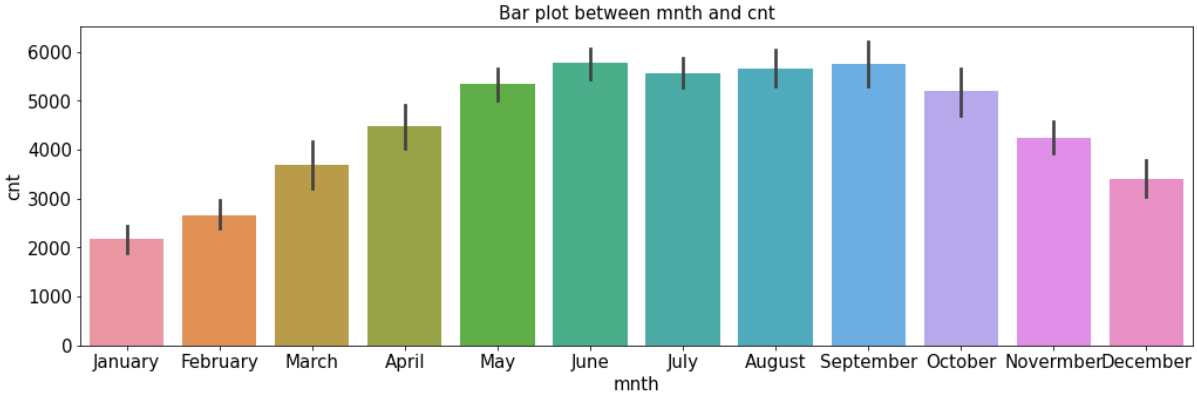
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

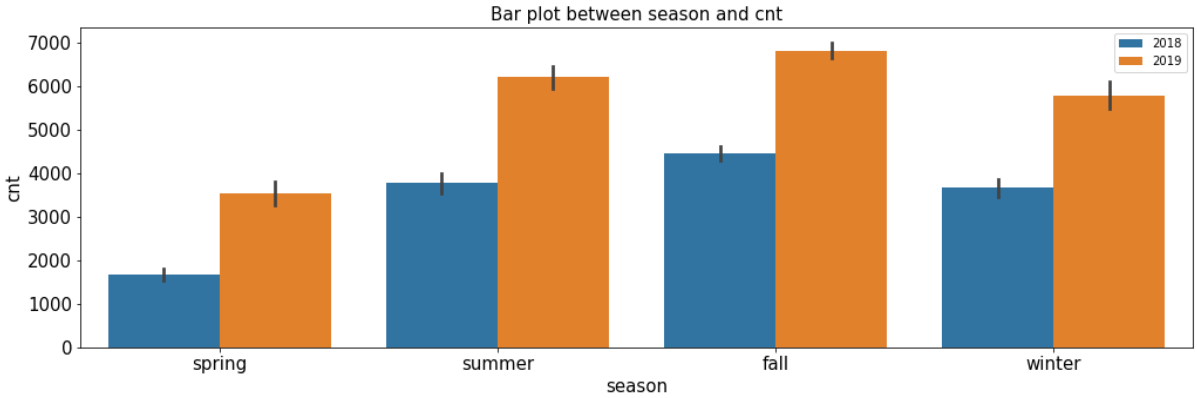
Ans: Several analysis have been made in the EDA slot and will be explained one by one in this assignment.

month vs Count:



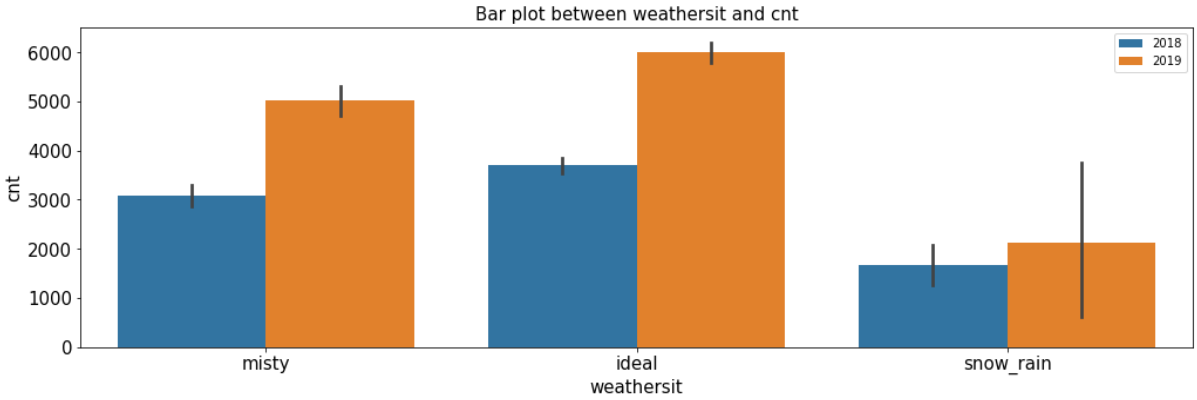
From the projected plot the analysis b/w the count and the month we can able to analyze that on June,July,august the most sales are happening in the yearly Basis. So the company can do more sales on these months

