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?

   The categorical variables are season, month, year, working day, weathersit, weekday and holiday

   - Season: We see that the dependant variable (cnt) which refers to the total number of rental bikes is highest during the fall season
   - Yr: Total number of bikes rented is highest in the year 2019
   - Month: The months August (8) and September (9) reported the higher sales of rental bikes
   - Working dayay: There is not much influence of working day on the sale of rental bikes
   - Weathersit: The bikes were rented much when the weathersit is clear_few clouds
   - Weekday: We do not see much much pattern with the weekday, as the no. of sales is almost constant across the weekdays
   - Holiday: The bikes were rented more when it was not a holiday

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

   The get_dummy function allows us to convert categorical data to indicator variables. If the categorical variable has 3 levels, we only need 2 indicator variables to represent it.

   For Ex: We have a categorical variable Gender with two values Male and Female, we can represent it using one indicator variable Female, where

   Female-0 indicated the Gender is Male

   Female -1 indicates Gender is Female

   Hence we use drop_first=True in get_dummies() function to drop the first Column(In our case Male)

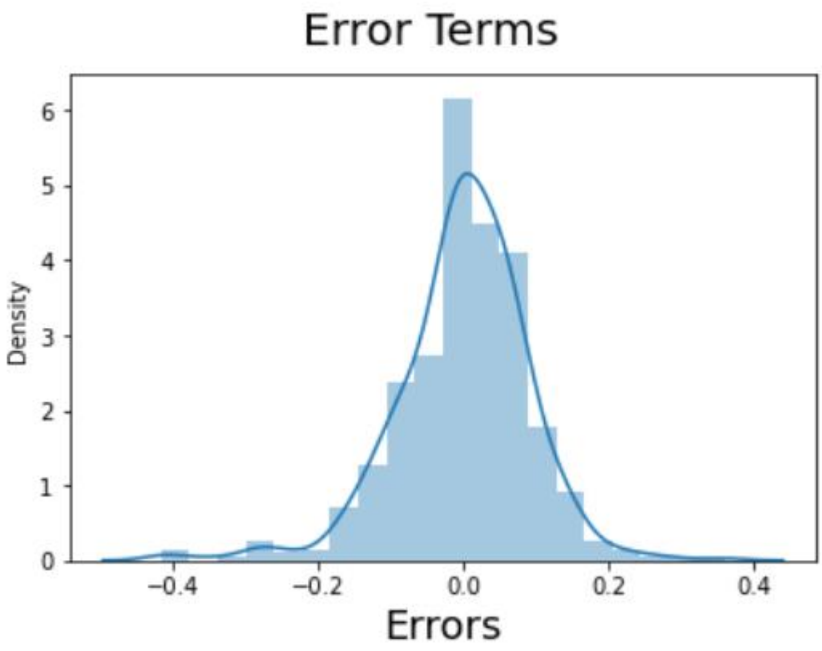
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Variable	Correlation
temp	0.63
atemp	0.63
Yr	0.57

Temp and atemp have the highest correlation with target variable cnt

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

   We did a residual analysis of the data and plotted the histogram of error terms to check if they are normally distributed which is the assumption if linear regression



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for the shared bikes?
   - atemp
   - Windspeed
   - Summer
   -

General Subjective Questions

1. Explain the linear regression algorithm in detail

   Simple linear regression explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points. The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1X$

   The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point are found by subtracting the predicted value of a dependent variable from the actual value of the dependent variable

   The strength of the linear regression model can be assessed using 2 metrics:

- R^2 or Coefficient of Determination
- Residual Standard Error (RSE)

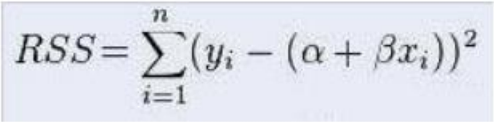
R2:

R2 is a number that explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general terms, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Mathematically, it is represented as $R^2 = 1 - (RSS / TSS)$

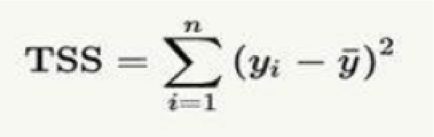
RSS (Residual Sum of Squares):

In statistics, it is defined as the total sum of errors across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.



$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

TSS(Total sum of squares): It is the sum of errors of the data points from the mean of the response variable. Mathematically, TSS is:

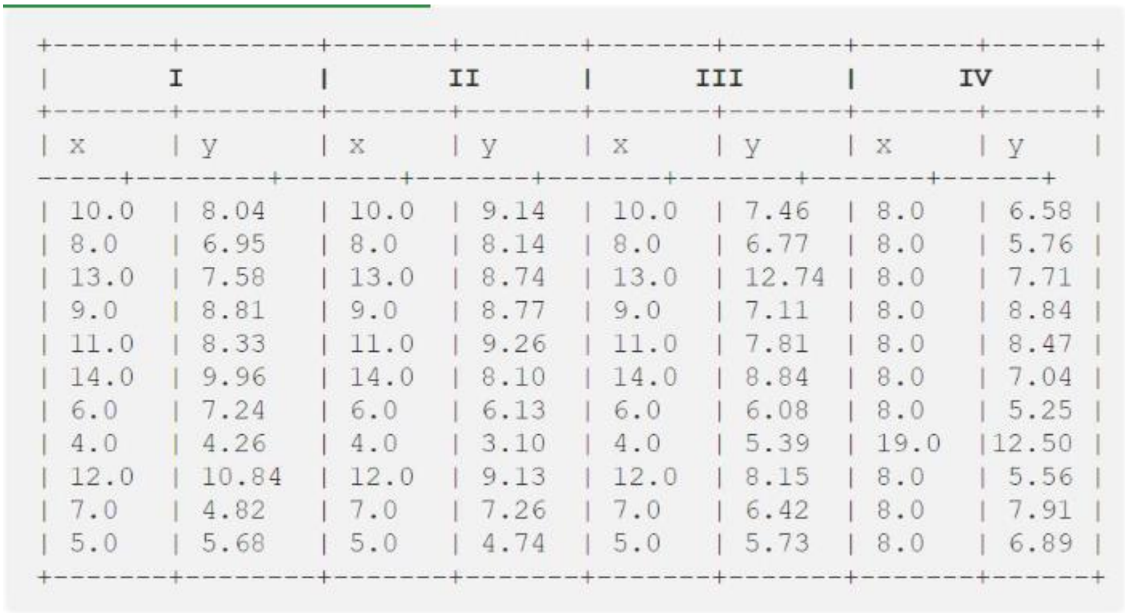


$$\mathbf{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

2. Explain the Anscombe's quartet in detail.

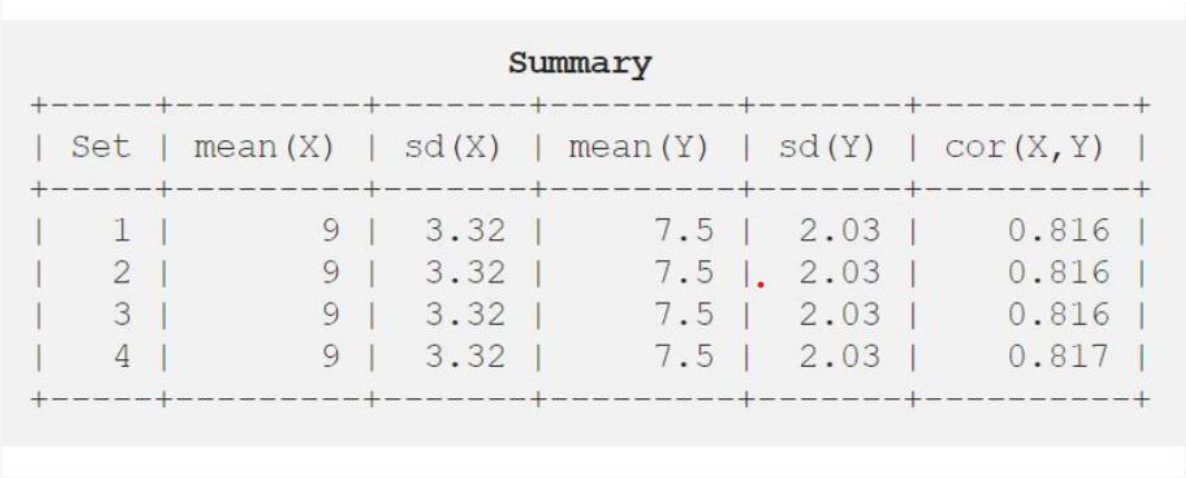
According to the definition given in Wikipedia, Anscombe's quartet incorporates four datasets that have nearly the same simple statistical properties, yet appear when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points Those 4 sets of 11 data points are given below.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

The table shows the mean, standard deviation, and correlation of those datapoints



```
                              Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R

   There are mainly three types of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

   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.

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Formula:

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

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

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

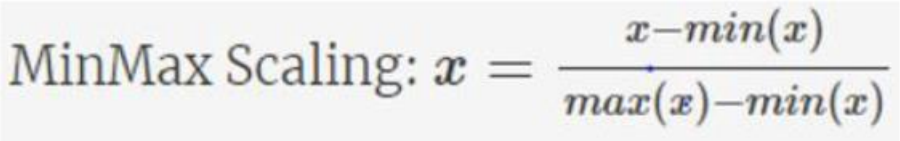
It is a step of data Pre-Processing that is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done then the algorithm only takes magnitude into account and not units hence incorrect modeling. 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.

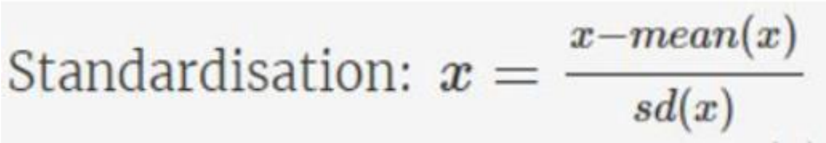
- Normalization/Min-Max Scaling:

    1. 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:
  i. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has a mean (μ) of zero and a standard deviation of one (σ).



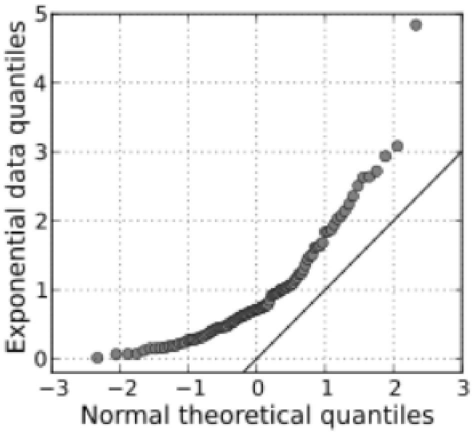
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

If there is a 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 leads 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.

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

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 falls below that point and 50% lies 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.