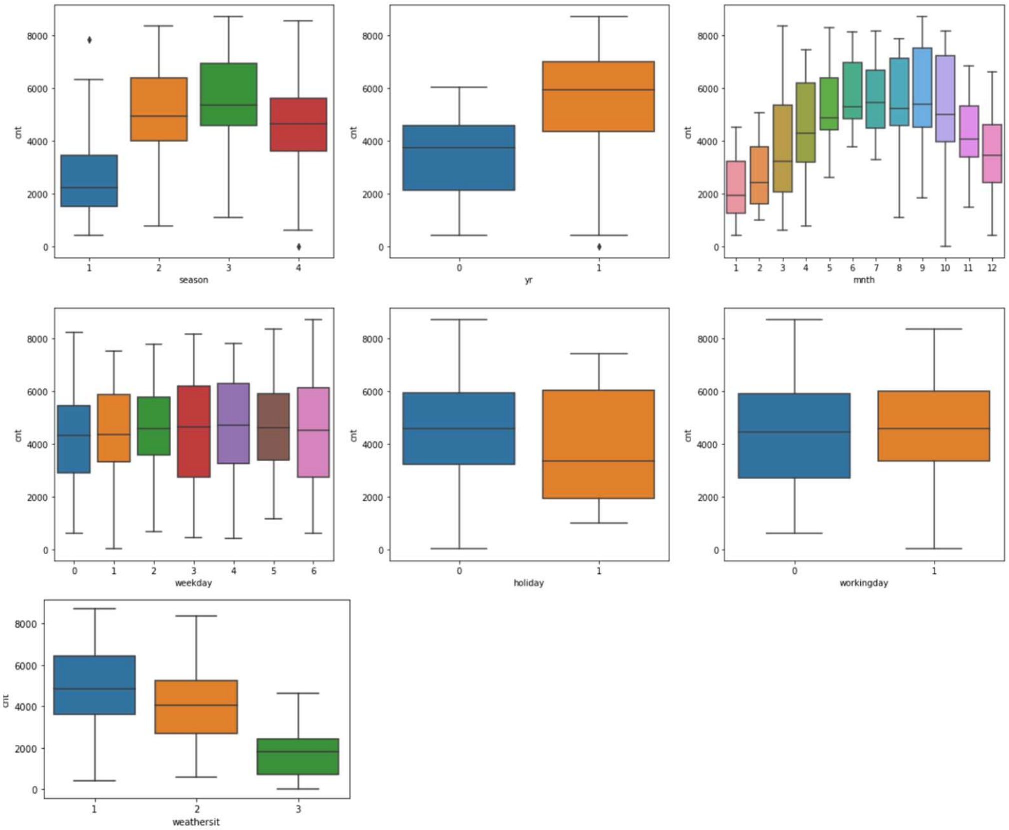
Assignment-based Subjective Questions

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

Answer: Categorical variable can be analyzed visually by using box plot.



Below are the inferences from the box plot.

- **Season** - The distribution of booking count for different season varies significantly. Specially for season 2 and season 3, the median is higher then season 1 and season 4. Therefore season can be a good predictor variable.
- **yr** - The distribution of booking count for 2019 is higher than 2018. So yr plays an important role in predicting the booking count.
- **mnth** - mnth shows a significant variation in booking count across different month. Hence mnth can be a good predictor variable.
- **weekday** - The distribution of booking count for different weekday values does not show any significant variation. Median is also almost same for all the day. Hence weekday might not be significant.
- **Holiday** - The median for booking count is more when it is not holiday. So definitely some trend is there and might be a significant variable.
- **Workingday** - The distribution of booking count is almost same for working day 1 and 0. Hence it might not be a significant variable in predicting the booking count.
- **weathersit** - The distribution of booking count for different weathersit varies significantly. Therefore, this is a significant variable.

