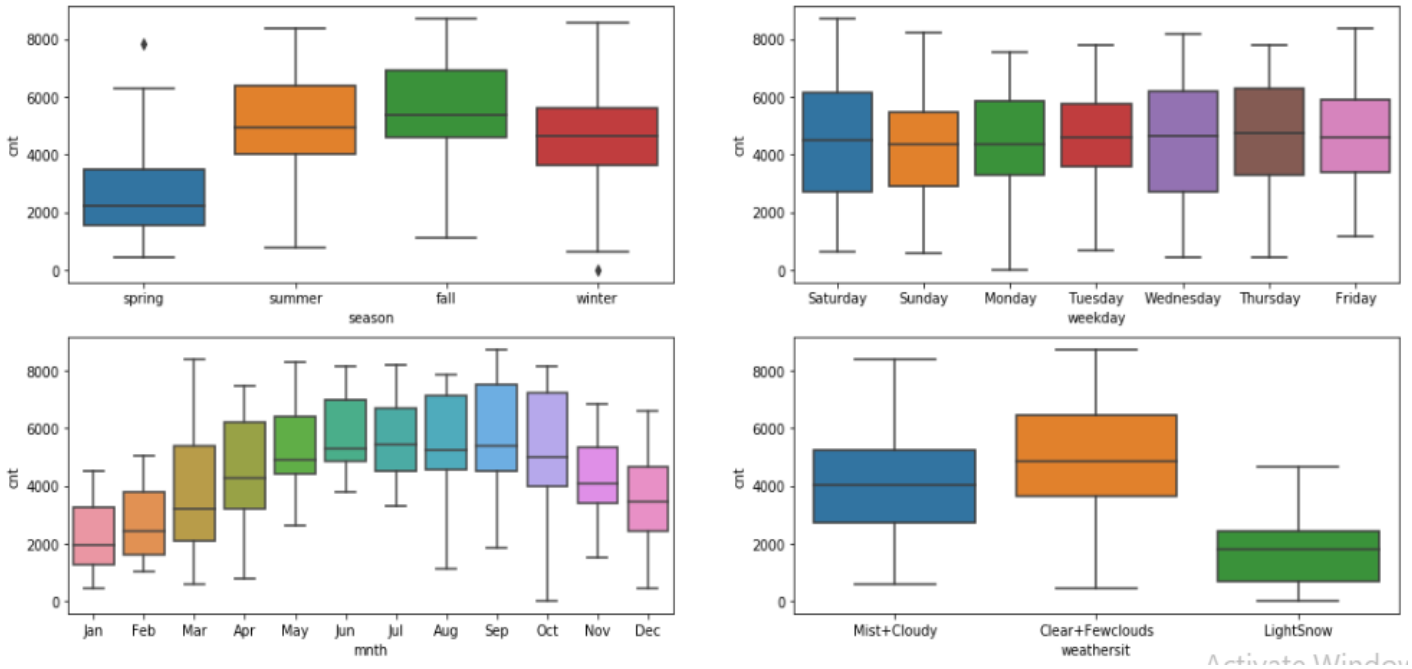
Assignment-based Subjective Questions

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

Effect of categorical variables on dependent variables can be explained as follows.

Demand of bikes in season fall is the highest while in spring it is the lowest. Demand is almost similar on all weekday but on Sunday it is lowest. In Jan & Feb demand is comparatively low but in Sep & Oct demand is high. Demand of bikes is low Lights snow weather situation & demand is high when weather is clear.



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

Because if we have n levels in categorical variable then we need only n-1 level so drop_first=True will drop 1 level from categorical variable & hence reduce complexity.

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

Temp & atemp

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

By drawing the distribution plot of residuals/error terms(y_train-y_train_pred) & if this plot is normally distributed around 0 then it follows the assumptions of linear regression.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

temp,
windsit(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
yr

General Subjective Questions

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

Linear Regression algorithm finds the relationship between dependent & independent variable in the form of

Y =B0+B1X1+B2X2+…..BnXn.

where y is dependent variables and X1,X2…Xn are independent variable.

B0 is intercept

B1,B2..Bn are coefficient.

Linear Regression Types:

1. **Simple Linear Regression:-**When number of independent variable is only 1.
2. **Multiple Linear Regression:-**When there is more than 1 independent variables.

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.

