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?

   Based on the analysis we can conclude that the variables “season”, “weathersit”, “mnth”(month) are the strong predictors.

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

   If any categorical variable has n levels, (n-1) dummy variables are required to indicate these n levels. These are denoted by 0 and 1. We can safely remove one of the n dummy variables. However, during coding we practice the best way of dropping the first one and hence “drop_first = True” is used.

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

   On analysing the pair-plot we see that variables “temp” and “atemp” have the highest correlation with the target variable “cnt”.

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

   The assumptions of Linear Regression on the model is validated based on Multicollinearity, Linearity of the model and residual analysis wherein we concluded that the error terms are normally distributed.

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 model, year, temperature and season are the top 3 features contributing significantly towards explaining the demand for the shared bikes.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

The linear regression algorithm is an algorithm used to build a model based on correlation between target and predictor variables which are linearly related to each other and thereby make inferences on target variable. Linear Regression is based on concept of straight line with equation:

$y = mx+c$

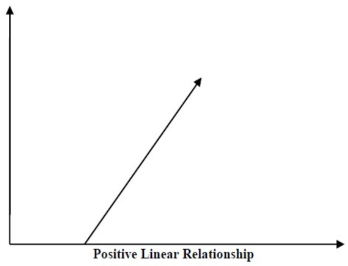
where, y is the target or dependent variable.

x is the independent or predictor variable

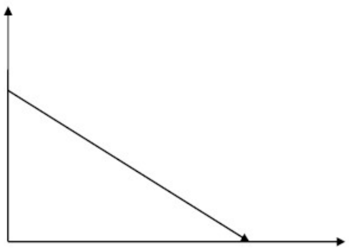
m is the slope which tells how well x is correlated with y

c is the constant which defines the value of y when $x=0$

There can be either positive or a negative correlation between target and predictor variables. Below are the graphs explaining the same:



Positive Linear Relationship



Negative Linear Relationship

There are 2 different types:

1) Simple Linear Regression – Target variable correlates to only one predictor variable
2) Multiple Linear Regression- Target variable correlates to more than one predictor variables.

Assumptions of Linear Regression are as follows:

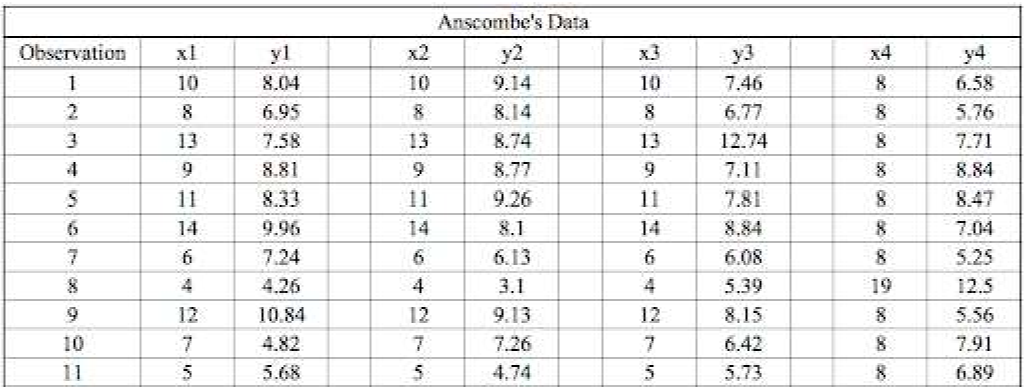
1) The dependent and independent variables are linearly related to each other.
2) The error terms are independent of each other.
3) The error terms follow the Normal distribution.

4) There is no collinearity between the predictor variables, i.e., multicollinearity is not present.
5) The residuals have constant variance.

2) Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is basically proposed to emphasize the significance of graphs over the statistical measures, to overcome that misconception that numerical calculations are better, and the graphical representation is just rough estimation.

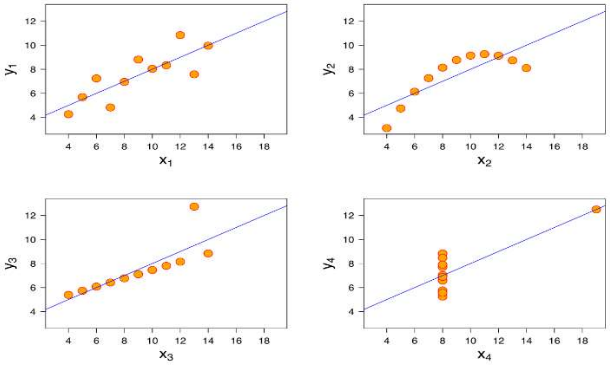
The data proposed by Anscombe consists of 11 different values with the same statistical measures but different graphical representations. Below is the table.



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

The graph for the above data is shown below:



In the first graph, the linear relationship between the 2 variables is seen.
In the second graph, there is some kind of obvious relation between x and y but not linear in nature.
In the third graph, the relationship is linear. But an outlier is observed.
In the fourth graph, a single data point is sufficient to increase the correlation coefficient and the rest of the variables doesn't define the model in any way.

3) What is Pearson's r?

Pearson's coefficient analysis the relationship between predictor and target variables. It determines:

1) The strength of the correlation
2) The direction of the correlation

This is measured by Pearson's correlation coefficient(r) and it varies between -1 and 1

- If r is between 0 and 0.1, it indicates no correlation.
- If r is between 0.1 and 0.3, it indicates less correlation.
- If r is between 0.3 and 0.5, it indicates medium correlation.
- If r is between 0.5 and 0.7, it indicates high correlation.
- If r > 0.1, it indicates very high correlation.

If the large values of one variable go along with large values of other variable or if the small values of one variable go along with small values of other variable, then the positive correlation exists.

Similarly, negative correlation exists when large value of one variable go along with small value of other variable and vice versa.

Pearsons Coefficient, r = cov(X,Y)/SD(X)*SD(Y)

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

Scaling is a technique used to bring all the values or magnitudes in the data set to an equal or comparable scale. If scaling is not performed the system

just assumes the bigger value as large and the smaller value as lower irrespective of the magnitude or values of the dataset.
For example, if the cost is measured in 50 cents and 10 Euros, 50 cents would be considered as higher than 10 Euros which is indeed not true. To avoid such kind of errors by the model, scaling is very much necessary. The two extensively used scaling techniques are:

1) MinMax Scaling
   The values are compressed in the range [0,1]
   Uses the min and max values of data.
   There is no probability that outliers will get involved in scaled data.

2) Standardized Scaling
   The values aren't compressed to any range.
   Uses the Standard Deviation and Mean for scaling.
   Sometimes, outliers may get involved in the scaled data.

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

   A large value of VIF indicates that multicollinearity is very high. The VIF value of infinite indicates that there is a perfect correlation between the variables. Statistically defining if there is a perfect correlation, R-Squared is 1 and therefore 1/1-R2 would be infinite.

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

   Q-Q plot, in other words quantile-quantile plot is used to specify the characteristic of residuals. The closer the points to the line in the graph, more normal is the distribution. So basically, this indicates the normality in the distribution.