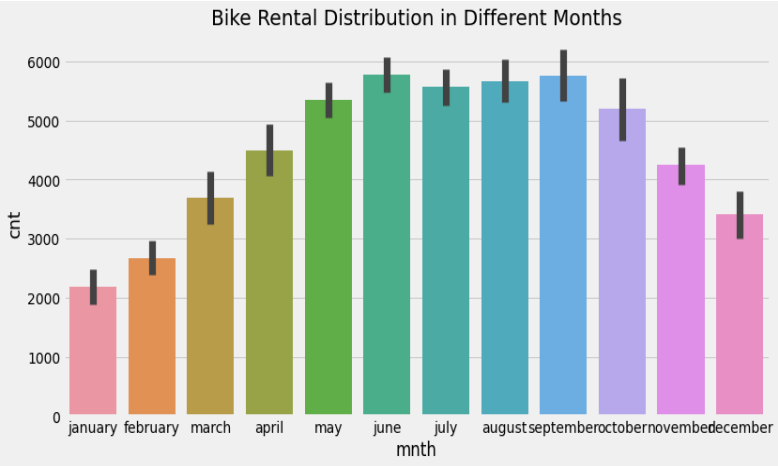
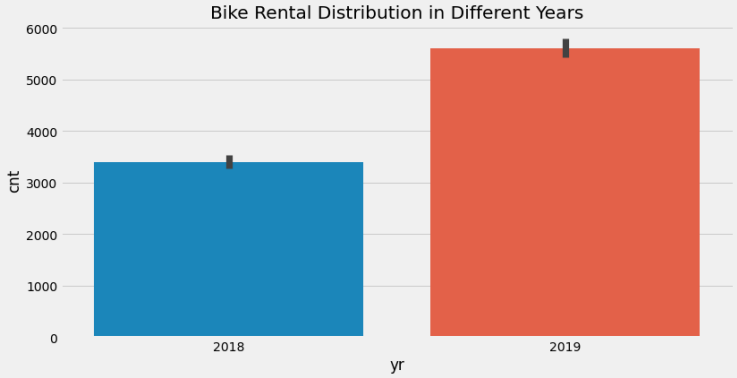
Assignment-based Subjective Questions

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

   Categorical values have effect on dependent variables. If we observe the below plot. The Rental demand is high on June and very low on January. This indicates that Categorical value month has effect on dependent variable Rental Demand. The Same influence is observed in second plot. It shows how year effects rental demand. Hence, we can infer that categorical values are correlated with dependent variables either negatively or positively.





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

   One-Hot Encoding is popular technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. One – Hot encoding builds on top of first variable and hence the outcome of one variable can easily be predicted with the help of the remaining variables. This leads to the problem of multicollinearity. Multicollinearity occurs where there is a dependency between the independent features. Multicollinearity is a serious issue in machine learning models like Linear Regression as it drastically swings the coefficients.

   So, it is important to always drop the first column or first feature during dummy variable creation

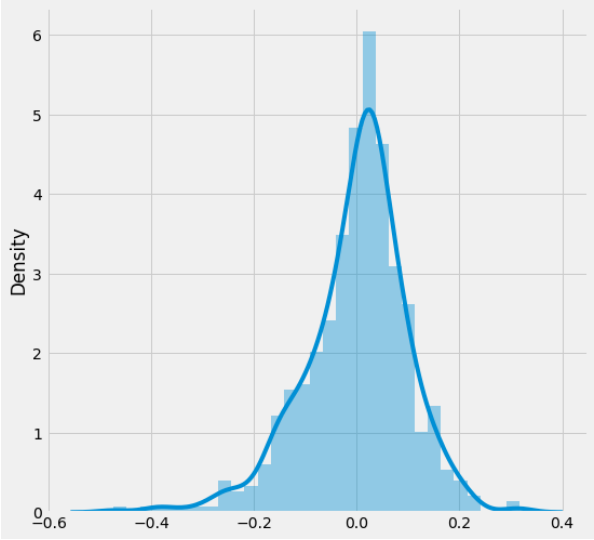
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? **(1 mark)**

