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?**

   We can infer that season plays a crucial role in the count of bike rentals, where spring is inversely proportional while winters is directly correlated with the count. Also the count is higher in the month of March, April, May, June, August, September and October

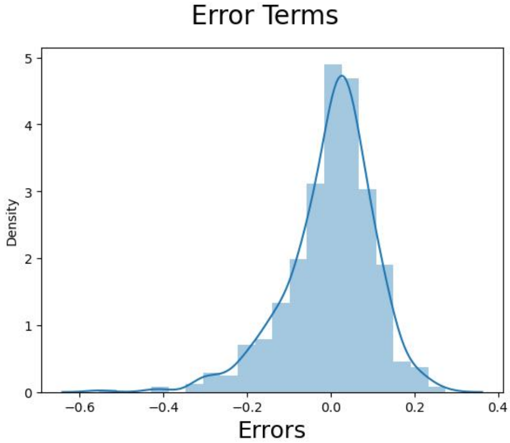
2. **Why is it important to use drop_first=True during dummy variable creation?**
   We use **drop_first=True** during dummy variable creation, to reducing the extra column created during dummy variable creation, as the type of details can be identified with just the remaining columns. So it also reduces the additional correlations created among dummy variables

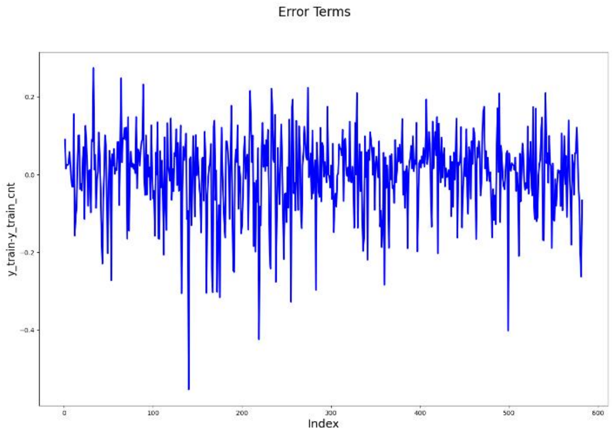
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable?**

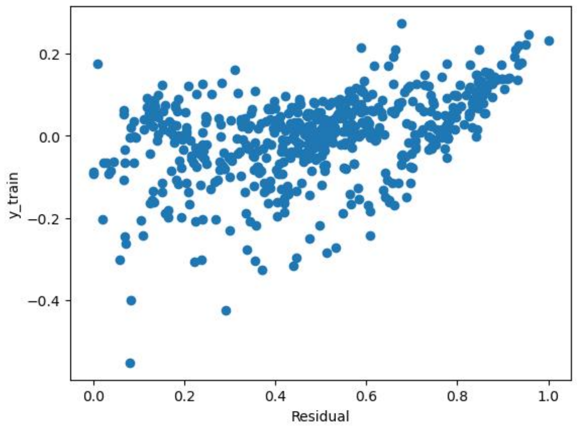
   Going by the pair-plot among the numerical variables, ”temp” and “atemp”, these both seem to have the highest correlation with the target variable. Also these both seem to have a direct mutual correlation as well.

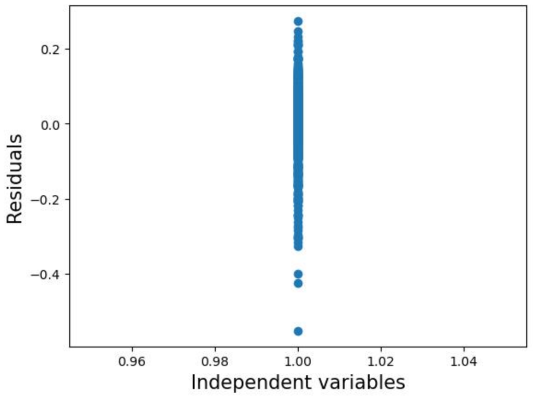
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   We Validated the assumptions of Linear Regression after building the model on the training set by doing Residual Analysis, checking the error terms are normally distributed, we Plotted a histogram with 20 bins of the error terms and found the inference to support the claims. It presented a uniformly distributed bell shaped graph.

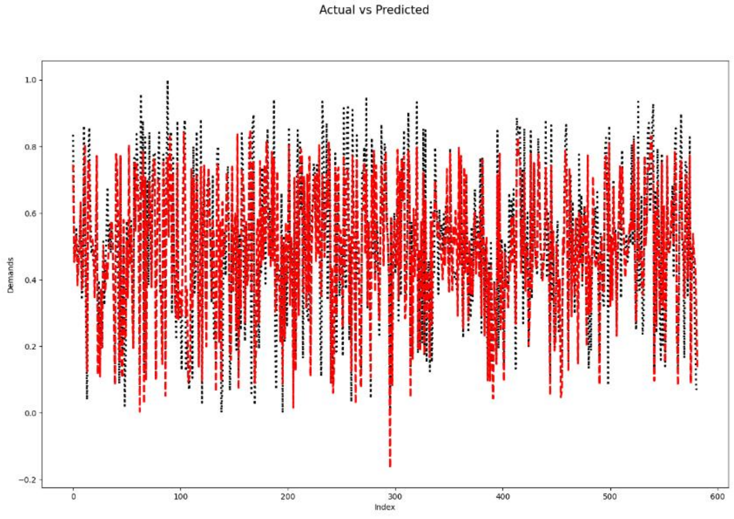
   Also we checked for visible pattern in the error terms by plotting scatter plot of residuals, in order to determine if they have a constant variance. It showed evenly distribution with very little distortion in the graphs. It can be inferred from the plots about all validity of assumptions of LinearRegression, the errors are uniformly distributed, having mean 0, and are independent of each other also the error terms have constant variance. On later stage we also did the cross validation.




Error Terms
y_train-y_train_cnt
0.2
0.0
−0.2
−0.4
0
100
200
300
400
500
600
Index


y_train
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
Residual


Residuals
0.2
0.0
−0.2
−0.4
0.96
0.98
1.00
1.02
1.04
Independent variables


Actual vs Predicted
Demands
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
100
200
300
400
500
600
Index

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

The top 3 features contributing significantly towards explaining the demands of the shared bikes are

Atemp [having direct proportion to cnt]

Light rain_Light snow_Thunderstorm in weathersit [having direct proportion to cnt]

&

mnth [March,April,May,June,August,September,October show to have positive influence in cnt]

cnt = 0.1168 + 0.2499yr - 0.0672holiday + 0.4594atemp - 0.1132spring + 0.0595winter - 0.2444(Light rain_Light snow_Thunderstorm) +0.0470March + 0.0251April + 0.0649May + 0.0576June + 0.0318August+0.0958September +0.0677*October

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a type of supervised learning method, it is used to estimate or predict real values depending on continuous variables, where what we need to predict is dependant variable and what we use to predict is the independent variables, few example of dependant variables are like cost of houses, total sales etc..

We establish the relationship or dependence between dependent and independent variables by fitting a best fit line.

This best fit line is also known as regression line and we represent it by a linear equation

Y= a *X + b.
In this equation:
Y is the Dependent Variable, a is the Slope, X is the Independent variable and b is the Intercept

There are a few assumptions of linear regression model;
There is a linear relationship between X and Y
X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.

Error terms are normally distributed with mean zero(not X, Y):
There is no problem if the error terms are not normally distributed if you just wish to fit a line and not make any further interpretations.

Error terms are independent of each other:
The error terms should not be dependent on one another.

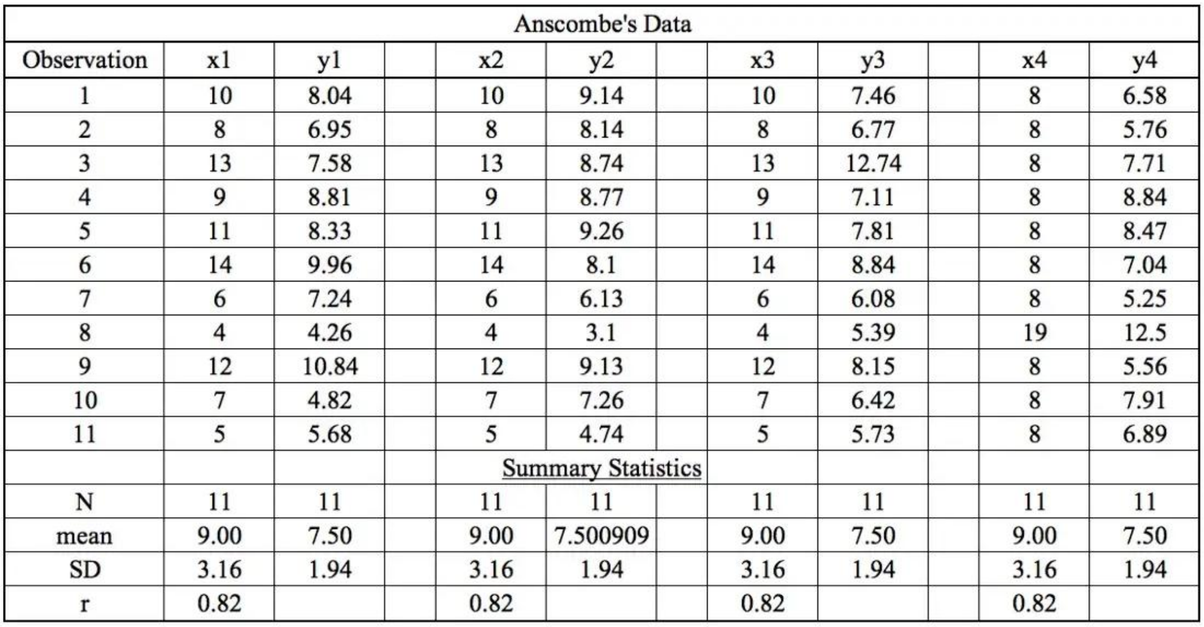
Error terms have constant variance i.e homoscedasticity
The variance should not increase or decrease as the error values change. Also, the variance should not follow any pattern as the error terms change.

2. Explain the Anscombe’s quartet in detail.

The Anscombe’s quartet is a group of four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

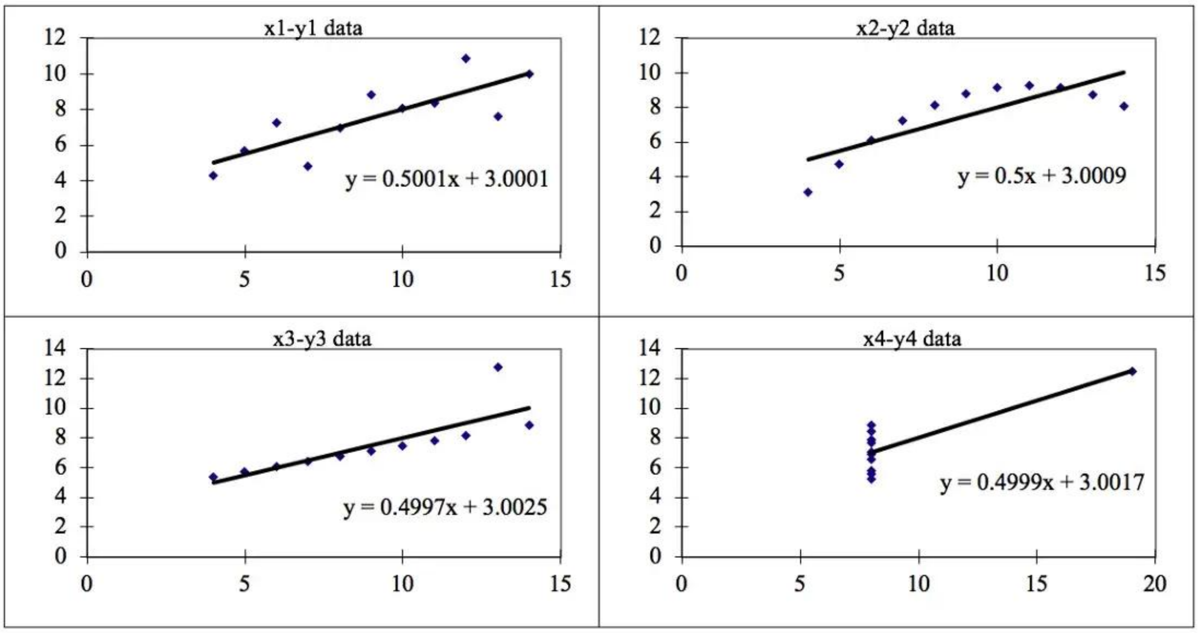
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

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. These four plots can be defined as follows:



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	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



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

3. What is Pearson’s R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the correlation between two variables. 1 represent complete positive correlation, -1 represent complete inverse correlation while 0 shows no correlation. The closer the value is to 0, the least correlated are the two variables.The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables. It is used to find the pairwise correlation of all columns in a dataframe. Any null values are automatically excluded. Any non-numeric data type column in the dataframe will be ignored.
Following are the Assumptions of Pearson correlation coefficient (r):
For the Pearson r correlation, both variables should be normally distributed like the Bell Curve’ or the ‘Gaussian Curve’.
Homoscedascity. for this the error terms should be the same across all values of the independent variables.
The variables have a linear relationship.
Each variable should be continuous.
There should be no significant outliers. This is because Pearson’s coefficient, is very sensitive to outliers and including outliers in our analysis can lead to misleading results.

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

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. We do this to transform our data so that it fits within a specific scale. Scaling is performed for making data points generalized so that the distance between them will be lower.
Normalized scaling or Min-Max Scaling is used to transform features and to normalize the data within a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1].
Normalization is useful when there are no outliers as it cannot cope up with them.

Standardized scaling is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score. X_new = (X - mean)/Std
Standardization can be helpful in cases where the data follows a Gaussian distribution. Standardization does not get affected by outliers because there is no predefined range of transformed features.

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is are extreme data point (outlier).
Standardised scaling will affect the values of dummy variables but MinMax scaling will not.
MinMax scaling scales in such a way that all the values lie between 0 and 1 using the formula:
x−min(x) / max(x) − min(x)
So if you have dummy variables, which can only take the values 0 and 1, you can notice that for the case of zero, the variable remains zero and for the case of 1, the variable remains 1.
On the other hand, the standard scaler scales in such a way that the mean of the dataset becomes zero and standard deviation becomes one. So this will clearly distort the values of the dummy variables since some of the variables will become negative.

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

The formula for VIF is given as: 1/(1− R2). When two independent variables are in perfect correlation then R2=1 and hence the denominator is 0 so the value of VIF becomes infinite. Thus, when the value of VIF is infinite, it actually means that the two variables are perfectly correlated with each other.

It might also mean that the corresponding variable might be expressed exactly by a linear combination of few other variables in the dataset.

To rectify this issue, we would drop one of these two variables from the data frame that cause perfect multicollinearity and then check the VIF value.

We never delete two or more than two variables at once even though all might have a very high VIF, since by removing one the VIF for the remaining variables might automatically go down. Hence the correct way is to remove the variables one by one and subsequently check.

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.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

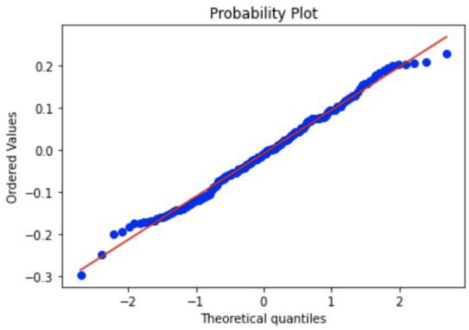
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variables.

It also 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.

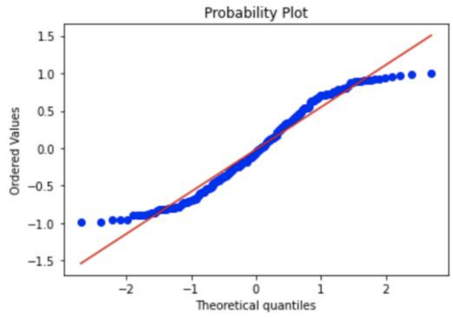
Advantages:
a) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
b)It can also be used with sample sizes

It is used to check following scenarios:-
If two data sets;
have similar tail behavior
have common location and scale
come from populations with a common distribution
have similar distributional shapes

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



The data points lie approximately in a straight line. So, we can conclude the data points are normally distributed



Most of the points, do not lie in a straight line, so we can conclude that the distribution is not normal.