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

The categorical variables in our data set as well as its impact on our dependent variable i.e.'Cnt' are as follows:

- Season - The maximum percentage of Bike share count occurs in the fall season (32%) followed by summer (28%), winter (25.5%) and spring (14.26%) season. The median and mean bike count follows the same order.

- Month - The maximum percentage of Bike share count occurs in the months of August (10.67%) followed by June (10.52%), September (10.51%) and July (10.48%) season. The median value of bike share count is highest for July (5446) whereas the highest mean value of bike share occurs in the month of June (5772). The months of January and February experience the lowest average bike share count.

- Holiday - Almost 97% of the total bookings happens when it is not a holiday with a median value of 4563.
- Weekday - The maximum percentage of Bike share count occurs on Thursday (14.82%) followed by Sunday (14.74%). The median and mean bike count follows the same order with values as (4676,4691) and (4590,4665) respectively.

- Weathersit - The maximum percentage of Bike share count occurs when the weather is clear (68.6%) followed by misty conditions(30.23%). There were no bookings that were made when the weather outside was too extreme. The median and mean bike count follows the same order.

- Working day – The bike share is higher for when it is neither a holiday nor weekend (69%).
- Year – We can see that there has been an increase in the number of bike share count from 2018 to 2019. The median for 2018 was 3740 bookings as against 5936 bookings in 2019.

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

The parameter drop_first = True allows us to remove an extra column while creating dummy variables for our categorical variables and helps us in reducing correlation among dummy variables. When we create dummies, if we do not drop the first column, we always run the risk of employing redundant variables in our model, which might hamper the accuracy and prediction power of our model.

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

The variable ‘atemp’ has the highest correlation of 0.630685 with the target variable followed by variable ‘temp’ with correlation of 0.627044.

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

There are a few assumptions that we need to make while building the model and must ensure that our final model does not violate these assumptions. The assumptions are as following:

- Linear Relationship between predictor and target variables, which can be validated through scatterplots or pair plots.
- The mean of error terms should be zero, which can be validated using a simple computation of mean of all the error terms.
- The error terms must be normally distributed which can be validated through plotting a kdeplot or a displot using seaborn library. QQ plots can also be used to see the distribution of error terms.
- Homoscedasticity i.e. a situation where the variance of the residual or the error terms in a regression model is constant. The error term should not vary much as the value of the predictor variable changes. This can be validated creating a scatterplot of dependent variable and the residual and ensuring that there are no identifiable patterns that can be found in the plot.

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

The top three features contributing significantly towards explaining the demand of the shared bikes are:

1. Temperature: A coefficient value of 0.406863 reflects that our predictor variable moves our dependent variable 0.4068 times in the positive direction with every 1-unit change in temperature.

2. Snow, Light Rain - A coefficient value of 0.2786 reflects that our predictor variable moves our dependent variable 0.2786 times in the negative direction with every 1-unit change in Snow, Light Rain. Bike rentals decreases when there is light snowfall or light rain.

3. Year - A coefficient value of 0.2450 reflects that our predictor variable moves our dependent variable 0.2450 times in the positive direction with every 1-unit change in Year.

Q1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning where the result is predicted by the use of known parameters which are correlated with the output. Linear Regression algorithm shows a best fit line or curve that passes through all the data points in such a way that the vertical distance between the data points and the best fit line is minimum which is also referred to as the cost function of linear regression.

There are two types of Linear regressions:

-Single linear Regression where we are trying to establish a relationship between one independent variable and one dependent variable.

- Multiple Linear Regression where we establish a relationship between one dependent variable and multiple predictor or independent variables.

There are various algorithms that can be used in linear regression:

-OLS(ordinary least square) : In this method, we are trying to minimize the sum of squares of all errors or residuals. Residuals are defined as the difference between the y-coordinates of the actual data and the y-coordinates of the predicted data.

- LAD(Least Absolute Deviation): The OLS method is affected by outliers. To negate the impact of outliers , we use least absolute deviation method, where we use the absolute values of our error terms instead of squaring them.

-Huber –M cost: This method employs the best qualities of both OLS and LAD. As soon as we reach a predetermined level, the algorithm switches from OLS to LAD. This method applies Least squares when the absolute residual is small and switches to LAD when the absolute residual becomes large.

Linear Regression Algorithm has two forms of optimisation methods for its cost function:

-Closed form solution: In this method we find the optimal value of theta in just one step by equating our cost function to 0.

-Iterative Form solution: Gradient Descent approach is used in this method to minimize the cost function by repeatedly decreasing the value of theta till our cost function becomes 0.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet refers to a group of four sets of 11 data-points that were created and published by a statistician named Frank Anscombe in 1970s. Below are the given sets:

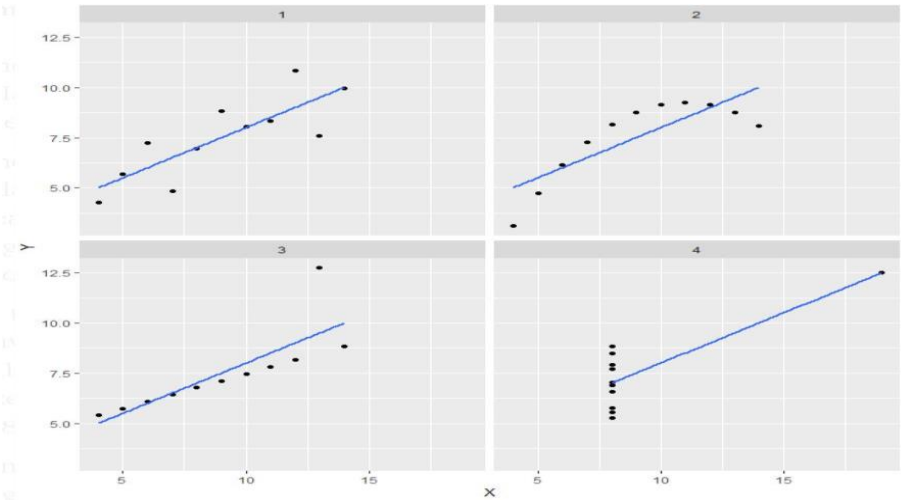
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The interesting thing about these datasets is that these sets all have the same summary statistics i.e. they have the same mean of X, same variance of X, same mean of Y, same variance of Y, same correlation, same slope of the line, same intercept of the line. The summary statistics is given below:

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

The summary statistics of all of these different datasets are the same but the overall picture is quite different. The underlying principle behind Anscombe's quartet is that summary statistics itself do not tell the whole story and it is always a good practice to plot your data. The following observations can be made if you plot the data-points:



In the first figure you can see a linear relationship between X and Y on a scatterplot. In the figure next to it, if you see clearly there seems to be a non-linear relationship between X and Y. The third figure shows X and Y to be in a perfect linear relationship except for one data point that happens to be an outlier. The fourth figure explains how one high leverage point is enough to produce a high correlation coefficient.

The quartet reflects on the importance of visualising a dataset graphically before analysing and the possibility of inadequacy of basic statistic properties for describing realistic datasets.

Q3. What is Pearson's R?

Pearson's R is a statistic that measures the strength of association between two continuous variables as well as the direction. Pearson's R is sometimes referred to as Pearson's Correlation Coefficient or bivariate correlation. Like any other correlation, it can take up any numerical value between -1 and 1. There are certain requirements that needs to be met in order to calculate Pearson's r:

-Scale of measurement should be interval or ratio.

-The association should be linear.

