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)

I can conclude some of the following effects of categorical variables on the dependent variable from my research conducted on categorical variables of the data set:

- Clear weather is more suitable.
- Hike on bike rental in the fall.
- Hike in the number of bikes rent from June to September
- Counts are higher on weekdays rather than weekends.

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

It reduces the correlations created among dummy variables. So, it helps in reducing the extra column created during dummy variable creation.For instance, you don't need both a male and female dummy if you have a variable gender. Only one's going to be fine. If male=1, then the individual is a male, and if male=0, then the individual is female.

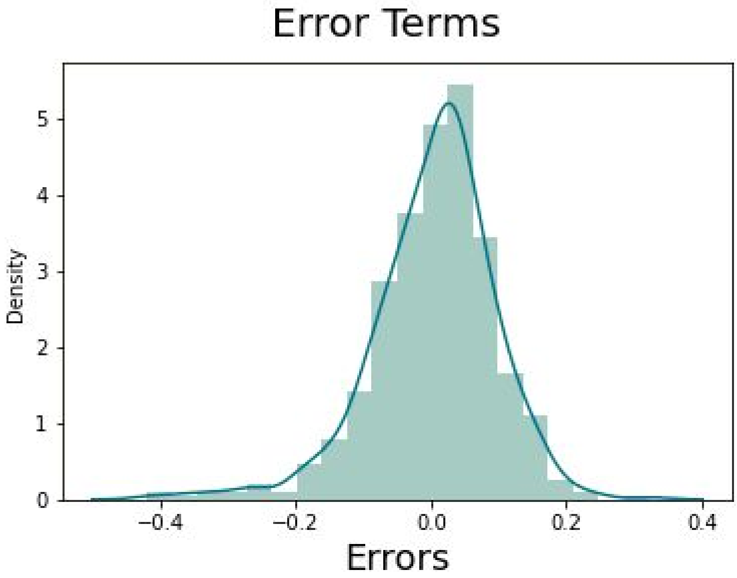
So, we can construct n-1 columns of dummy variables for a variable with n levels to better describe the variable. For this purpose, to construct n-1 columns for the n level variable, it is necessary to use 'drop first=True'.

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

‘temp’ has the highest correlation with the target variable ‘cnt’.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The assumption that we make on the training data set after constructing the linear regression model is that error terms are normally distributed. We have done the residual analysis in support of that. Residual is the error or discrepancy between the real target variable or the value of y and the model's expected value of y.We can see from the histogram below that the residuals are usually Assigned. Hence, our Linear Regression assumption is true.



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

As per our final Model, the top predictor variables that influence the cnt are:

1. Temp: A coefficient value of ‘0.6006’ indicated that a unit increase in temp variable increases the cnt numbers by 0.6006.
2. Snow: A coefficient value of - ‘0.2896’ indicated a unit increase in the Snow variable decreases the cnt numbers by 0.2896 units.
3. Yr: A coefficient value of ‘0.2382’ indicates that a unit increase in yr variable increases the cnt numbers by 0.2382 units.

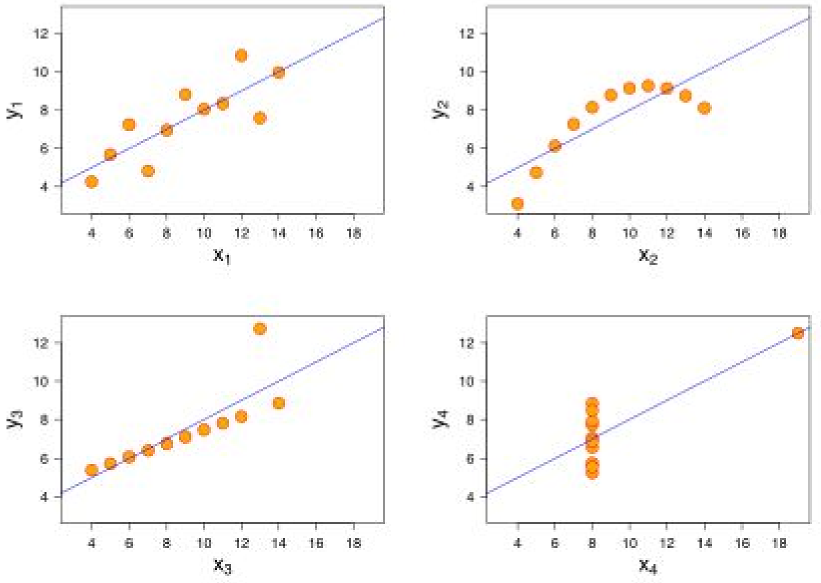
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

2. Explain 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 analyzing it and the effect of outliers on statistical properties.



- The first scatter plot (top left) is a simple linear relationship, corresponding to two variables correlated where y could be modeled 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 not linear, and the Pearson correlation coefficient is not relevant.
- In the third graph (bottom left), the distribution is linear but should have a different regression line. 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?

Pearson *r* correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. For example, in the stock market, if we want to measure how two stocks are related to each other, Pearson *r* correlation is used to measure the degree of relationship between the two. The point-biserial correlation is conducted with the Pearson correlation formula except that one of the variables is dichotomous. The following formula is used to calculate the Pearson *r* correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature Scaling : Most Important step to generalize your models. (Never skip this step)
Why Should We Use Feature Scaling?

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

In the formula, you can see Feature value X, affecting the Gradient Descent step size.

- The difference in ranges of features *(area of apartment will have enormous value compared to the number of rooms in the apartment)* will cause different step sizes for each feature.
- To ensure that the gradient descent moves smoothly towards the minima and that the gradient descent steps are updated at the same rate for all the features, we scale the data before supplying it to the model.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.In formula of VIF = 1 / (1 – R2) if R2 is equal to 1 then the denominator will become 0. As the denominator becomes 0 therefore it will be infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

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.