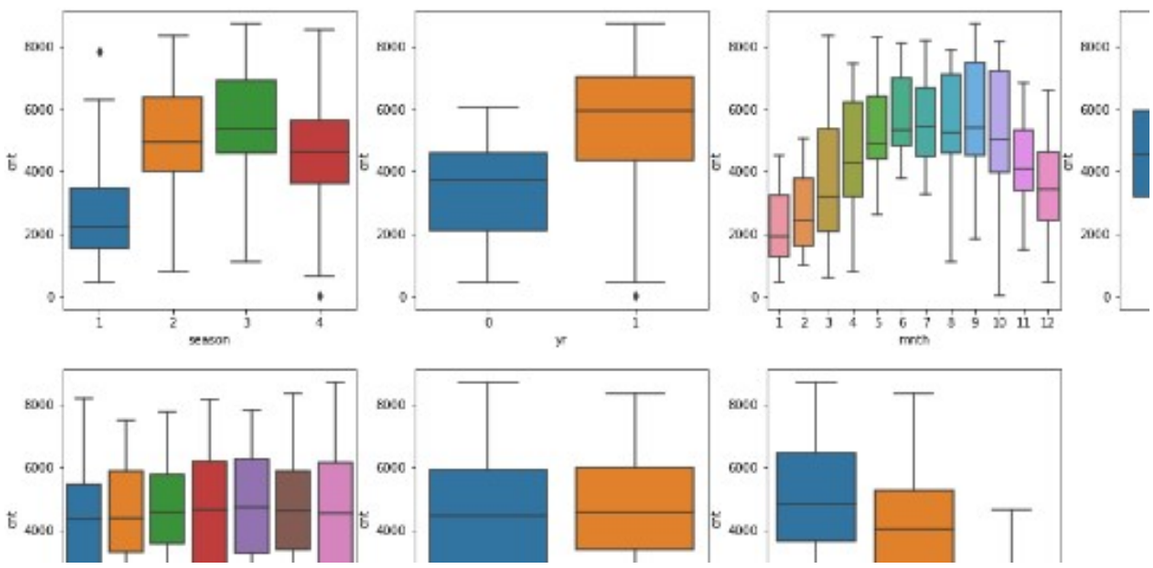
Assignment-based Subjective Questions

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

From above plot analysis we can say that :

- season 'fall' has high number of bookings
- year 2019 has high number of bookings as compared to 2018
- In the mid of year from Jun to Sep booking are increased as compared to other months in year
- If it is not holiday, than bookings are higher
- In the Clear weather bookings are higher
- Weekday and working day variables have no effect on booking.



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

While encoding categorical variables, information of n categories can be explained using n-1 dummy variables. For example Lets take weather situation example from our assignment

When we encode it with dummy variable we get like below:

weathersit	clear	snow	mist
clear	1	0	0
mist	0	0	1
snow	0	1	0

If we remove one column, information of that column can be obtained from other two. If we remove clear column, then other two having zeros means, clear is 1.

Therefore 1st column can be dropped during Dummy variable creation

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

Temp and atemp are highly correlated with target variable

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

There are below Assumptions in Linear Regression:

- Linear relationship between X and Y – As we analyze through plots, there is linear relationship between few variables and target variable
- Error terms are normally distributed with mean zero
- Plot between target and error residuals shows that error terms are not dependent on each other
- Error terms have constant variance(homoscedasticity)

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

Top 3 features affecting demand of shared bikes are:

- Temperature (0.378393): Temperature have high coefficient 0.378393, which means that with raise in Temperature demand is increasing.
- Year(0.234756) : 2019 as more number of bookings and it has positive coefficient, means with increasing year demand will increase
- Rain (-0.293938): Weather Rain is negatively affecting bike booking, which means that when it is rainy weather, people does not book for bike.

General Subjective Questions

1. **Explain the linear regression algorithm in detail**

Regression is type of machine learning algorithms which predict target based on input data. Target in regression is continuous in nature.

Linear Regression is that type of regression in which assumes linear relationship between independent variable and target and find out relation between variables

Mathematically, we can write linear regression equation like Y = mX+c

Where Y is target variable which model will predict,
X is independent variable which model will use for prediction,
m is slope of regression line
c is value of y when x = 0 (intercept)

There are 2 types of Linear Regression:

I. Simple Linear Regression : When there is only one independent variable to predict target variable, this is simple linear regression
II. Multiple Linear Regression: When more than one independent variable is used in prediction of target variable it is called multiple linear regression

Difference between models's predicted output and actual target is called residuals errors. Model attempts to explain relationship between dependent and independent variables by minimizing this error.

