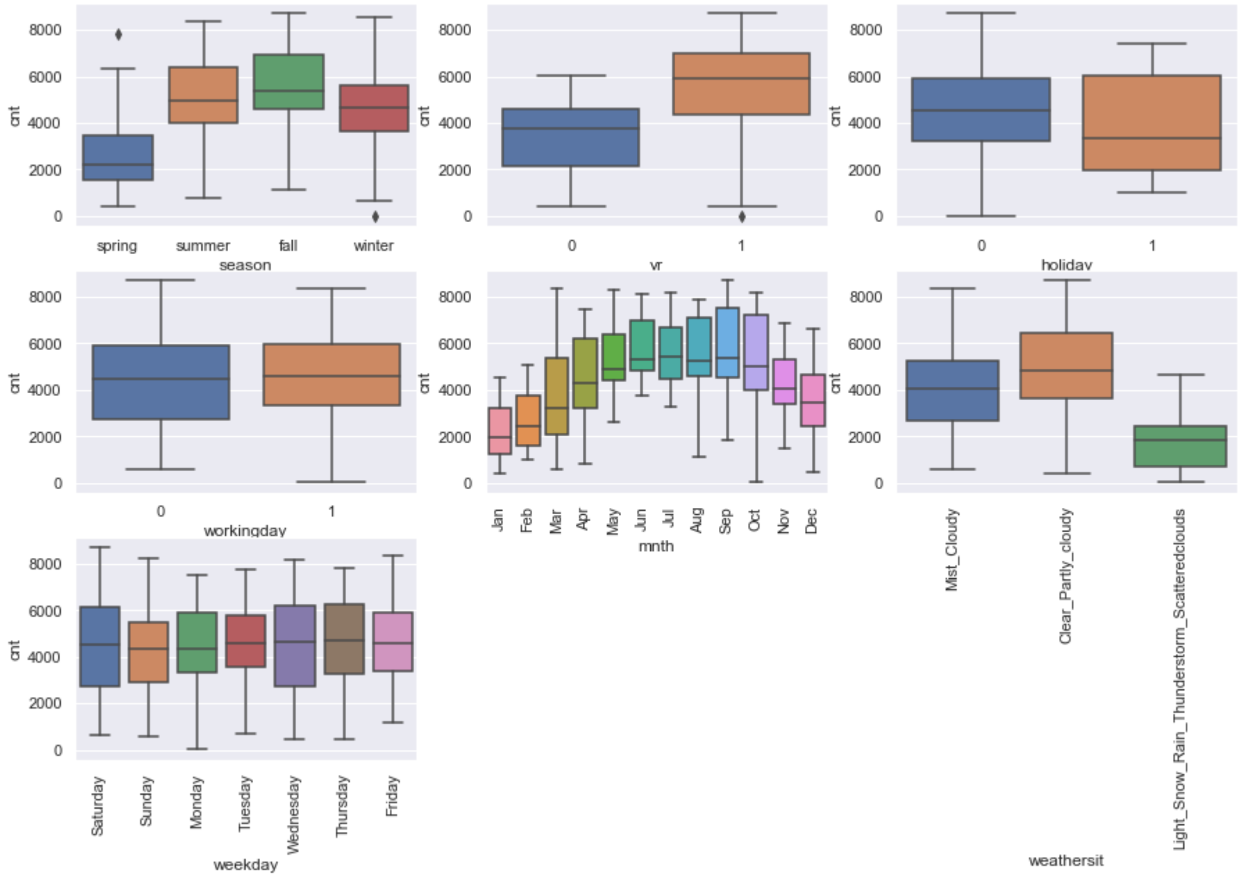
Linear Regression Subjective Questions Assignment

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?

Answer: Below is the analysis of categorical variables with dependent variable cnt i.e total count of rental bikes.

