Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Technically all variables provided in the dataset were numeric in nature except for the date variable. However, most of them were not continuous variables. Meaning those variables could acquire only certain values and not any possible numeric variables. Variables that are used to describe something are referred as categorical variables. In the available dataset, following were the categorical variables:-

Dteday	Season	Yr	Mnth
Holiday	Weekday	Workingday	weathersit

Most of these variables have direct relationship with the dependent variable (cnt). Interesting point to note is a specific value of these categorical variable can have higher or a different impact on the target variable as compared with other values of the same variable.

For example, based on the heatmap in the assignment; it can be inferred that most of the months have small corrleation with cnt but border months of year (November, December, January, February) have a negative relation with cnt.

Note: Since to create a model all the categorical variables have to be converted to numeric categorical variables, we have not changed the table values to the actual category and then again changed them to number using dummy variable method. Rather we have directly used dummy variable method on the numeric values and renamed the columns to the correct category.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Dummy variables are created to convert a categorical variable to numeric Boolean of 0 and 1.

Pandas take each value under the categorical variable and assign it to a column in a dataframe. It names the column on the value of the cell. At a particular cell, the values is assigned as 0 when the actual value of the cell does not match the column header and 1 when the actual value of the cell matches the column header.

Hence, Pandas by default creates ‘N’ columns if there are ‘N’ unique values of a variable.

Now, if we see all the cells can either have a value of 0 or 1 as described above. Also, if one column is True (1) for a given cell, no other column can be true (0) for that value. Hence, out of all the columns created there will only be 1 column to have a value of 1 and rest of the columns will have a value of 0.

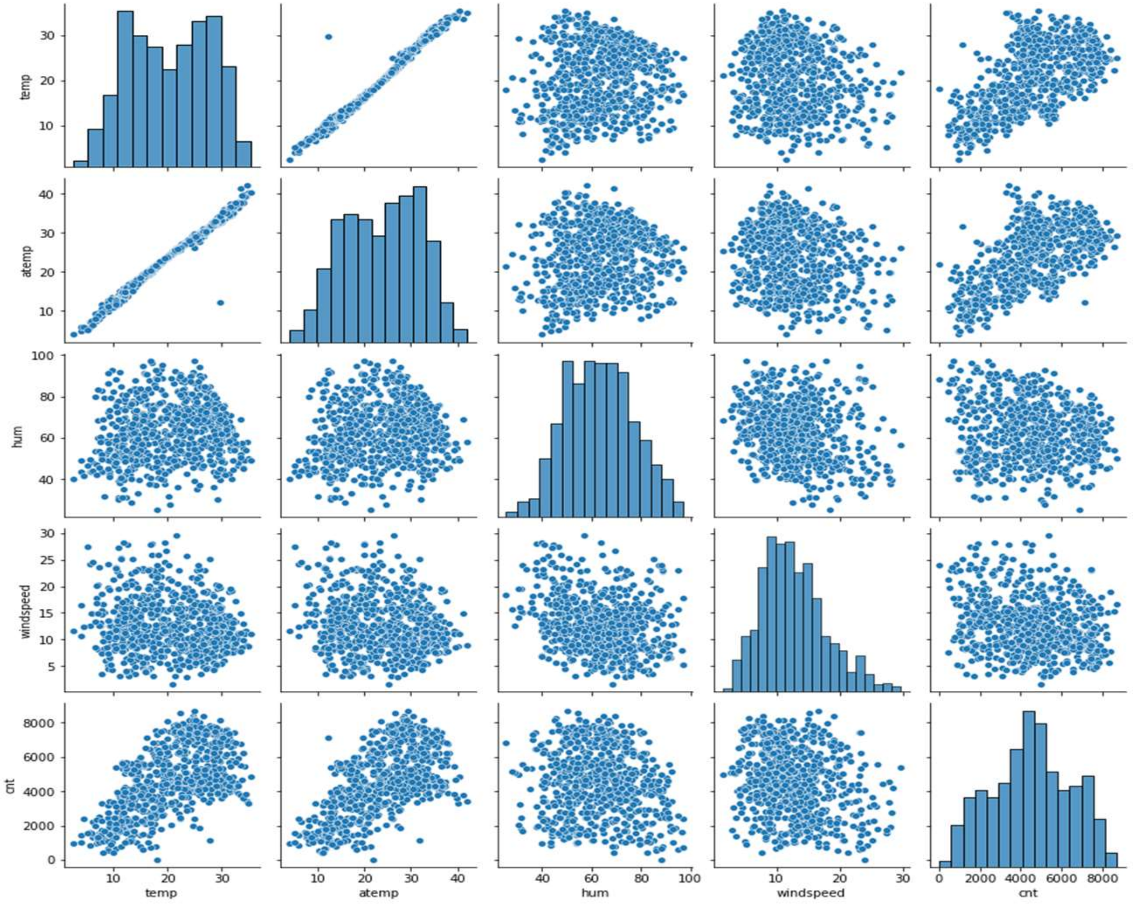
Based on the above observation, it can be inferred that we don’t need to actually define ‘N’ columns for ‘N’ unique values. Our work can be done with ‘N-1’ columns itself. If all the columns for that value are 0, system can easily assume a value of 1 of the column not defined. Similarly, if one of the column is already 1, then the column not defined will have a value of 0.

