Linear Regression Assignment-based Subjective Questions

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

- Season - Season 3 (fall) has the highest booking rate whereas season 1 (spring) has the lowest compared to other seasons, Hence the rental company should plan their services accordingly.
- Year - Year 2019 was significantly better than the year 2018 in terms of rental counts
- Month - We see the month 5,6,7,8,9,10 have high mean compare to other months, which mean these months have good booking rate
- Holiday - we see maximum booking when it is not a holiday
- Weekday - we don't see a strong pattern in booking irrespective of the day of week
- Working day - we don't see a strong pattern in booking irrespective of whether it is a working day or not
- Weather situation - Weather situation (1) Clear, Few clouds, Partly cloudy, Partly cloudy is most optimal for rentals

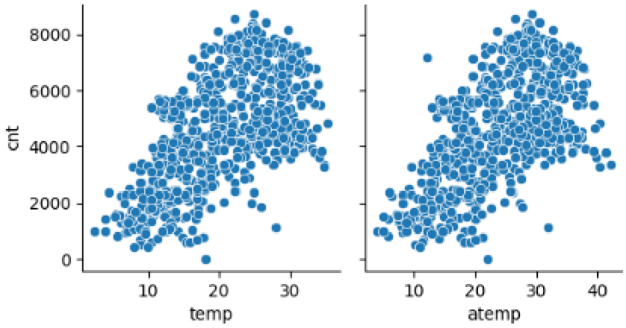
Why is it important to use drop_first=True during dummy variable creation?

- Your model must be built in a way that only relevant features must be included, which make your model light and improve interpretability. When creating dummies for categorical variables the data can be explained by n-1 features for n categories within that feature, for example. If season have 3 categories within it, you will need 2 dummy feature as value 0 for Season 1 and Season 2 indicates its Season 3

Season 1	Season 2	Season 3
1	0	0
0	1	0
0	0	1

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

- From the pair plot we can see that atemp and temp are highly correlated with the target variable



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

- You must make sure that none of your features have VIF and PE value greater than your threshold
- You must also check the R-squared and Adjusted R-squared so that the difference between them is minimal which mean all the features used are significant
- Once you have built your model, you must perform Residual Analysis to check if your error is normally distributed
- You can then make predictions on your test set and evaluate the r2_score for the actual test and predicted test values which will help us understand the efficiency of model on test data

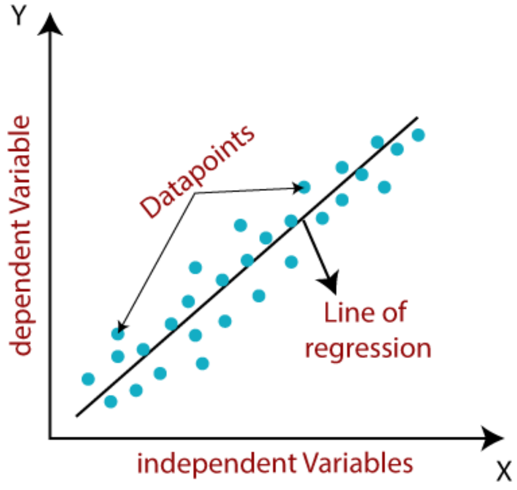
Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes

- The top three features contributing significantly are
    - Temperature
    - Weathersit (3) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - Year