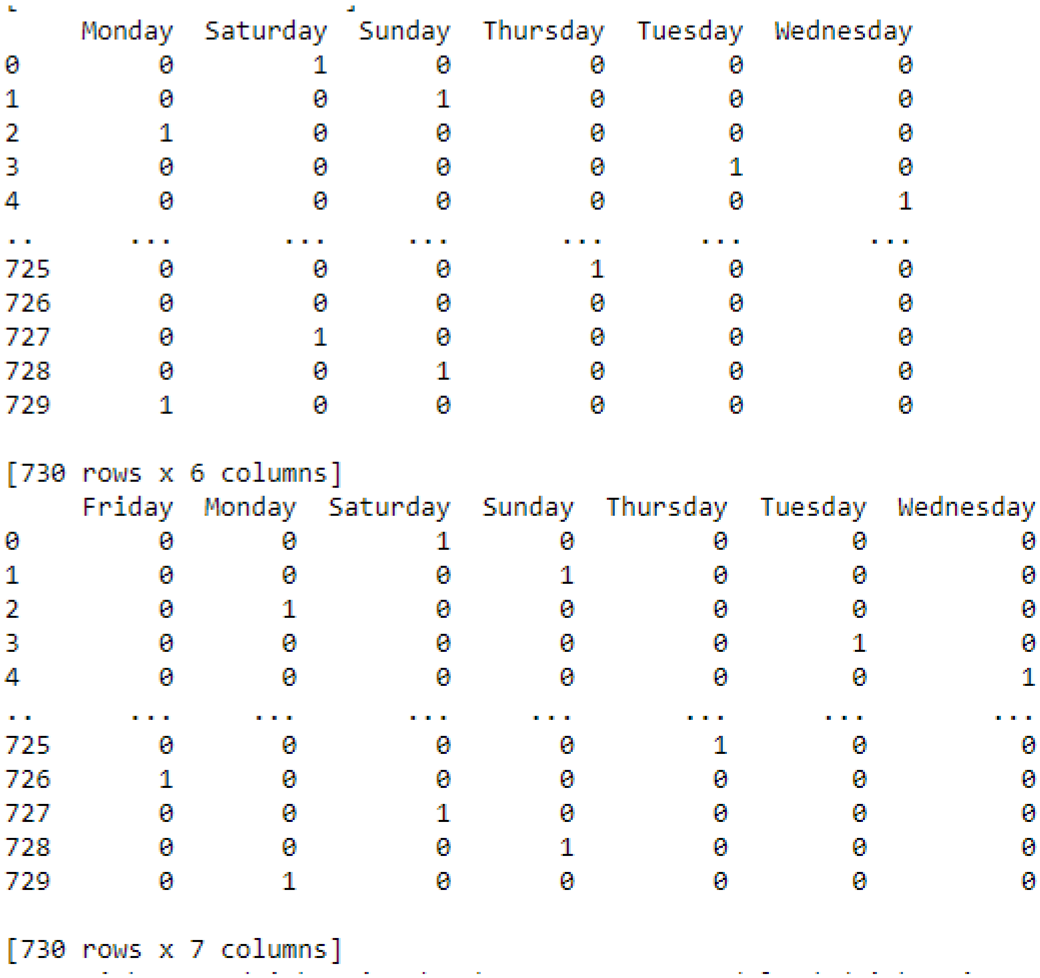


On the basis of above plot we can infer that:

1. People rent more bikes in the fall season and less in the spring season.
2. People rent more bikes in year 2019 than in 2018

3. People rent more bikes in the Clear or Partly cloudy weather and less in light snow or thunderstorm situations.
4. People rent more bikes on Non holiday days

2. Why is it important to use drop_first=True during dummy variable creation?
Answer : It is important to use drop_first=True during dummy variable creation as it creates one column less which was present in the dataset.So after merging with main dataset it will not create duplicate columns
Below is dummy variable created of column weekdays with drop_first=True and without drop_first=True



```
     Monday  Saturday  Sunday  Thursday  Tuesday  Wednesday
0         0         1       0         0        0          0
1         0         0       1         0        0          0
2         1         0       0         0        0          0
3         0         0       0         0        1          0
4         0         0       0         0        0          1
..      ...       ...     ...       ...      ...        ...
725       0         0       0         1        0          0
726       0         0       0         0        0          0
727       0         1       0         0        0          0
728       0         0       1         0        0          0
729       1         0       0         0        0          0

[730 rows x 6 columns]
     Friday  Monday  Saturday  Sunday  Thursday  Tuesday  Wednesday
0         0       0         1       0         0        0          0
1         0       0         0       1         0        0          0
2         0       1         0       0         0        0          0
3         0       0         0       0         0        1          0
4         0       0         0       0         0        0          1
..      ...     ...       ...     ...       ...      ...        ...
725       0       0         0       0         1        0          0
726       1       0         0       0         0        0          0
727       0       0         1       0         0        0          0
728       0       0         0       1         0        0          0
729       0       1         0       0         0        0          0

[730 rows x 7 columns]
```

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: temp and atemp are highly correlated with target variable cnt.

.4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: To validate the assumption of Linear Regression Model,we have to check following parameters :

1. Linear Relationship- There should be a linear relationship between model and predictor variables.
2. Homoscedasticity-residual value analysis
3. Absence of Multicollinearity- there should be insignificant multicollinearity among predictor variables
4. Independence of residuals- There should be no autocorrelation
5. Normality of Errors - Error terms should be normally distributed

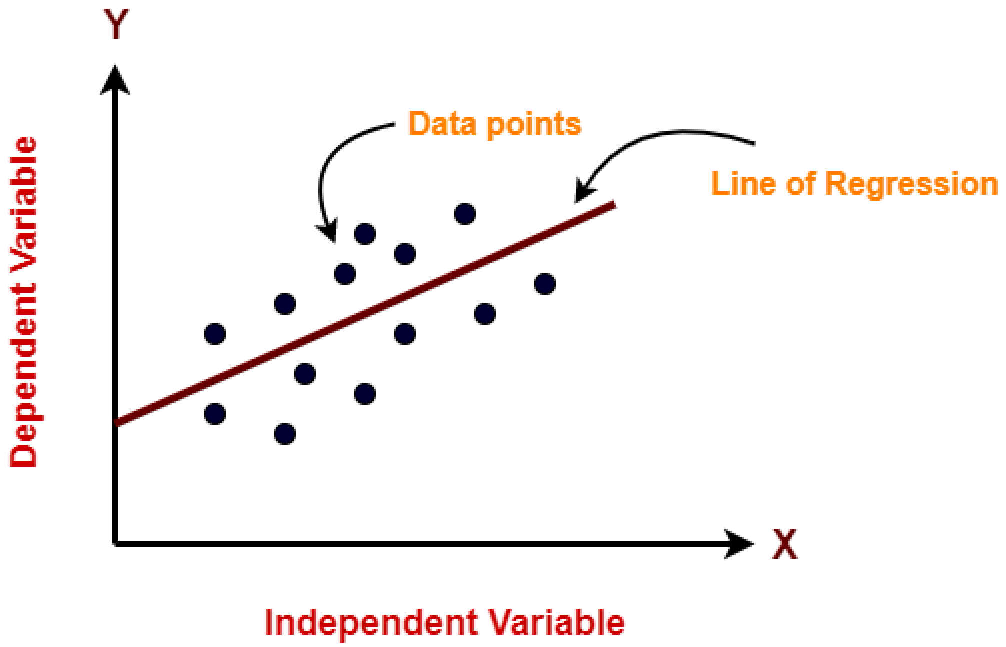
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: Based on the final model following are top 3 features contributing toward explaining the demand of the shared bikes ::

1. Temp - Temperature
2. Yr- Year
3. Working day

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Answer: Linear regression algorithm is a supervised machine learning algorithm.
In it , the output variable which we have to predict is a continuous variable like marks secured by student or runs scored by batsman in cricket .
In Linear regression past data with label is used for building the model
It explains the relationship between dependent and independent variables using straight line.

Below image explains Linear regression algorithm nicely



In the above image the relation between Independent variable , which is on x axis is shown with dependent variable which is on y axis.
Line of Regression is best fit line of linear regression model .Main objective of this algorithm is to find this line which fits all its data points linearly

Linear regression algorithm is simple to implement.

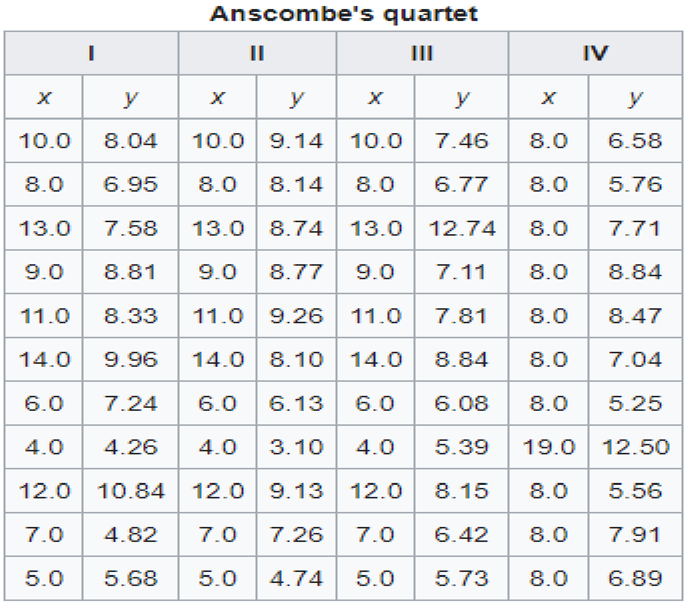
There are two types of Linear regression:

1. Simple linear regression -
   Simple linear regression explains the relation between dependent variable and independent variable using straight line.
   The standard equation of the regression line is given by the following expression:
   $Y = \beta_0 + \beta_1 X$

   Here Y is dependent variable, X is independent variable Bo is intercept and B1 is slope

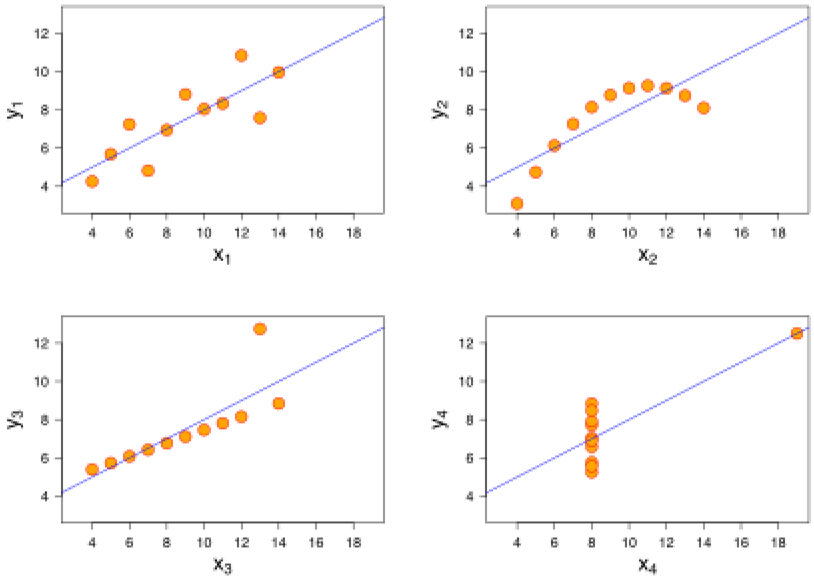
2. Multiple linear regression -
   Multiple linear regression explains the relation between one dependent variable and multiple independent variables
   The main objective of this algorithm is to find . the linear equation which nicely get the value of Y variable for different values in X
   It help us to to understand the change in dependent variable when ther is change in independent variable

2. Explain the Anscombe’s quartet in detail.
Answer: Anscombe’s quartet is a group of four dataset which seems statistically identical but it appeared very differently when plotted.
It basically shows the importance of data visualization before implementing model



Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



As we can see, the value of X is mostly identical for all four quadrants. But when it is plotted the real difference in their distribution is visible
In first graph x1 and y1 are linearly related
In second graph the relationship between x2 and y2 is non linear
In third graph there is linear relationship with outliers
In the fourth graph there is high leverage point (outlier) which produce high correlation coefficient whereas other points shows no relationship between variables.

3. What is Pearson’s R?
Answer: Pearson’s R also known as Pearson’s correlation coefficient is a statistical measure that calculates linear correlation between two variables. Its value lies between -1.0 to 1.0
1.0 correlation value signifies strong positive correlation
Mathematically Pearson’s R is the covariance of the two variables divided by the product of the standard deviation of variables

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer: Normalization of data with help of independent variables within a particular range is called scaling .
Scaling is done to make all variables of same magnitude so that it became easier measure their relatability and create good model
Difference between normalized scaling and standardized scaling

Normalized Scaling	Standardized Scaling
It scale data in range of 0 and 1	It replaces data with a z-score . It scales data between standardized normal distribution which has mean value zero and standard deviation one.
X_new = (X - X_min)/(X_max - X_min)	X_new = (X - mean)/Std
It scales value between -1 to 1	It is not bounded to certain range
It is affected by outliers	It is ver less affected by outliers
It is often used when we dont know about the distribution	It is used in case of normal /Gaussian distribution

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: VIF is infinite in case of perfect correlation between independent variables .In this case we get R2 value 1.It indicates that the corresponding variables may be expressed exactly by linear combination of other variables whose VIF is also infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: Quantile-Quantile plots , plot the quantile of a sample distribution against quantile of a theoretical distribution
It helps to determine probability distribution type followed by a dataset. Like normal, uniform, exponential.
Interpretation in Q-Q plot :
-all point of quantiles lies on or close to straight line at an angle of 45 degree from x-axis
-y-quantiles are lower than the x-quantiles.
-x-quantiles are lower than the y-quantiles.
-all point of quantiles lies away from the straight line at an angle of 45 degree from x-axis