Now consider a case where there are high number (say 10) of categorical columns and each column having 10 unique values. Hence, a total of 100 columns will be created if we create 1 column for 1 value. If we take the N-1 approach, work will be done only in 90 columns. Hence, 10 less columns to analyze.

Hence, to create only ‘N-1’ columns instead of ‘N’ columns, drop_first=True option was introduced in the dummy variable creation command.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Looking at the pair plot, it can be inferred that temp and atemp (feeling temperature) has highest correlation with the target variable.

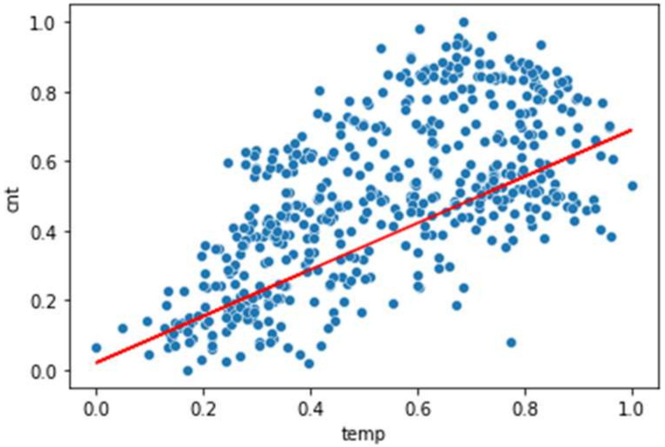


Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

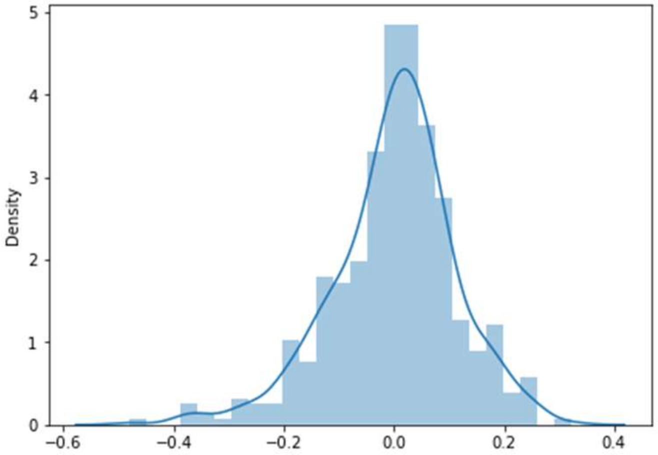
The following assumptions were made for the linear regression model for the bike rental:-

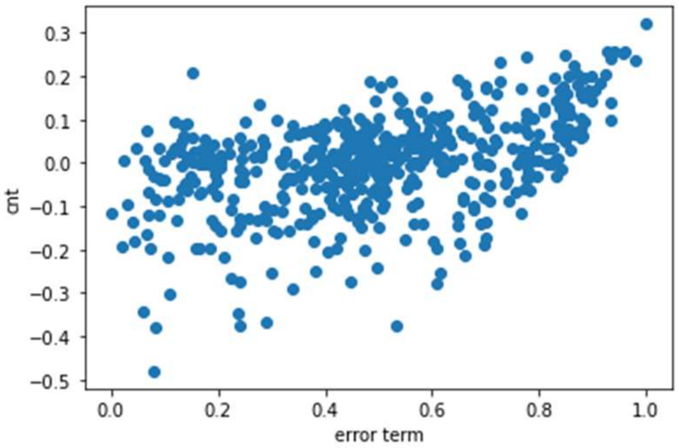
1. There is a linear relationship between X and y
2. Error terms are normally distributed with mean zero
3. Error terms are independent of each other
4. Error terms have constant variance

The first assumption was validated using the scatter plot after model building against the numeric variable.



Error terms are normally distributed with mean zero. This was verified by the residual analysis. See the plot here which is normally distributed around the center 0.