Linear Regression General Subjective Questions

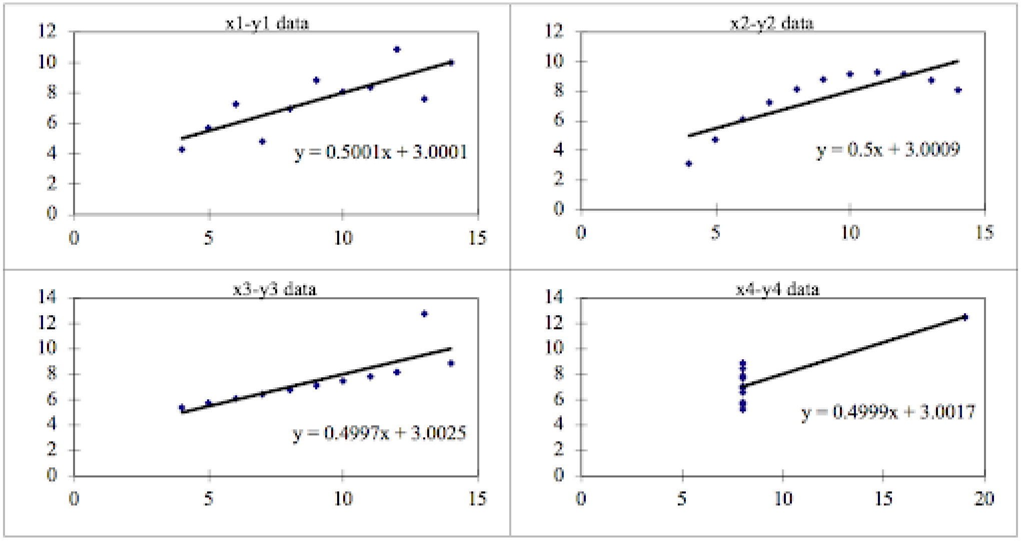
Explain the linear regression algorithm in detail

- Linear regression predicts the value of a dependent variable based on the value of other variables which shows some sort of linear relationship with the dependent variable. The variables which are used for predictions are called the independent variables.
- It estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression will try to fit a straight line or surface that minimises the gap between predicted and actual output values. Methods like the "least squares" method are used to discover the best-fit line for a set of paired data. You then estimate the value of the dependent variable from the independent variable.



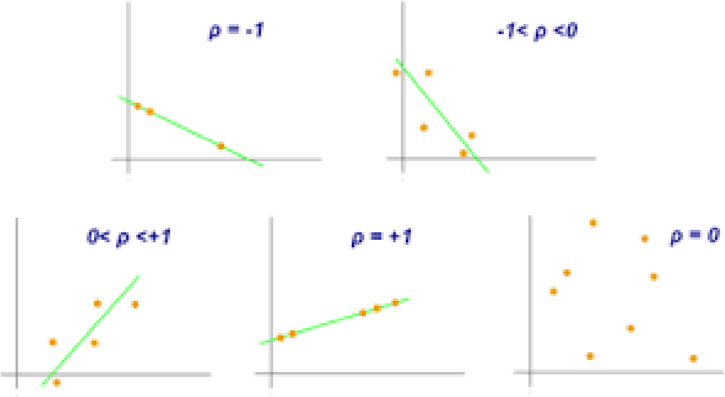
Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualise the data on scatter plots.
- Anscombe's quartet tells us about the importance of visualising data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.



What is Pearson’s R?

- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- Below are some degrees of correlation which is widely used to identify the strength of relationship
  - Perfect: If the value is near ± 1, then it is said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
  - High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
  - Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
  - Low degree: When the value lies below + .29, then it is said to be a small correlation.
  - No correlation: When the value is zero.



What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

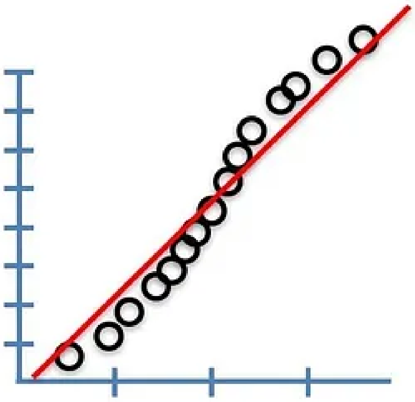
- Feature scaling is a method used to normalise the range of independent variables or features of data. In data processing, it is also known as data normalisation and is generally performed during the data preprocessing step
- Feature scaling is usually performed to help optimise the regression process, It makes the flow of gradient descent smoother and helps algorithms reach the minimum of cost function quickly.
- There are 2 widely used scaling technique normalised scaling and standard scaling which have some major differences
- Normalisation scales in a range of [0,1] or [-1,1]. Standardisation is not bounded by range.
- Normalisation is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.
- Normalisation is considered when the algorithms do not make assumptions about the data distribution. Standardisation is used when algorithms make assumptions about the data distribution.

