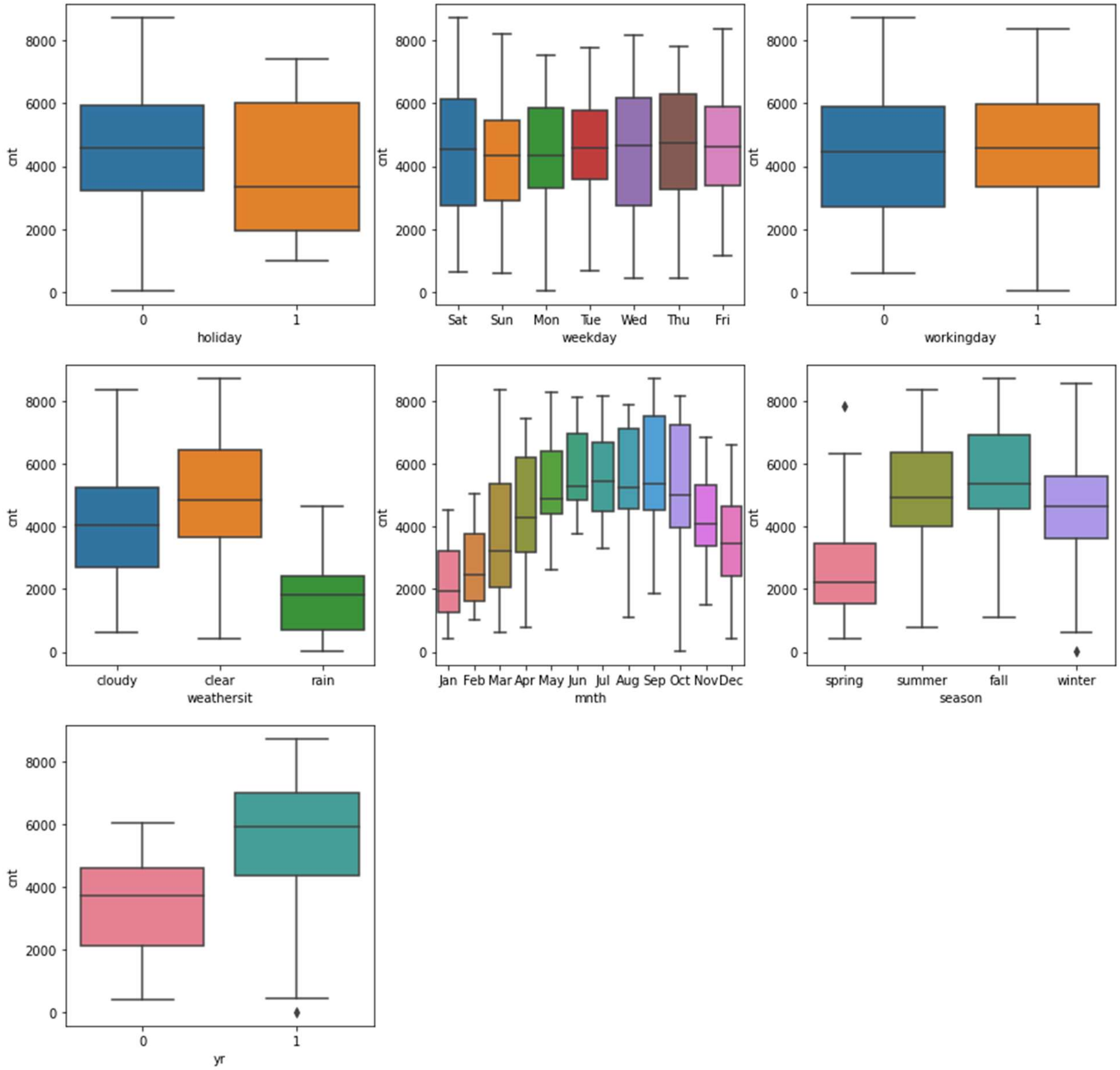
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

A) Here is the pictorial representation of the categorical variables with respect to target variable cnt:



- Year (1:2019) shows higher bike rentals. It clearly indicate the growing trend of the bike rental business.

- There is a spike in bike rentals on Clear weather, while there is a high dip for rain days. There was no rentals for the heavy rain days. Obviously people tend to use bikes in good pleasant weather.

- September shows the highest rental. followed by July , august and october. Rentals were less in January, february and December, when the weather is harsh.

- Season graph shows most rental took place in fall when the weather is pleasant.

- Distribution on month and season looks similar.

- Bikes were rented more on non-holidays than holidays as shown by its median.

- Within a week, more number of rentals were seen for Saturdays followed by Wednesdays and Thursdays.

- On average there is no much difference in the renting pattern of working days and non-working days (including weekend and holidays)

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

A) If drop_first=True will not create a dummy table for the first categorical value. Thus managing to create n-1 tables when there are n distinct values for the categorical value.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A) registered shows the highest correlation with target variable cnt. In a way registered can also be considered one among the target in a different sense. The other independent variable temp also shows a high correlation with cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A) Assumptions of Linear Regression are:

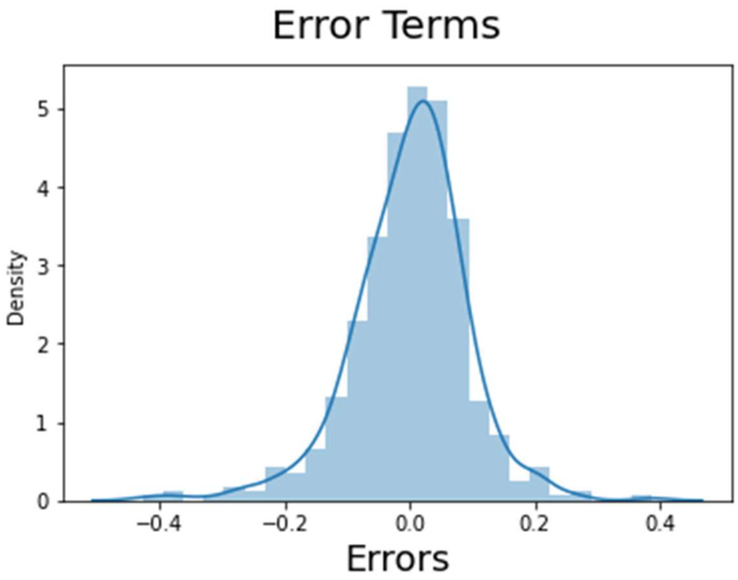
1. Linear relationship between X (independent) and Y(dependent) variables

2. Error terms are normally distributed (over y i.e y_train - y_train_pred)

3. Error terms are independent of each other (We have eliminated correlation of independent variable time to time)

4. Error terms have constant variance (homoscedasticity)

To check this we have calculated the errors from y_train - y_train_pred (predicted) and have plotted a distplot over it.



The graph shows a normal distribution with mean 0 and a constant variance. This shows that the model has a good fit.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A) We have temp, weathersit_rain and yr in top three significant predictors of demand of shared bikes in our model

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a machine learning algorithm based on supervised learning.

It is used in predicting a dependent variable value (y) based on a given single or multiple independent variable(s) (x).

This regression technique finds out a linear relationship between x (input) and y(output).

When we have one independent variable we call is simple linear regression, while for using multiple variable it is termed as multiple linear regression.

It works by fitting a regression line y=B0 + B1X for Simple linear regression (SLR)

While it would fit a hyperplane for multiple linear regression (MLR) with y=B0 + B1X1 + B2X2 …..BnXn

Where B0 is the intercept on y axis when X is zero.

B1,B2….Bn are the coefficient of the independent variables X1,X2…Xn.

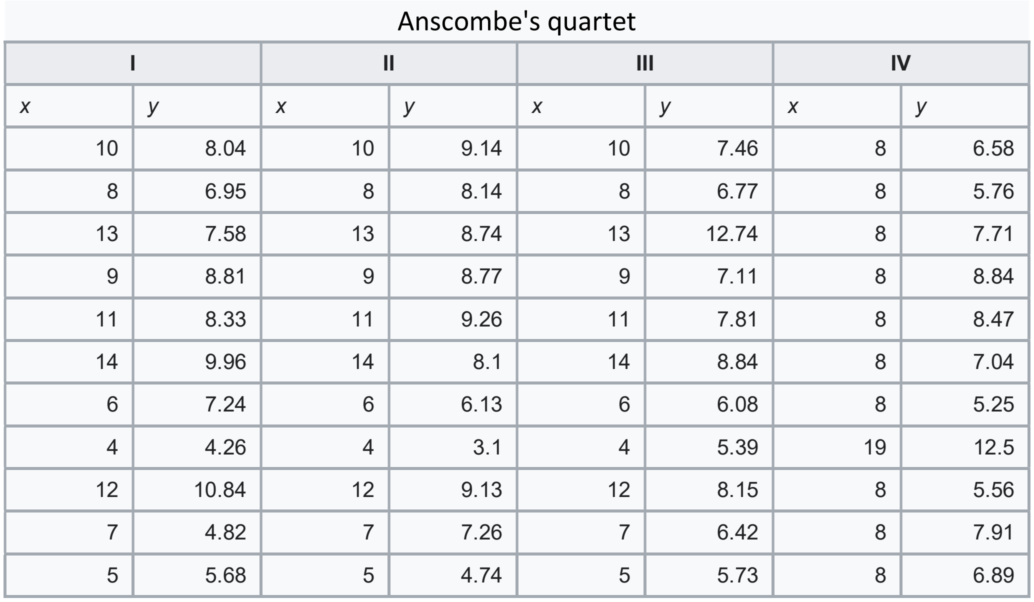
The best fitting line or hyperplane is constructed by Gradient decent method by minimizing the expression of Residual Sum of Squares (RSS) which is equal to the sum of squares of the residual for each data point in the plot.

