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?

Let us consider categorical variable ‘weathersit’ on the target variable ‘cnt’ and analyse its effect.

We did EDA and saw that the relationship between the categorical variables and the target variable.
It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being around 50000.

Similarly, certain inferences could be made for categorical variables ‘season’ and ‘yr’ as well. Also, during model building on inclusion of categorical features such as yr, season etc, we saw a significant growth in the value of R-squared and adjusted R-squared.

This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

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

drop_first=True is used as it helps in reducing the extra columns created during dummy variable creation. Hence, it reduces the correlations created among dummy variables which would lead to incorrect models being created or unnecessary columns being dropped.

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

‘Registered’ has highest correlation of 0.9 with the target variable ‘cnt’. After it was eliminated, both ‘yr’ and ‘temp’ have same highest correlation of 0.6 with the target variable ‘cnt’.

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

Below are the assumptions of Linear Regression and how we validated them:

a) **Linearity:** Pair-wise scatter plots were drawn to visualize the linear relationship on plots.

b) **Homoscedasticity:** To verify homoscedasticity, residual plot as plotted and we verified that the variance of the error terms was constant across the values of the dependent variable.

c) **Independence/ Absence of Multicollinearity:** We used VIF method to check for multicollinearity and eliminated the features showing multicollinerity.
Apart from this, we made a correlation matrice to check for relationship between various independent variables.

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

Based on the final model, the Top 3 features contributing towards explaining the demand of shared bike are:

a) Weathersit_mist

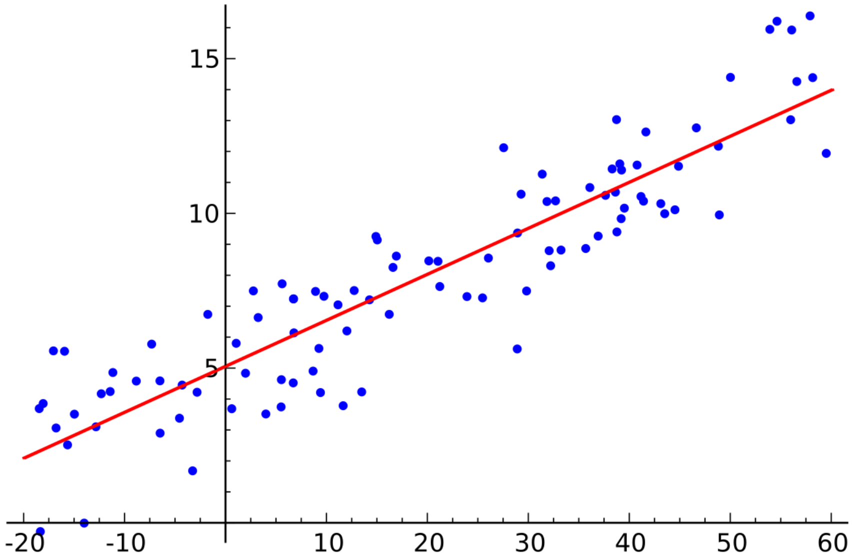
b) Season_spring

c) Mnth_september

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.



The red line in the above graph is referred as the best fit line.
The equation of the line is:

Y = B_0 + B_1 *x

The motive of the linear regression algorithm is to find the best values for B_0 and B_1

Cost Function

The cost function helps us to figure out the best possible values for B_0 and B_1, which can be found out by converting it to a minimization problem where the aim is to minimize the error between the predicted and actual value.

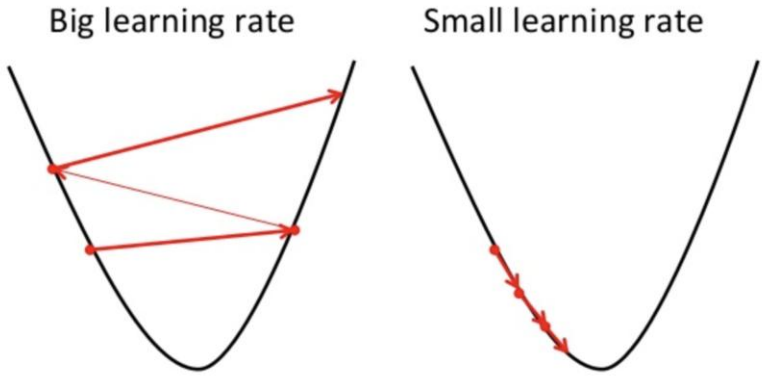
$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function. Using this MSE function, we change the values for B_0 and B_1 so that the MSE value represents the minima.

Gradient Descent

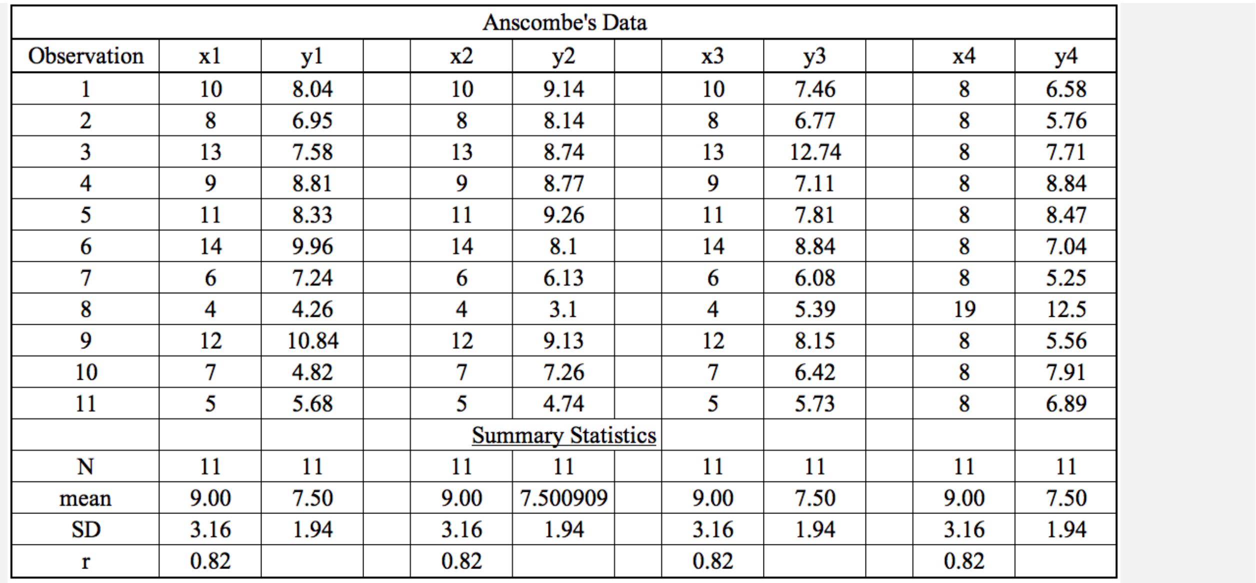
Gradient Descent is the method where we update the values of B_0 and B_1 to reduce the Cost Function. We take iterations to change the value of B_0 and B_1 to reduce the cost function. Here we use a learning rate to determine the pace of the iterations. This is used to determine how fast the algorithm is able to converge on the minima.
Below is an illustration to explain Gradient Descent.



2) Explain Anscombe’s quartlet in detail.

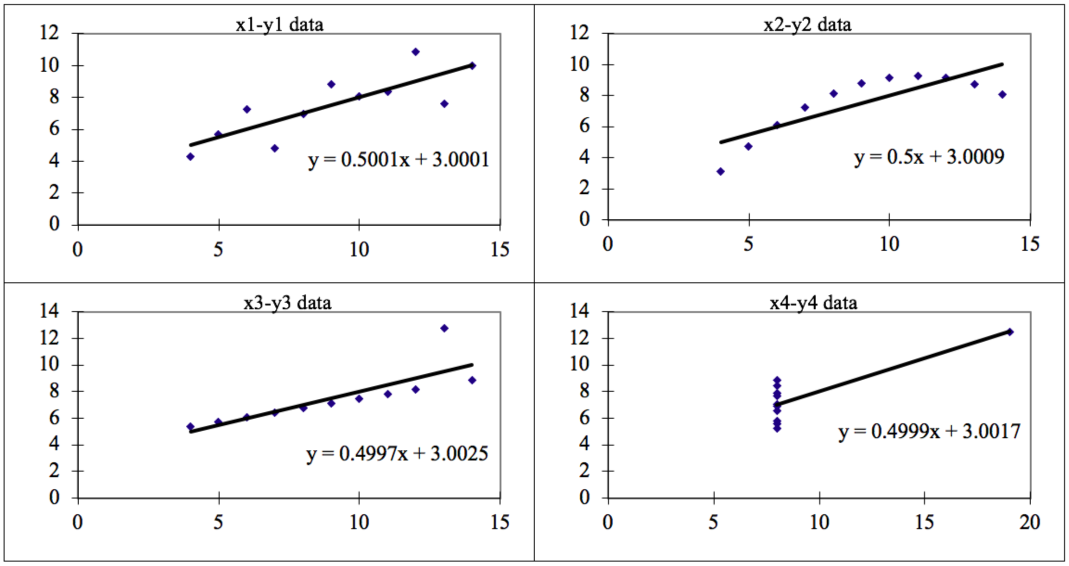
Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate
the importance of plotting the graphs before analyzing 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 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	

We see that the statistics for the 4 datasets are similar and would easily fool the Regression model, but when we plot them while visualizing, we see that all of them are different.



The four datasets, after visualization, can be described as:

1) Dataset 1: this fits the linear regression model pretty well.
2) Dataset 2: this could not fit linear regression model on the data and the data is non-linear.

3) Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
4) Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3) What is Pearson's R?

Pearson's R is also referred as Pearson's Correlation Coefficient (PCC). It is a measure of linear correlation between 2 sets of data. It is the covariance of two variables divided by the product of their standard deviations.

The result always has a value between -1 and 1.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Pearson's R cannot tell the difference between a dependent and independent variable. It also does not give us any information about the slope of the line. It tells us only if there is a relationship.

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

Feature Scaling is a method used to normalize the range of independent variables or features of data. It is an important step in the preprocessing of data.
Some machine learning algorithms do not work correctly without standardized data. If one of the features has a wide range of values, the distance will be affected by that particular feature.
Hence all the features should be scaled/ normalized so that each feature contributes proportionately to the final distance.

In **normalized scaling**, the values are reduced to the range of 0 to 1. It is also known as Min-Max Scaling

$$X^{'} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

In **standardized scaling**, the values are centered on 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

$$X^{'} = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature. In these cases, the values are not restricted to a particular range.

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

Infinite VIF is shown for cases in which the features show perfect correlation.

In the case of perfect correlation, we get R-square =1, which lead to 1/ (1 – (R-square)) = infinity.

To solve this problem, we can drop one of the variables from the dataset which is causing this perfect multi-collinearity.

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

Q-Q Plot refers to quantile – quantile plot. It is a graphical technique for determining if two data sets come from populations with a common distribution.

The **q-q plot** is a **plot** of the quantiles of the first dataset vs the quantiles of the second dataset.

Q-Q plot is helpful in cases where we have received taining and test data separately and we want to know if both of them are from populations with same distribution. It can also point out presence of outliers.