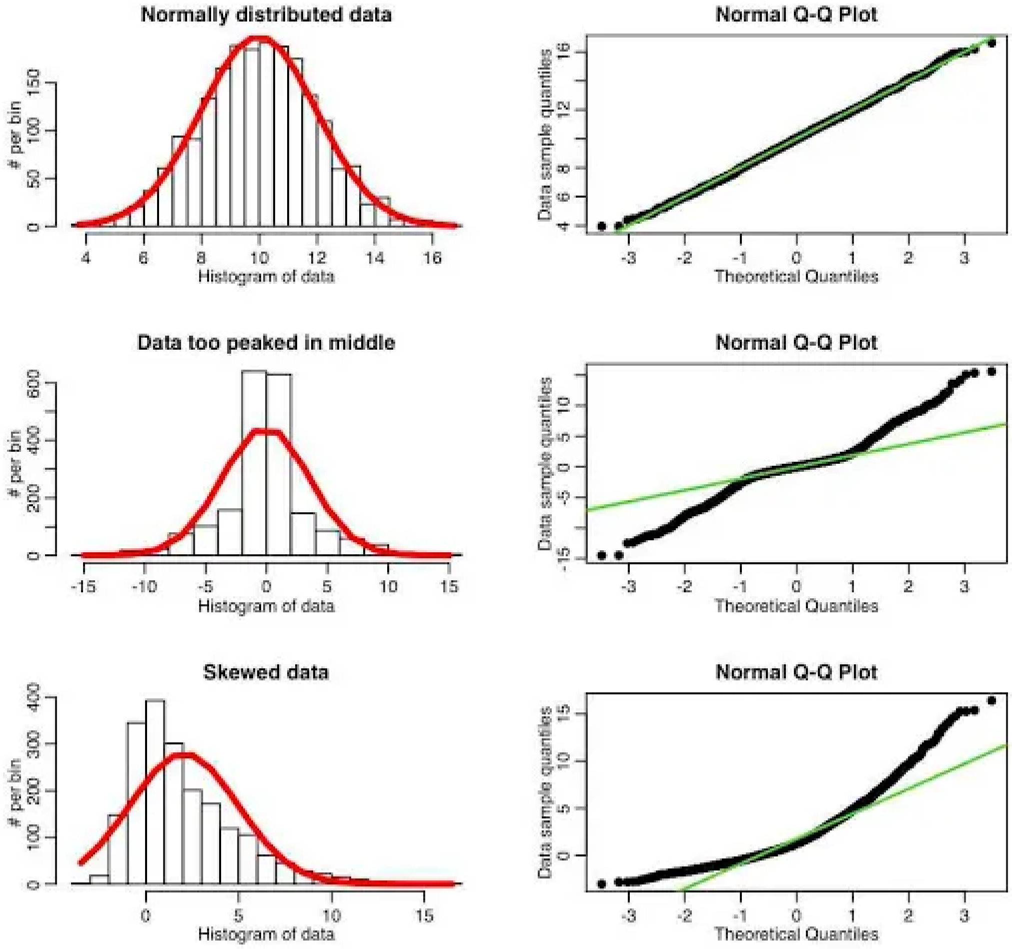
You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- VIF may have infinite values when two independent variables have a perfect relationship among them (collinearity), you must exclude one of the independent variables from your model to resolve this issue.
- In other words an infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model

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

- InStatistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare 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.




Normally distributed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Data too peaked in middle
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Skewed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles