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 :

There are total 6 categorical variables in our dataset.

Inference about effect of categorical variables on the dependent variable "cnt" are:

1. **season** : category 3 i.e.,fall have highest number of bike hiring.
2. **month** : month may, june, july, aug, sept have higher number of booking i.e., greater than 4000.
3. **holiday** : Majority of booking are done when there is no holiday. Hence, this column can be bias towards no holiday and cannot be use for prediction.
4. **weekday** : All days are having close trend.
5. **workingday** : Majority of the bike booking were happening in 'workingday' with a median of close to 5000.
6. **weathersit** : Category 1 have highest no of bike booking, category 2 have second highest no of booking while category 3 have lowest.

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

Ans :

By using **drop_first=True,** it will reduce one column that will get create while creating dummy variables and due to reduction in one column the correlation created between each variables will get reduce.

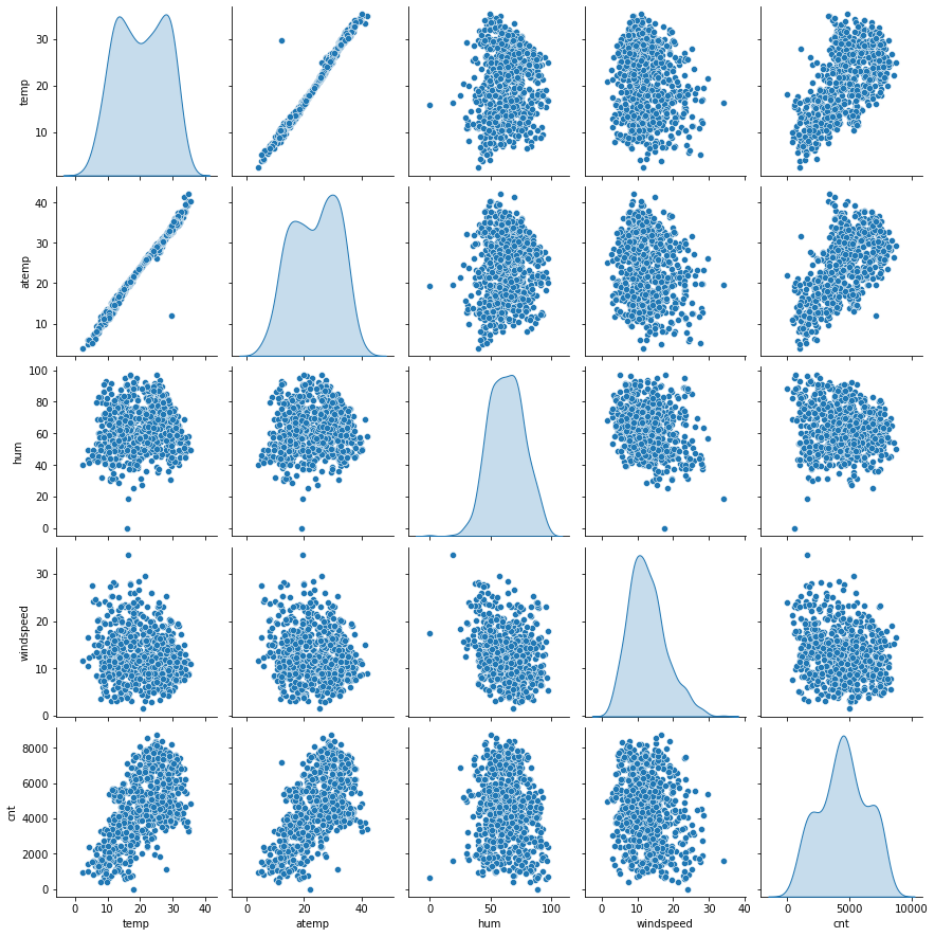
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans:
From the pair-plot, we can conclude that **'temp'** and **'atemp'** features have highest positive correlation with target variable 'cnt'

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

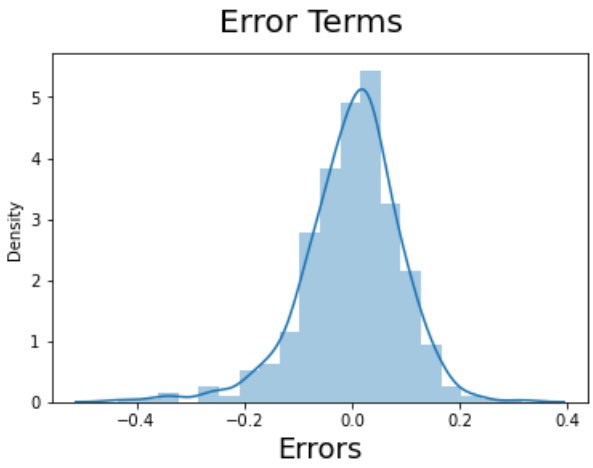
Ans:

1. By plotting the pairplot, we can know the relationship between variables with target variable whether it is linear or not. Hence, the assumption that "**There is a linear relationship between X and Y**" can be validate.



In our model, temp and atemp has linear relation with target and thus, this assumption is valid.

2. By Residual Analysis we can know whether the error terms follows the normal distribution curve or not. Hence, assumption that "**Error terms are normally distributed with mean zero" can be validate.**



In our model, the residual analysis plot is shown above which follows normal distribution. Thus, this assumption is valid.

3. By finding the VIF value for each variable we can know whether multicollinearity is present between variables or not. We can say that If VIF<5, there is no multicollinearity present and if vif > 5, multicollinearity between variables is present. Hence, the assumption that "**No Multicollinearity between the predictor variables"** can be validate.

   In our model, the VIF value of all features is below 5. Thus this assumption is valid.

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

Ans:

Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **Temp** with 0.5436 as coefficient value
2. **weathersit_3** with -0.2936 as coefficient value
3. **yr** with 0.2333 as co coefficient value

General Subjective Questions :

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

Ans:

- Linear regression is supervised 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 weight, sales, salary, age, product price, etc.

- Linear regression algorithm shows a linear relationship between a dependent or target variable y and one or more independent x variables, hence called as 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.
- The linear regression model provides a sloped straight line representing the relationship between the variables.

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression**

  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**

  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

- There can be positive linear relation or negative linear relation between variables.
- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
- For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values.
- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.
- The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization**.** It can be achieved by R-squared method.

Assumptions of Linear Regression

1. Linear relationship between the features and target:
2. No multicollinearity between the features.
3. Normal distribution of error terms.
4. No autocorrelations.

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.
- These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.
- Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can describe the four data sets as:

- **Data Set 1**: fits the linear regression model pretty well
- **Data Set 2**: cannot fit the linear regression model because the data is non-linear
- **Data Set 3**: shows the outliers involved in the data set, which cannot be handled by the linear regression model
- **Data Set 4**: shows the outliers involved in the data set, which also cannot be handled by the linear regression model
- Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3.What is Pearson's R?

Ans:

- Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation.
- It is a statistic that measures the linear correlation between two variables. It has a numerical value that lies between -1.0 and +1.0.
- It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
- The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables.

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

Ans:

What is scaling?
Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

Why is scaling performed?

- Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.
- The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Example:

- Suppose we have two features of weight and price, as in the below table. The "Weight" cannot have a meaningful comparison with the "Price." So the assumption algorithm makes that since "Weight" > "Price," thus "Weight," is more important than "Price."
- So these more significant number starts playing a more decisive role while training the model. Thus feature scaling is needed to bring every feature in the same footing without any upfront importance.

Normalization:
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution.

Standardization:
Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Ans:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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.
- A quantile is a fraction where certain values fall below that quantile.
- For **example**, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
- If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.