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 variables yr has maximum correlation with the target variable cnt. We have seen that the month Aug and Sept has also shown an increase in cnt count with holiday and rain having a negative coefficient

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

➔ If you 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 then the person is a male and if male = 0 then the person is female. However if you have a category with hundreds of values, not dropping the first column. That will make it easier for the model to "see" all the categories quickly during learning (and the adverse effects are negligible).

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

➔ atemp has the highest correlation followed by yr column

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

➔ By checking if the **error terms** are also normally distributed and by plotting y_test and y_pred to understand the spread and calculating the r2_score

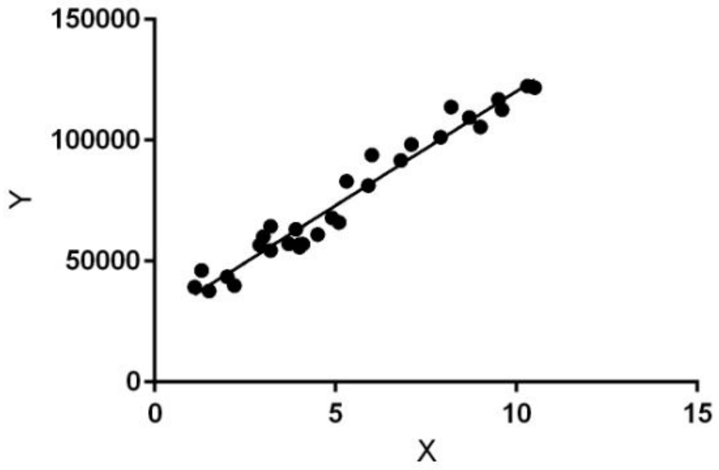
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

➔ Yr and atemp explains the demand of the shared bikes significantly along with rain with negative coefficient

General Subjective Questions

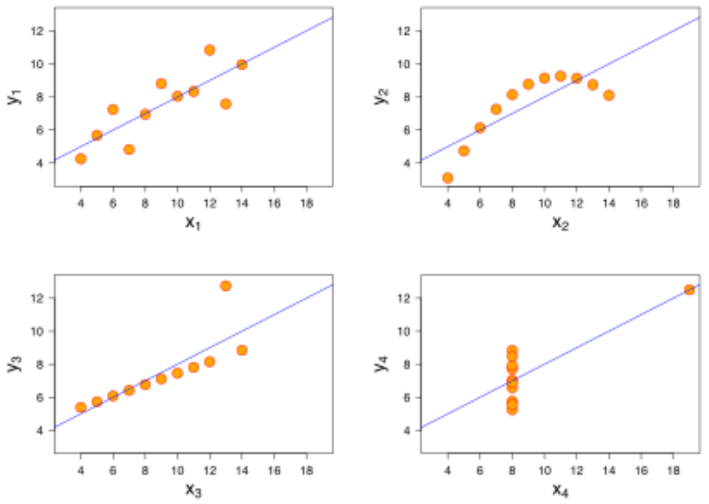
1. **Explain the linear regression algorithm in detail.**

➔ Linear regression is a type of supervised learning algorithm, commonly used for predictive analysis. As the name suggests, linear regression performs regression tasks. It is mostly used for finding out the relationship between variables and forecasting.
For example, you might use linear regression to see if there is a correlation between height and weight, and if so, how much – both to understand the relationship between the two, and predict weight if you know height. Statsmodels vs scikit-learn appear to be the two most popular libraries for modelling linear regression



2. **Explain the Anscombe's quartet in detail.**

➔ 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.



a) 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.
b) 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.
c) 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.
d) 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?**

➔ The Pearson correlation coefficient is a very helpful statistical formula that measures the strength between variables and relationships. In the field of statistics, this formula is often referred to as the Pearson R test. When conducting a statistical test between two variables, it is a good idea to conduct a Pearson correlation coefficient value to determine just how strong that relationship is between those two variables. In order to determine how strong the relationship is between two variables, a formula must be followed to produce what is referred to as the coefficient value. The coefficient value can range between -1.00 and 1.00. If the coefficient value is in the negative range, then that means the relationship between the variables is negatively correlated, or as one value increases, the other decreases. If the value is in the positive range, then that means the relationship between the variables is positively correlated, or both values increase or decrease together.

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

➔ Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling

Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\textbf{Standard Deviation}}$$

The terms normalization and standardization are sometimes used interchangeably. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

➔ If there is perfect correlation, then VIF = infinity. 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). For a better model the VIF value should be less than 5 or even 2 depending upon the model.

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

➔ 1. 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.

2. 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.

3. 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.