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)

The categorical variables in our dataset are: 'season', 'mnth', 'weekday', and 'weathersit'. Looking at our correlation table, we can infer:

a. Weak positive correlation for the months of June, July, August, May, October, and September. Slightly negative correlation for the months of January, February, March, November, December.
b. Strong negative correlation with the Spring season, and slightly positive correlation with the summer season.
c. Slightly negative correlation with cloudy and light rain & snow weather.
d. Weak positive correlation for days of Tuesday, Wednesday, Thursday, Saturday and slight negative correlation for days of Sunday & Monday.

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

"drop_first=True" tells the function to return (n – 1) columns for 'n' category values that a column can hold. This aids us in a few ways:

a. It reduces the number of features
b. It does not lead to information loss, and removes redundant information.

The reason why information is not lost, is due to the fact that not belonging to (n – 1) category values automatically implies belonging to nth category value. For example, 'seasons' is a categorical column whose values are either fall, winter, spring, or summer. One of the dummy columns can be dropped (for instance, 'fall'), and it can still be calculated as all seasons that are not 'winter', 'spring', or 'summer' would be 'fall'. Hence, number of features is reduced, and no information is lost.

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

'atemp' seems to have the highest correlation with our target variable 'cnt'. We can infer a somewhat strong positive linear correlation between the two variables.

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

The following are the assumptions of linear regression and their validation methods:

a. Assumption of a Linear Relationship between the feature & target variables
   This assumption is validated, as our final model has a very good r2-score, with coefficient p-values well below 0.05.
b. Error terms are normally distributed with a mean of 0.
   We plot the error terms in and our assumption is validated as the mean is 0 and the fit appears to be normal.
c. Error terms are not dependent on each other or X or y.
   We can see from the plots obtained in [49], that there is no visible pattern which validates our assumption that error terms are independent.
d. Variance in error terms is constant, i.e. homoscedasticity.
   This assumption can be checked by plotting as we did in [50], a plot of residuals vs the fitted values. This plot displays no discernible pattern, and from that we can conclude that it is homoscedastic.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes can be determined by looking at output in [49]. They are namely:

1. 'yr' which refers to year has a t-value of 28.315
2. 'atemp' which refers to the temperature, and has a t-value of 20.94
3. 'season_winter' which is a dummy column for variable 'season' with value 'winter', and has a t-value of 11.938.

These t-values correspondingly have the lowest probabilities associated with them. Therefore, implying that they contribute significantly towards explaining the demand of shared bikes.

General Subjective Questions

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

There are two types of linear regressions, Simple & Multiple. These are statistical models that can predict a continuous numerical value (target variable), based on one (Simple) or multiple independent feature variables. The outcome of these models are a linear line which has coefficient values for each of the relevant feature variables. The predicted values lie along this line. Once, the relevant data has been collected, we perform a fit using either OLS or another statistical method. OLS stands for Ordinary Least Squares method. Let us look at the equation of a multiple linear regression line and understand the parameters we are trying to estimate through this method:

$$y = \beta_0 + \chi_1\beta_1 + \chi_2\beta_2 + \chi_3\beta_3 + \ldots + \chi_n\beta_n + \varepsilon$$

In this equation,

β_n refers to the coefficients that our model would predict.

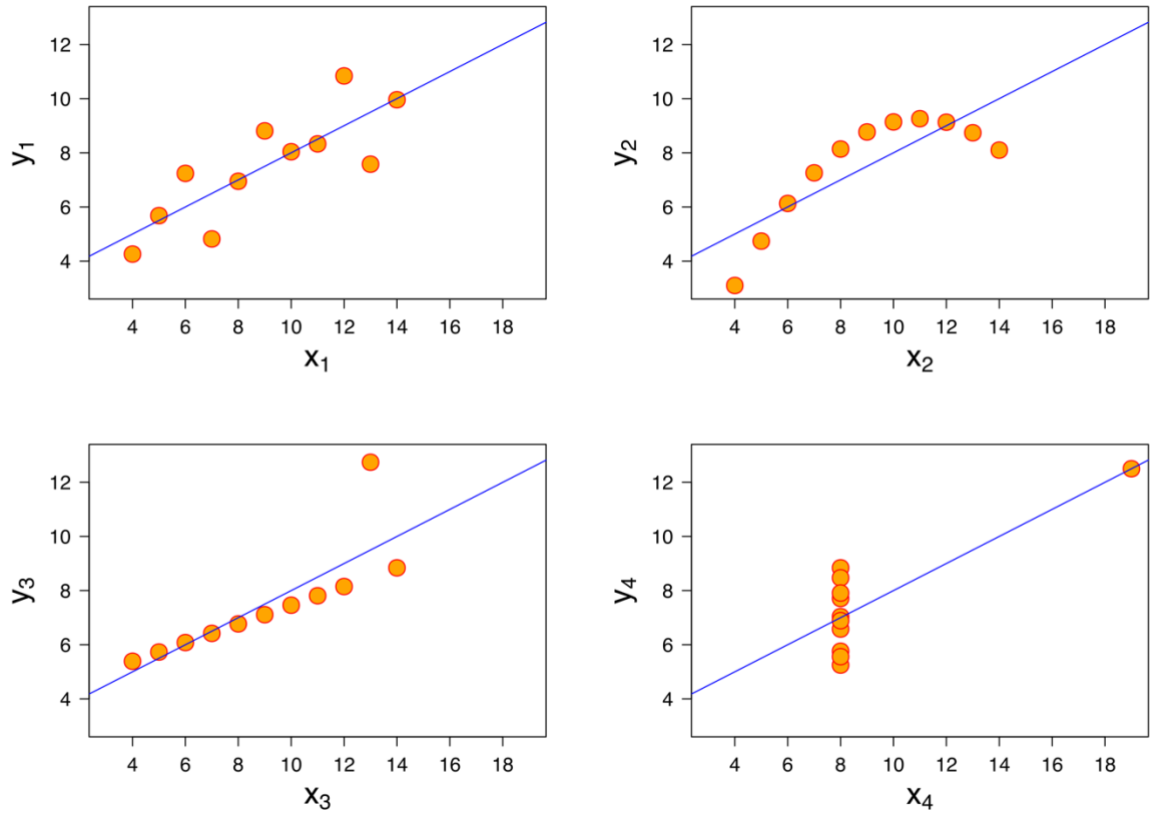
χ_n refers to the feature vector, and

ε is the error term

Firstly, the collected data is divided into training dataset, and testing data sets. It is important to keep the testing data aside, and never use it for training purposes. We will then run our Multiple Regression Model on the training data set in which it will 'learn' the relationships by using OLS method. In the OLS method, the goal is to find a line, or a plane, or a hyperplane such that the sum of the squared differences between the observed values, and the predicted values from the hyperplane is minimized. To summarize, OLS is a method to find the line of best fit for a data set by minimizing the sum of squared distances between the line and data points. Once, our coefficients have been estimated by this method, we can then use the test dataset to test the accuracy of our model which is also known as the r2-score. Once, we are satisfied with the model accuracy, we can use the model to predict values for unknown data.

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

Anscombe's Quartet refers to four distinct datasets that were published in a paper by Francis Anscombe in 1973. These four distinct data sets have the same summary statistics – mean, standard deviation, variance, regression line, and coefficient of determination. However, these datasets are qualitatively very different. Anscombe wanted to illustrate that basic statistical properties of a dataset were not enough, and that it is equally important to look at the data visually. In the image below, we can see that the four datasets designed by Anscombe are visually very different, but share the same descriptive statistical properties.



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

Pearson's R refers to the Pearson Correlation Coefficient. It's value lies between -1 and 1. It is a measure of linear correlation between two sets of data. Mathematically, it is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where,

n = Sample Size

The correlation coefficient is output in python by the .correl() function, as in [ ]. It is invariant under separate changes to two variables in scale and location. This implies that transformations of the two variables of the form a + bx and c + dy for (x, and y) will yield the same correlation coefficient. Furthermore, correlation of (x, y) is the same as correlation of (y, x). A high number indicates a strong high similarity, and a score near zero indicates no correlation. Consequently, a value of 1 indicates a perfect match, and a value of -1 indicates a perfect negative correlation.

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

Scaling is a process to transform the data which has multiple units and magnitudes across columns to a 'scale' that is comparable to each other. In the absence of feature scaling, greater values in magnitude tend to get higher weightage by the machine learning algorithm than values with lower magnitudes. The units of these values are not taken into consideration, and hence scaling is done to prevent this from happening. Scaling also improves the algorithmic efficiency of the machine learning algorithms such as early convergence in gradient-descent as compared to non-scaled dataset. It is important to note however, that in multiple linear regression, scaled dataset and unscaled dataset would yield same regression model in terms of correctness and evaluating metrics. Overall, scaling helps remove bias if there is one.

Normalized Scaling

In Normalized Scaling, we transform the feature values by subtracting each value with the mean value of the column data, and then divide the result by the difference between the minimum and maximum value. Normalized Scaling in python is done through Normalizer class of sklearn.preprocessing module.

Standardized Scaling

In Standardized Scaling, we transform the feature values by subtracting the mean value from each data point, and then dividing it by the standard deviation. This method is based on the central tendencies and variance of the dataset. Post this scaling method, we obtain a normal distribution with mean of 0 and standard deviation of 1. In Python, the corresponding module of sklearn.preprocessing provides StandardScaler class for implementing this scaling method.

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

To investigate why value of VIF is infinite at times, let us have a look at the mathematical formula for calculating VIF:

$$VIF_k = \frac{1}{1-r_k^2},$$

Where r_k^2 is the goodness of fit of the linear model for x_k based on all other variables.

We can see that, a value of 1 for r_k^2 would lead to the denominator being equal to 0, consequently the fractional value would henceforth be infinite. This would be the case when there is a perfect correlation between the feature variable and the target variable.

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

Q-Q plot refers to Quantile-Quantile Plot. There are two primary uses of a QQ plot:

a. It is used to determine if two datasets belong to the same underlying hypothetical probability class. For example, in the case of linear regression we would see if the training dataset, and the testing dataset, both come from the same family of probability function. The way to do this is to take the dataset with less number of elements, and match them with the same n-quantiles of another dataset. Then these are plotted against each other, and it is observed whether the points lie on a straight line. Affirmative observation leads to the conclusion that they belong to the same underlying probability distribution.
b. It can also be used to determine if a dataset comes from a particular probability distribution family. For instance, when the n-Quantiles of the dataset are plotted against n-quantiles of a theoretical probability distribution, we can infer if the points lie on a straight linear line that the dataset comes from that probability distribution.