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

Six categorical variables are

- Month
- Weekday
- Season
- Weathersit
- Holiday
- Workingday

Box plots were used to analyze the effect on dependent variable 'cnt'

inferences:

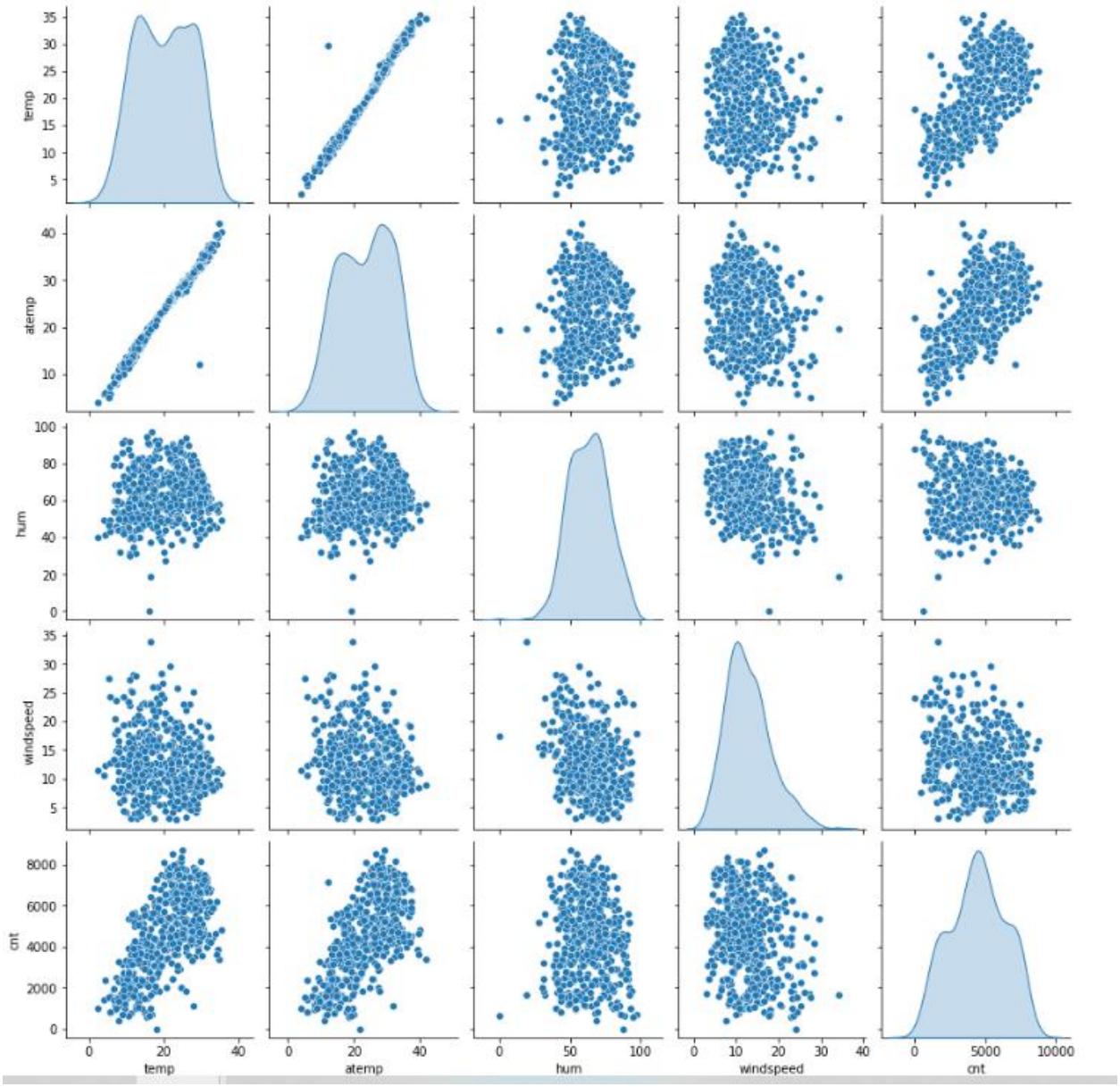
- **season**: 30% of the bike booking happened in season3 with a median of over 5000. This was followed by season2 & season4 with 27% & 25% of total booking. Hence season is a good predictor
- **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. Hence month can also be a good predictor of 'cnt'
- **weathersit**: Weathersit is not showing any trend in the boxplot hence can't be a good predictor
- **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
- **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

- drop_first=True helps in reducing the extra column created during dummy variable creation.
- Hence it reduces the correlations created among dummy variables

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

- Using pair plot it can be analyzed that 'cnt' has linear relationship with 'temp'
- Using pair plot it can be analyzed that 'cnt' has linear relationship 'atemp'

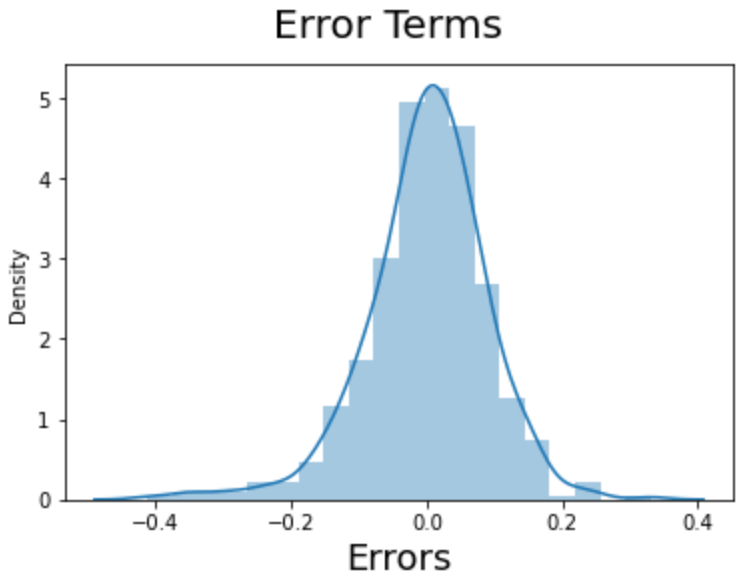


4. *How did you validate the assumptions of Linear Regression after building the model on the training set?*

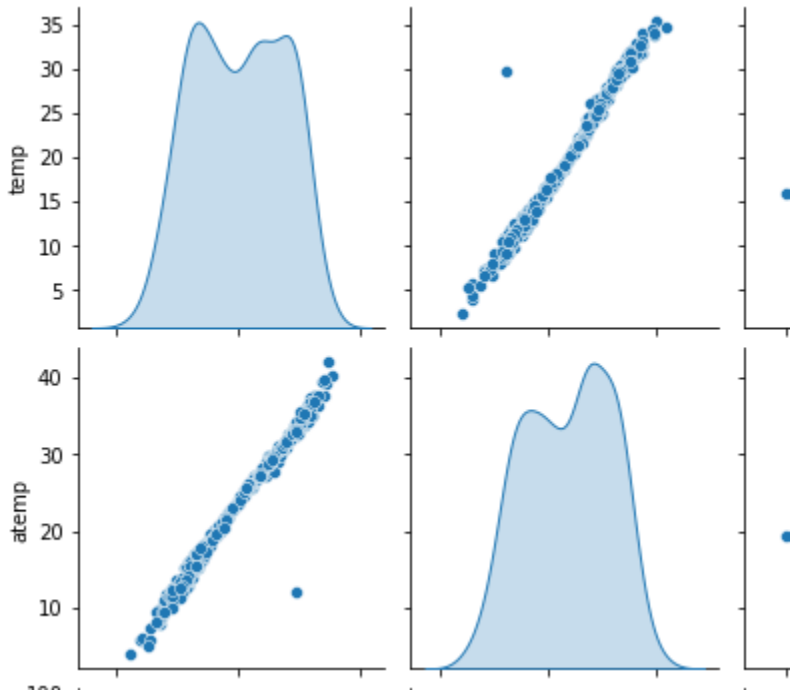
1. Error terms are normally distributed with mean zero (not X, Y)

From the histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid. Hence our assumption for Linear Regression valid.

```
Text(0.5, 0, 'Errors')
```



2. There is a linear relationship between X and Y
   Using the pair plot, we could see there is a linear relation between 'temp' and 'atemp' variable with the predictor 'cnt'



3. There is No Multicollinearity between the predictor variables
   From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5 (apart from temp but it can't be removed since it's an important variable)

	Features	VIF
2	temp	6.10
1	workingday	4.04
3	windspeed	3.49
0	yr	2.02
4	season_2	1.81
8	weekday_6	1.69
6	mnth_8	1.58
9	weathersit_2	1.54
5	season_4	1.51
7	mnth_9	1.30
10	weathersit_3	1.08

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 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of '0.5209' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5209 units.
- **Weather Situation 3 (weathersit_3)** - A coefficient value of '-0.2869' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
- **Year (yr)** - A coefficient value of '0.2328' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2328 units.

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

- Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
- Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + b$$

- Here, Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is the slop of the regression line which represents the effect X has on Y

- b is a constant, known as the YY-intercept. If X = 0,Y would be equal to bb.
- Furthermore, the linear relationship can be positive or negative in nature
- Following are the assumptions of linear regression

Furthermore, Linear regression is of two types Simple Linear Regression and Multiple Linear Regression

Multi-collinearity − Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.

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

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Source - https://medium.com/analytics-vidhya/anscombes-quartet-an-importance-of-data-visualization-856b3d1bd403

8. *What is Pearson's R?*

Pearson's correlation (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression.

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

- Normalization/Min-Max Scaling brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

- Whereas Standardization replaces the values by their Z scores. It brings all of the data into standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**). **sklearn.preprocessing.scale** helps to implement standardization in python.

- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

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 (which show an infinite VIF as well).

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

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.

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. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.