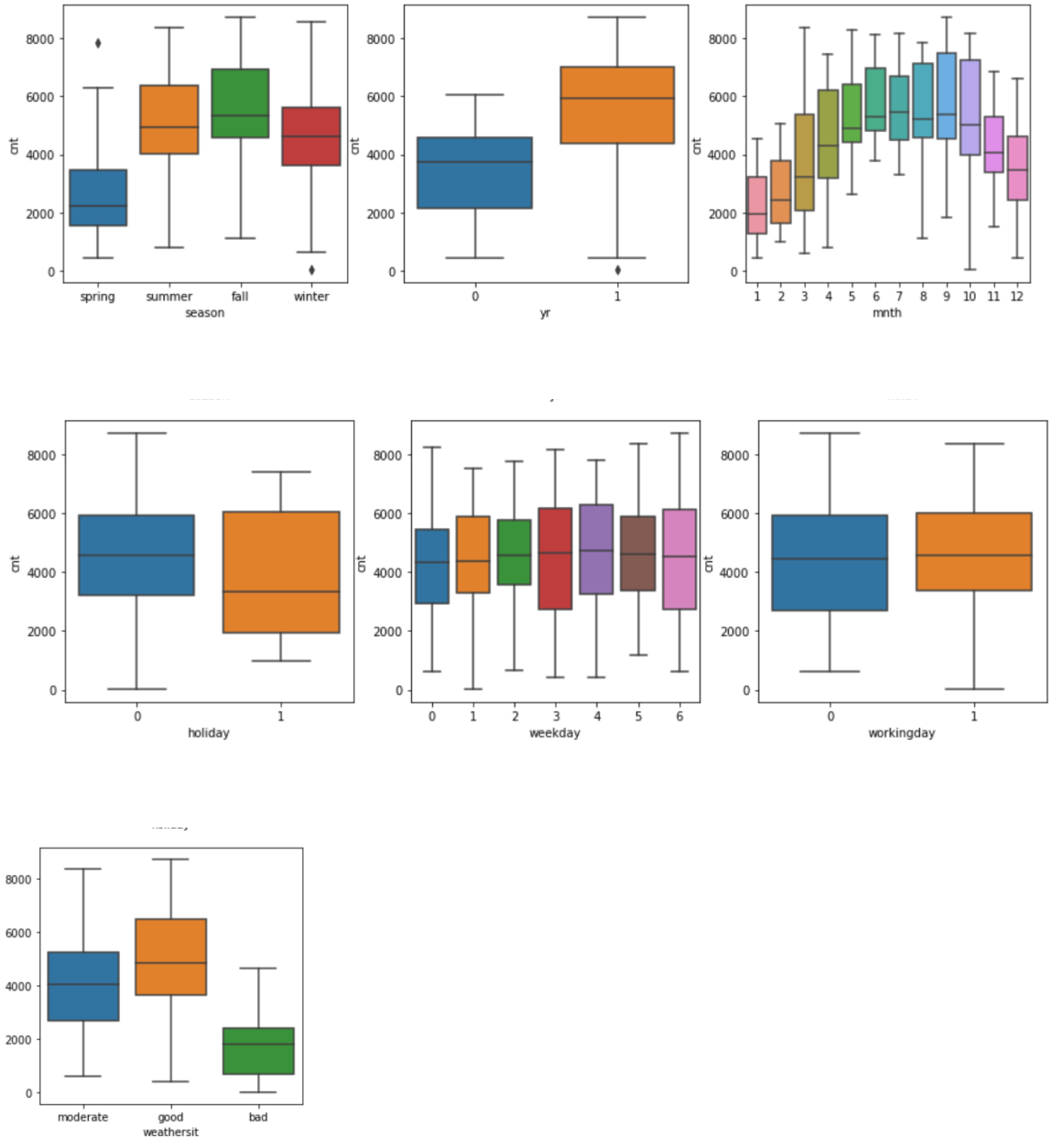
Assignment-based Subjective Questions

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 7 categorical variables i.e.
'season','yr','mnth','holiday','weekday','workingday','weathersit'

```
# Boxplot for categorical variables to see demands
vars_cat = ['season','yr','mnth','holiday','weekday','workingday','weathersit']
plt.figure(figsize=(15, 15))
for i in enumerate(vars_cat):
    plt.subplot(3,3,i[0]+1)
    sns.boxplot(data=bikesales, x=i[1], y='cnt')
plt.show()
```



- If you see if season is Fall then bike demand is getting increased where as in winter and in spring it is getting decrease.