Season vs Count with hue as year:



In the summer and the fall seasons more rides are happened that too year after year the growth in between that months also have been increased.

Weather situation vs Count with hue as year:



People usually tend to rent bikes when the weather is Ideal or Mist(with some visibility) we observe that the count goes up both for 2018 as well for 2019 for both weather conditions

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

ANS: Let's break this down step by step by first learning how dummies are made. There is no harm in keeping that extra category; it is just there for the convenience of our model. We will explain this better with the help of an example. When we use pandas' get dummies() function to create dummies, it creates "K" dummies. The reason we use drop first = True is to drop one level of category that is beneficial for our model, so it can converge or rather learn at a faster rate.

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

Ans: The dependent variable has the highest correlation with temp and atemp, although we later omitted atemp due to its high correlation with temp.

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

Ans: The following are the four assumptions of linear regressions; I will first describe each one before moving on to how I validated it.

1. There should be some sort of linear link between X and Y, as we have seen between temperature and count, as well as a modest linear relationship between windspeed and count, where count was the variable we were aiming for.

2.Error terms should be normally distributed. We used our training set to plot the residuals, which are simply the error terms (y train - y train pred). y train pred contains the values predicted by our model using the training set. We plotted the residuals using a histogram, which revealed that the error terms were normally distributed around zero, supporting our second assumption.

3.Error terms should be independent - To test this, we plotted our residuals on the X axis and the y train values that were used to train our model on the Y axis, and saw that there was no relationship between them. We also saw that the points were randomly distributed, which was sufficient to support our third assumption.

4. We plotted the residuals with the expected values and saw a linear trend since the projected value (line) was sufficiently close to the actual points. Error terms should have constant variance.

5)Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANS: The top three features contributing significantly are

1.Temp

2.Ideal_weather_Conditions

3.summer_season

General Subjective Questions

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

ANS: Linear regression is a type of machine learning algorithm that is based on supervised learning, meaning we already know the outputs we can compare our projected values with and simultaneously improve our model. Based on other values, this regression model predicts a value (the dependent value) (independent variable). It seeks to mathematically fit the best line through the provided collection of data points; we may say that the line represents the anticipated value and the other set of points represents the actual values.

Linear Regression is of two types:

1) **Simple Linear regression (Y = mX + C):** It has one dependent value and one independent value
2) **Multiple Linear regression (Y = C + m1X1 + m2X2 +.....+ mnXn) :** It has one dependent variable and multiple independent variable.

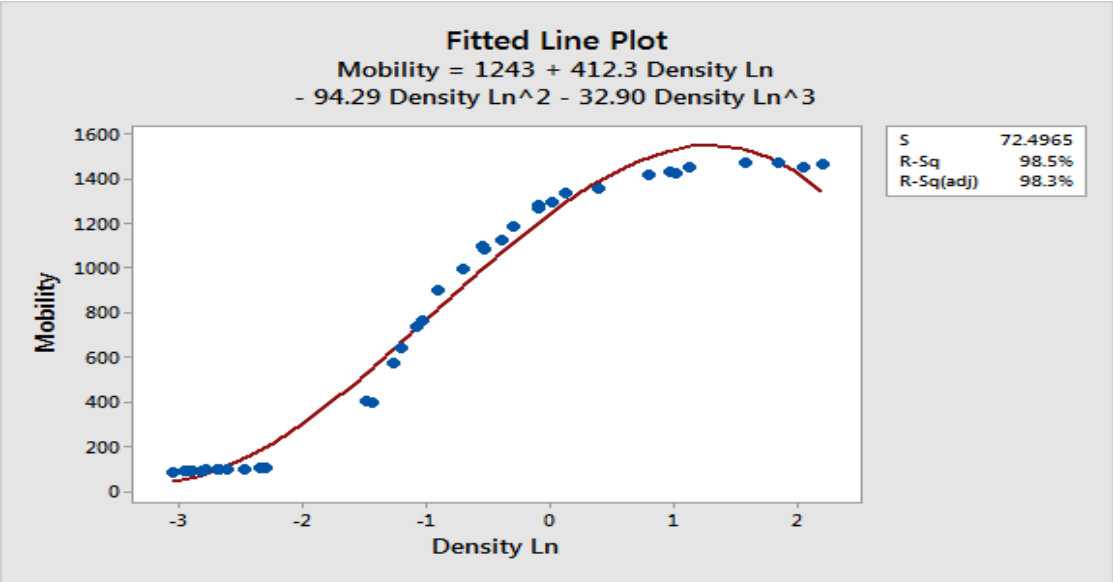
Y – The dependent variable
X or X1, X2.....Xn – Independent variables
m or m1,m2...mn – Slope of the line
C – The constant or the Y intercept

