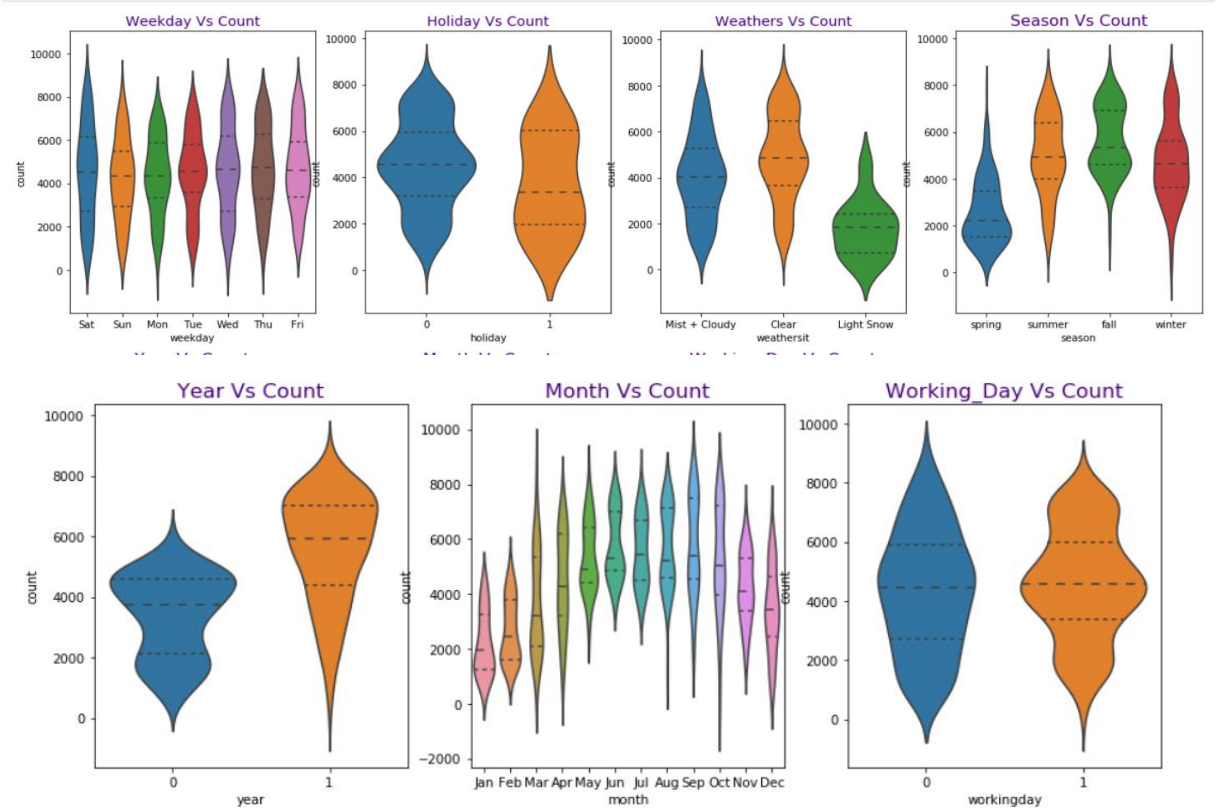
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Given day dataset the categorical variables used in this dataset are

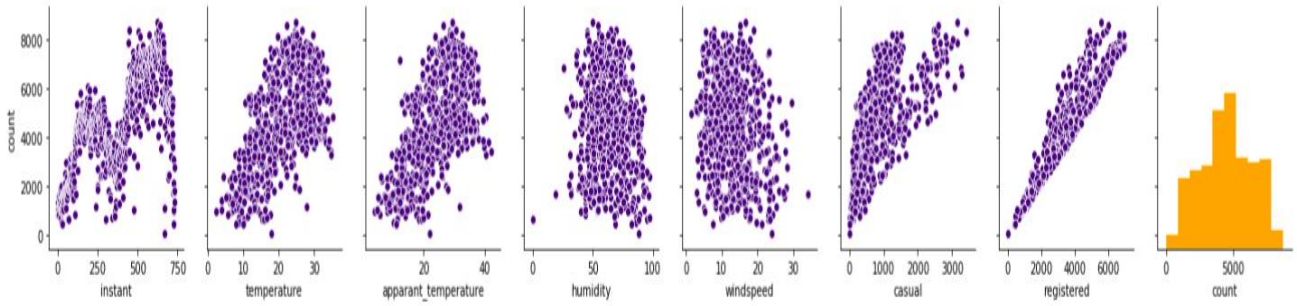
- Weather sit: Bikes Rental count is more when there is clear weather.
- Season: More Bike Rental count is observed during summer and fall season.
- Weekday: Bike rental count is kind of similar during all weekdays.
- Holiday: From the analysis it is seen that during holidays less bike rentals are taken.
- Year: Count is more in year 2019 than 2018.
- Month: September month has more bike rental count.



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

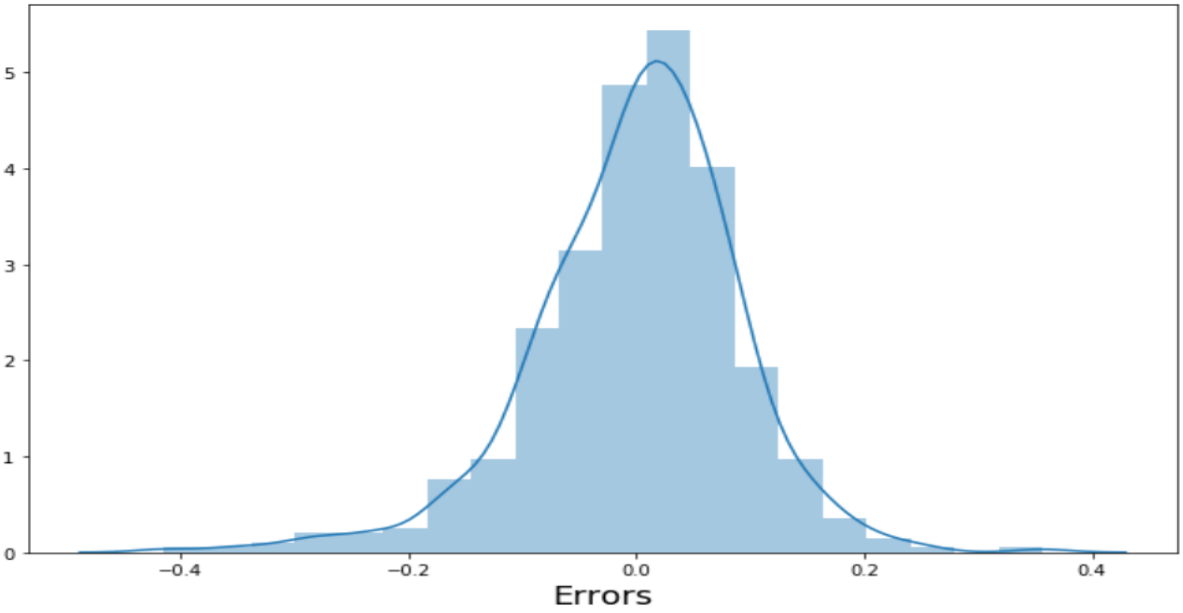
- This is an important step considered during dummy variable creation, which reduces the extra column created during dummy variable creation there by reducing the redundancy among the variables.
- One more important point is that it also reduces multicollinearity among the variables otherwise which would have created unwanted noisy data and would have adversely affected during model building process.

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



By observing the above pairplot segment of numerical variables with **count** target variable, it is seen that casual, registered, temperature, apparent_temperature variables have linear correlation. During the process of analysis, casual, registered and apparent temperature variables are dropped. Thus, resulting **temperature** has the highest correlation with target variable **count**.

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



We need to check on residuals after the model is built, where these residuals must be normally distributed having mean = 0. The same is obtained in the current model building where when a distribution plot is plotted, we see a normal distribution. The same has been depicted in the above graph.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top factors that are contributing significantly towards demand of shared bikes are as below: -

temperature	0.490988
year	0.23357
winter	0.081741
Sep	0.076846

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression algorithm is the supervised machine learning technique which is part of regression analysis. Its one of the simplest regression analysis technique which is used to interpret how the independent input variables have influence over the dependent target variable.

By the name linear, it means the variables on x and y axis are linearly correlated. In regression we define best fit line by the relationship between dependent and independent variables. In this method dependent variable is the function of coefficients, independent variables and error terms.

There are two types of linear regression mainly

- **Simple Linear Regression:** Is the type of linear regression where there is one target variable and one independent variable.
  Is represented by $y = mx + c$
  m: where m is slope
  c: is the intercept

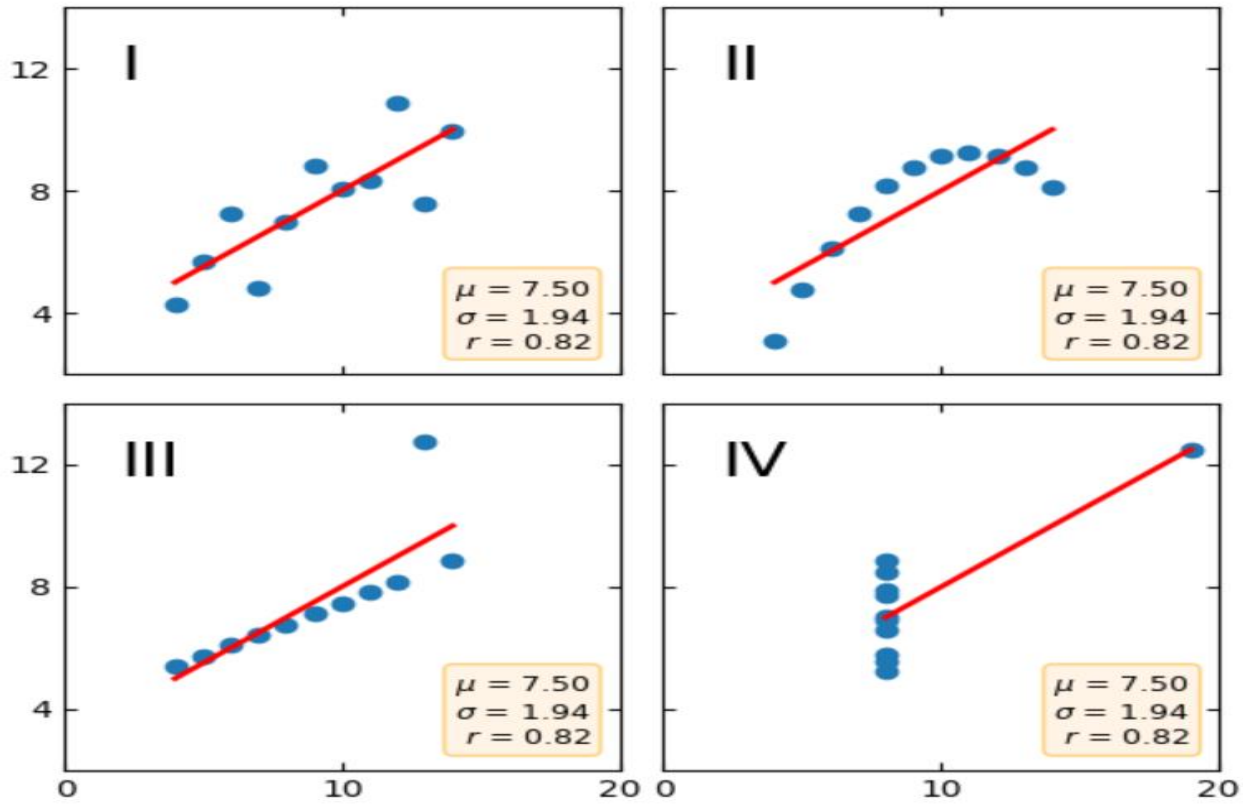
- **Multiple Linear Regression:** Is the type of linear regression where there is one target variable and many independent variables.
  Is represented by $y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3$ ----- βnXn.
  $\beta0$: is the intercept
  $\beta1$: is co-efficient of X1
  $\beta2$: is co-efficient of X2
  $\beta3$: is co-efficient of X3 and so on …

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet mainly comprise of four datasets which have identical descriptive statistics, but they have different distributions and appear differently when visualized via graph.

We find below mentioned observations with reference to the graph below.

- **The First Quadrant:** is a scatter plot, which seems to be a simple linear relationship, corresponding to y modelled as gaussian with mean linearly dependent on x.
- **Second Quadrant:** it is observed that data is not distributed normally, it is not linear. Thus, general regression and corresponding co-efficient of determination would be more appropriate.
- **Third Quadrant:** it is observed that distribution is linear but should have different regression line. The calculated regression is offset, by the one outlier which will influence enough to lower the correlation coefficient.
- **Fourth Quadrant:** here we see how one high leverage point is enough to produce high correlation co-efficient, though other points don't show up any relation between variables.



3. What is Pearson's R? (3 marks)

Pearson's R which is also referred as Pearson correlation coefficient in statistics. It is the measure of linear correlation between two sets of data. It is the ratio between covariance between two variables and product of there standard deviation which lies between the values -1 to +1

Say:

- If r is between 0 and 1 then the data is perfectly linear with positive slope.
- If r is between -1 and 0 then the data is perfectly linear with negative slope.
- If r =0 then there is no linear association between two sets of data.

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

- Scaling is one of the preprocessing steps applied to normalize the data, that is bringing all the data to comparable magnitude, so that model building becomes easier.
- Initially when dataset is obtained it comprises of data which is distributed with different magnitude, scale, units. Thus, without scaling steps the data won't be brought to uniform magnitude, that's the reason scaling is being done.
- The difference between normalized and standardized scaling is
    - **Normalized Scaling:** It brings all the data in range of 0 and 1.
    - For this we use *sklearn.preprocessing.MinMaxScaler* in python
    - MinMax Scaling: x = x – min(x) / max(x) – min(x).

    - **Standardized Scaling**: this is a method which replaces values by their Z scores. That is, it brings all of data into standard normal distribution which has mean zero and standard deviation as one.
    - For standardized we use *sklearn.preprocessing.scale* in python.
    - Standardization**:** x = x -mean(x)/ sd(x)

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

- VIF – Stands for Variance Inflation Factor which is measure of multi-collinearity in any given dataset.
- If there is a perfect correlation, then we observe that VIF is infinity. That is in this case RSquare is one and
    - VIF = 1 / 1 – Rsquare
    - which will be 1/ 1-1 = 1/0 = infinity.
- Thus, to solve such situation we usually recognize and drop variables which are causing VIF as infinity.

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

- Q- Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- This helps in linear regression when we have received training and test data set separately and then to confirm we use using Q-Q plot to check if both the data sets are from populations with same distributions.
- A quantile is a fraction where certain values fall below that quantile.
- The purpose of Q Q plots is to find out
    - If two sets of data come from the same distribution.
    - To check if the two data sets come from a common distribution,
    - To check if the points will fall on that reference line.