Bike Sharing Demand Prediction Assignment
Subjective Questions

Assignment-based Subjective Questions

Question 1:

•From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

>The categorical variables from the dataset and their effect on the dependent variable are as follows:
1. season - Spring has least demand, while Fall, Summer and Winter has high demands (in same order).
2. year - There is a very high year on year increase in demand.
3. month - May through October has very good demand while January has the lowest demand.
4. holiday - On holidays, the demands are less but variance is more.
5. weathersit - Clear and moderate (mist and cloudy) weathers, the demands are good. The demand is least in Light Snowy and light rain weather. There are no demand on extreme (Heavy Snow & Rain) weather condition.
6. weekday - The demands have been approximately similar for all weekdays. There is only slight difference.
7. workingday - The demands have been similar for working and non-working day, though for non-working day the variance is more.

Question 2:

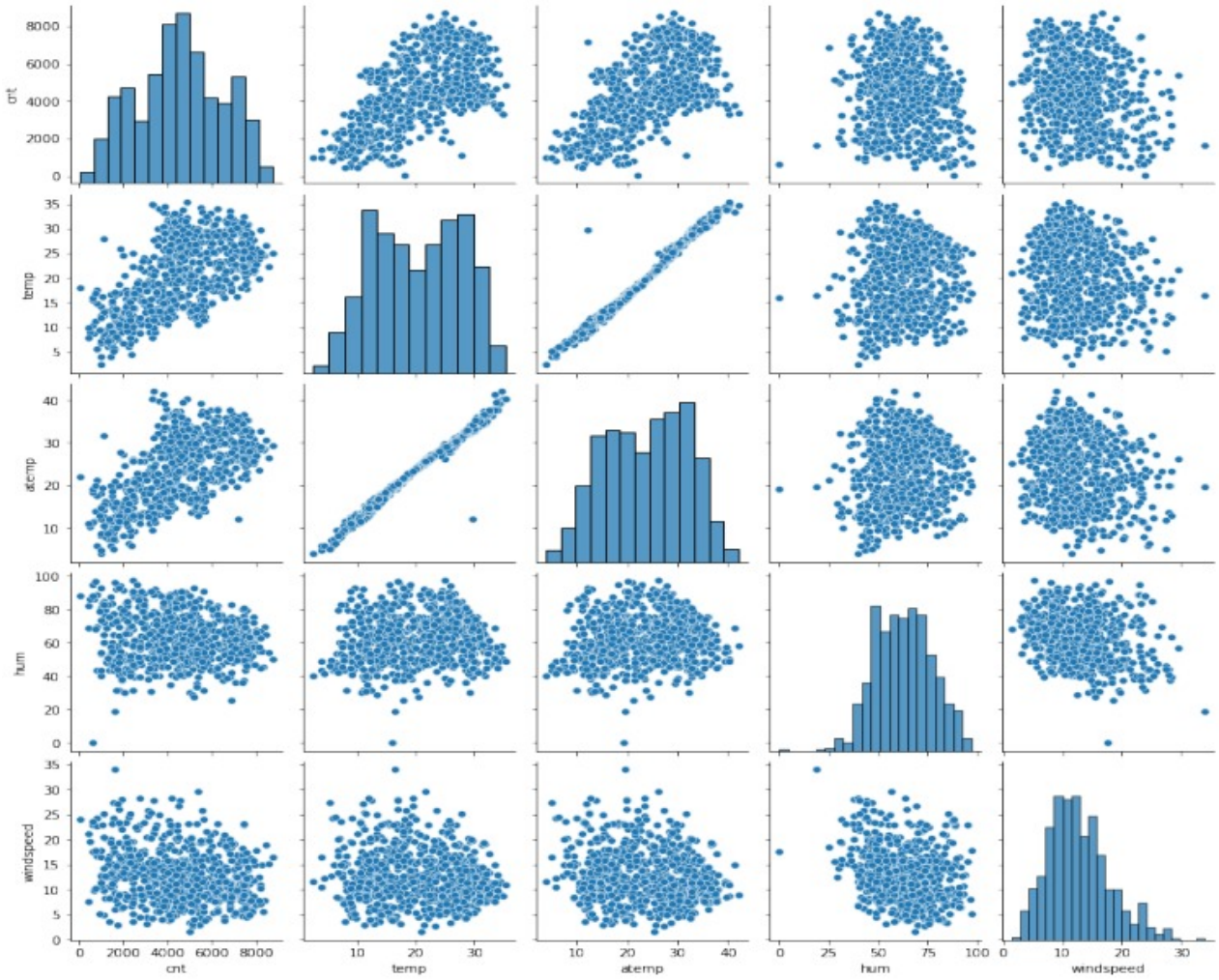
•Why is it important to use drop_first=True during dummy variable creation?

