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?

There are seven categorical variables in the dataset. Those are: season, weathersit, year, month, holiday, weekday, workingday

Visualized these variables using boxplots and noticed the following

- season: Number of bikes shared is less in spring season whereas fall has maximum count.
- weathersit: Clear weather has highest number where as light_snow rain has minimum number and 0 in number when weather is Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- year: year wise bike sharing trend is increased. So, in 2019 has more number when compared to 2018
- month: Bikes has been borrowed more in the mid of the year and in increasing trend from May to October months and then slowly started decreasing.
- holiday/weekday/workingday: During holidays count is less compared to working days.

After creating dummy variables for the categorical variables and building the model, these variables show considerable amount of influence on bike sharing count.

The coefficients of the variables related to above categorical variables used in the final model are as follows

a. year 0.2354
b. holiday -0.0970
c. season_spring -0.1162
d. season_winter 0.0480
e. month_Sep 0.0700
f. weathersit_Light_snowrain -0.2885
g. weathersit_Misty -0.0786

year, weather situation, seasons, holidays effect the target variable more compared to others.

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

If you don't drop the first column using drop_first=True while creating dummy variables then dummy variables will be corelated and multicollinearity issue arises.

So, it is very important to drop redundant columns.

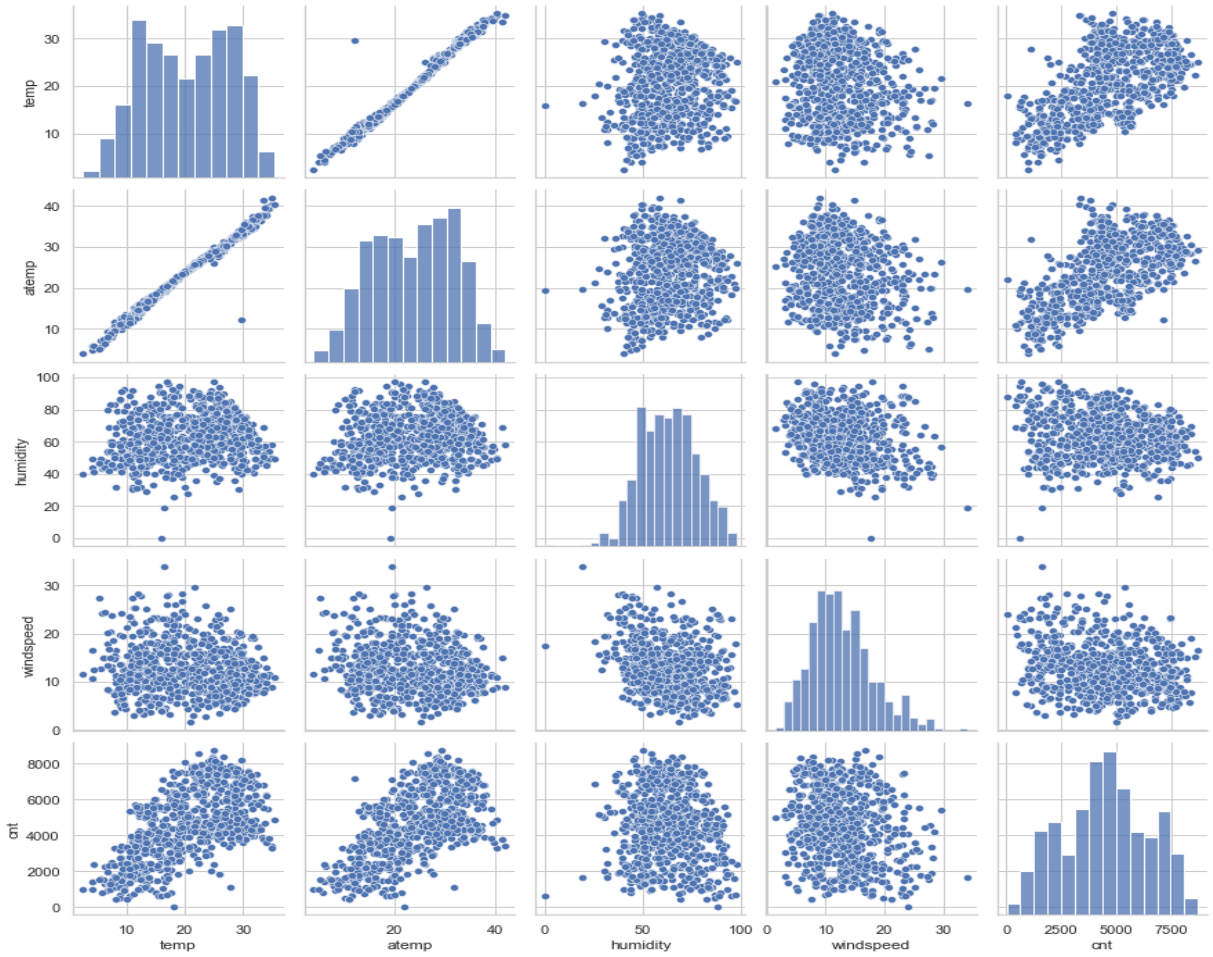
for E.g., in a dataset you have three different types of values for a column named "State"

