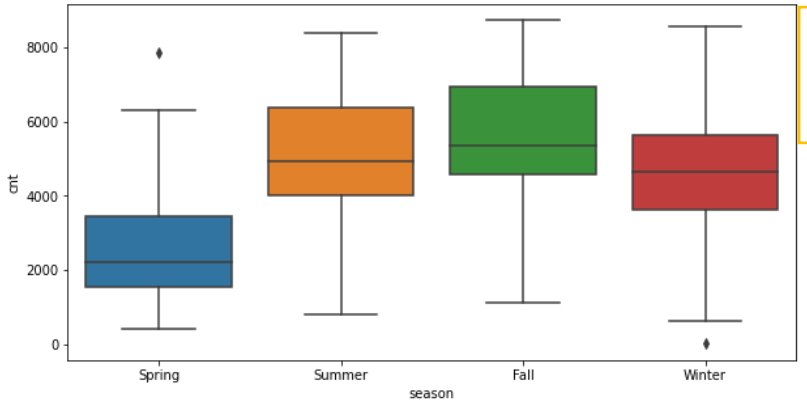
Linear Regression Subjective Questions Submission

Name: Rajib Biswas

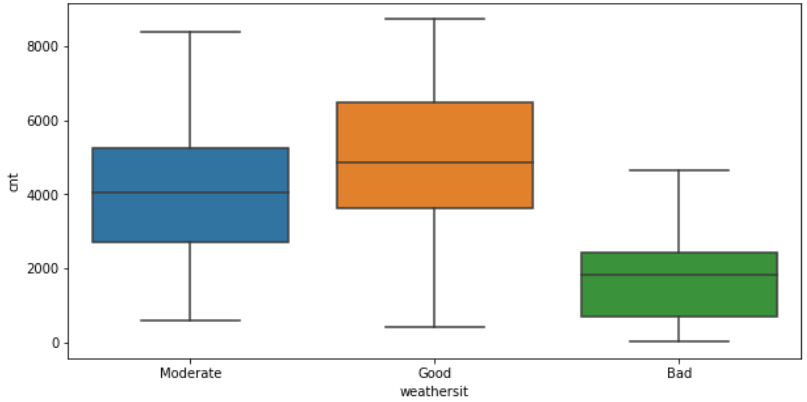
Date: 21/04/2021

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

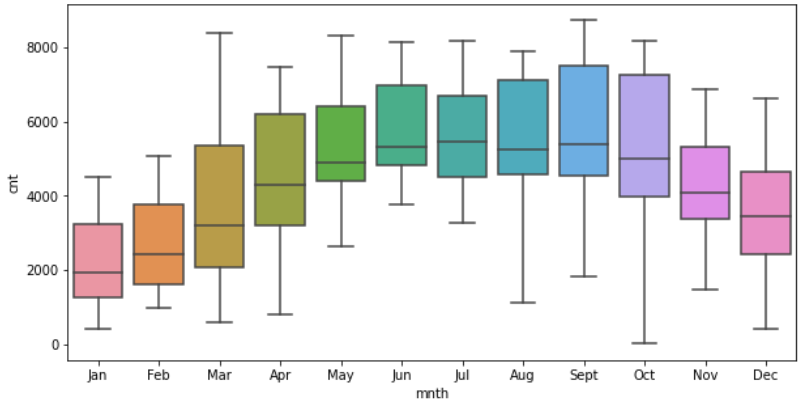


"Fall" followed by "Summer" are the key seasons for the business

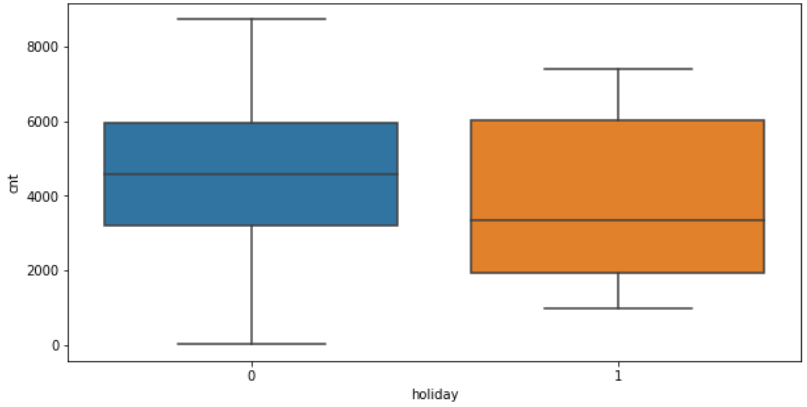


"Good weather" sees increase in business.