>Dummy variables are created from the categorical variable in the data set so the number of dummy variables created depends on the values or category of the variable. AND therefore, the original categorical variable is highly correlated with the newly created one dummy variable.
Number of dummy variables = number of categories – 1
Therefore, if we do not drop the original categorical variable, the model may be unfavorable affected due to multicollinearity and the effect will be stronger when the cardinality is smaller. Iterative models may have difficulty with convergence and feature selection or elimination be distorted. So we always drop the original categorical variable

Question 3:

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

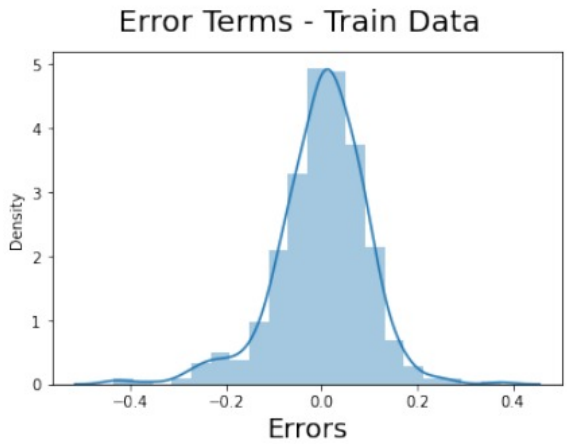
>Looking at the below pairplot, we find that temp and attempt are 2 variables which are highly correlated with target variable ‘cnt’.



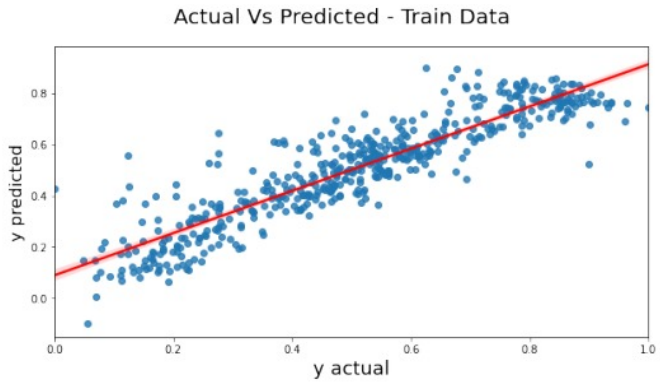
Question 4:

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

> One of the assumptions for a linear regression model is a residual distribution (error terms) are normally distributed and centered around 0. This can be verified by plotting a distribution plot of the distribution of residuals (or errors). The chart below shows that the residual distribution is normally distributed and centered around 0.0



Another assumption for linear regression is that the variance should not increase or decrease as the error changes and the variance should not follow any pattern with a change in error terms. Based on the graph below for the error terms, we can say that the variance does not increase or decrease and does not follow any pattern. It is homoscedastic in nature.



Question 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 3 features contributing significantly towards explaining the demand of shared bikes are as follows:
1. temp – (positive coefficient)
2. yr – (positive coefficient)
3. weathersit_Light_Snow_Rain - (Negative coefficient)

General Subjective Questions

Question 1:

•Explain the linear regression algorithm in detail.

>Linear regression is a machine learning algorithm based on supervised learning. It is performing regression task. Regression models the target predictive value based on the independents variables. It is mostly used to find out the relationship between variables and forecasting. Different regression models differ in the type of relationship between them dependent and independent variables, consider and the number of independents variables used.
Linear regression performs the task of predicting the value of the dependent variable(Y) based on a given independent variable (x). Thus, this regression technique detects a linear relationship between x (input) and y (output). Hence the name linear regression.
Linear regression is based on the popular equation: "y = mx + c"
It assumes that there is a linear relationship between the dependent variable(Y) and predictor(s)/independent variable(x). In the regression we calculate the best fit line which describes the relationship between the independent and dependent variable.
Regression is performed when the dependent variable is of continuous data type a
The predictors or independent variables can be of any data type, such as continuous, nominal/categorical etc. The regression method tries to find a line of best fit that shows the relationship between the dependent variable and the predictors with the smallest error.
In regression, the output/dependent variable is a function of the independent variable a coefficient and error term.
Regression is generally divided into simple linear regression and multiple linear regression.
1. Simple linear regression: SLR is used when the dependent variable is predicted only by one independent variable.
2. Multiple Linear Regression: MLR is used when the dependent variable is predicted by multiple independent variables.
The equation for MLR will be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \ldots n.$$

β1 = coefficient for variable X1
β2 = coefficient for variable X2
β3 = coefficient for variable X3 and so on... β0 is the intercept (constant term)

