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)
   Ans: Season - We can notice a positive trend in the number of customers in 2 - Summer, 3 - Fall and 4 - Winter seasons .This indicates, season can be a good predictor for the dependent variable.
   Year - The overall business shows a increasing trend in their user base year on year
   Month - Similar to the season trend, there is a postive trend in the months of summer, fall and winter.This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
   Holiday : On holidays, the users show a wider spread in the counts. On normal days, the users are more than holidays .On an average there is high demand in bikes during non holidays
   Weekday : Weekdays or weekends do not show any specific trend here.
   Weathersit : Clearer weathers show a postive trend in the number of bike users - 1: Clear, Few clouds, Partly cloudy, Partly cloudy - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans : drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans:We checked if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression), by plotting the histogram of the error terms .We observed that the error terms are normally distributed.
   we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.
   Using the pair plot, we could see there is a linear relation between temp variable with the predictor 'cnt'.
   There is a linear relation ship between final model and predictor variables
   There is no visible pattern in residual values, thus homoscedacity is well preserved
   From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Ans: As per our final Model, the top 3 predictor variables that influences the bike

booking are:

- **Temperature (temp)** - A coefficient value of '0.5180' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5180 units.
- **Weather Situation 3 (cloudy)** - A coefficient value of '-0.2225' indicated that, w.r.t Weathersit 1, a unit increase in Weathersit 3 variable decreases the bike hire numbers by 0.2225 units.
- **Year (yr)** - A coefficient value of '0.2347' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2347 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.Mathematically, we can write a linear regression equation as:y = a + bx

   There are 2 types of linear regression:

   1. Simple Linear Regression
   2. Multiple Linear Regression

   Simple Linear Regression: It is a type of linear regression model where there is only independent or explanatory variable. For e.g., the above scatter plot follows a simple linear regression with age being an independent variable is responsible for any change in height (dependent variable).
   Multiple Linear Regression: It is similar to simple linear regression but here we have more than one independent or explanatory variable.
   Steps to be followed in Linear Regression Algorithm:

   - Reading and understanding the data
     - Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization
     - Cleaning and manipulating data to make it up to the standards that exploratory data analysis can be performed by treating null values if any, updating to necessary formats, changing data types if needed, removing unwanted rows or columns etc.
   - Visualizing the data (Exploratory Data Analysis)
     - Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.
     - Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.
   - Data Preparation

    - Converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be represented during model building in order to contribute to the best fitted line for the purpose of better prediction.

- Splitting the data into training and test sets

    - Splitting the data into two sections in order to train a subset of dataset to generate a trained (fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets. Generally, the train-test split ratio is 70:30 or 80:20.

    - Rescaling the trained model: It is a method used to normalize the range of numerical variables with varying degrees of magnitude. For e.g. height or bmi or age are of different magnitude and units or some feature may have values in 10000s while feature may contain values in the magnitude of 10s or 100s, then the contribution of each feature for the dependent variable will be different

- Building a linear model

    - Forward Selection: We start with null model and add variables one by one. These variables are selection on the basis of high correlation with target variable. First we select the one, which has highest correlation and then we move on to the second highest and so on.

    - Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p- values).

    - RFE or Recursive Feature Elimination is more like an automated version of feature selection technique where we select that we need “m” variables out of “n” variables and then machine provides a list of features with importance level given in terms of rankings. A rank 1 means that feature is important for the model, while a rank 4 implies that we are better off, if we don’t consider the feature.

- Residual analysis of the train data:

    - It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0.

- Making predictions using the final model and evaluation:

    - We will predict the test dataset by transforming it onto the trained dataset

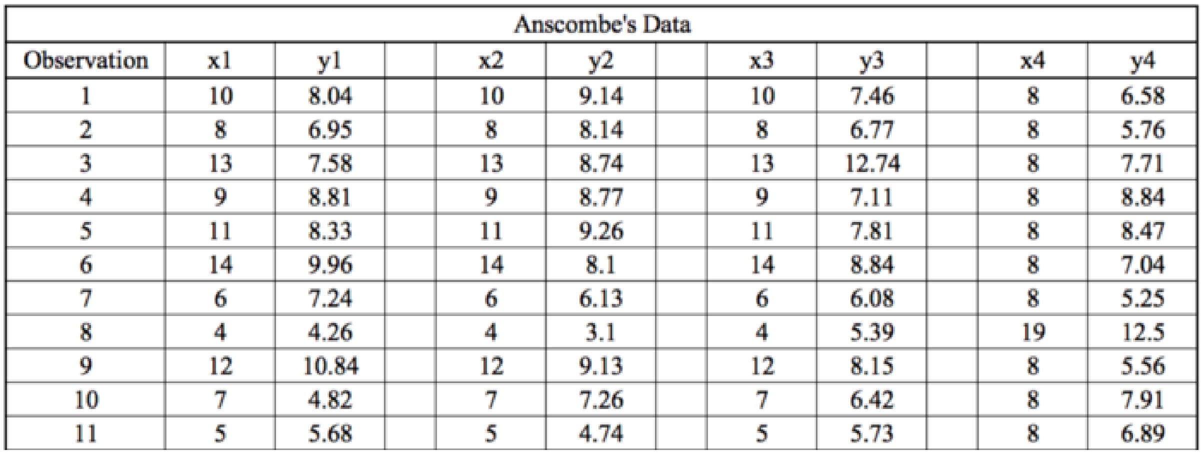
    - Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score. A difference of 2–3% between r2_score of train and test score is acceptable as per the standards.

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

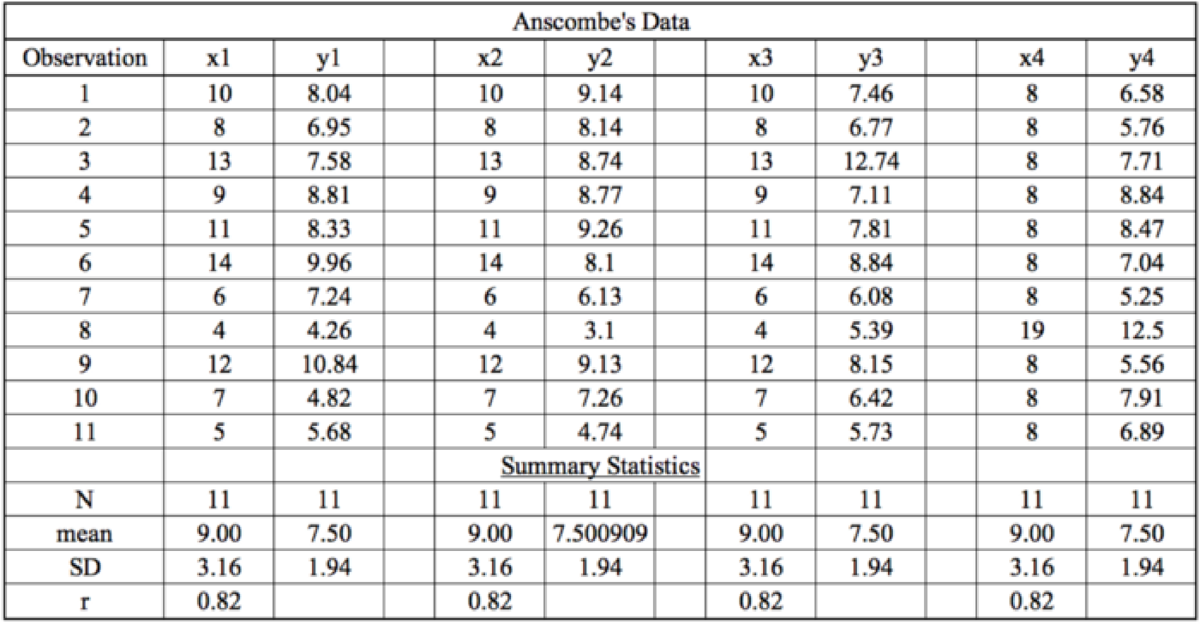
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which 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 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:



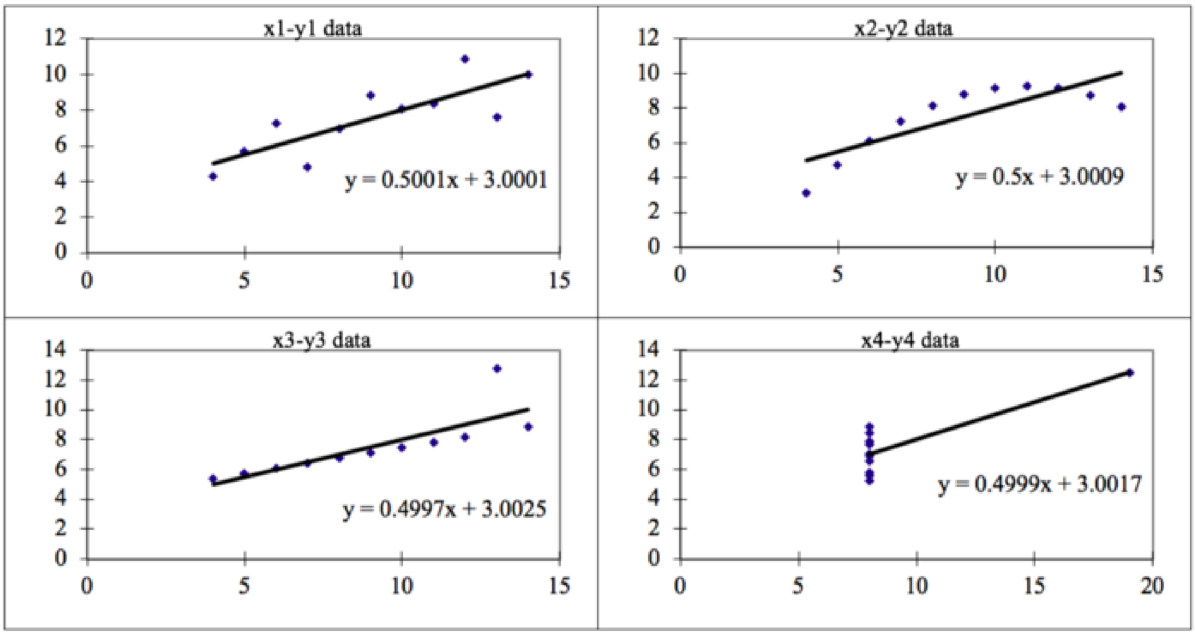
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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

The statistical information for all these four datasets are approximately similar and can be computed 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:



The four datasets can be described as:

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? (3 marks)

Ans : In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r**,** the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**.** It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Using the formula proposed by Karl Pearson, we can calculate a **linear relationship** between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

1. Scale of measurement should be interval or ratio
2. Variables should be approximately normally distributed
3. The association should be linear
4. There should be no outliers in the data

Pearson correlation coefficient formula:

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

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

   Ans: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

   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.

   Normalization/Min-Max Scaling:

   It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

   $$\text{MinMax Scaling: } x = \frac{x-min(x)}{max(x)-min(x)}$$

   Standardization Scaling:

   Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

   $$\text{Standardisation: } x = \frac{x-mean(x)}{sd(x)}$$

   sklearn.preprocessing.scale helps to implement standardization in python.

   One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

   Ans: 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.
Ans: Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.
Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.
We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.
Now we have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are not Normally distributed.
If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.
Q-Q plots are also used to find the Skewness of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution we want to know on the y-axis then we see a very peculiar shape of a Normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is left-skewed (or *negatively skewed*) but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer till to its right and it is right-skewed (or *positively skewed*).
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.
Few advantages:

   1. It can be used with sample sizes also
   2. 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.

It is used to check following scenarios:
If two data sets

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.Below are the possible interpretations for two data sets.

- **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.
- **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.
- **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

statsmodels.api provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.