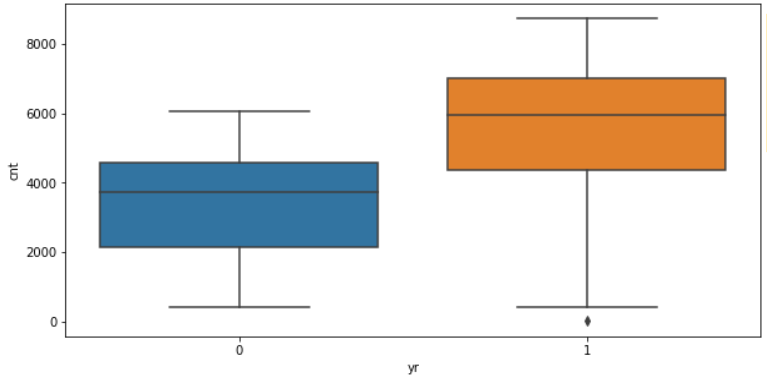
"Holidays" results in fall of business.



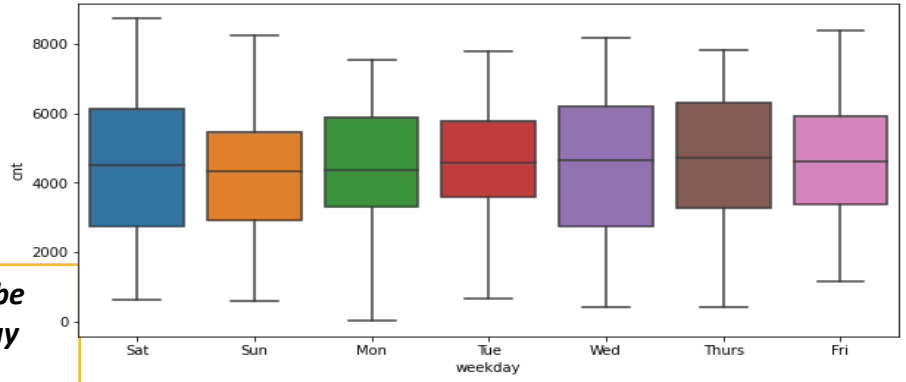
"September" has the highest demand-post that business falls due to harsher weather conditions.



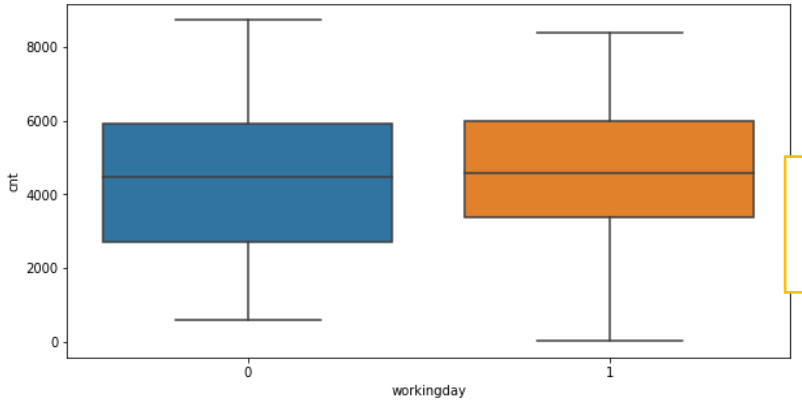
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



Business has grown significantly in 2019 in comparison to 2018.



No Clear pattern can be inferred from weekday variable..



No Clear pattern can be inferred from working day variable..

Q2. Why is it important to use "drop_first"=True during dummy variable creation?

