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?

From the above categorical analysis we can see that:
- *The median count of bike rentals is the lowest in spring and highest in the fall*
- *The median count of bike rentals in higher during June - Sept while it is the lower during December and January*
- *Bike rentals have increased in 2019 as compared to 2018, this indicates an upward trend.*
- *Bike rental demands are higher during weekdays and low during the holidays.*
- *Bike rental demand is relatively the same on all days of the week.*
- *Bike rental is high when the weather is clear, it starts to reduce when it is misty. When it is Snowing bike rentals are the lowest while no one takes bike rentals where there is heavy snow.*

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

It is important to use_drop=True during dummy variable creation as it reduces the number of columns by 1, hence it reduces the correlation in the dummy variables.

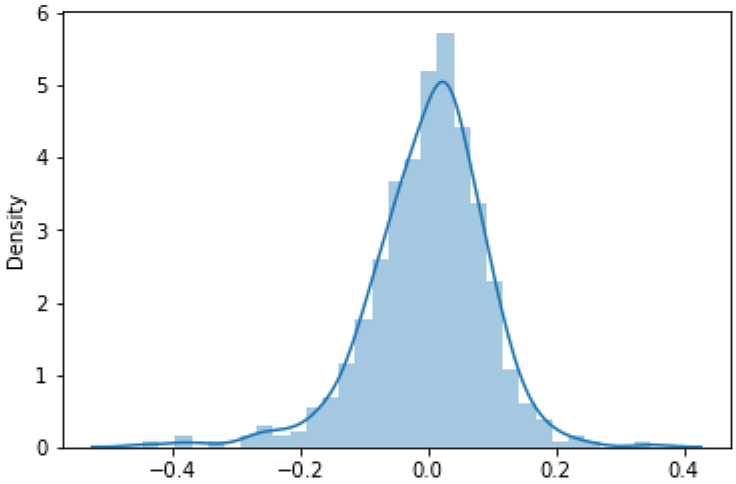
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Looking at the pair plot we see that 'temp' and 'atemp' has the highest correlation the target variable 'cnt'

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

We validated the Linear Regression model by plotting the error terms. This should be normally distributed and centered around 0.

```
res = y_train - y_train_pred
sns.distplot(res)
plt.show()
```



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

   Top 3 features are -
   1. *temp with coeff = 0.436175*
   2. *yr with coeff = 0.234850*
   3. *weathersit_Light Snow = -0.288561*

General Subjective Questions

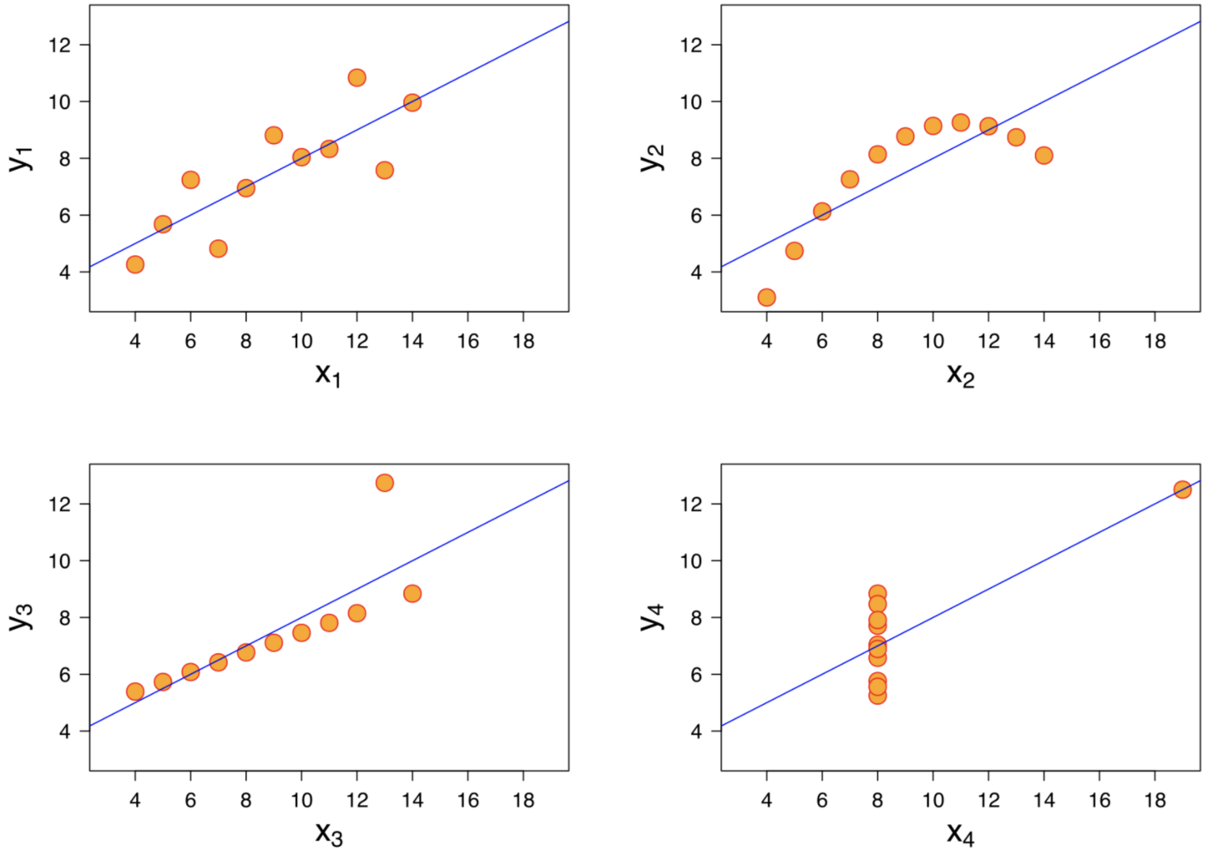
1. Explain the linear regression algorithm in detail.

   Linear Regression is a supervised Machine Learning Algorithm that is used in the prediction of numeric variables. It assumes that there is a linear relationship between the dependent variable - y and the predictor variables - x.
   This is depicted by the equation y= mx + c.
   There are two type of Linear Regression -
   - *Simple Linear Regression - when the dependent variable is predicted by using one variable.*
   - *Multiple Linear Regression - when the dependent variable is predicted using multiple independent variables*
     The equation is denoted by -

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet, was constructed by Francis Anscombe in 1973, 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.
It was constructed to demonstrate both the importance of graphing data before analyzing it, and the effect of outliers and other influential observations on statistical properties.



- *The first scatter plot (top left) appears to be a simple linear relationship,*
- *The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear.*
- *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?

   Pearson’s correlation coefficient, R, measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship

   Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

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.

   Scaling is of two types - Normalized scaling and Standardized Scaling

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

The Variance Inflation Factor (VIF) provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
〖VIF〗 =1/(1-R^2)

If there is perfect correlation, then VIF = infinity. This is because the said independent variable can be perfectly explained by other independent variables and the R square value be 1.
So VIF = 1 / (1-1) = 1/0 = infinity.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is used to compare the shapes of distributions.A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight.

It is used to find out that if the -

- *Two data sets come from populations with a common distribution?*
- *Two data sets have a common location and scale?*
- *Two data sets have similar distributional shapes?*
- *Two data sets have similar tail behavior?*