The linear regression algorithm updates m and C to find the best line for those data points, or, to put it another way, it tries to find the best line for a given problem.
It estimates the precision of the model with "N" features using the residual sum of squares and the total sum of squares; we decide which features to include or leave out based on our domain expertise; in other words, it relies on minimizing the difference between the actual value and the predicted values.

This brings us to the issue of overfitting and underfitting. Overfitting occurs when a model memorises data rather than learning it; in this

situation, there are too many variables and the model becomes overly complex, failing to generalise. Underfitting occurs when a model is simply too simple and lacks features, causing the model to fail to generalise. Additionally, we must verify certain linear regression assumptions because failing to do so could cause our model to make major mistakes and stop learning from or generalising from the dataset. There are some assumptions to validate in linear regression on the training set and violating any of these can introduce some serious errors in our model

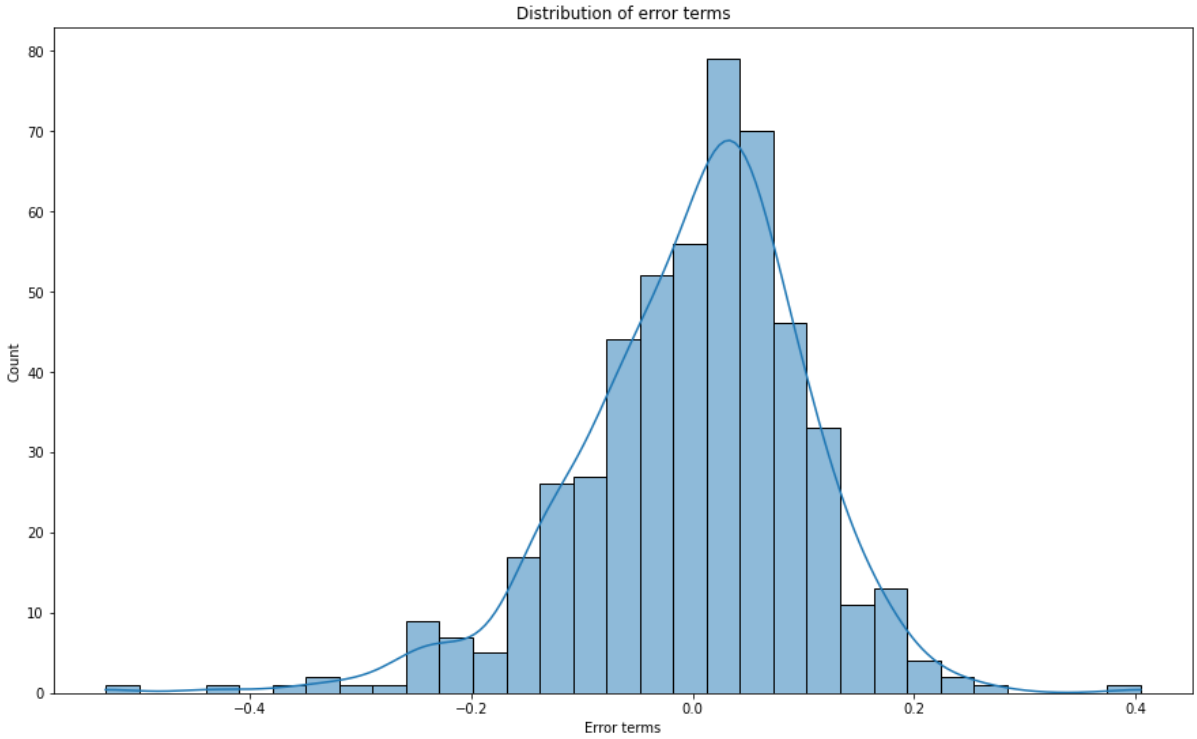
1. **There should exist some linear relationship between X and Y**:
   There is no point in utilizing this algorithm if there doesn't exist any



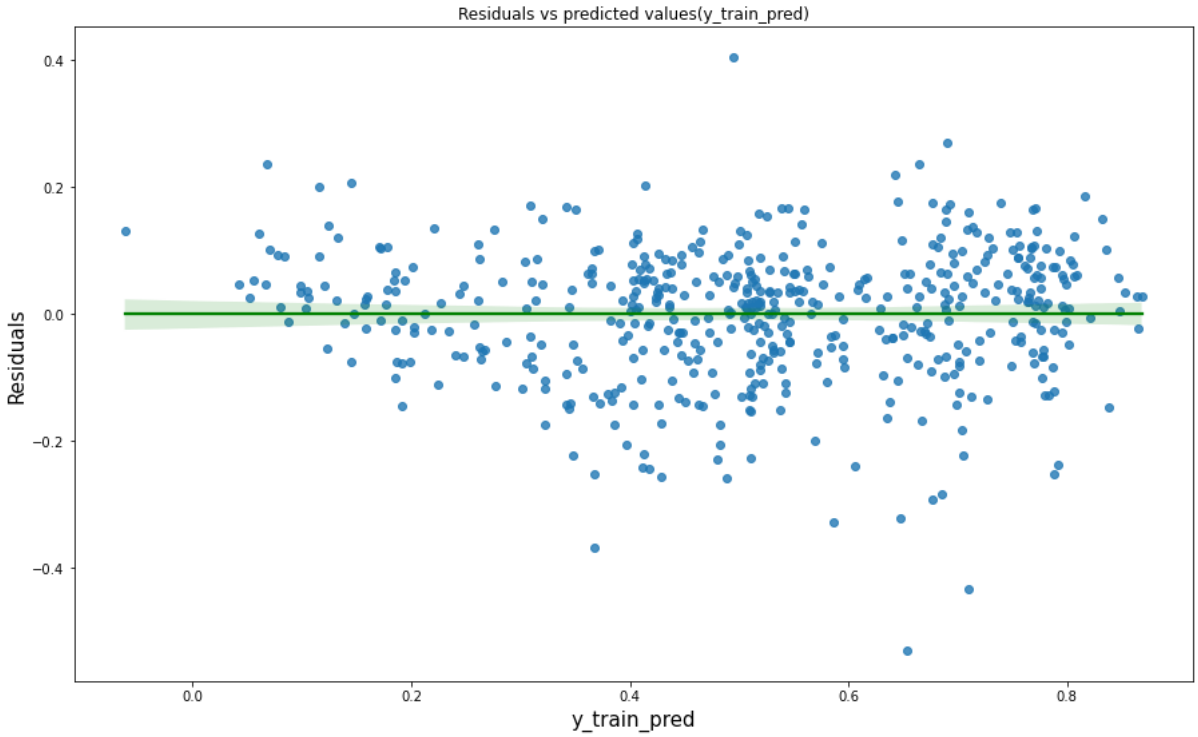
linearity between X and Y

The Blue line indicates a linear relationship and the red one a non-linear

2. **Error term should be normally distributed around zero**:



3. Errors have constant variance



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

ANS: Anscombe's quartet, according to Wikipedia, consists of four data sets with very identical descriptive statistics (Mean, standard deviation etc). It was created by a statistician by the name of France Anscombe who sought to demonstrate the importance of charting the data rather than drawing conclusions just from descriptive statistics. A dataframe with Anscombe's quartet and its descriptive statistics is provided below;

```
In [63]: numerical_vars = ["temp", "hum", "windspeed", "cnt"]

         train_df[numerical_vars] = scaler.fit_transform(train_df[numerical_vars])

In [64]: train_df[numerical_vars].describe()
```

Out[64]:

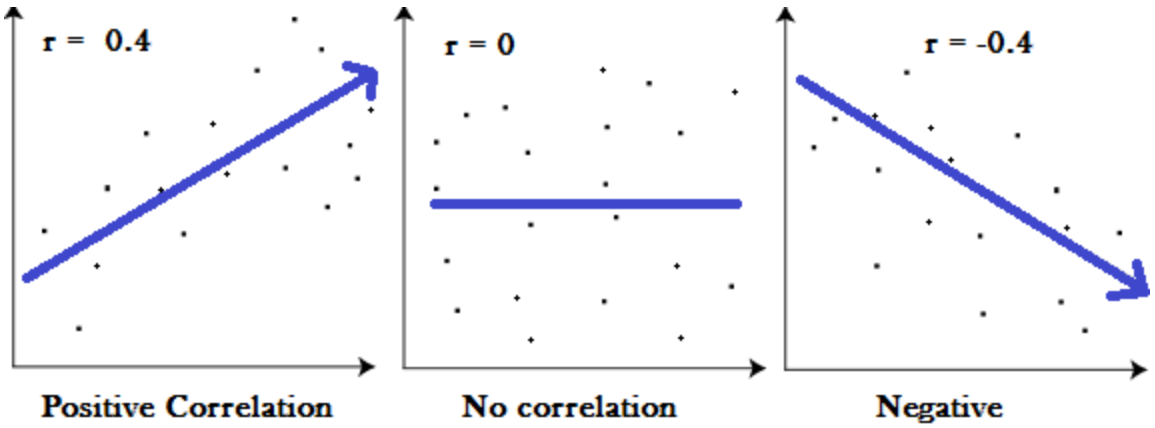
	temp	hum	windspeed	cnt
count	510.000000	510.000000	510.000000	510.000000
mean	0.537262	0.650369	0.320768	0.513620
std	0.225844	0.145882	0.169797	0.224593
min	0.000000	0.000000	0.000000	0.000000
25%	0.339853	0.538643	0.199179	0.356420
50%	0.540519	0.653714	0.296763	0.518638
75%	0.735215	0.754830	0.414447	0.684710
max	1.000000	1.000000	1.000000	1.000000

Modeling the Data ¶

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

ANS: The Pearson's R, also known as the Pearson's correlation coefficient, provides the summation of the strength of the linear correlation between two variables. It does this by taking the product of the individual sums of X and Y and dividing it by their standard deviation.

Pearson's r can take values ranging from -1 to 1 and can be categorized as Positive Correlation, Negative Correlation and No Correlation.



From the first graph we can see a Positive correlation(greater than zero) where one variable goes the other goes up as well

From the second graph we can see a Negative correlation(less than zero) where one variable increases the other decreases

The third graph depicts No correlation(equal to zero) where the points are scattered everywhere and no visible pattern can be identified

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

ANS: In the pre-processing stage of machine learning, scaling, also known as feature scaling, is done to standardise the data. This involves grouping the data into a range of similar values so that the larger values, for instance, are not given more weight. Since it is evident that the weight measured in pounds will never be larger than the weight measured in kilogrammes, feature scaling is used to correct this discrepancy. However, feature scaling is not used in linear regression because there is only one independent variable. The following are the differences between standardised scaling and normalized(MinMax) scaling:

Normalized scaling	Standardized scaling
Transforms the data in the range of [0,1] or [1,-2]	Transforms the data by removing the mean and scaling the data about mean being 0 and SD 1
It used when we do not known the distribution of the features in our data	It is used when we are certain of the distribution of data(normal distribution)
It is sensitive to outliers	It is not as sensitive to outliers compared to normalized scaling
It is calculated by subtracting the data in individual column by the minimum value and diving it by the range of that column X – Xmin/Xmax - Xmin	It is calculated by subtracting the data in individual column by the mean of that column divided by

	the standard deviation of that column X – Mean/ SD

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

ANS : 1/(1-R2) is the formula for VIF (Variance Inflation Factor). Multicollinearity is quite low when VIF is below 5; when it is between 5 and 10, one should explore the variable, and anything beyond 10 should be immediately deleted.

The value in the denominator approaches zero, indicating extremely high correlation between independent variables, and must be dropped after careful consideration of P-values because changing or dropping even one variable affects the VIF of other variables. This is why the VIF is "inf" as the R2(R square) approaches 1, i.e., the dependent variables are able to explain the variance of independent variables brilliantly.

6)What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANS : The Q-Q plot, also known as a quantile-quantile plot, is a probability plot that uses the quantiles of (x,y), where x stands for the first dataset's quantile and y for the second dataset's quantile. Two probability distributions are compared (x,y)

Plotting the points on the X,Y axis allows us to establish whether the two datasets share a common distribution. If so, the points will (roughly) sit on the line with slope 1, or the identity line, which forms a 45-degree angle with the X-axis.

Let's use the distribution of error terms or residuals from our bike sharing assignment as an example. Linear regression assumes that residuals are normally distributed; however, this assumption can only be approximate because the model that is built cannot be 100% linear. Therefore, a line is drawn at a 45-degree angle from the X-axis to fit normally distributed data points (residuals), and we then check to see if the points are on the line or around the line. If we see such a pattern