- If year 2019 then rental bikes demand is high.
- If you see month impact it is in line what we see in season i.e. July, Aug, Sept, Oct having high demand.
- There is not much impact of weekday as well.
- In case of holiday is 0 median is high so it means more people take bike on rents during working day.
- Weather situation is bad resulting in low bike rentals.

Q2: Why is it important to use drop_first=True during dummy variable creation?

Ans: To get rid from **Dummy Variable Trap.** The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use one-hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a dummy variable trap. So, the regression models should be designed to exclude one dummy variable.

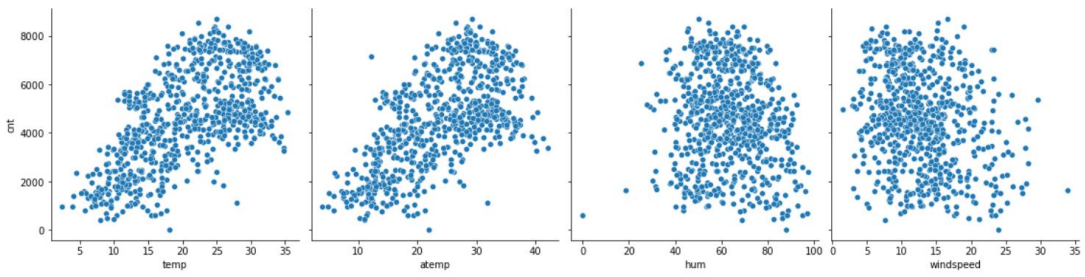
How to identify it : if we see inf in VIF output as well as R2 as 1.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans :

```
In [11]: #Draw pairplots for continuous numeric variables using seaborn
         plt.figure(figsize = (15,30))
         sns.pairplot(data=bikesales,x_vars=['temp', 'atemp', 'hum','windspeed'], y_vars='cnt',size=4, aspect=1, kind='scatter')
         plt.show()
```

```
<Figure size 1080x2160 with 0 Axes>
```



Temp and atemp both having high correlation with output variable cnt.

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

```
In [145]: y_train_pred = lr.predict(X_train[cols])

In [146]: #Print R-squared Value
          from sklearn.metrics import r2_score
          print(r2_score(y_train,y_train_pred))
          print(r2_score(y_test,y_test_pred))
```

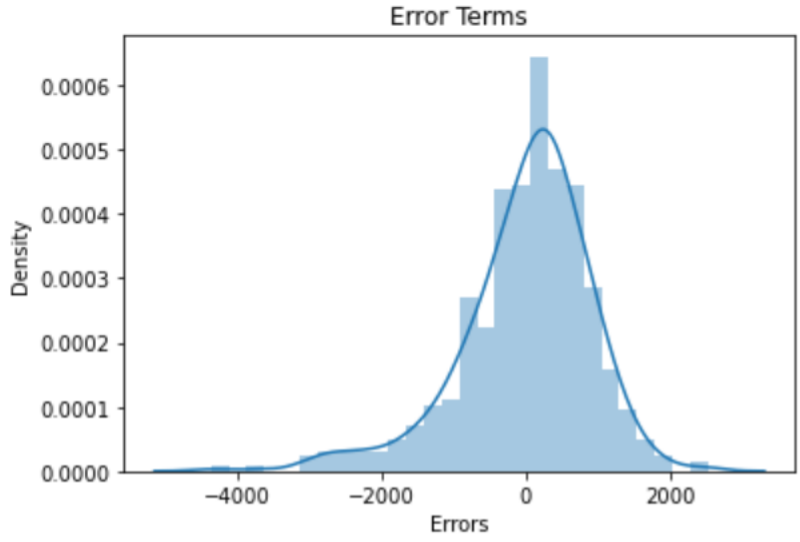
```
0.7798173997884894
0.7530732958122833
```

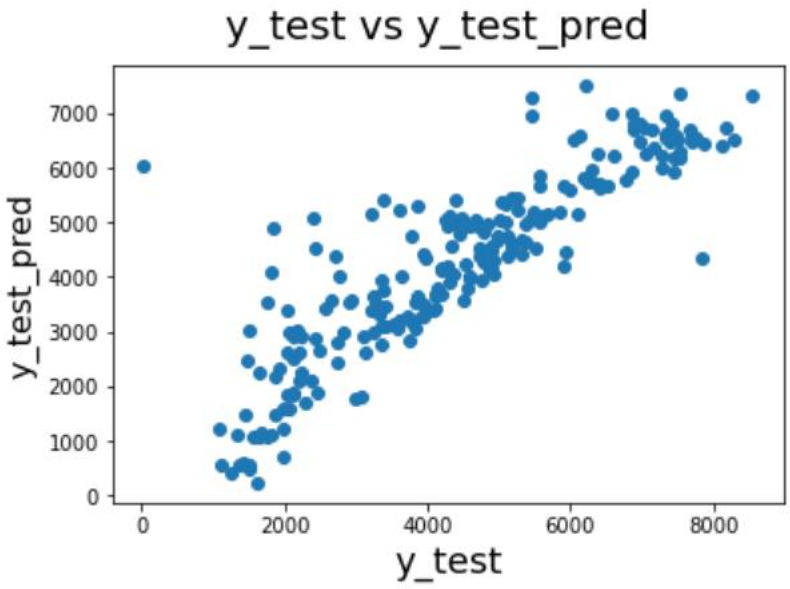
By using the r2 score on test data set we verify whether it is nearly equal to train r2 score or not.

We also checked error distribution which is normally distributed so seems to be correct.

```
    sns.distplot(act-pred)
    plt.title('Error Terms')
    plt.xlabel('Errors')

plot_res_dist(y_train, y_train_pred)
```





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

Ans : yr and season_winter

	coef	std err	t	P>|t|	[0.025	0.975]
const	708.1649	301.920	2.346	0.019	114.983	1301.347
yr	2038.9837	81.438	25.037	0.000	1878.983	2198.985
temp	133.1542	9.878	13.480	0.000	113.747	152.561
holiday	-592.5539	249.555	-2.374	0.018	-1082.855	-102.253
season_spring	-653.2314	209.282	-3.121	0.002	-1064.409	-242.054
season_summer	480.7224	135.074	3.559	0.000	215.342	746.103
season_winter	873.9456	161.209	5.421	0.000	557.218	1190.673
mnth_sept	725.4990	152.811	4.748	0.000	425.271	1025.727
weathersit_moderate	-560.9028	85.844	-6.534	0.000	-729.560	-392.246

General Subjective Questions

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

Linear regression analysis is **used to predict the value of a variable based on the value of another variable**. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: When the number of independent variables is 1
- Multiple linear regression: When the number of independent variables is more than 1

The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:

- Differentiation
- Gradient descent method

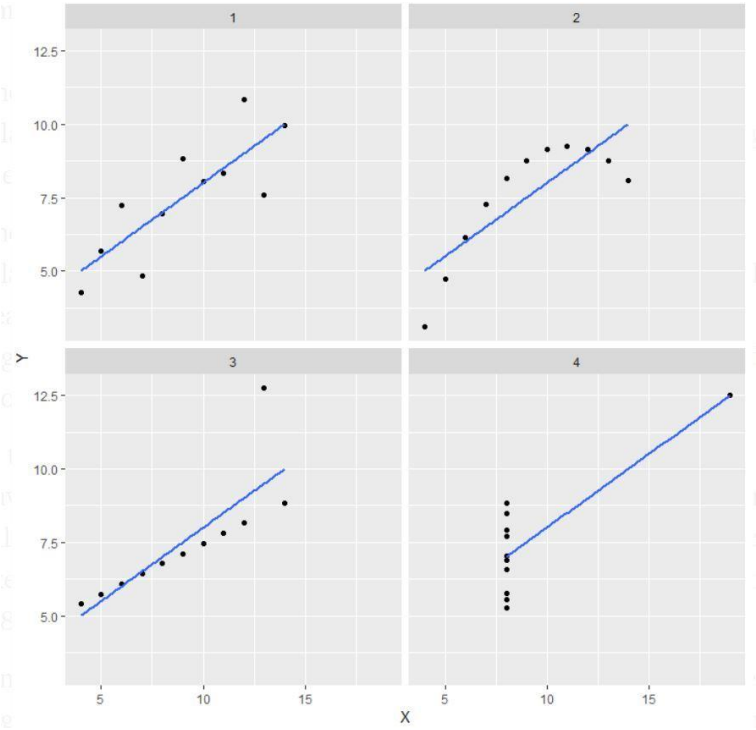
The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

- RSS: Residual Sum of Squares
- TSS: Total Sum of Squares

Q2 : Explain the Anscombe's quartet in detail.

Ans : 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.

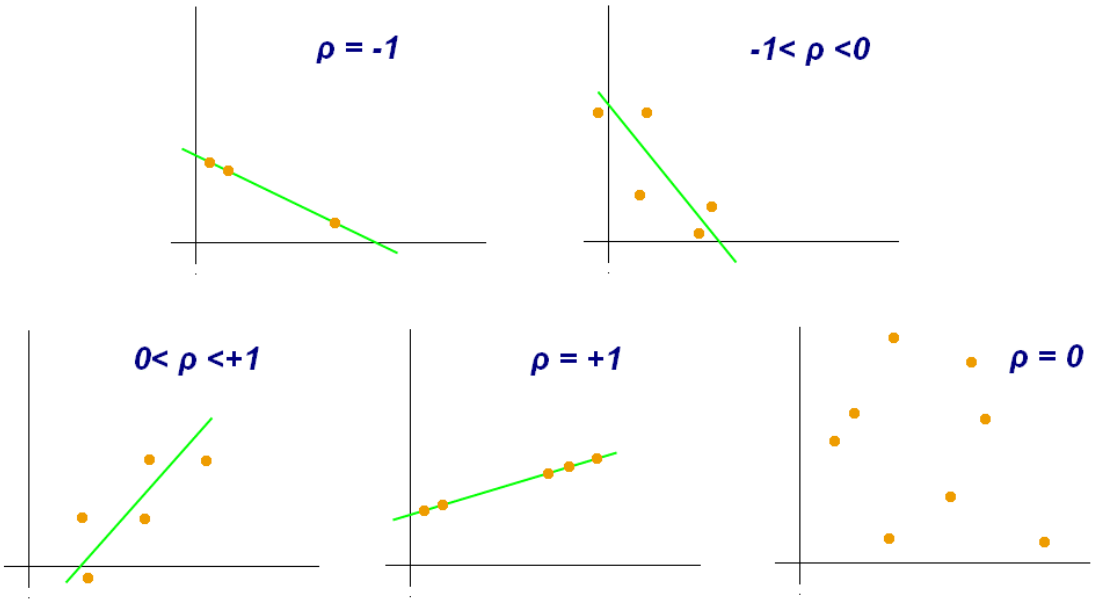
Simple understanding:
Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient

Q: . What is Pearson’s R?
Ans : In statistics, the Pearson correlation coefficient also known as Pearson's r — is a measure of linear correlation between two sets of data.
It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).



Q: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans :

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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

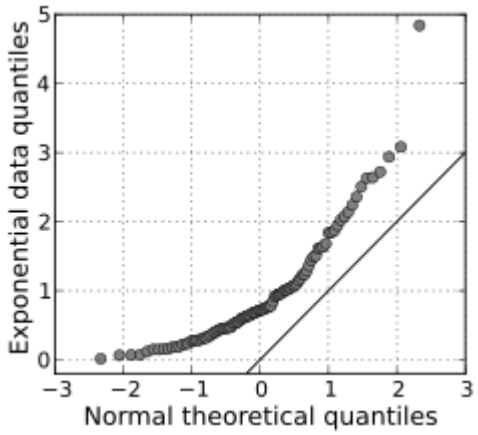
Q: **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

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 (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. 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. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
A Q Q plot showing the 45 degree reference line:



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.