State: CA,FL,NY

While creating dummy variables for this column only two variables are sufficient. If one variable is not CA and not FL then automatically it belongs to NY. So, there is no need to use third variable to represent/identify it as NY.

We can say that, when you have a categorical variable with say n levels, then the idea of dummy variables creation is to build n-1 variables.

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

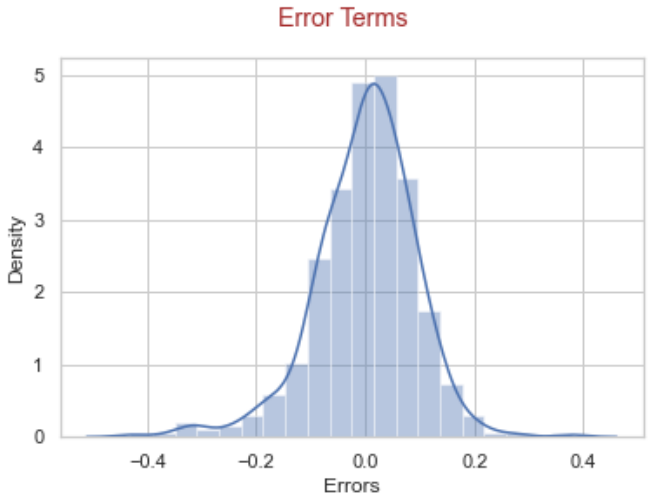


temp and atemp are highly correlated with target variable (cnt) and these two variables are correlated with each other.

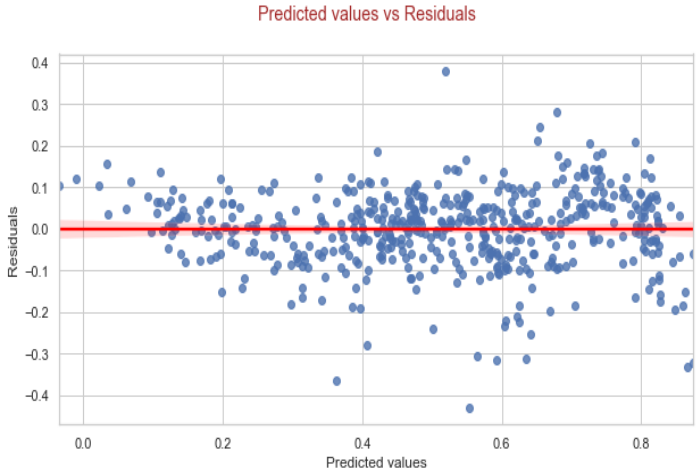
So, while building the model will use either of them (used temp in this code) to avoid Multicollinearity.

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

A. By performing Residual analysis on the error terms and visualizing it by plotting in the graph noticed that errors are normally distributed and mean is 0.



B. There is a linear relationship between the variables. Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor 'cnt'.
C. While building the model, closely monitored VIF and p values to avoid multicollinearity.
D. Checked the variance of error terms which is constant(Homoscedasticity)



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

Based on coefficients of final model the top 3 features which explain variability of the demand are:
temp (temperature→ A coefficient value 0.4078 indicated that a unit increase in temp variable increases bike rentals by 0.4078 units
year→ A coefficient value 0.2354 indicated that a unit increase in year variable increases bike rentals by 0.2354 units
weathersit_Light_snowrain → A coefficient value -0.2885 indicated that a unit increase in this weather situation variable decreased bike rentals by 0.2885 units.

So its suggested to consider these variables at most importance while planning. The other features are windspeed, season_spring etc.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression a Machine learning technique based on Supervised learning method which is used to predict a continuous/numeric output variable.

It explains the relationship between dependent (output variable) and independent(predictor) variables using a straight line.

First to check whether a linear regression is suitable for any given data, a scatter plot can be used. If the relationship looks linear, we can go for a linear model.

There are two types of linear regression models:

1. **Simple linear regression**: SLR is used to predict the dependent variable using one independent variable. Mathematical equation is

$$Y = \beta_0 + \beta_1 X$$

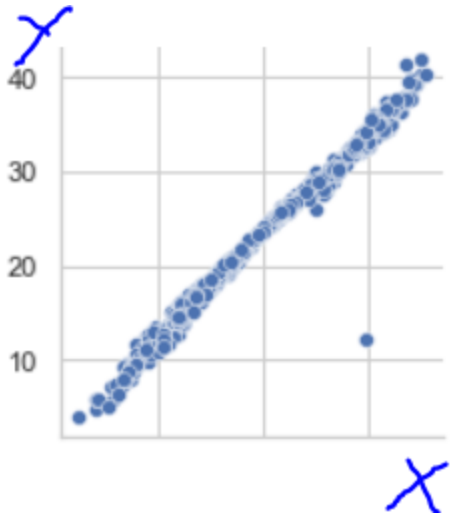
β_o ==> intercept
β_1 ==> slope
Y ==> dependent variable
X ==> independent variable
The straight line is plotted on a scatter plot

2. **Multiple linear regression**: MLR is used to predict the dependent variable using multiple independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

β_o ==> intercept/constant
β1 ==> coefficient for X1 variable
β2 ==> coefficient for X2 variable
βp ==> coefficient for Xp variable
€ ==> model error (how much variation there is in our estimate of y)
In case of MLR, model now fits a hyperplane instead of line.

See the following plot for example



Here we could see that there is a linear relationship between X and Y variables

Once we found that linear relationship exists then we will try fitting a line (best-fit line) in a way that Residual Sum of Squares is minimum.

Residuals for any data point is calculated by subtracting predicted value of dependent variable from actual value of dependent variable. Total Sum of residuals needs to be minimum to say that model is a good fit. Statistical formula for RSS as follows

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

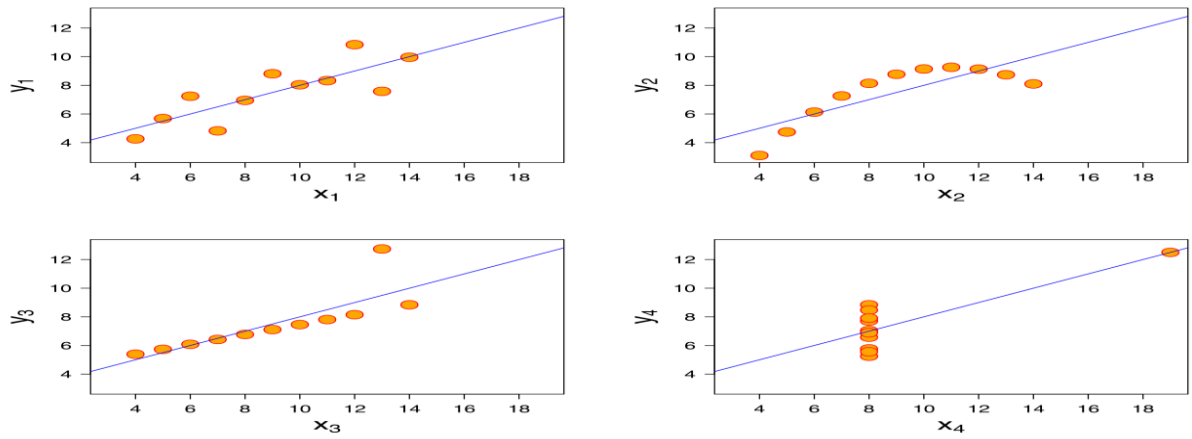
y_i is the i^{th} value of the variable to be predicted.
x_i is the i^{th} value of the explanatory variable
α estimated value of the constant term
β estimated value of the slope coefficient
i value varies from 1 to n

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet explains the importance of plotting the graphs before analyzing and model building and the effect of other observations in statistical properties.

Basically, it consists of four datasets which are nearly identical in simple descriptive statistics yet very different in the datasets which have misguided the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Check the below image which explains the same



All four sets are identical when examined using simple summary statistics, but vary a lot when graphed. The four datasets or plots explains the following

Graph1/Dataset1- Fits the linear regression model pretty well.
Graph2/Dataset2- couldn't fit the linear regression model and the data is non linear which confirms that linear regression is incapable of handling any other data.
Graph3/Dataset3- shows the outliers involved present in the dataset which cannot be handled well by linear regression
Graph4/Dataset4- shows the outliers involved present in the dataset which cannot be handled well by linear regression
So, we shouldn't run a regression without having a good look at data.

3. What is Pearson's R?

Pearson's R is a statistical measure used to determine the strength of association between two variables. This correlation coefficient will always be between -1 and 1

+1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

- Negative value implies that variables are negatively correlated with each other i.e. Both tend to different directions. If one variable value decreases then the other variable value increases and vice versa.
- Positive value implies that variables are positively correlated with each other i.e. Both the variables tend to change in same direction. If one variable decreases other also decreases and if one variable increases other also increases.
- Zero value shows no relationship between the variables.

Mainly Pearson's R correlation coefficient designed for linear relationships and might not be a good measure if the relationship between variables is non linear.

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 data of independent variables to be in a particular range for easy interpretation. This is performed during data preprocessing stage.

Scaling is performed because multiple features in datasets comes in different scales, units and range.

If scaling is not done then a Machine learning algorithm assumes that features with greater values are more significant than the features with lower values without considering units of values.

To solve this, we need to bring all the values of the variables with in the same range.

Scaling just effect the coefficients and no other values like P-values, F-statistic, R-squared etc. will get impacted.

Two common types of scaling used are

1. **Standardization-** Standardization brings all the data in to a standard normal distribution with mean 0 and Standard Deviation(SD) 1

   Standardization x=(x-mean(x))/SD(x)

2. **Normalization/MinMax Scaling**- MinMax scaling brings all the data in the range of 0 and 1

   MinMax Scaling x=(x-min(x))/(max(x)-min(x))

But normalization loses some information in the data, especially about outliers, this can be a disadvantage when compared to standardization.

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

Variance Inflation Factor (VIF) explains how well one independent variable is explained by all other independent variables combined.

This helps in detecting Multicollinearity between the independent variables.

The formula for VIF is

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

R2i is the coefficient of determination.

R squared value shows how much of variance is been explained by the model and fit between the dependent variable and the other independent variables.

R squared value lies between 0 and 1. If there is perfect correlation between the variables R squared value becomes 1. So, if this tends to 1,VIF becomes 1/(1-1)=1/0 which leads to infinite.

So, if there is a perfect correlation between the variables VIF value becomes infinite.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential.

With Q-Q Plots (Quantile-Quantile plots we can compare two probability distributions by plotting their 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% lies above it.

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.

If the two distributions being compared are similar, then the points in 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.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in linear regression and we can verify this using Q-Q plots
- Skewness of distribution