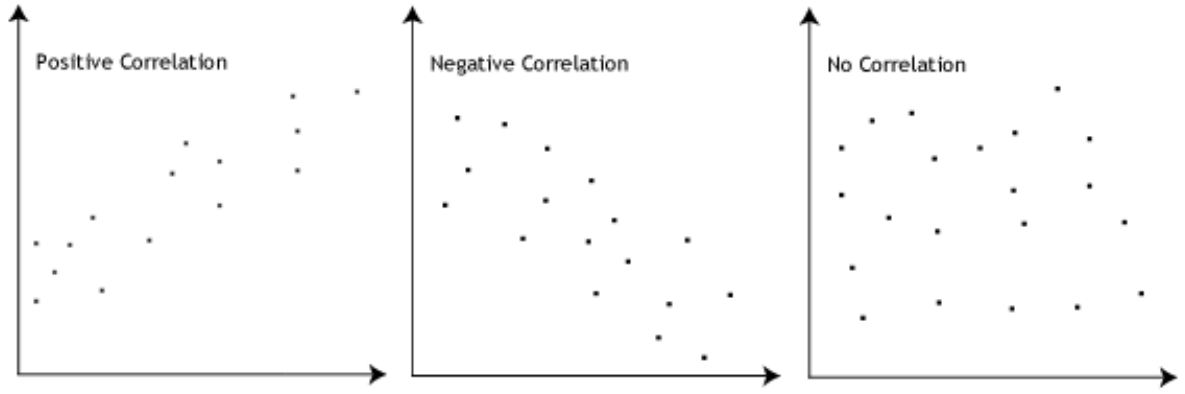
-absence of outliers in the dataset.

-Variables should be approximately normally distributed.

Formula for calculating Pearson's r:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearson's r is basically a normalised measurement of covariance such that the result always lies between -1 and 1.



The first figure shows a positive correlation meaning that both variables tend to change in the same direction. The second figure shows a negative correlation meaning that both variables tend to change in different directions. The third figure indicates that there is no linear relationship between the variables.

Q4. 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 in which independent variables or features of the dataset are either normalised (bound within a range between 0 and 1) or standardised(with mean 0 and

variance as 1) in order to make different attributes similar in some context for possible comparisons.

Scaling is performed in order to ensure that the magnitude of a particular feature does not skew or impact the outcome of a machine learning model. A machine learning algorithm simply works on a number and does not know what that number represents. For e.g. a machine learning algorithm will treat a weight of 10 grams and a price of 10 rupees as same as it can't distinguish between the two units. Scaling is must for algorithms that calculate distances between data as the interpretation of the model changes with the change in unit. Scaling does not change either p-values or model accuracy, it only changes the coefficients of the predictors.

The most common feature scaling techniques are standardisation and normalisation. We use normalisation when we want to bound our values within a range typically [0,1]. Standardisation does not bound variables within a range and transforms the data to have zero mean and a variance of 1, essentially making them unit less. Normalisation is generally better suited to handle outliers as the maximum or minimum data-point in our set remains between 0 and 1.

The formula for normalisation is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In normalisation the normalised value is obtained by subtracting the data point with the minimum value divided by the difference between the maximum and minimum data point.

The formula for standardisation is :

```
X_new = (X - mean)/Std
```

The standardised value is obtained by subtracting the value with mean and dividing by the standard deviation.

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

VIF or Variance Inflation Factor assesses how much the variance of an estimated regression coefficient increases if the predictor variables are correlated.

If no factors are correlated , then VIF will be 1. A VIF between 5 and 10 indicate high correlation among the predictor variables.

If the VIF goes beyond 10 , it is safe to assume that the regression coefficients are poorly estimated due to multi-collinearity.

A VIF on infinity therefore indicated a perfect correlation between two independent variables. When there exists perfect correlation, the value of R2 becomes 1 which results in VIF being infinity.

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

Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. They are used to find the type of distribution for a random variable. It can be normal ,Gaussian or exponential distribution. In Q-Q plots we plot the theoretical quantiles on the x-axis and the ordered values for the random variable on the y axis. The shape and distribution of these ordered pairs on the X-Y axis tells us whether there is a relationship between the variables or not. If all the points plotted on the graph perfectly lies on a 45 degree line then we can say that the distribution is normal.

Q-Q plot can also be used to find the skewness of a distribution. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not ,then we can say that the distribution is left skewed. If the bottom end falls on the straight line and the upper end deviates from the line, then we can say that the distribution is right skewed.