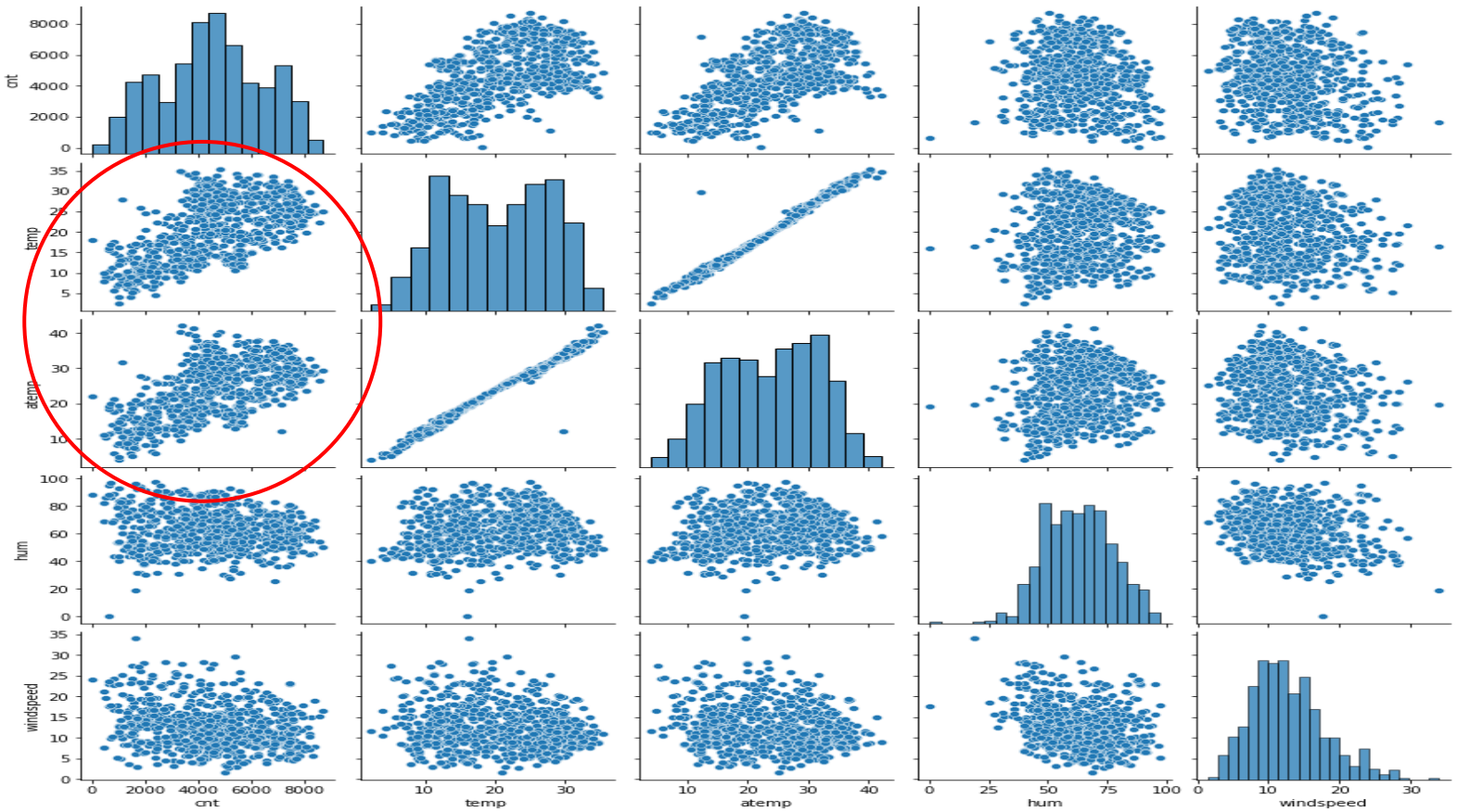
Answer –

"drop_first"=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

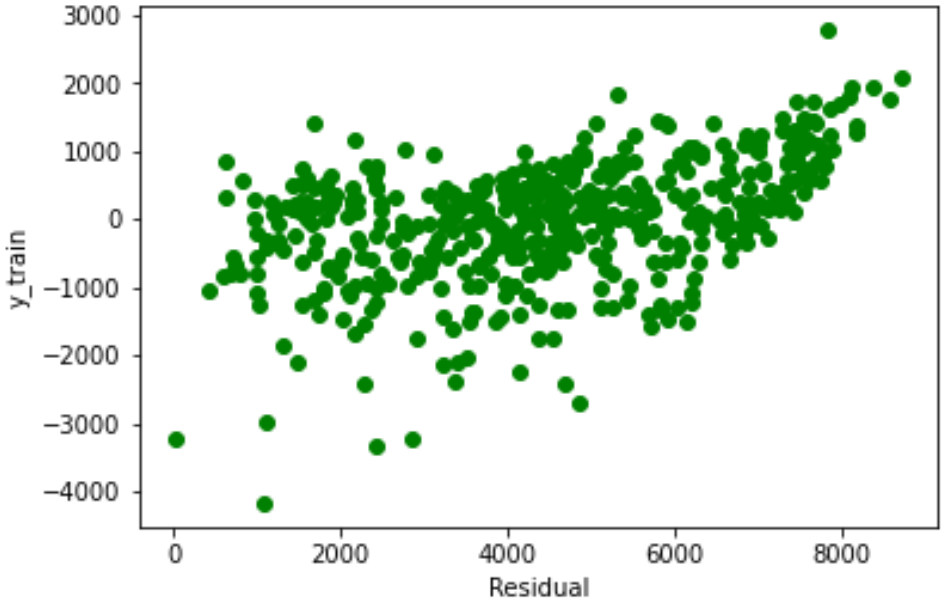
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer – ***"temp"*** and ***"atemp"*** has the highest correlation with the target variable.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