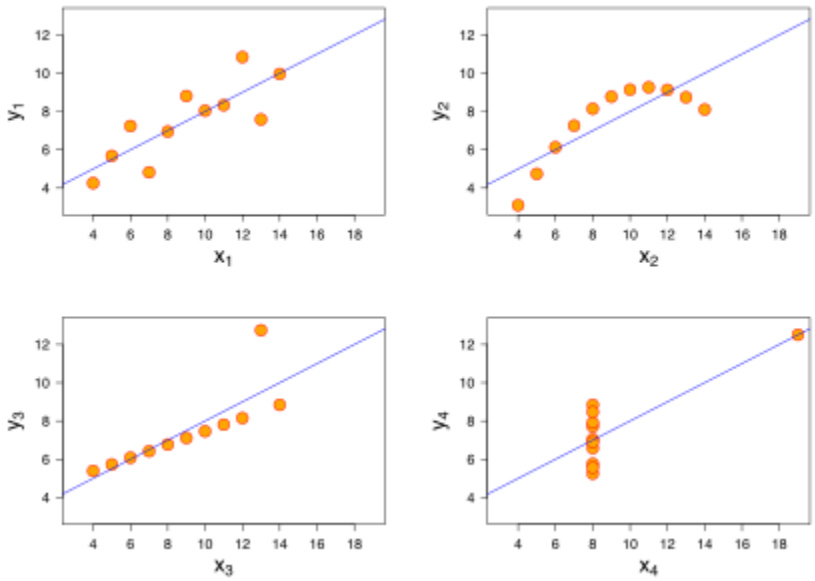
$$\sum_{i=0}^{n}(y - B0 - BiXi)\text{^}2$$

Gradient descent is an optimization algorithm used to find the values of the parameters (coefficients) of a function (f) that minimizes a given cost function (cost).

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

Anscombe's quartet explains that even though Set of 4 data set have identical simple descriptive statics like mean,variance,correlation coefficient,line of best fit, they have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

It demonstrates both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

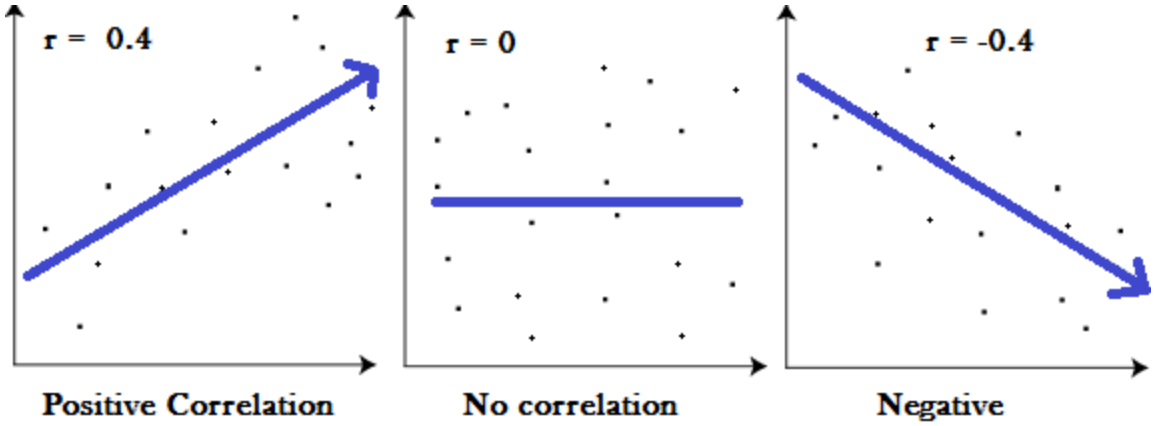
3. What is Pearson's R? (3 marks)

Pearson's R(also called Pearson Correlation Coefficient) measures the strength of linear relationship between two variables.Its value lies between -1 to 1.

R=1(perfect positive correlation which means as x increases y also increases)

R=-1(perfect negative correlation which means as x increases y also decreases)

R=0(no correlation which means data points are randomly distributed and there is no relation between x & y)



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

It is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

1**. Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x=(x-\text{mean}(x))/\text{sd}(x)$$

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x=(x-\min(x))(/\max(x)-\min(x))$$

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

It happens because of perfect correlation between variables where R^2 is 1 and hence VIF is infinity. And so there is very high collinearity.

$$VIF = \frac{1}{1 - R^2}$$

So when R^2 is 1 VIF is infinity

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

(3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Q-Q plot used to check if

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