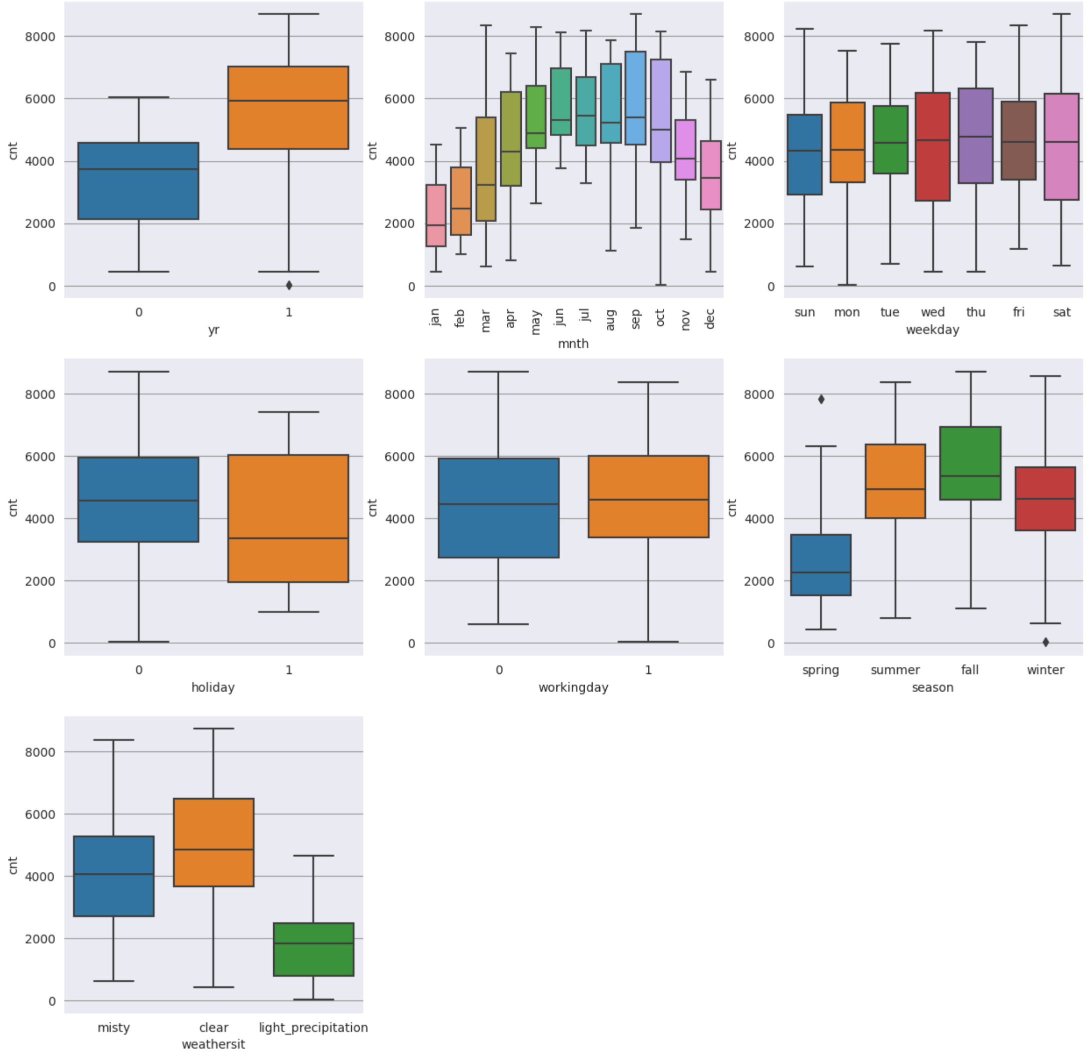
Assignment-based Subjective Questions

Ques 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans – We can see in below image that –

- Year 2019 got maximum variation in cnt
- Month “Sep” has highest counts among other months (this is in our top feature also)
- Number of holidays are less, hence non-holiday has variety of counts of bike users.
- Whether it’s a working day or not, it has no major effect
- Fall season’s count mean is more than any other season.
- Most of the times, it’s clear weather (we can see counts in data visualization section in notebook) and hence it has the most variation in cnt



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

Ans – Let’s take an example of season column from bike dataset, there are four types of seasons available, i.e., fall, spring, summer and winter. When we create dummy variables, it looks like below –

Fall	Spring	Summer	Winter
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

From above table, we can clearly see that even if we drop first column, i.e. “fall” column, even then information about “fall” column can easily be derived. Ex. –

Spring	Summer	Winter
0	0	1
0	1	0
0	0	0

When all three seasons are 0, then it directly implies that fall season value is 1. And when any other season value is 1 that means fall season value is 0. Hence there is no information loss when we drop first/any column out of dummy variable creation.

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

Ans – Correlation values –

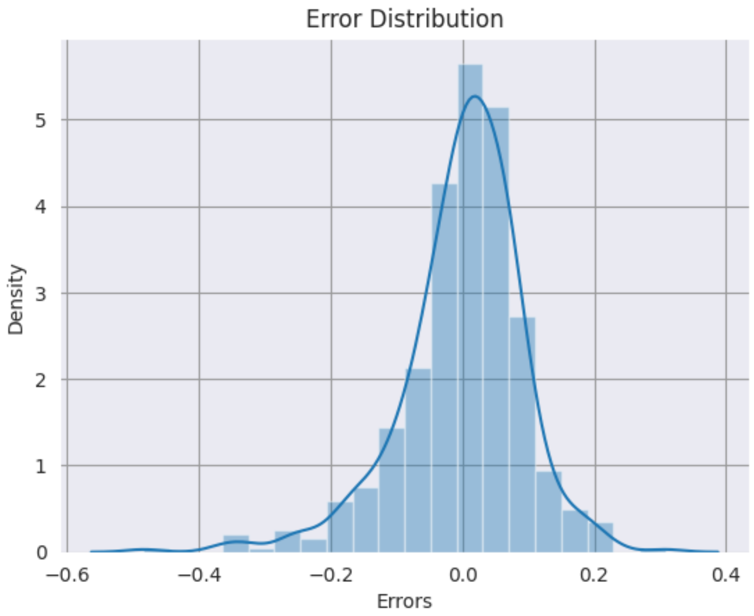
- Cnt and Registered – 0.95
- Cnt and Casual – 0.67
- Cnt and temp – 0.63
- Cnt and atemp – 0.63

That means if we take “registered” and “casual” columns in account, then highest correlated column with “cnt” is “registered”. If we don’t consider these two columns (with fact that cnt = registered + casual), then highest correlated columns with “cnt” are temp and atemp with value of 0.63.

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

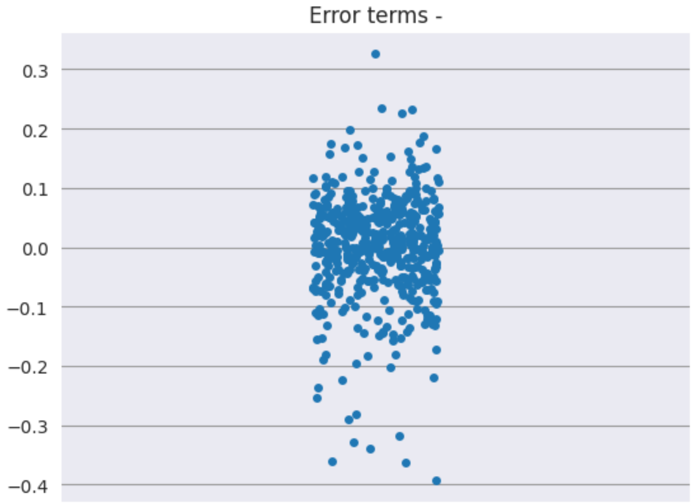
Ans – Assumptions –

1. Linear Relationship between X and Y.
2. Error terms are normally distributed, not X and y. This, I have checked by plotting residuals of train data and we can clearly see below that error terms are normally distributed.

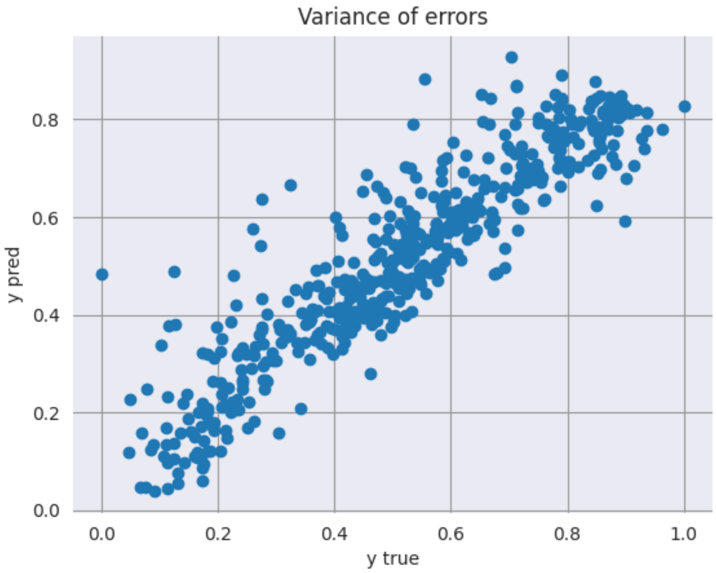


3. Error terms are independent of each other. I have checked this by plotting a strip plot of residuals and thus found that all the errors terms are distributed around 0 and there is no pattern. I have checked the mean also by –

```
round((y_train - y_train_pred).mean(), 2)
answer is 0.0
```



4. Error terms have constant variance (homoscedasticity). This, I have checked by plotting scatter plot between y_train and y_train_pred, here we can clearly see that variance of error terms is constant.



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

Ans -

Code - `lm.params.sort_values(ascending=False)`

Output –

```
temp 0.541569
const 0.296008
yr 0.221255
sep 0.077670
winter 0.067541
may 0.058154
sat 0.050984
mar 0.044904
workingday 0.033424
jul -0.077591
spring -0.097189
windspeed -0.175958
hum -0.289417
```

From above result, we can clearly see that temp, yr and sep (const is added by us) are top features. Because all the features are scaled on same scale, and we know that these coefficients provide a crude base for feature importance.

General Subjective Questions

Ques 1. Explain the linear regression algorithm in detail.

Ans – Linear Regression explains the linear relationship between 1 dependent/target variable and 1 or more independent variables.

Suppose we have 1 independent variable x and dependent variable y, So, we can explain the linear relationship between both as –

$$y = mx + c$$

Here m = slope and c = intercept.
We derive the best fit line between X and y by minimizing the residual sum of squares, which is –

RSS = sum((yi – yi_pred)^2) for i from 1 to n

There are a lot of optimization algorithms which helps in learning the best values of m and c, e,g. – gradient descent.
To identify how good our model is – we have R-squared and F-statistic scores
R-squared determines the strength of best-fit line and the value explains the amount of variance in y explained by model.
F-score determines how good overall model is.

Assumptions in Linear Regression –

1. Linear Relationship in X and y
2. Error terms are normally distributed (not X, y)
3. Error terms are independent of each other (not like time series)
4. Error terms have constant variance (homoscedasticity)

There is **multiple linear regression** also, where we have more than 1 independent variables.

y = b0 + b1*x1 + b2*x2 + b3*x3 + … + bn* xn

- Interpretation of the coefficients –

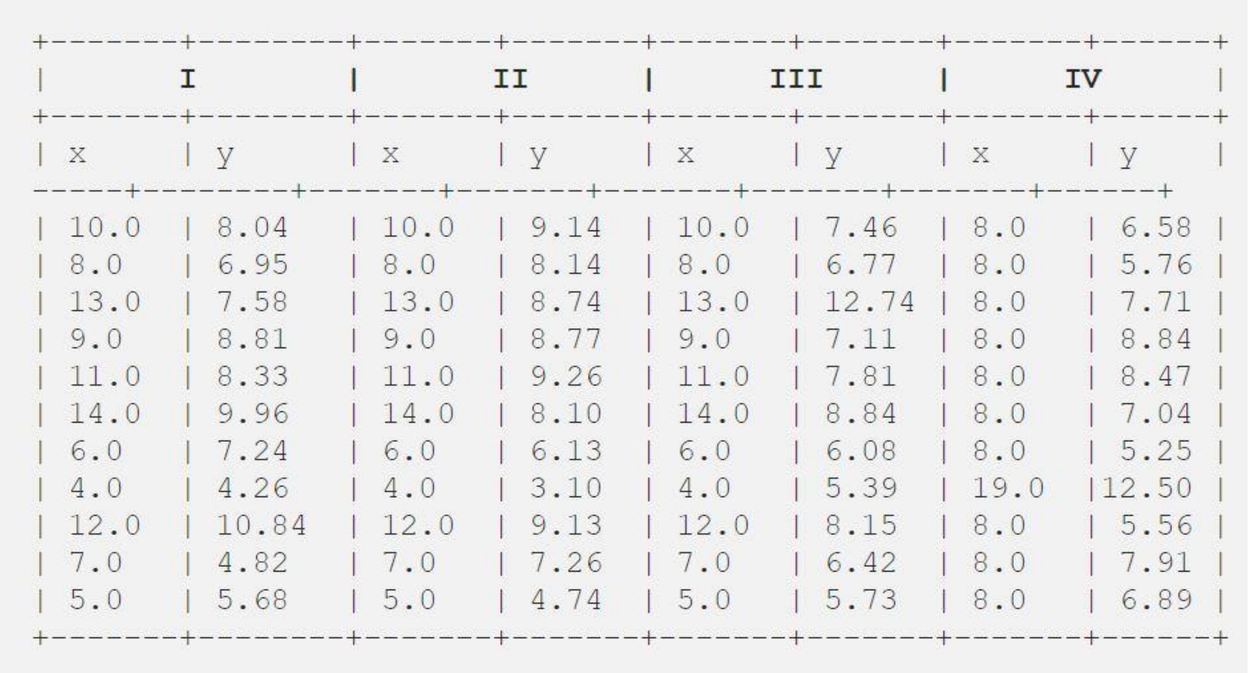
Change in the mean response E(y), per unit increase in the variable when other predictors are held constant.

- Here, model fits hyperplane, instead of line. Coefficients still obtained by minimizing RSS (mentioned above)
- Assumptions are same as above.
- But there are a lot of new concepts in picture, Feature Scaling, Detecting and dealing with multicollinearity, Handling categorical variables using dummy variable concept, Feature selection (manual + automated approach)
- We have dealt with all these concepts in notebook in detail, hence not putting here again.

Ques 2. Explain the Anscombe's quartet in detail.

Ans – Anscombe's quartet explains the importance of data visualization before analyzing and model building, and shows the drawbacks of depending only on statistics.
It comprises of 4 datasets with almost same simple descriptive statistics in terms of means, variance, r-squared, correlations.

Below are the 4 datasets -

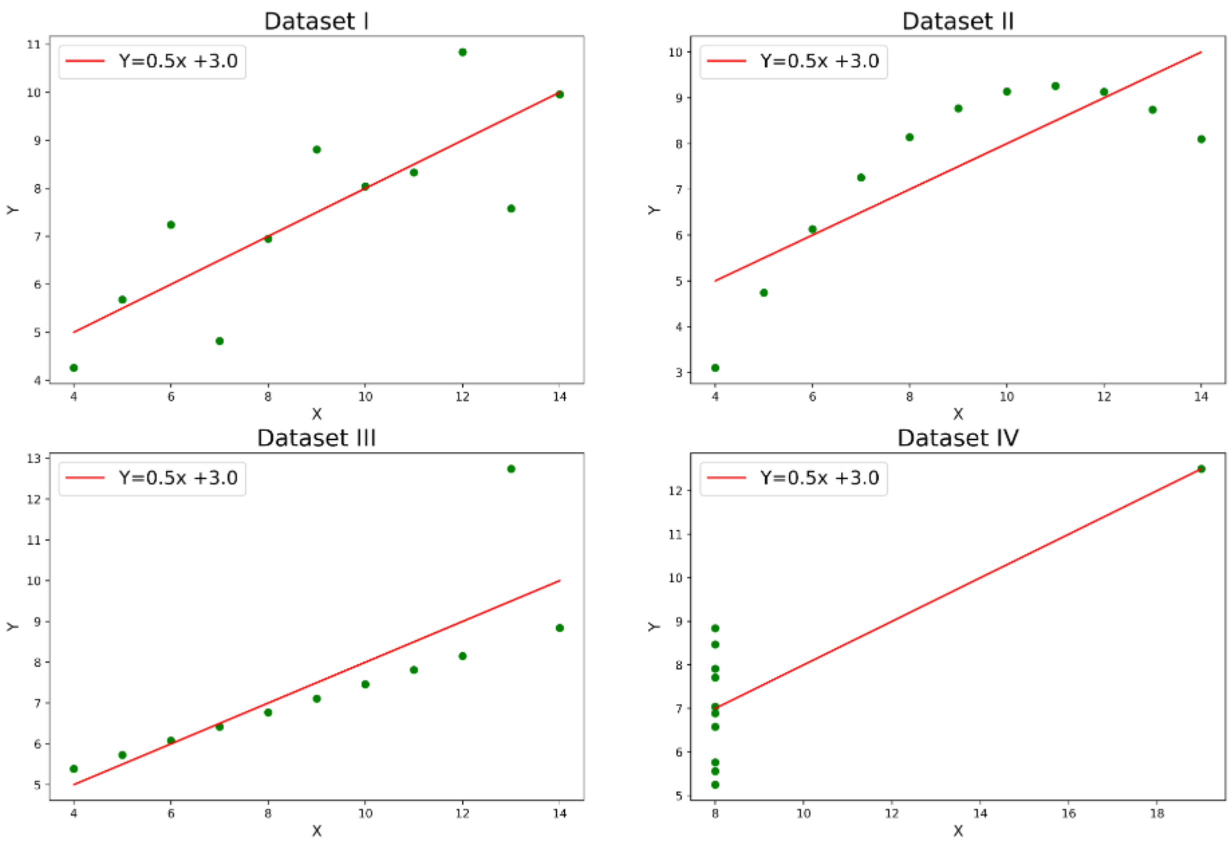


```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

Statistics –

```
                                  I         II        III        IV
Mean_x                        9.000000  9.000000  9.000000  9.000000
Variance_x                   11.000000 11.000000 11.000000 11.000000
Mean_y                        7.500909  7.500909  7.500000  7.500909
Variance_y                    4.127269  4.127629  4.122620  4.123249
Correlation                   0.816421  0.816237  0.816287  0.816521
Linear Regression slope       0.500091  0.500000  0.499727  0.499909
Linear Regression intercept 3.000091  3.000909  3.002455  3.001727
```

We can clearly stats results are almost similar for all datasets. But if we see dataset we can see the problem, for example, in 4th dataset, all x-values are same (8), except one outlier (19) which is causing mean of x = 9. In other datasets, it is varying. But again, we can't see actual patterns everytime from our eyes, we need EDA for that, so let's see how these 4 datasets looks like in visualisation –



Wow! We did not expect such a variety in datasets, hence it proved that we should not depend only on statistics, we must always do EDA before model building.
Note – Images and stats results are taken from GeeksforGeeks.

Ques 3. What is Pearson's R?

Ans – Pearson's R – way to measure correlation. It measures the strength and direction of relationship between two variables. Range of this coefficient is [-1, 1]. -1 value signifies negative relationship, 0 value – neutral and +1 value signifies positive relationship. Formula →

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

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

Ans – Scaling is a process of normalizing the range of features in a dataset.
Feature scaling may or may not improve model accuracy, but it has significant effect in understanding features quickly and improve optimization process by making gradient

descent flow smoother and thus attaining minimum of cost function as quickly as possible.

Difference between normalized scaling and standardized scaling –

Normalized scaling / MinMaxScaler –

- brings feature range in [0, 1] or in [-1, 1] if there are negative values in feature.
- formula : (X – Xmin)/(Xmax – Xmin)

Standardized scaling / StandardScaler –

- brings feature mean at 0 and standard deviation at 1.
- Formula : (X – mean)/(std_deviation)

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

Ans – We know that VIF of a variable signifies how much that variable is described by other independent variables (not target). Higher the value of VIF, the more the degree of multicollinearity. So, when a variable is linear combination of other variables/features (not target), the VIF tends to infinity.

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

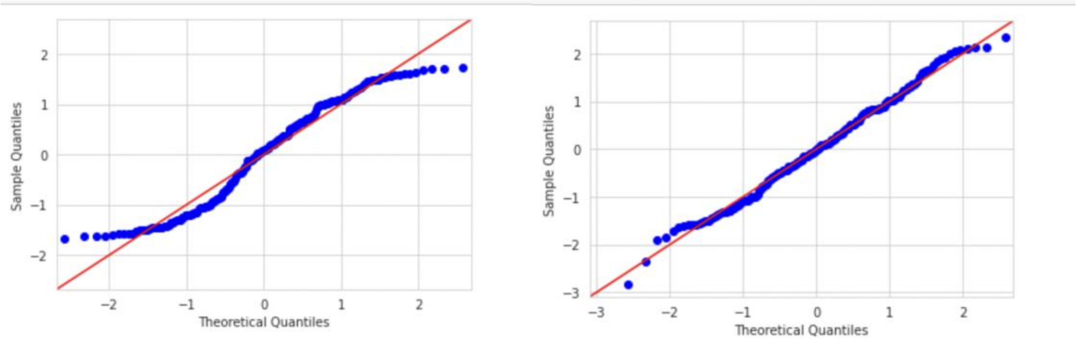
Ans – Q-Q stands for Quantile-Quantile.., hence Q-Q Plot plots the quantiles of a sample distribution against quantiles of a theoretical distribution. Using this plot, we can identify if our data follows any type of probability distribution, i.e., normal, uniform, etc.

Use and Importance of Q-Q plot –

- It helps in determining whether two datasets are of same distribution or not
- We know that in regression we assume that residuals follows a normal distribution, but Q-Q plot helps us to verify that.
- It also helps in determining skewness in data. If left side of plot is deviating from line, we say that it is left-skewed, same for right side as well.

Let's see examples of various q-q plots explaining probability distributions –

Here, 1st plot is for uniform distribution and 2nd plot is when data follows normal distribution.



Note – Images are taken from Analytics Vidhya's blog