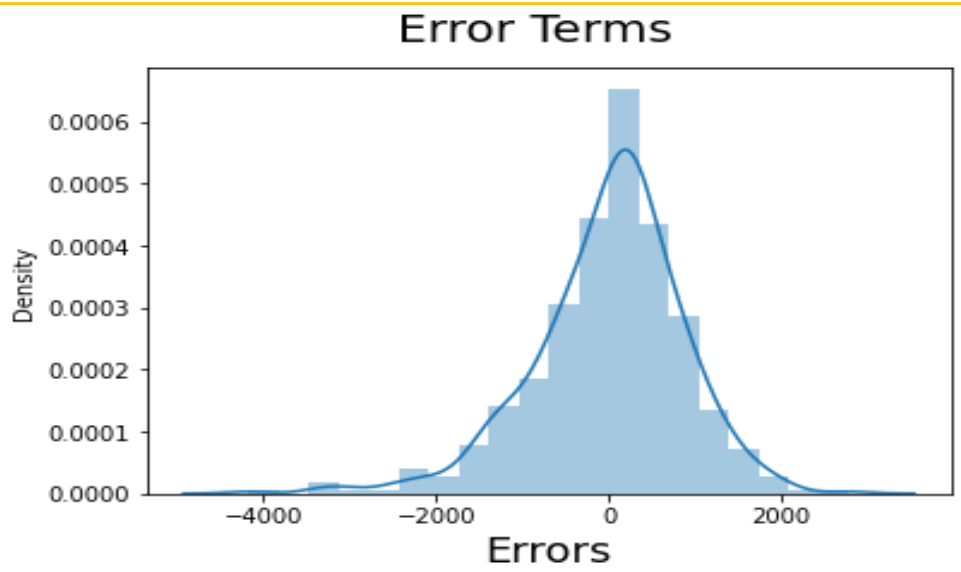
Linearity Check - Assumption 1 of Linear Regression



Observations-

1. We can see clear linear relationship between the X and y variables in the data.

Error terms are normally distributed with mean zero - Assumption 2 of Linear Regression

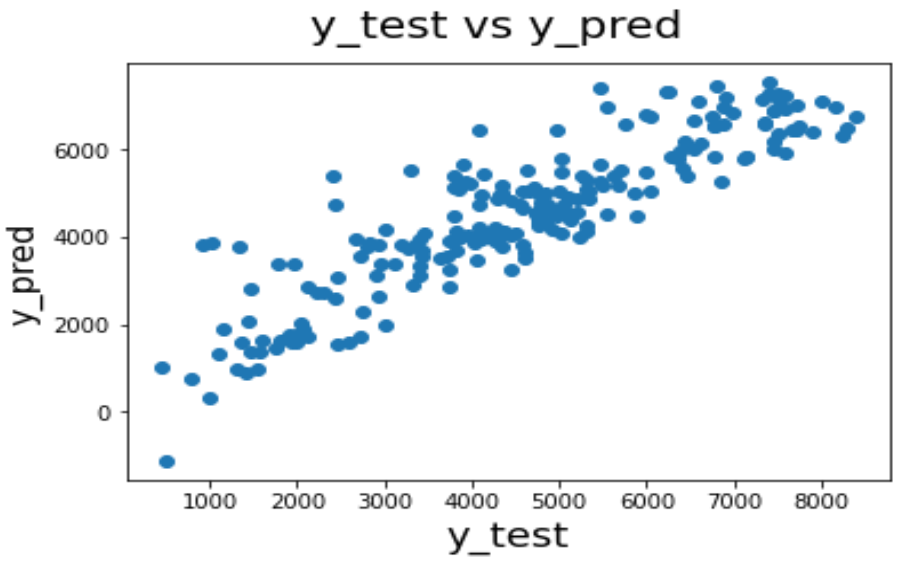


Observations-

1. Errors are normally distributed with mean 0 - which meets one of the key assumptions of Linear Regression

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

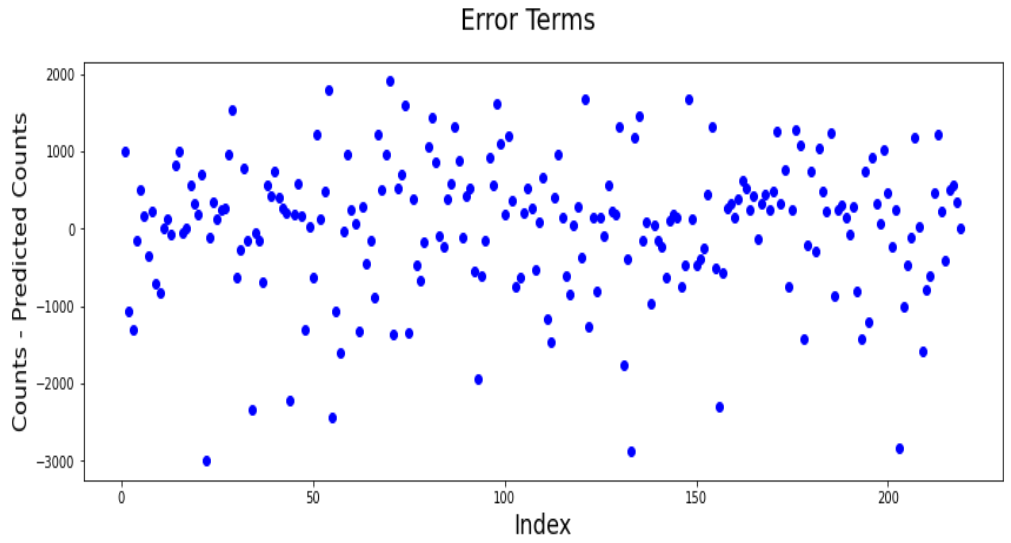
Error terms have constant variance (homoscedasticity) – Assumption 3 of Linear Regression



Observations-

1. The variance of the residuals (error terms) is constant across predictions & does not vary much as the value of the predictor variable changes - we can conclude the error terms are homoscedasticity in nature

Error terms are independent of each other – Assumption 4 of Linear Regression



Observations-

1. It is evident that the error terms are randomly distributed and doesn't have any pattern. This means that the errors of the response variables are not correlated. Presence of correlation in error of response variables reduces model's accuracy.

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

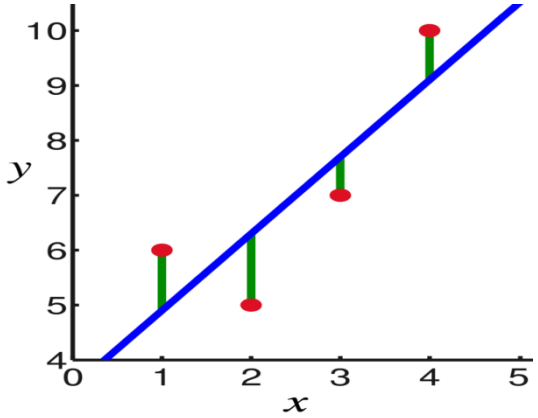
Answer – ***"Temperature", "Season"*** and ***"Weather"*** has significant contribution towards explaining the demand of the shared bikes.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

- A linear regression model attempts to explain the relationship between a dependent and 1 or multiple independent variables using a straight line.
- Linear regression uses the fact that there is a statistically significant correlation between two variables to allow you to make predictions about one variable based on your knowledge of the other.
- Based on the number of independent variables and the type of relationship between the independent and dependent variables regression techniques can be classified into 2 types –
  - **Simple Linear Regression algorithm** - *The standard equation of the regression line is given by the following expression*

y = β0 + β1.X – where y is the Dependent variable ,X is the independent variable,β0 = Intercept and β1 is the slope of the straight line. ***Note : It means if we increase 1 unit of X ,y increases by β1.***



In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (y) and an independent variable (x).

Q1. Explain the linear regression algorithm in detail.

- **Multiple Linear Regression algorithm** - *The standard equation of the regression line is given by the following expression*

$y = \beta 0 + \beta 1.X1 + \beta 2.X2\ldots..+ \beta p.Xp+\epsilon$

– where y is the Dependent variable ,X1,X2…XP are the independent variables

- β0 = Intercept
- β1, β2… β2p is the slope of the straight line for **"p" variables**

Note : It means if we increase 1 unit of X1 ,y increases by β1 – provided all other predictors are held constant.

Important Considerations

Best Fit Line - The line in the plot

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable

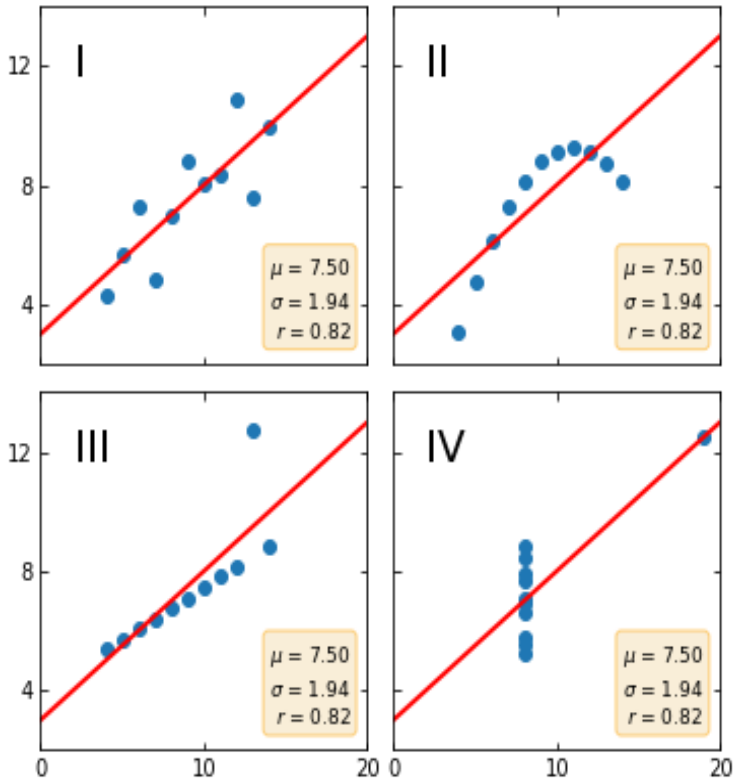
The **strength of the linear regression** model can be assessed using 2 metrics:

1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of 4 datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. Each dataset consists of eleven (x,y) points.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



Explanation of the 4 scatter plots

I. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

II. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

III. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated by the dot far away from that line.

IV. Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

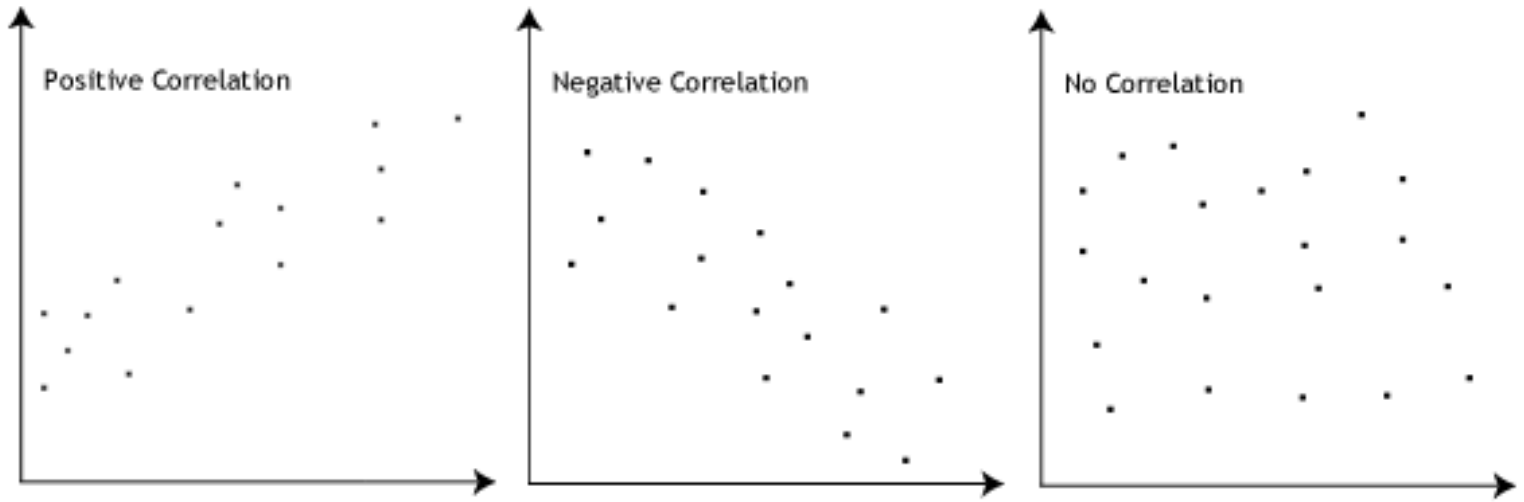
Q2. Explain the Anscombe's quartet in detail.

Conclusion

1) While all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analysed with a linear regression at all because it's a curvature.
2) Conversely, the top left graph probably should be analysed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing your data before analysing it.
3) Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets - if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.
4) Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analysing data, and statistics about a data set do not fully depict the data set in its entirety.

Q3. What is Pearson's R?

1) **Pearson's r** is a statistic that measures the linear correlation between two variables.
2) Pearson's r ranges from −1 to 1.
3) A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which *Y* increases as *X* increases.
4) A value of −1 implies that all data points lie on a line for which *Y* decreases as *X* increases.
5) A value of 0 implies that there is no linear correlation between the variables.



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

What is scaling?

Scaling (also known as **data normalization**) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

Why is scaling performed?

- Before running a model, such as regression (predicting a continuous variable) or classification (predicting a discrete variable), on data, it is almost always required to do some pre-processing. For numerical variables, it is common to either *normalize* or *standardize* the data available to get a best fit model.
- The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things —but for a model as a feature, it treats both as same.
- Thus feature scaling is needed to bring every feature in the same footing without any upfront importance.
- Feature scaling is essential for machine learning algorithms that calculate **distances between data**. If not scale, the feature with a higher value range starts dominating when calculating distances.
- Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization. For example, the majority of classifiers calculate the distance between two points by the distance. If one of the features has a broad range of values, the distance governs this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

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

What is the difference between normalized scaling and standardized scaling?

- All *normalization* means is scaling a dataset so that its minimum is 0 and its maximum 1. **It is also known as Min-Max scaling.**
- ***Standardisation*** is slightly different; it's job is to centre the data around 0 and to scale with respect to the standard deviation.

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

Answer –

- The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation. ***VIF = infinity shows there is perfect correlation between independent variables.***
- In the case of perfect correlation, **we get R2 =1, which lead to 1/(1-R2) infinity**.
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).
- VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others.

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

What is a Q-Q plot?

- When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot.
- Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution etc.
- 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.

Explain the use and importance of a Q-Q plot in linear regression.

It is used to check following scenarios for 2 given datasets-

- If two data sets come from populations with a common distribution
- If two data sets have common location and scale
- If two data sets have similar distributional shapes
- If two data sets have similar tail behaviour

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

Explain the use and importance of a Q-Q plot in linear regression.

Interpretations for two data sets.

1. All point of quantiles lie on or close to straight line at an angle of 45 degree from x – axis. It indicates that two samples have similar distributions.
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.
3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.
4. **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis