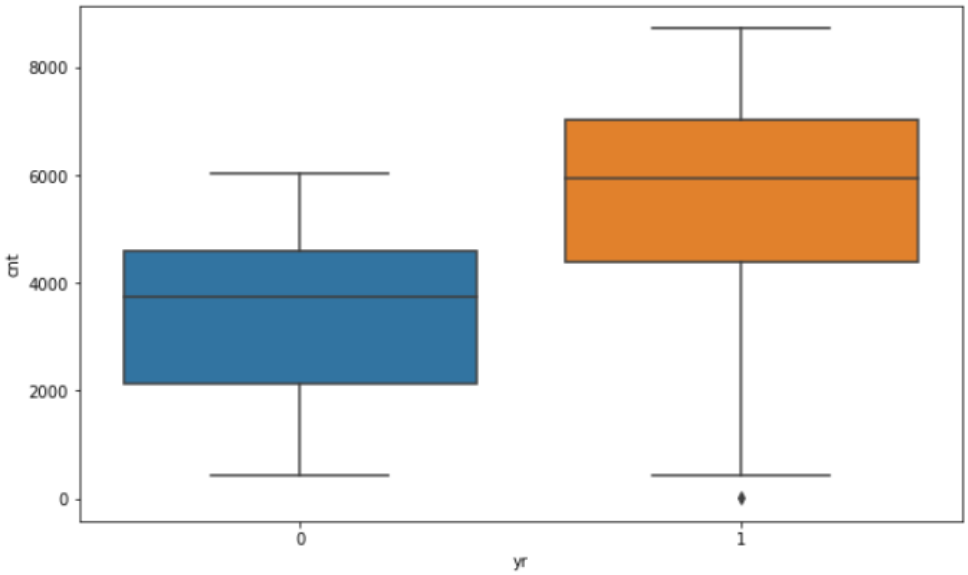
Assignment-based Subjective Questions

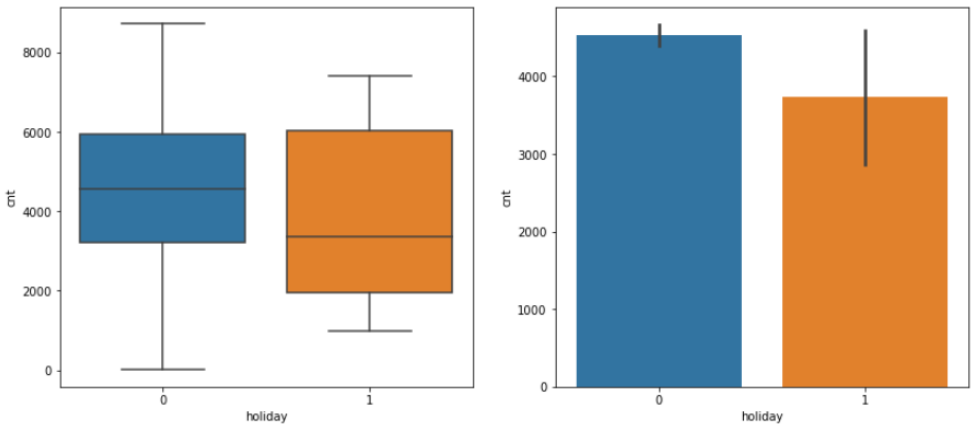
1. **. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans. Let's observe one variable at a time with the dependent variable 'cnt'

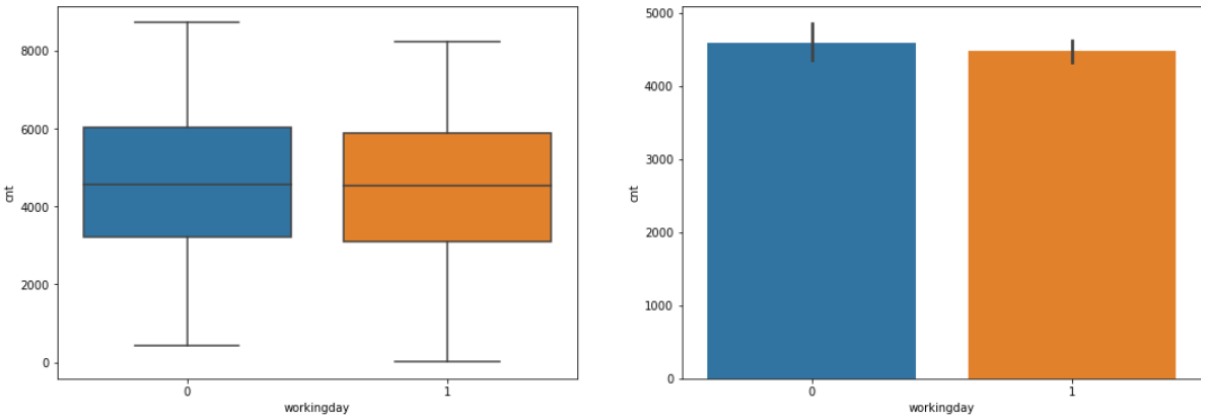
1) **'Season' variable** – We can see from the heat map in the jupyter notebook that in the seasons only spring shows the best correlation with the 'cnt' variable. It is affecting the 'cnt' varibale negatively. All the other season don't affect the dependent variable too much.
2) **'Yr' variable –** Just looing into the below box plot we can conclude that in 2019 which is here represented by '1' the demand of the bike drastically increases compare to 2018 (represented by '0').



3) '**holiday' variable** : We can say that on holidays(represented by 1) demand is low compare to days when it was not holiday(represented by 0).



4) **'workingday' –** We can see that there is no significant change in demand in working and non working days.

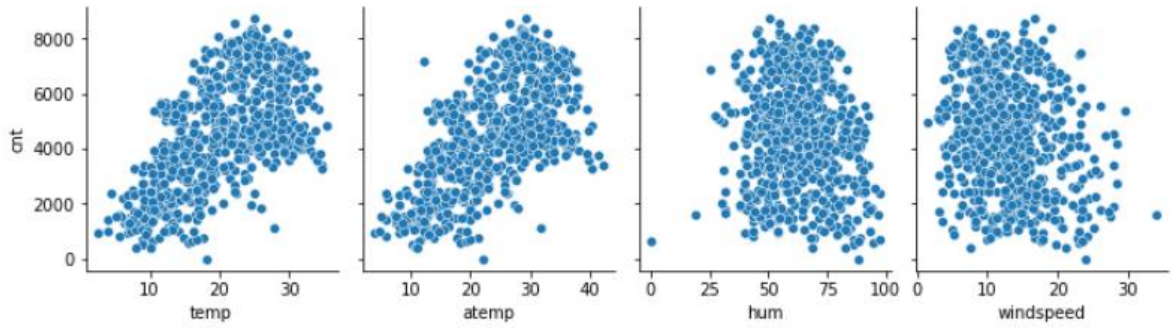


2 . Why is it important to use drop_first=True during dummy variable creation?

Ans. Because using this statement we can drop the first column of our dummy varibales column as we only need (n-1) dummy variable in n categorical column.

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

Ans. 'temp' and 'atemp' both have the highest correlation of 0.64 and 0.65 respectively with the target variable 'cnt' . you can refer below plot.



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

Ans. By calculating the R_squared score on test set using the following code we can validate our assumptions on the test set and how much it was reliable for our predications.

```
In [721]: #calculating the R-squared score on the test set

          from sklearn.metrics import r2_score
          r2_score(y_test, y_pred)

Out[721]: 0.8216461699732155
```

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

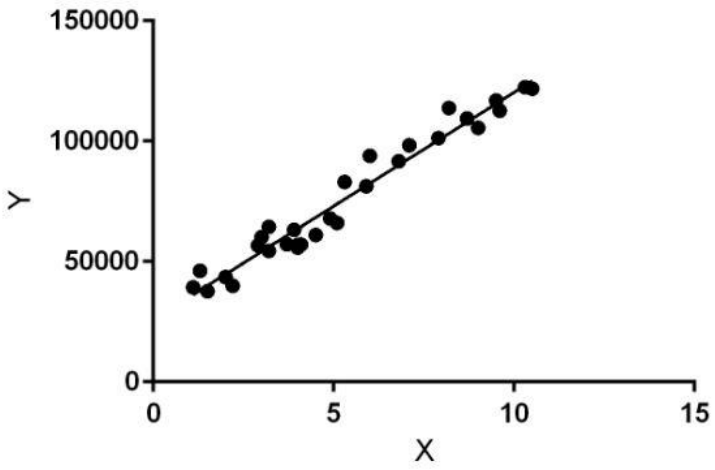
Ans. By looking into the coefficients below three features contributing significantly explaining the demand of the shared bikes –

1) Temp
2) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
3) Yr

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised machine learning algorithm. The algorithm models a target prediction value based on the independent variables. It performs a task to predict a dependent varibale(y) with respect of one or many independent variables(X). It fits a line between the dependent variables which represents linear relationship between them. Hence, It is said linear regression. For example in the figure below, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.



The line can be modelled based on the linear equation given below-

Y=a0+ a1x1 + a2x2……+ anxn + e

Here c = coefficient

X = independent variables or predictor variables

Y = dependent varibale or target variable

E = error

A0 = constant

A1…an = coefficient

The motive of the linear regression is to find the best values of the coefficient and constant. To find out the best values we use the cost function which is explained below.

Cost Function –

The cost function helps us to find the best fit line by minimising the difference or error between predicted and actual value.

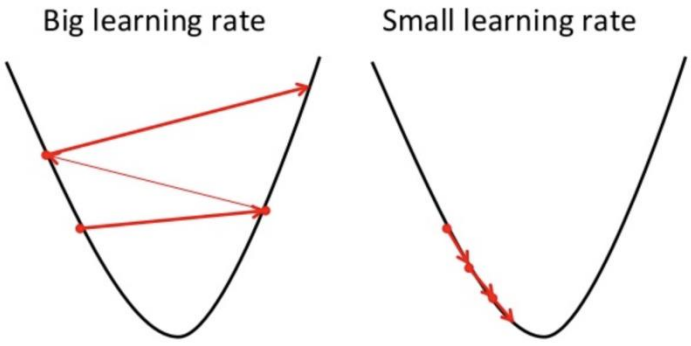
$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

We choose the above formula to minimise the error. The difference between the predicted values and the ground truth provide the error difference. Then we square the error difference and sum over all the data points and divide that value by the total number of data points. Therefore, the function is also known as mean squared error (MSE) function.

Gradient Descent –

Gradient descent is a method of updating a_0 and a_1 to reduce the cost function(MSE). The idea is that we start with some values for a_0 and a_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.



To draw an analogy, imagine a pit in the shape of U and you are standing at the topmost point in the pit and your objective is to reach the bottom of the pit. There is a catch, you can only take a discrete number of steps to reach the bottom. If you decide to take one step at a time you would eventually

reach the bottom of the pit but this would take a longer time. If you choose to take longer steps each time, you would reach sooner but, there is a chance that you could overshoot the bottom of the pit and not exactly at the bottom. In the gradient descent algorithm, the number of steps you take is the learning rate. This decides on how fast the algorithm converges to the minima.

To update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to a_0 and a_1.

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)^2$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i) \implies \frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i) \cdot x_i \implies \frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

$$a_0 = a_0 - \alpha \cdot \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