We also have some assumptions for the correctness of the regression model. That is the independent variables have relation with the dependent variable. The error formed by predicted dependent variable with the actual variable is normally distributed with mean 0 and a constant variance.

2. Explain the Anscombe's quartet in detail. (3 marks)

A) Anscombe's quartet were constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. This was to counter the belief that "numerical calculations are exact, but graphs are rough."

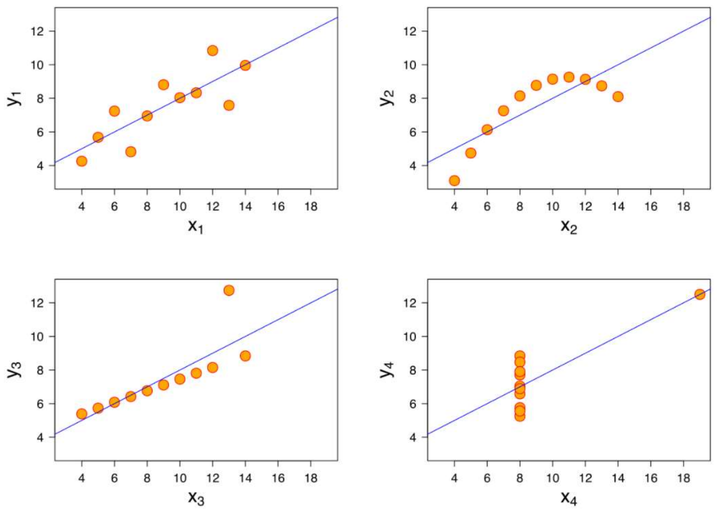
The quartet comprises four data sets that have nearly identical simple descriptive statistics (mean and standard deviations), yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.



Anscombe's quartet

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

When plotted in graph would look like this:

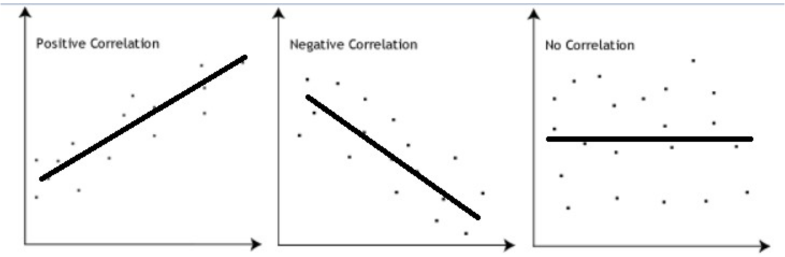


3. What is Pearson’s R? (3 marks)

A) The Pearson's R or Pearson product-moment correlation coefficient or Pearson correlation coefficient is a measure of the strength of a linear association between two variables usualy denoted by r.

It can take a value between +1 to -1. Where +1 denoted perfect positive corelation, -1 represents perfect negative coorelation while 0 deotes no coorelation between the variables.

Here is a pictorial representation of some ositive , negative and no or zero corelation.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

It is generally observed in datasets that data values could range in magnitude, units and range it could be enormously large or even at time very minute. Carrying out the numeric and statistical calculation on it would sometimes get cumbersome with the effect of rounding offs. Most of the data analysis algorithm only take the magnitude into account.

Scaling or feature scaling is a process in which we normalize these variables so that their values fall into a range easy enough to run the algorithms. Also it is seen that the calculations and algorithms runs faster on scaled data. The variables and its values could also be interpreted easily with scaling.

Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

We can scale the features/variables using two popular method:

1. Standardized scaling or Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

X_new = (x - mean(x)) / sd(x)

New values would range in between -1 and +1

2. Normalised scaling or MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

X_new = (x – min(x)) / (max(x) – min(x))

New scaled values would range in between 0 and 1

It is also noted that scaling only affects the coefficients while other parameters like t-statistic, F statistic, p-values, R-square remains same.

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

(3 marks)

A When we have some high correlated features in the independent variables, then the VIF would sometime reach infinite. As its corresponding R-squared value is high and almost equal to 1.

By formula VIF = 1/(1-R^2) , when R^2 is 1 then VIF = infinity

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

(3 marks)

A) Quantile-Quantile (Q-Q) plot helps in assessing if two sets of data are from the same theoritical distribution or from a population with common distribution.

It works by plotting the quantiles of the first data set against the quantiles of the second data set.

This tool will be handy in scenarrios of linear regression when training and test data sets are given seporately. We will have to confirm that both are from one population with same distribution.

We can clearly find out, If two data sets:

- come from populations with a common distribution

- have common location and scale

- have similar distributional shapes

- have similar tail behavior