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 the dataset that helps us decide about the demand count of bikes are - season, mnth, weekday, weathersit, year, holiday, workingday.

All these categorical variable are initially studied as boxplot in the python notebook and the results convey the effect of these variables on the dependent variable

- *season - We see that spring season sees the least demand count, and Fall accounts to most sales (highest 75% quantile). Fall and Summer almost go neck to neck in the demand created on bikes*

- *mnth - Similar to the season, the mnth where the highest demand is seen is mostly in June - November. We already are familiar with this, because we know that the highest count happened in the seasons Summer and Fall. Hence, month provides similar effect on the count as seasons*

Some variables like season_spring, mnth_Jul, weathersit_light_snow_rain_thunderstrom, weathersit_mist_cloudy,season_winter have negative coefficient , it means in these seasons bike count will be less whereas variables like season_winter , mnth_Sep have positive coefficient and count will be higher.

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

It is important to use **drop_first=True** because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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

The numerical variable temp seems to have the highest correlation.

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

1. Linearity: The relationship between independent and the mean of dependent is linear.
2. Homoscedasticity: The variance of residual is the same for any value of independent variable.
3. Independence: Observations are independent of each other.

4. Normality: For any fixed value of dependent and independent is normally distributed.

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

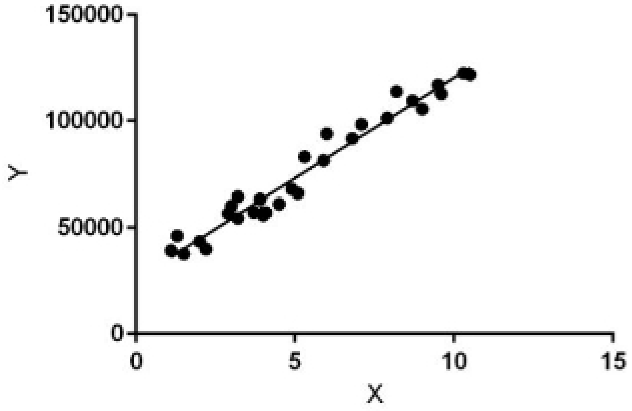
Following are features that was found to be contributing significantly towards explaining the demand of the shared bikes having positive relation.

- `temp`
- `season_winter`
- `mnth_Sep`

General Subjective Questions

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

Linear regression is one of the very basic forms of machine learning where a model is trained to predict the behaviour of data based on some variables.



Here two variables which are on the x-axis and y-axis respectively are linearly correlated, it means whenever X is increasing Y is also increasing, there can even be a linear downward relationship.

Mathematically, we can write a linear regression equation as:

y = mx+c

Where

m = slope or coefficient of x

c = intercept

x = independent variable

y = dependent variable

While training the model followings are given as:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best c and m values.

Once the θ1 and θ2 values are determined the best fit line can be get. So for the input value of x the value of y will be predicted.

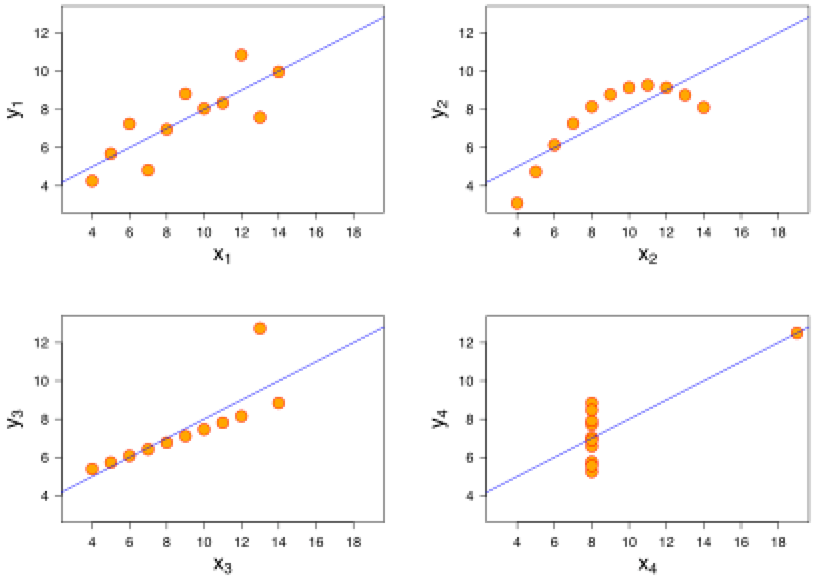
Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the m and c values, to reach the best value that minimize the error between predicted y value and true y value.

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

Cost function (J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value and true y value.

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

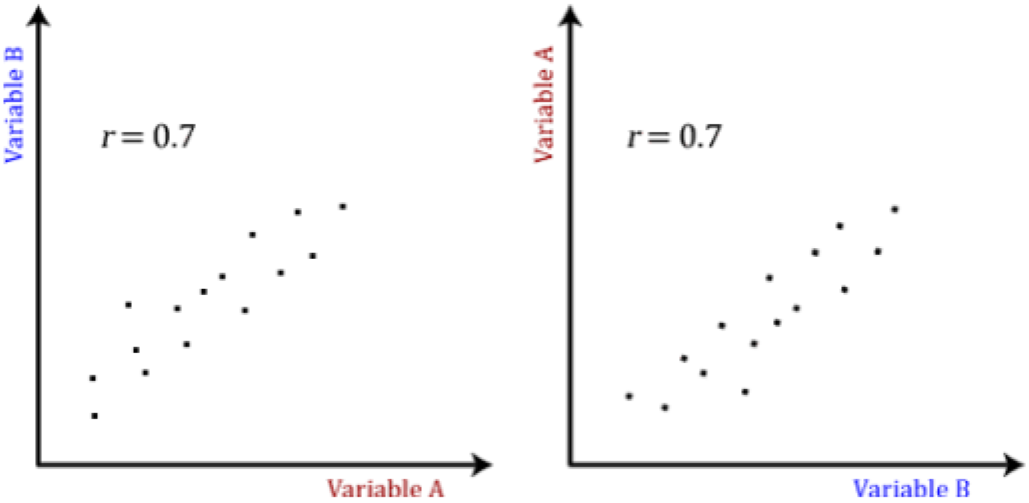
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

3. What is Pearson’s R? (3 marks)

Pearson's r or alternatively Pearson correlation coefficient , also referred to as the Pearson product-moment correlation coefficient PPMCC, the bivariate correlation,[1] or simply as the correlation coefficient[2]) is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally. For example, you might want to find out whether basketball performance is correlated to a person's height. You might, therefore, plot a graph of performance against height and calculate the Pearson correlation coefficient. Lets say, for example, that *r* = .67. That is, as height increases so does basketball performance. This makes sense. However, if we plotted the variables the other way around and wanted to determine whether a person's height was determined by their basketball performance (which makes no sense), we would still get *r* = .67. This is because the Pearson correlation coefficient makes no account of any theory behind why you chose the two variables to compare. This is illustrated below:



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

Scaling

It is a personal choice about making the numbers feel right, e.g. between zero and one, or one and a hundred. For example converting data given in millimeters to meters because it's more convenient, or imperial to metric.

Scaling is performed because most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference in scaling and normalisation

In normalization you are changing the **shape of the distribution** and in scaling you are changing the **range of your data**. Normalizing is a useful method when you know the distribution *is not* Gaussian. Normalization adjusts the values of your numeric data to a common scale without changing the range whereas scaling shrinks or stretches the data to fit within a specific range.

Scaling is useful when you want to compare two different variables on equal grounds. This is especially useful with variables which use distance measures. For example, models that use Euclidean Distance are sensitive to the magnitude of distance, so scaling helps even the weight of all the features. This is important because if one variable is more heavily weighted than the other, it introduces bias into our analysis.

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

A large value of **VIF** indicates that there **is** a correlation between the variables. 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

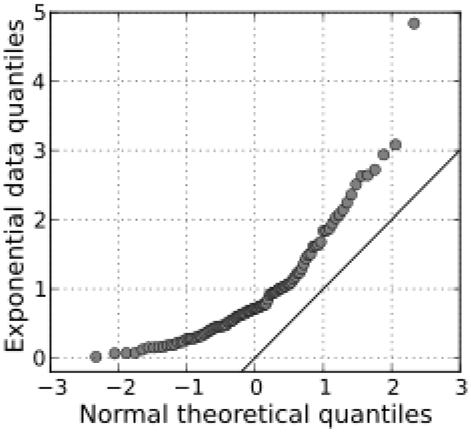
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This 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.

Moreover, Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.