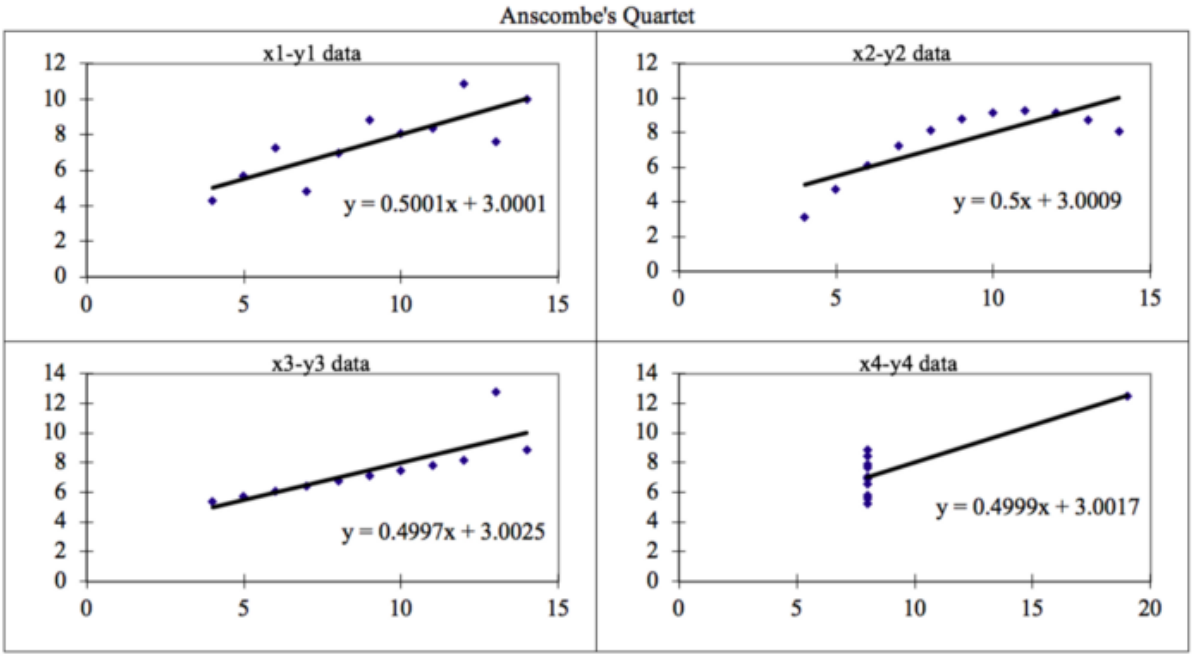
The partial derivatives are the gradients and they are used to update the values of a_0 and a_1. Alpha is the learning rat.. A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that you could overshoot the minima.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations,

which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



The four datasets can be described as:

1. Dataset 1: this fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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?

Ans. Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

There are two types of Correlation –

Positive linear relationship: If one varibale increases then other also increases

Negative linear relationship: If one variable increases then other variable decreases and vice versa.

Pearson Correlation Coefficient formula –

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Ans. Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- 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?

Ans. 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.

Ans. *Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common 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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

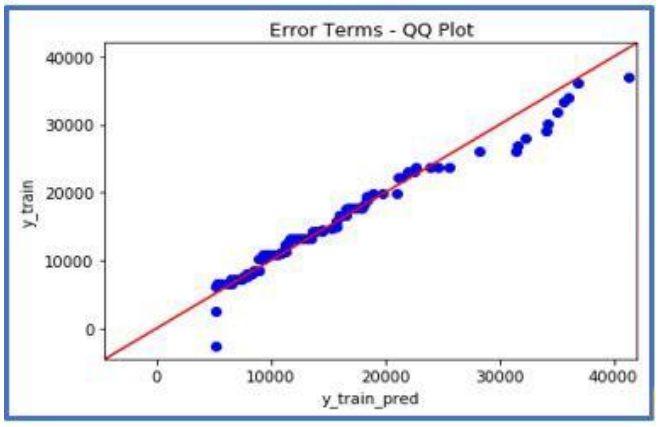
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

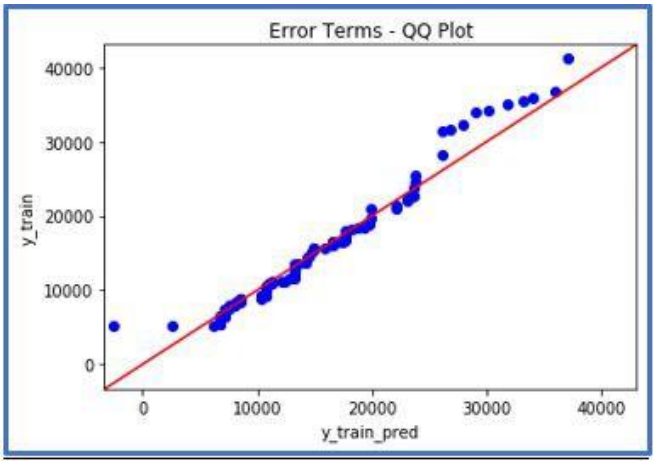
Below are the possible interpretations for two data sets.

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*