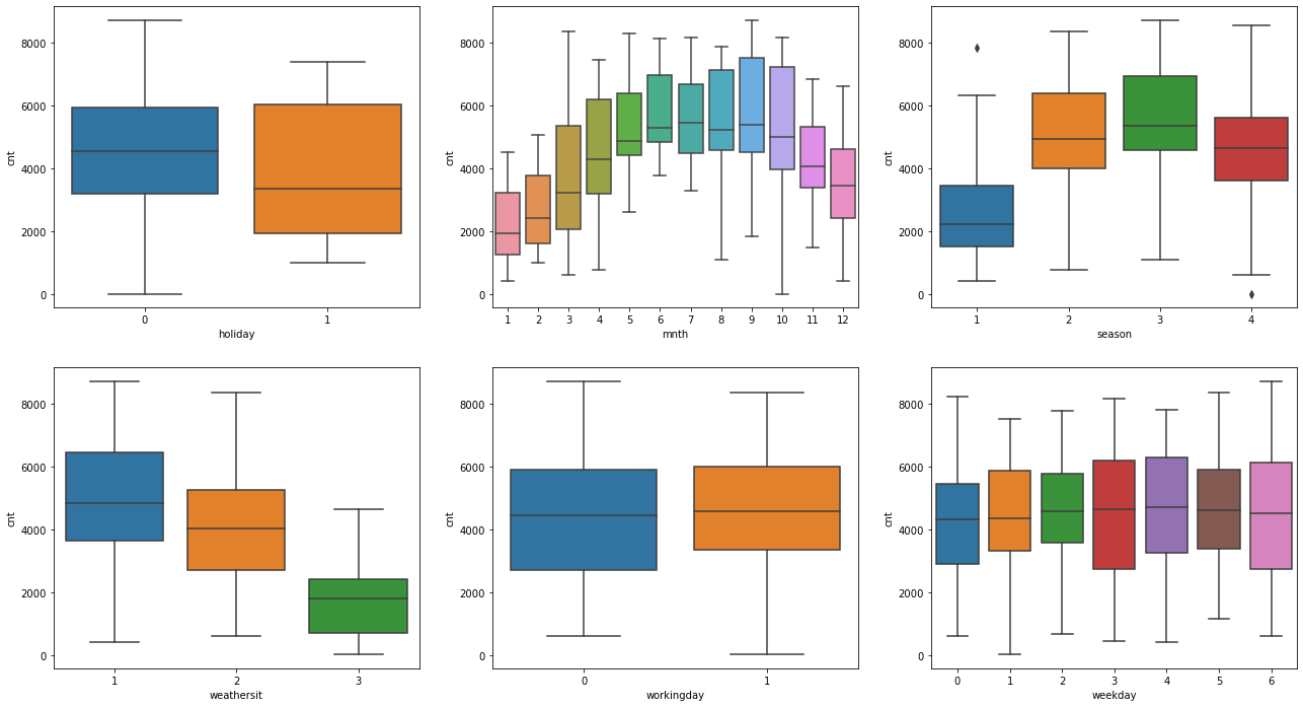
Name : Roshan Choudhary

Batch : ML C 36

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?**



Using Boxplot to study predictors effect on the dependent variable ('cnt') .

The inference that we could derive were:

- **holiday**: About 97.62% of the bike registrations happened on a non-holiday which may also means that the data was mostly recorded during non-holidays. Hence, holiday CANNOT be used as a good predictor for the dependent variable.
- **mnth**: About 10% of the bike registrations happened in the months 5,6,7,8 and 9 with a median of about 4000 registrations per month. This indicates, month has valid trends for registrations and can be a good predictor for the dependent variable.
- **season**: About 32% of the bike registrations happened in the season3 with a median of about 5000 registrations per month. Season2 & season4 had 27% & 25% of total registrations.This indicates, season has valid trends for registrations and can be a good predictor for the dependent variable.
- **weathersit**: About 67% of the bike registrations were happened during 'weathersit1 with a median of about 5000 registrations. Weathersit2 had 30% of total registrations. This indicates, weathersit has valid trends towards the bike registrations can be a good predictor for the dependent variable.
- **workingday**: About 69% of the bike registrations happened in 'workingday' with a median of about 5000 registrations. This indicates, workingday can be a good predictor for the dependent variable.
- **Weekday:** About 13.5%-14.8% of total registrations on all days of the week having their independent medians ranging from 4000 to 5000 registrations. The variable can or cannot turn out to be a good predictor

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

Answer: drop_first=True is mainly used to drop out one level of that particular categorical variable in order to avoid redundancy and add only valuable fields to the model. It helps to discount the collinearity that is being introduced while adding dummy creation.

For example: If there' a column which denotes gender as male-student, female-student and we want to create dummy variables for this column. If we have 0 for male-student and 1 for female-student, then only by having one level data would help us to get the entire story. A single column can have 0 or 1 based on the gender of the student. Having 0 would automatically a male student and thereby no additional column is needed to explicitly mention the same.

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

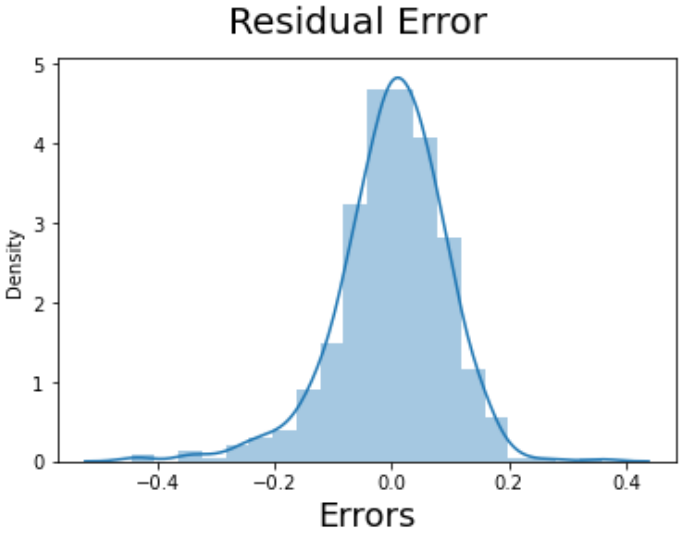
Answer: 'temp' and 'atemp' has very high correlation with target variable 'cnt'.

1. Corelation between temp and atemp is 0.99
2. Corelation between temp/atemp and cnt is 0.63

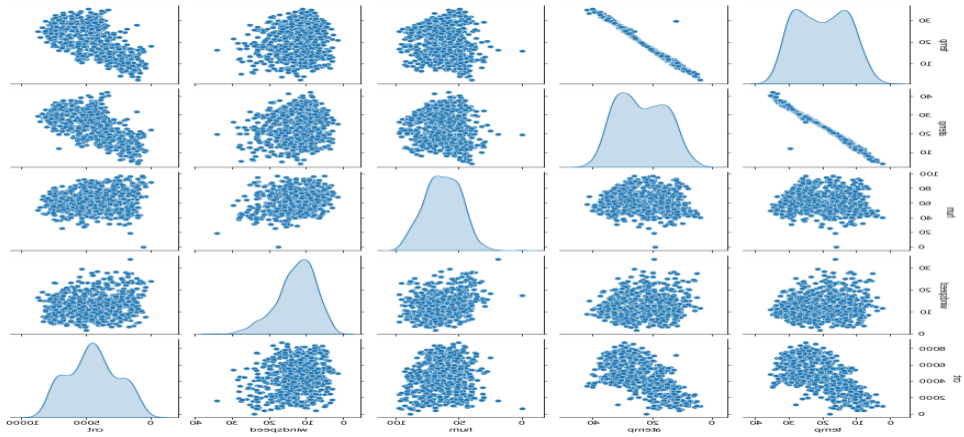
4.How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

1. **Checked if Error terms are normally distributed with mean zero by plotting the residual** (res = y_train-y_train_pred) distplot



2. **Checked Linear relationship between X and Y**
   Using a pair plot between predictors and dependent variable and studying the pattern forming from various datapoints



3. **To Make sure there exists no multicollinearity between the predictor variables**

- This was mainly done using the variance inflation factor and removing attributes having high VIF values.

	Features	VIF
2	temp	4.14
4	season_4	1.98
0	yr	1.94
3	season_2	1.85
7	mnth_10	1.67
5	mnth_8	1.55
9	weathersit_2	1.45
6	mnth_9	1.32
8	weekday_6	1.15
10	weathersit_3	1.06
1	holiday	1.03

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

1. atemp – feels like temperature has highest positive correlation
2. windspeed – has highest negative correlation
3. yr – year attribute too has correlation as the demand shot up sharply in 2019 compared to 2018.

General Subjective Q&A

Q1. Explain the linear regression algorithm in detail.

Answer: In lay man term, Linear Regression is basically a mathematical technique to predict/estimate the future values in the data, given the variables in the dataset depicts linear relationship with the target variable that is to be predicted.

This is a form of supervised learning which means there exists a Target Variable and based on the historical patterns between predictors (non-target variables) and Target variables, the mathematical equations is formed to determine the best fit line passing through these data points which can be consumed to predict the target variable for unseen data.

Equation of Linear regression Line is : Y = mX + C

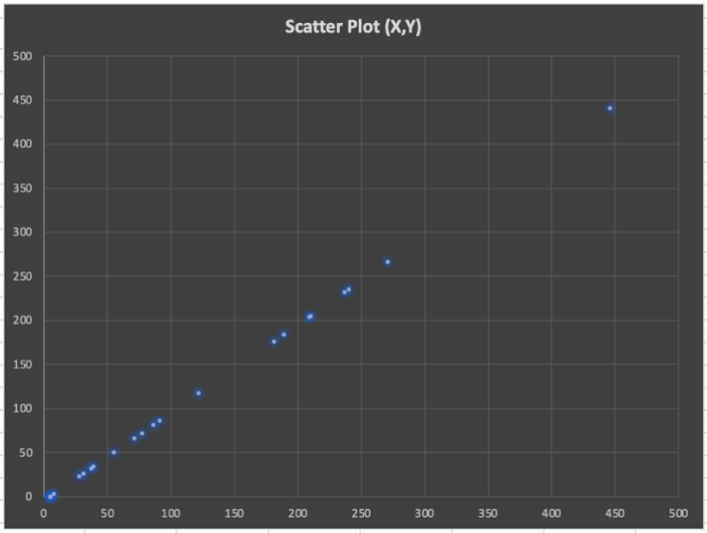
Where Y = Target Variable (To be Predicted)

m = slope of the line (change in y for change in x / (dy/dx))

C = Intercept (Measure of line cutting Y-Axis)

To understand better, let's plot a random X having m as slope for which Y would be mX

m	X	C	Y=mX+C
21	0	5	5
23	8	5	189
4	8	5	37
1	3	5	8
29	8	5	237
13	2	5	31
41	5	5	210
50	1	5	55
23	1	5	28
39	3	5	122
17	2	5	39
38	7	5	271
27	3	5	86
12	6	5	77
49	9	5	446
7	0	5	5
11	6	5	71
34	6	5	209
47	5	5	240
44	4	5	181
48	0	5	5
43	2	5	91



Here you would observe a perfect straight line when Y and X are plotted against each other. This is because each value of X is perfectly able to calculate the values of Y. This is called as simple linear regression (SLR).

When we have more than more Xs say x1, x2, x3… xn we use multiple linear regression where all the method remains same however there are few assumptions that we hold while feeding data into linear regression.

Such assumptions are

1. Assumption 1: Linear Relationship between predictors and Target Variable
2. Assumption 2: Independence of errors or no lags
3. Assumption 3: Homoscedasticity or no heteroskedasticity exists
4. Assumption 4: Normal distribution of variables

The goodness of fit of linear regression is generally obtained by minimising the errors which is predicted values vs original values (Yorig – Ypred), these are also called as Residuals.

Best fit line is measured using R-square. Which implies the variance in Y that the model is able to explain using various values of Xs.

R-square can also be a inflated number hence adj. R-square is used along which measures redundancy of variables in the model. For good model R square and Adj R-Square should be a closer value.

R square = 1 – RSS/TSS

Where RSS = Residual sum of squares

TSS = Total sum of squares

Adjusted_r2 = 1-(1-r2)*(n-1)/(n-p-1)

Where n = samples

p = predictors

Model evaluation also utilises AIC and BIC values however those are mainly used while comparing 2 or models. The Beta co-effecients (m or slope above) acts as the weights to various predictors and helps providing the strength as well as directions of independent variable in contributing for target variable.

Example: yr: A coefficient value of '0.2315' indicated that a unit increase in yr variable, increases the bike registrations by 0.2315 units.

Q2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed.

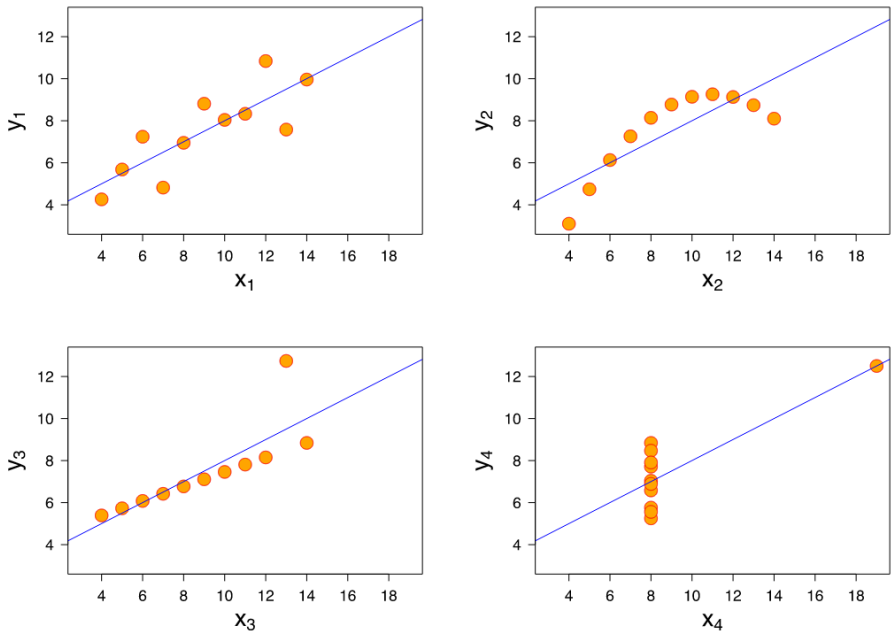
For example: Each dataset consists of eleven (x,y) pairs as follows:

I		II		III		IV	
x	y	x	y	x	y	x	y
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

All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

Q3. What is Pearson's R? Answer: It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations, thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

Pearson's R = COV(X,Y) / SD(X) * SD(Y)

Note: Perason's R is capable to measure only the linear co-relation and ignore any other combinations

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

Answer: Scaling is a method to bring uniformity in the data and get the values into particular range without disturbing the statistical measures of the data.

Scaling is a practise in statistics/modelling where a particular variable having different units or having elevated values are brought within a particular range. This helps the variable to normalize.

Normalized scaling : It mainly gets the data into the range of 0 to 1 and the formula is as below

Norm Scaling (X)= X – min(X) / max(X) – min(X)

Standardized scaling: This mainly makes use of Z scores and replaces the values against it. It generally brings data into a standard normal distribution getting mean to 0 and at 1 standard deviation.

Stand scaling(X) = X – mean(X) / std(X)

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

Answer: VIF is a check for multicollinearity within dataset and VIF = 1 / (1-R^2)

VIF = infinity generally implies a very high multicollinearity hence indicating a say perfect co-relationship between predictors.

This happened in my experience when using dummy variable for a categorical variable, the drop_first = True was not used and the variable that needed to be dropped still existed causing a perfect relationship with other levels of that particular variable depending upon the cardinality.

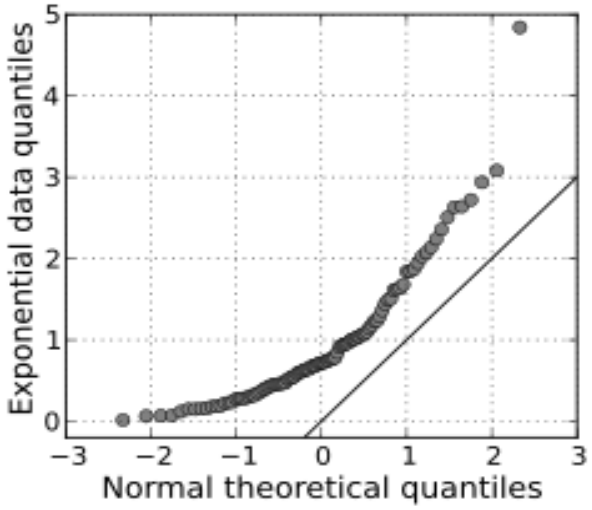
This was rectified once I was able to drop one level of information from dummy variable.

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

Answer: Q stands for Quantile. Wherein a quantile is a fraction below where certain values falls below that point. For example median value is a 50% quantile where 50% of the values lie below that median value and 50% of it lies above the median value.

Here Q-Q plot is nothing but Quantile – Quantile plot where 2 quantiles are plotted against each other. This is mainly done to find out if the 2 given data come from the same distribution.

A 45 degree angle is plotted on the Q-Q plot. If the two data sets come from the same distribution, the points will fall on the 45degree reference line.



- 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