Error terms are independent & have a constant variance; can be verified from this graph. Here except for right tail the distribution is random.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Following are the most impacting features effecting the bike rental demand:-

- Temperature – This has a positive impact on rentals meaning increase in temperate will increase the demand.
- Year – This has a positive impact on the rental meaning demand in 2019 rose as compared to 2018 considering all the other factors as constant.
- Weather (Rains) – This has a negative impact on rentals meaning in the rainy weather the demand for bike rentals decreases.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear corresponds to straight line.
Regression algorithms are the predictive analytics models designed to predict value of a continuous numeric target variable.

Hence, a simple linear regression algorithm is a predictive model that attempts to explain the relationship between a dependent variable and an independent one using a straight line.

In real world however a dependent variable can be influenced by multiple independent variable. Hence, linear regression algorithm in case of multiple independent variable will be explaining the relationship between the dependent variable and one of the independent variable using a straight line keeping all the other independent variables constant or zero. This way all the affecting variables will be evaluated by keeping others constant.

A linear model can be defined as follows:-

$y = mx + c$ where y is dependent and x is independent variable. Note that the power to x is 1 making it linear equation.

In case of multiple independent variables the model is defined as:-

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots\ldots + m_nx_n$$

Following assumptions are made while making a model:-

1. There is a linear relationship between X and y
2. Error terms are normally distributed with mean zero
3. Error terms are independent of each other
4. Error terms have constant variance

While creating multiple linear regression model, following points should be taken care of:-

- Overfitting – While making the model adequate number of variables should be selected. In case of selecting many variables, the model will fit the train dataset exceptionally well but there are high chances that it will not work well with test dataset.

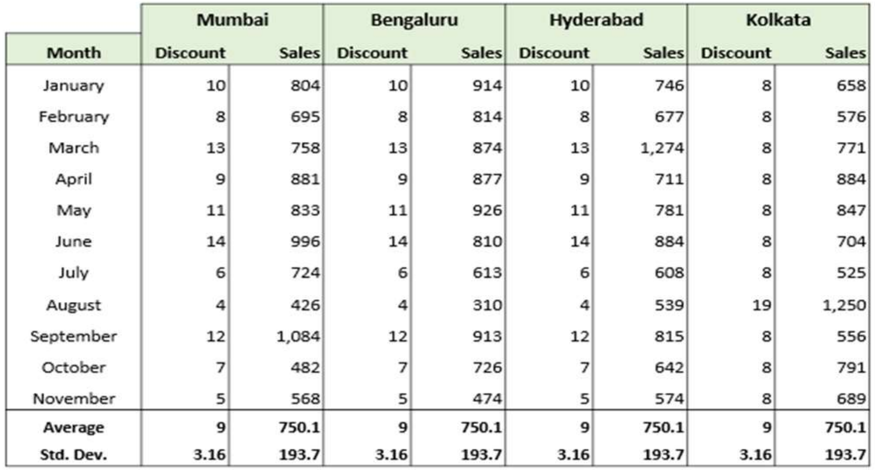
- Multicollinearity – In case there is a relation between independent variables, its best to ignore one of them and simplify the model.

- Feature Selection - Selecting an optimal set from a pool of given features, many of which might be redundant, is important while defining a model.

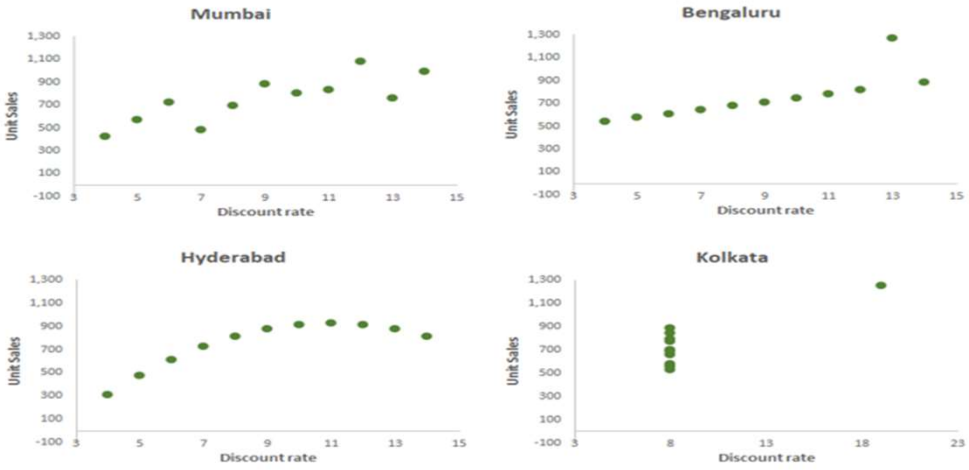
Q2. Explain the Anscombe's quartet in detail.