Assumptions in Linear Regression Model:

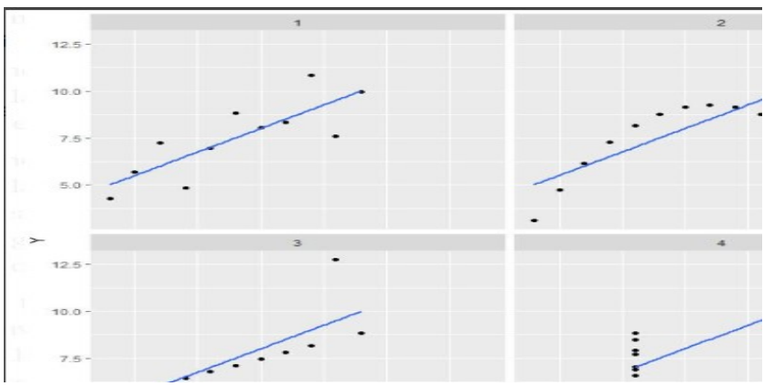
I. Linear relationship between X and Y : X and Y should display some linear relationship. Otherwise, there is no use of fitting a linear model between them.
II. Error terms are normally distributed with mean zero
III. error terms should not be dependent on one another
IV. Error terms must have constant variance (homoscedasticity), that is variance should not increase or decrease as error values change.

Evaluation in Linear Regression : Below matrices can be used in Linear Regression Evaluation

I. Root mean square error(RMSE)
II. Mean Square Error(MSE)
III. R-Squared

2. Explain the Anscombe's quartet in detail

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties (mean, std deviation and correlation), yet appear very different when graphed. Each dataset consists of eleven (x,y) points.



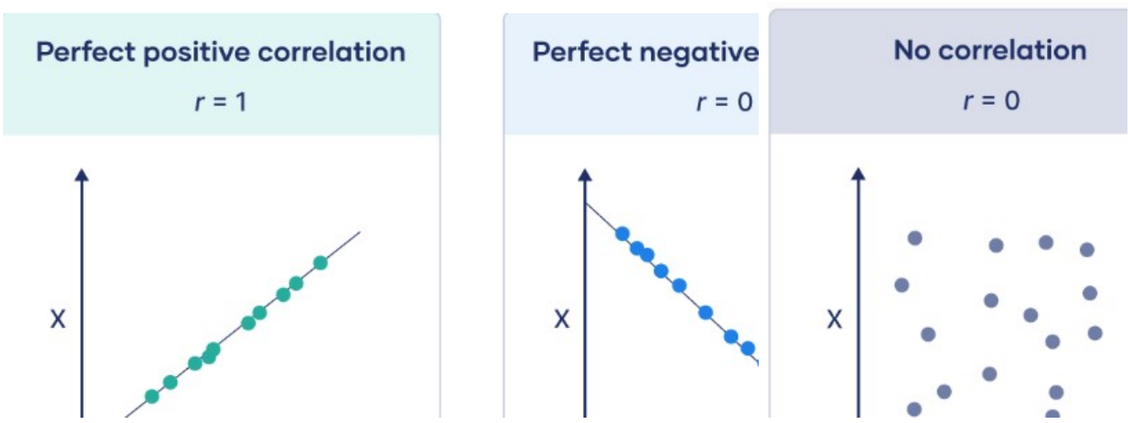
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets

3. **What is Pearson's R?**

Pearson's R is a measure of 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.

The Value of correlation coefficient is between 1 and -1. Here,

- -1 indicates a strong negative relationship
- 1 indicates strong positive relationships
- And 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?**

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Formula for is VIF = 1/(1-r2)

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two 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).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

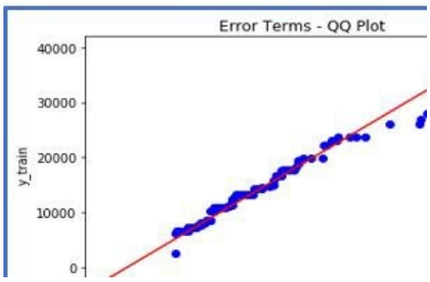
The purpose of Q Q plots is to find out if two sets of data come from the same distribution

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

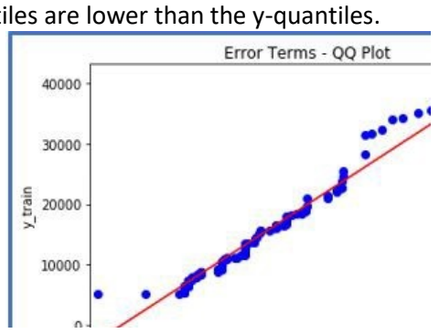
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis