, Y = covariance (X, Y)/sigma(X). sigma(Y)

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables.

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

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. Since the range of values of raw data is spread widely, objective functions do not work properly, So Scaling is used.

ML algorithms uses gradient descent as an optimization technique which requires data to be scaled.

To understand the difference between Normalized scaling and standardized scaling, mentioning the explanation of the two:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

X' = X-Xmin/Xmax-Xmin

Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

X'=(X-mean)/standard deviation.

Difference between Normalization and Standardization is that Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution whereas Standardization is better approach when we know that distribution of our data does follow a Gaussian distribution. Also, unlike normalization,

standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

To address this question, let us see the formula of VIF, which is given by:

VIF = 1 / (1 – R2)

We already know that R2 is the coefficient of determination of the overall model, its values range from 0 to 1, higher the value of R squared denotes that regression model perfectly fits the data and it means the feature is correlated with other features.

VIF is the measure to check the multicollinearity between different features.

According to this formula, when R2 is 1, VIF will become infinite. So, an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which will show an infinite VIF as well).

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

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus, the line is a parametric curve with the parameter which is the number of the interval for the quantile.

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.

QQ plots are used to fit a linear regression model, check if the points lie approximately on the line, and if they don't, residuals aren't Gaussian and thus errors aren't either. This implies that for small sample sizes, you can't assume that estimator is Gaussian either, so the standard confidence intervals and significance tests are invalid.