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?

Ans: The effect we can draw from the dependent variable is when we start looking into the module using linear regression, we tried to find the p-values and multicollinearity from the given dataset. But what we observed was the p-values were significant and when we went for VIF values some of the factors were more than 5 which considered as not efficient. So we have to drop those values and tried to build the model. We dropped some of the values like humidity, season, const and month which were showing more VIF values.

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

Ans: It is important to use drop_first=True during dummy variable creation because here we are storing the data of the given dataset in the new variable. Hence in order to not to make the dataset complicated by look we can drop the previous data, as we have a copy of that in dummy variable.

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

Ans: Bike rental counts show a positive correlation with temp and atemp and therefore bike rental counts increase at higher temperatures and vice-versa.

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

Ans: We will be using the **Mixed Approach** for model building i.e., firstly we will select 15 variables by using the **Automated Approach of RFE** and then using **Manual Approach for removing variables one by one based on the P-values and VIF values**. We will be using the **Linear Regression function from SciKit Learn** for its compatibility with RFE (Recursive Feature Elimination which is a utility from sklearn)

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

Ans: The three most significant variables affecting the demand for shared bikes are:

- **temperature**
- **year**
- **month September**

These features are having positive coefficients and an increase in them is going to result into an increase in the demand for shared bikes.

General Subjective Questions

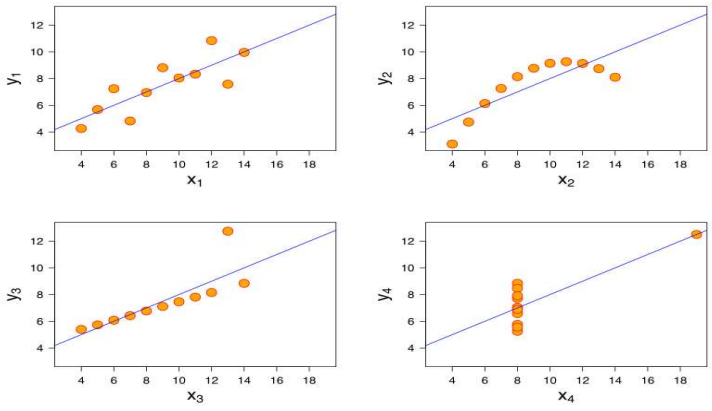
1. Explain the linear regression algorithm in detail.

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.

Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

2. Explain the Anscombe's quartet in detail.

Ans: **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.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

Ans: Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. **Pearson's** correlation (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's *R* first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.

Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

S_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance.

Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The PPMC is not able to tell the difference between dependent variables and independent variables. For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, as a researcher you have to be aware of the data you are plugging in. In addition, the PPMC will not give you any information about the slope of the line.

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

- 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 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: The **higher** the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to **high**, with values of 10 or more being regarded as very **high**. These numbers are just rules of thumb; in some contexts a **VIF** of 2 could be a great problem.

Secondly, how VIF is calculated? The Variance Inflation Factor (**VIF**) is a measure of colinearity among predictor variables within a multiple regression. It is **calculated** by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

Keeping this in consideration, why is Vif infinite?

If there is perfect correlation, then **VIF** = **infinity**. A large value of **VIF** indicates that there is a correlation between the variables. If the **VIF** is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

What does infinite VIF mean?

The user has to select the variables to be included by ticking off the corresponding check boxes. 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).

What is known is that the more your **VIF** increases, the less reliable your regression results are going to be. In general, a **VIF** above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above. Sometimes a high **VIF** is no cause for concern at all.

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

Ans: 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.

Few advantages:

a) It can be used with sample sizes also

b) 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 —

i. come from populations with a common distribution

ii. Have common location and scale

iii. Have similar distributional shapes

iv. Have similar tail behaviour

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.