Assignment –based Subjective Questions

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

Answer:

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.

If the dataset have a small number of dummies, i suggest removing the first dummy. For example, if you have a variable Gender, you don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female.

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

- Highest correlation is between *registered* and *cnt.* With the target variable.
- As per the pair-plot, the *temp* vs *atemp* has high correlation(without target variable)

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

Linear Regression is one of the most important model to deal with.
Linear Regression have key assumptions: Linear relationship, Homoscedasticity, Absence of Multicollinearity, Independence of residuals (absence of auto-correlation) and normality of errors.

The validation can be done with respect to each assumption.

- The **linearity** is validated by plotting pair wise scatter plot, it is easy to visualize a linear relationship on a plot. We have plotted pair-plot to validate the same.
- **Homoscedasticity** means that the residuals have constant variance. This assumptions is validated by looking at the variance of error terms is constant across the dependent variable.
- **Multicollinearity** refers to the fact that two or more independent variables are highly correlated. Pair wise correlations using heatmaps could be first step to identify potential relationships between various independent variables, and next could be variance inflation factor VIF. Multicollinearity can be fixed by performing feature selection: deleting one or more independent variables. As multicollinearity is reduced, the model will become more stable and the coefficients' interpretability will be improved.
- **Independence of residuals** could mean that the linearity of the relationship is not respected or that variables may have been omitted. Auto-correlation would lead to spurious relationships between the independent variables and the dependent variable. This can be validated by, variables should be further fine-tuned and added to the model.
- **Normality of Errors** If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased. One could verify that the other assumptions are respected (i.e. homoscedasticity, linearity) to treat the large outliers.

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

Answer:

The top features contributing the final model based on the coefficients obtained.

Temp indicated that a unit increase in temp variable increases the bike hire numbers by 0.5982 Units.

weathersit_Partly_cloudy- indicated that, a unit increase in Weathersit_partly_cloudy variable decr eases the bike hire numbers by -0.2322 units.
Year (yr) −A coefficient value of 0.2284 indicated that a unit increase in yr variable increases the bike h ire numbers by 0.2308 units.

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   Answer:

   Firstly categorical variables in regression analysis are divided to Dummy variables through dummy coding. In dummy coding method it is observed that the test of significance of a given regression coefficient is equivalent to a test of difference between the mean of the group associated with the regression coefficient and the mean of the reference group.
   It is important to convert categorical variables into dummy to introduce them to regression analysis. It is not possible to calculate the R^2 and to understand the correlation between them. The effect on dependent variables which takes less time to obtain the results.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression Algorithm is a **Machine Learning** based on supervised learning.

Before knowing what linear regression is, let us get ourselves accustomed to regression. Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.
There are types of linear regression

1. Simple Linear Regression
2. Multiple Linear Regression etc.(mentioned as of now)

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation(**y=ax+b**). The motive of the linear regression algorithm is to find the best values for a and b.

In multiple linear regression two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables. In both cases there is only a single dependent variable.

The next important concept needed to understand linear regression is **Gradient descent**. Gradient descent is a method of updating a and b to reduce the cost function (MSE). The idea is that we start with some values for a and band then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

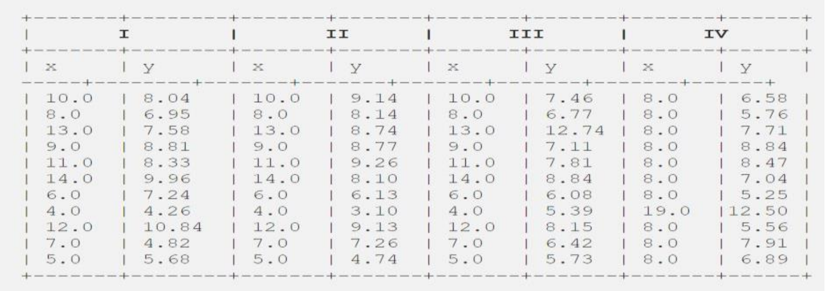
We have two choices, we can either use the **scikit learn** library to import the linear regression model and Use it directly or we can write our own regression model based on the equation.

We use scikit learn to import the linear regression model. We fit the model on the training data and predict the values for the testing data. We use **R2-square** to measure the accuracy of our model.

Conclusion:

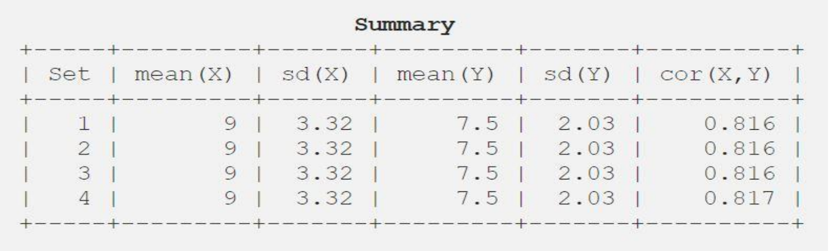
Linear Regression is an algorithm that every Machine Learning enthusiast must know and it is also the right place to start for people who want to learn Machine Learning as well. It is really a simple but useful algorithm. I hope this article was helpful to you.

2. Explain the Anscombe's quartet in detail. (Source: Geeks for geeks)
   Answer: According to the definition given in Wikipedia **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different 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 analysing it and the effect of outliers on statistical properties.
   Simple understanding:
   Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

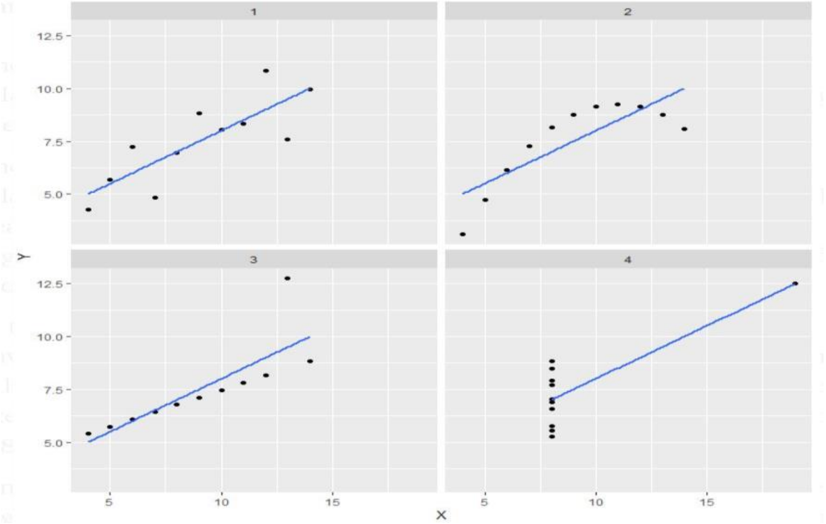


```
+-------+---------+--------+--------+---------+--------+--------+--------+
|       I         |        II       |       III        |       IV        |
+-------+---------+--------+--------+---------+--------+--------+--------+
| x     | y       | x      | y      | x       | y      | x      | y      |
-----+--------+--------+--------+--------+--------+--------+--------+
| 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   |
+-------+---------+--------+--------+---------+--------+--------+--------+
```

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.



```
                          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 |
+-----+----------+-------+----------+-------+----------+
```



Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

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?

- As the title suggests Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. You'll come across Pearson r correlation
- There should be **no significant outliers**. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient.
- Each variable should be **continuous** i.e. interval or ratios for example weight, time, height, age etc
- The two variables have a **linear relationship**.
- The observations are **paired observations.** That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

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

Answer:

What is Scaling?
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.

Why is Performed?
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.

Normalized Vs Standardized:

Normalized or MinMaxScaling:
It brings all of the data in the range of 0 and 1. From **sklearn.preprocessing import MinMaxScaler** helps to implement normalization in python**.**

MinMaxScaling(x) = x-min(x)/max(x)-min(x)

Standardized 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 (sd) one (**σ**)
sklearn.preprocessing.scale helps to implement standardization in python.

Standardization(x) = x-mean(x)/sd(x)

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?
Answer:

The more your **VIF(Variance Inflation Factor)** increases, the less reliable your regression results are going to be.
If there is perfect correlation, then **VIF** = **infinity**. A large value of **VIF** indicates that there is a correlation between the variables. In other words **infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer:
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.
statsmodels.api provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.
Advantages:

- it can be used with sample sizes also
- 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.