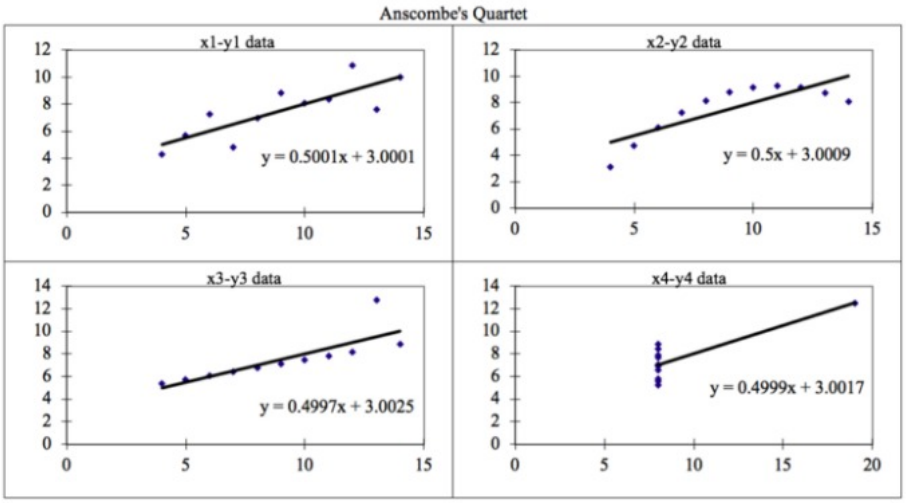
Question 2:

•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.
Anscombe's Quartet is the model example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties.
It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

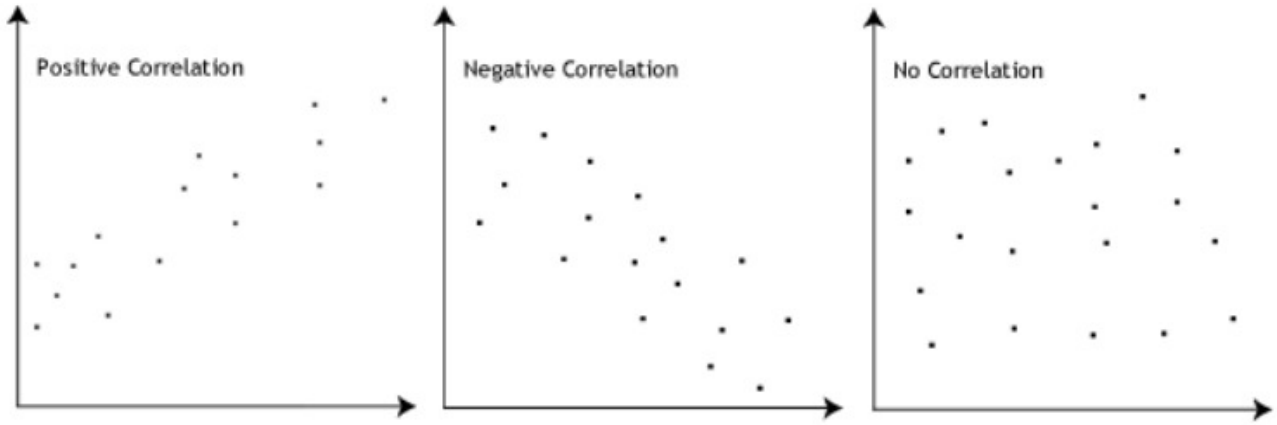
Each graph plot shows the different behaviour irrespective of statistical analysis. However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Question 3:
•What is Pearson's R?

> In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. 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.
r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association



Whenever we discuss correlation in statistics, it is generally the Pearson correlation coefficient. However, it cannot capture non-linear relationships between two variables and cannot distinguish between dependent and independent variables.
Pearson's correlation coefficient is the covariance of two variables divided by the product of their standard deviations. The definition form includes "product moment", i.e. the average (first moment about the origin) of the product of mean modified random variables; hence the product-moment modifier in the name.
Pearson's correlation coefficient is named after Karl Pearson. He formulated a correlation coefficient from a related idea by Francis Galton in the 1880s

Question 4:

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

> Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It 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 is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks. It brings all of the data in the range of 0 and 1. In sklearn.preprocessing, MinMaxScaler helps to implement normalization in python.

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

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. 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. 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). sklearn.preprocessing.scale helps to implement standardization in python

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

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Question 5:

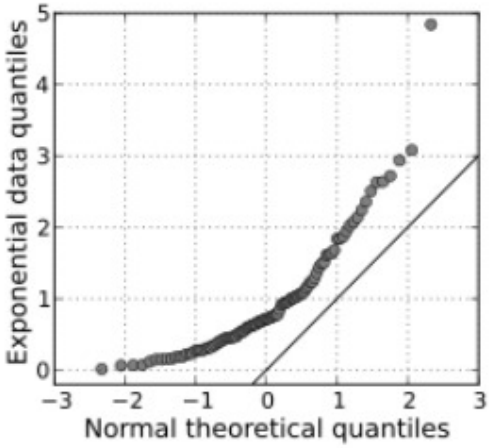
•You might have observed that sometimes the value of VIF is infinite. Why does this happen?

> VIF - the variance inflation factor - VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.
(VIF) =1/(1-R_1^2).
If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results"infinity

Question 6:

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

> Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
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



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a normal quantile-quantile (Q-Q) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.
Few advantages:
• It can be used with sample sizes also
• Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets -
1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behaviour