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

Answer: When we have categorical variable with 'n' levels, we can create 'n-1' dummy variables by using "**drop_first = True"** in pandas **get_dummies** function. If one level is dropped, still all the levels of the categorical variable can be explained.

For example: If Gender has two levels: Male and Female, it can be explained by a single dummy variable say "Male". For "Male" value is 1 and for Female value is 0.

The linear equation with dummy variable will be:

$Y = \beta_0 + \beta_1$ x Male

For Female (Male = 0), $Y = \beta_0$

For Male (Male = 1), $Y = \beta_0 + \beta_1$

If we keep both the dummy variable for Male and Female then equation will be:

$Y = \beta_0 + \beta_1$ x Male + β_2 x Female

For Female, $Y = \beta_0 + \beta_2$

For Male, $Y = \beta_0 + \beta_1$

This unnecessarily makes the equation complex when it can be defined in a simpler equation.

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

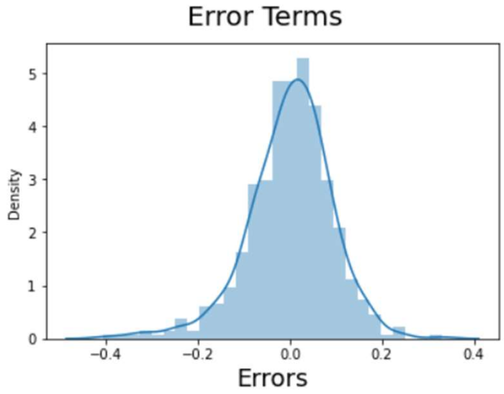
Answer: "temp" and "atemp" variable shows the linear relationship with target variable "cnt".

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

Answer: Following assumptions of Linear regression were verified after building the final model.

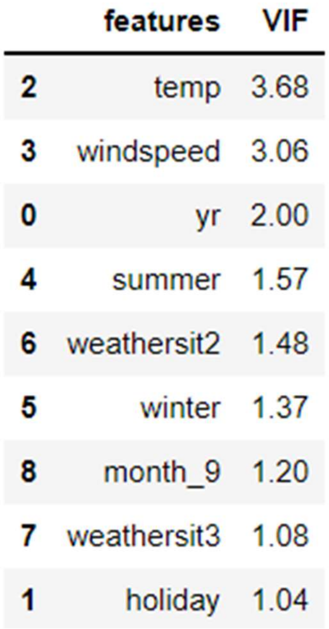
- Error terms should be normally distributed with mean equal to 0

  This can be verified by plotting the histogram of error terms. The histogram plot clearly shows the error terms are normally distributed with mean 0.



- There is no multicollinearity between independent variables.

  By checking the VIF values for the variables, we can say that there is no multicollinearity between independent variables as the VIF values are less than 5.



	features	VIF
2	temp	3.68
3	windspeed	3.06
0	yr	2.00
4	summer	1.57
6	weathersit2	1.48
5	winter	1.37
8	month_9	1.20
7	weathersit3	1.08
1	holiday	1.04

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer: As per the final model, the Equation for our best fitted line is:
cnt = 0.1259 + (yr × 0.2329) - (holiday × 0.0987) + (temp × 0.5480) − (windspeed × 0.1532) + (summer × 0.0881) + (winter × 0.1293) − (weathersit2 × 0.0784) − (weathersit3 × 0.2829) + (month_9 x 0.1012)
Based on above equation, "temp", "weathersit3" and "yr" are the most influential variables for bike booking.
weathersit3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions

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

Answer: Linear regression is a machine learning algorithm based on supervised learning. In linear regression explains the linear relationship between dependent variable and independent variables. If number of independent variable is one then its called simple linear regression and if the number of independent variable is more than one then its called multiple linear regression.

In linear regression, we try to identify the best fit line for the given scattered plot. To identify the best fit line, we find out the residuals and try to find out the RSS for a given line passing through scatter plot. Then best fit line is identified by minimizing the residual sum of squares(RSS).

The equation of the best fit line is Y = B0 + B1 * X. Here Y is dependent variable, X is independent variable, B0 is intercept of the line and B1 is the linear regression coefficient of X.

With the help of linear regression model, we can predict the value of Y for any given value of X.

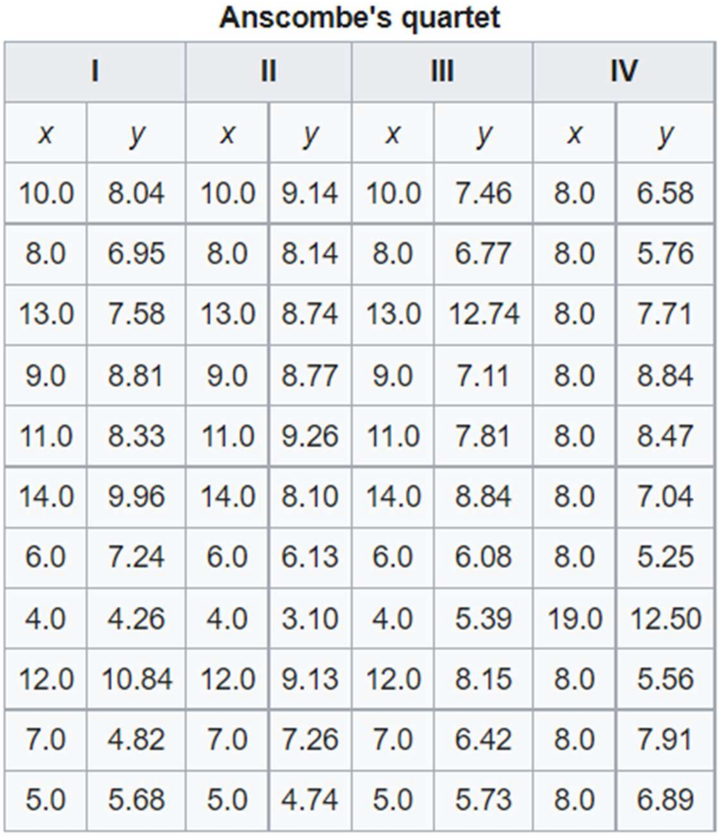
A regression line can be positive linear relationship or negative linear relationship.

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

Answer: Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but they have very different distributions and appear differently when plotted on scatter plots.

These four datasets have same statistical information such as mean and variance. But when visualized on graph, the distribution of data points varies significantly.

This tells us about the importance of visualizing the data before applying various algorithms to build models. Visualizing the data points can help in identify the various anomalies and patterns present in the. The Linear Regression can be only be considered a fit for the data if there is a linear relationship between them.



Anscombe's quartet

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

Image Courtesy: Wikipedia

For all four datasets: (Data Courtesy – Wikipedia)

Mean of x = 9

Mean of y = 7.5

Sample variance of x = 11

Sample variance of y = 4.125

When these models are plotted on a scatter plot, all datasets generate a different kind of plot.

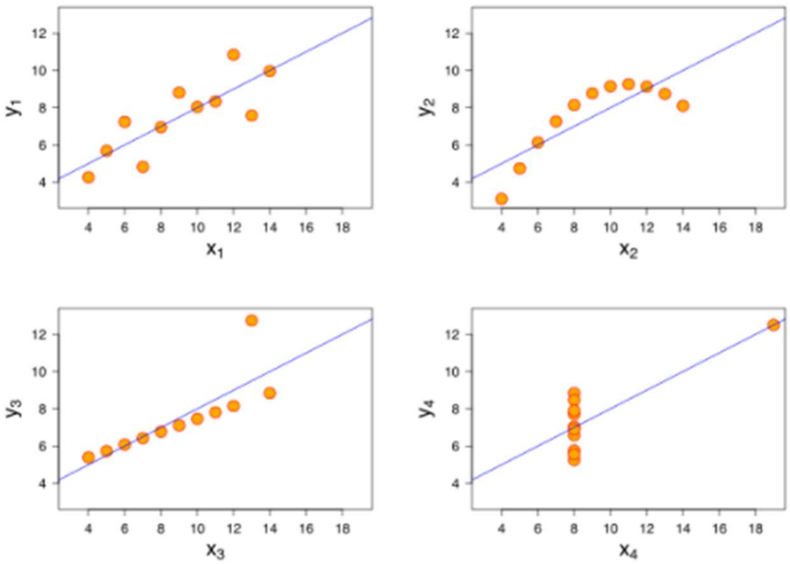


Image Courtesy: Wikipedia

The four datasets can be described as:

Dataset 1: this fits the linear regression model well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

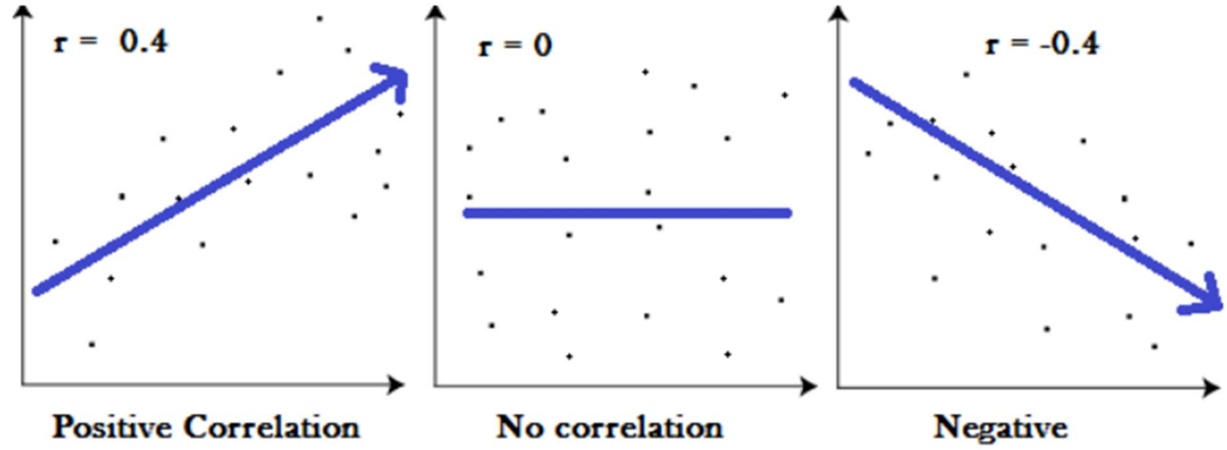
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R? (3 marks)

Answer: Pearson's correlation or Pearson's R is a correlation coefficient commonly used to measure the 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.

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



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

Answer: Variables in a dataset has different scales of values, some are small integer values, and some are large integer/float values. Scaling is a method to standardize the variable in a fixed range. scaling is performed on the variables of the dataset so that they have a comparable scale. If we don't have comparable scales then some of the coefficients obtained after fitting the regression model might be very large or very small compared to the other variable coefficients. This may give a false information that one variable has more influence on the target variable as compared to other variables. So, it is advised to use standardization or normalization so that the coefficients obtained are all on the same scale.

There are two common ways of rescaling:

- Normalization (Min-Max scaling) : Compresses the data between 0 and 1. It also takes care of any outliers present in data.
  Normalized value of X = (X – X_{min}) / (X_{max} - X_{min})
- Standardization: Converts the data such as the mean is 0 and standard deviation is 1.
  Standardized value of X = (X - µ) / sigma

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: If there is a perfect correlation between two independent variables then VIF is infinity. For perfect correlation, R^2 value is 1. Putting this in VIF formula 1/(1-R^2), vif value becomes infinity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.
We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is normal distributed or not, on the y-axis. This gives straight line like structure from each point plotted on the graph. If the points are not falling on the straight line, then it does not have normal distribution.