Often it is seen that equations looking very similar to each other provides a very different view upon plotting the datapoints. A statistician Francis Anscombe was the first one to demonstrate the importance of plotting data and hence this anomaly is named after him. He demonstrated the importance of plotting data before analyzing it and the effect of outliers on statistical properties. He took datasets to explain this difference.

Example – The below table shows sales & discount data of a year provided by 4 branches of a retail store in 4 different cities of India. It is interesting to see that Average & Std Deviation of Sales & Discount is same across branches and hence initial inference can be that all the 4 branches are performing similarly. But upon plotting the data into a scatter plot, a totally different picture can be seen.



	Mumbai		Bengaluru		Hyderabad		Kolkata	
Month	Discount	Sales	Discount	Sales	Discount	Sales	Discount	Sales
January	10	804	10	914	10	746	8	658
February	8	695	8	814	8	677	8	576
March	13	758	13	874	13	1,274	8	771
April	9	881	9	877	9	711	8	884
May	11	833	11	926	11	781	8	847
June	14	996	14	810	14	884	8	704
July	6	724	6	613	6	608	8	525
August	4	426	4	310	4	539	19	1,250
September	12	1,084	12	913	12	815	8	556
October	7	482	7	726	7	642	8	791
November	5	568	5	474	5	574	8	689
Average	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**
Std. Dev.	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**

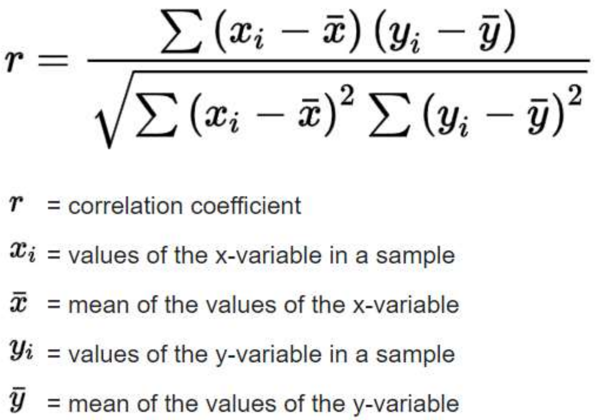


Source: Upgrad session on The Necessity of Data Visualisation

Q3. What is Pearson’s R?

Correlation coefficients are used to measure how strong a relationship is between two variables. It ranges from -1 to 1. A negative value implies that increase in one variable will cause a decrease in other while a positive Correlation coefficients means increase in one will cause increase in other variable and vice-versa.

There are several types of correlation coefficient, but the most popular is Pearson correlation coefficient, also called Pearson’s *R*. It is widely used for linear regression. It is derived 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

Source : wikipedia

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

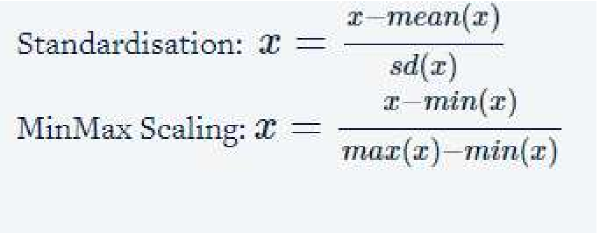
In the data used for modelling, there are various variables. Variables for amount may in range of thousands and millions while variables for temperature will be in range of 10-50. At the same time variables for length will be in range of 1-1000 kms and so on.

For a better modelling it is advisable to have all these variables in same range. This will be very useful while plotting these variables against each other. The process to bring these variables on a same scale is called Scaling.

So we can defined scaling as a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Important point to note is that scaling affects only the coefficients but not the effectiveness (R-Squared, P-value and other stats) of the model.

There are various ways to scale data. Standardisation and MinMax scaling are among widely used methods. Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1. MinMax scaling, on the other hand, brings all the data in the range of 0-1. Formula used for these are:-



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

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

Source: Upgrad - Dealing with Categorical Variables (for the snip of formula)

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. It helps to calculates how well one independent variable is explained by all the other independent variables combined. Technically it is defined as:-

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

Here R_i^2 is R-squared if the ith variable. Now, the R^2 can be defined as follows:-

$$R^2 = 1 - \frac{RSS}{TSS}$$

Here RSS is Residual sum of square and TSS is Total sum of squares. In case of a perfect linear regression model, all the predicted value and the actual values will be same making RSS to be 0. Hence, R^2 will be 1 (100%). This will lead the VIF value to be infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plots or Quantile-Quantile plots are plots of two quantiles against each other. It helps us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It also helps to determine if two data sets come from populations with a common distribution.

A quantile is a fraction where certain values fall below that quantile when arranged in ascending order. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The plot is often used for liner regression since it provides flexibility of sample size. It can also help is simultaneously testing various distributional aspects for example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.