   There is high correlation between target variable and Temperature

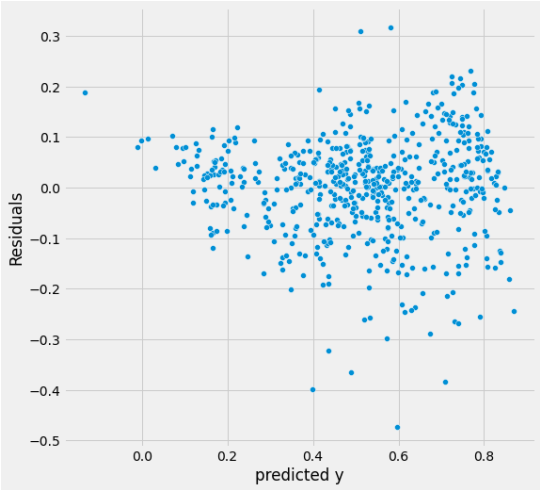
4. How did you validate the assumptions of Linear Regression after building the model on the training set? **(3 marks)**

   The following assumptions are validated:

   1. Error Terms are normally distributed with zero



   2. Error Terms are Independent to each other
   3. Error Terms have equal 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)**

   Top 3 features contributing significantly:
   1. Atemp
   2. Yr -year
   3. Light_snow

   The p values and VIF are significant also these high high co-efficient values

General Subjective Questions

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

   Linear Regression Algorithm is a machine learning algorithm based on supervised learning. In Linear regression, we train a model to predict the influence of independent features on a certain feature. In general, linear regression attempts to model relationship between independent variables and a dependent variable.

   When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables. Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares.

   There are four assumptions associated with a linear regression model:

   - Linearity: The relationship between X and the mean of Y is linear.
   - Homoscedasticity: The variance of residual is the same for any value of X.
   - Independence: Observations are independent of each other.
   - Normality: For any fixed value of X, Y is normally distributed.
   - Multicollinearity: Collinearity is not accepted

   The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

   The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

   $$y = B0 + B1*x$$

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

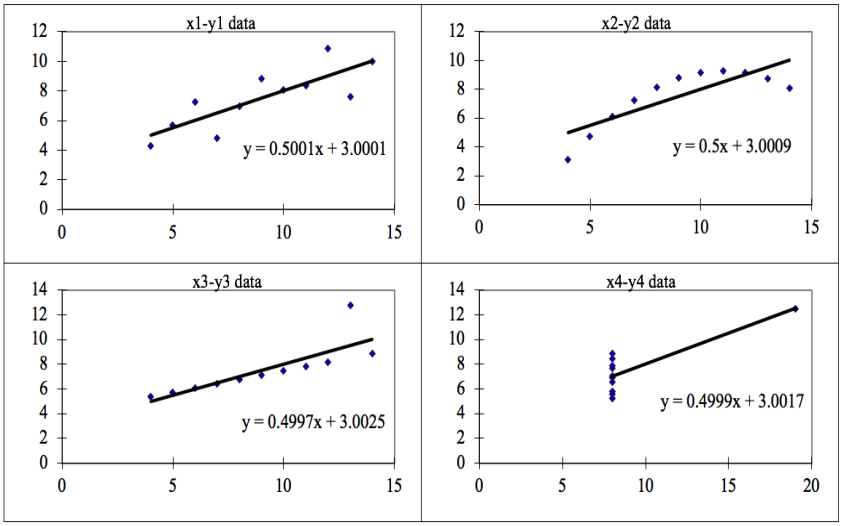
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.

Consider the below dataset and observe Mean, Standard Deviation and Correlation

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Every dataset has similar Mean, Standard Deviation and Correlation

Now we will plot the scatter plots and we observe that, they have very different distributions and appear differently when plotted on scatter plots.



Only the First Plot fits as linear regression and all others violate the linear regression assumptions 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

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

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression.

In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.

A value of ± 1 indicates a perfect degree of association between the two variables.

As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.

The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship.

For the Pearson r correlation, both variables should be normally distributed (normally distributed variables have a bell-shaped curve). Other assumptions include linearity and homoscedasticity. Linearity assumes a straight-line relationship between each of the two variables and homoscedasticity assumes that data is equally distributed about the regression line.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

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

Feature Scaling 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.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

Secondly, how VIF is calculated? The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression.

It is calculated as:

$$VIF = 1 / (1 - R2)$$

If R2 is very high nearly equal to 1 then (1-R2) is a value nearly equal to 0

If (1-R2) is very small neatly equal to 0 then 1/(1-R2) tends to a very big value nearly infinite

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare two probability distributions by plotting their quantiles against each other.

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution. we can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.

We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1) on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis.

If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution