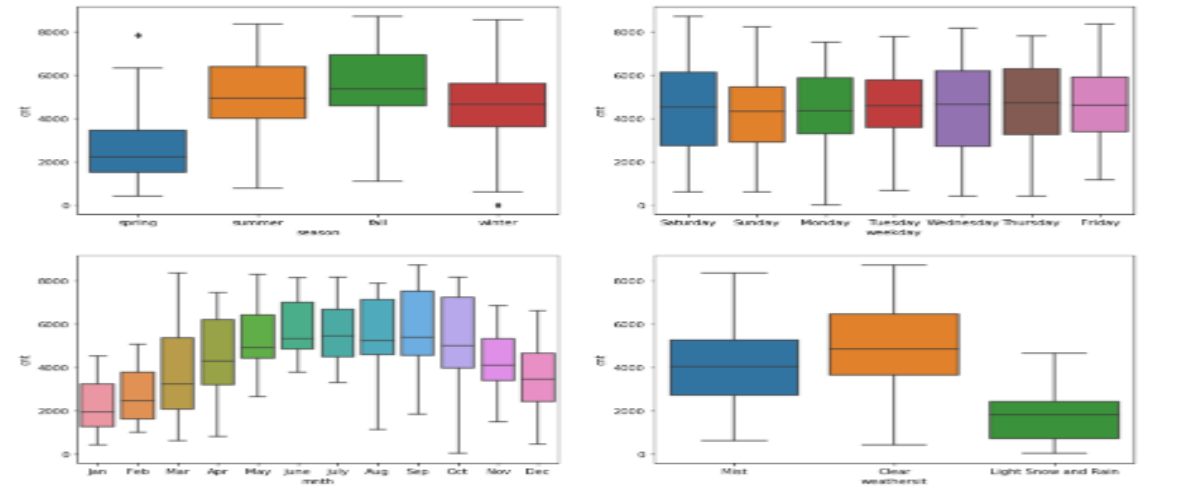
Assignment-based Subjective Questions

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

Ans:

Visualizing categorical variables using box plot

```
plt.figure(figsize=(15, 12))
plt.subplot(2,2,1)
sns.boxplot(x = 'season', y = 'cnt', data = df)
plt.subplot(2,2,2)
sns.boxplot(x = 'weekday', y = 'cnt', data = df)
plt.subplot(2,2,3)
sns.boxplot(x = 'mnth', y = 'cnt', data = df)
plt.subplot(2,2,4)
sns.boxplot(x = 'weathersit', y = 'cnt', data = df)
plt.show()
```



From the above figure we see that people tend to rent bikes when the weather is clear and the season “fall” is the top season to rent bike.

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

Ans: Whenever we have n categories we create dummy variables for n-1 categories hence we use the keyword drop_first=True to drop the first column. We can identify the value of first column with the help of other dummy variables. Also it helps in efficiency i.e. if we can explain the variable with help of other variables why to use them in the analysis hence we drop the first variable.
For eg if we have 4 blood groups named O+,A+,B+ and AB+ the dummy variables will be as follows.

Serial No	AB+	B+	AB+
0	1	0	0
1	0	1	0
2	0	0	0
3	0	0	0

Here if we see the 4th row we have all 0’s for all the 3 blood groups which indicate that the blood group of 4th row is A+.

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

Ans: Among the variables temp, hum and windspeed, temp has highest correlation with target variable(cnt)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - Linear relationship between X and Y variables. We have observed this using the pariplot in our analysis.
    - No multicollinearity for this we have used VIF method in the analysis to overcome multicollinearity and removed those features whose VIF>=5.
    - Error terms are normally distributed or not – We have plot distplot by plotting residuals which is difference between y_train – y_train_pred. We can also use Q-Q plot for the same
    - Error terms are independent and have constant variance(homoscedasticity) – Validated this assumption based on the scatter plot taking y_train_pred on X axis and residuals on y axis. If there is no pattern as such it shows that they have constant variance and independent to each other.

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

Ans: The final model obtained from the analysis is the eight model which is as below
cnt = 0.2331 * yr + 0.5499 * temp + 0.0563 * workingday + 0.0874 * summer + 0.1318 * winter + 0.0972 * Sep + 0.0677 * Saturday - 0.1552 * windspeed - 0.2880 * Light Snow and Rain - 0.0813 * Mist

The top 3 features contributing for the demand in shared bikes based on the absolute values are “temp”, “Light Snow and Rain” and “year” variable .

If we don’t consider the absolute value then we have temp, winter and Sep as 3 features contributing the demand of shared bikes.

The warmer the temperature there are more likely for people to rent a bike and its pleasant for people to rent a bike in winter where relying on public transit might delay and not sure when they will be able to reach the destination. One unit increase in Sep will lead to 0.0972 increase in count. There is also negative coefficients associated with few features link windspeed, light Snow and Rain and mist which is obvious for the decrease in demand when windspeed is high or the weather is Mist or light and snow

General Subjective Questions

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

    Linear regression is a machine learning algorithm based on supervised learning. It performs the task to predict the dependent variable(y) based on given independent variables (X).

    Application of regression is

    - to find the effect of independent variables on target variables
    - To find change in target variable with respect to one or more independent variables.

    Simple linear regression equation

    Y=m+cx where m is intercept of y line and c is the slope of the line and x is the independent

variable

Multiple linear regression

Y=m+c1x1+c2x2+….+cnxn

Where x1 …xn is the n independent variables and c1…cn are n slopes

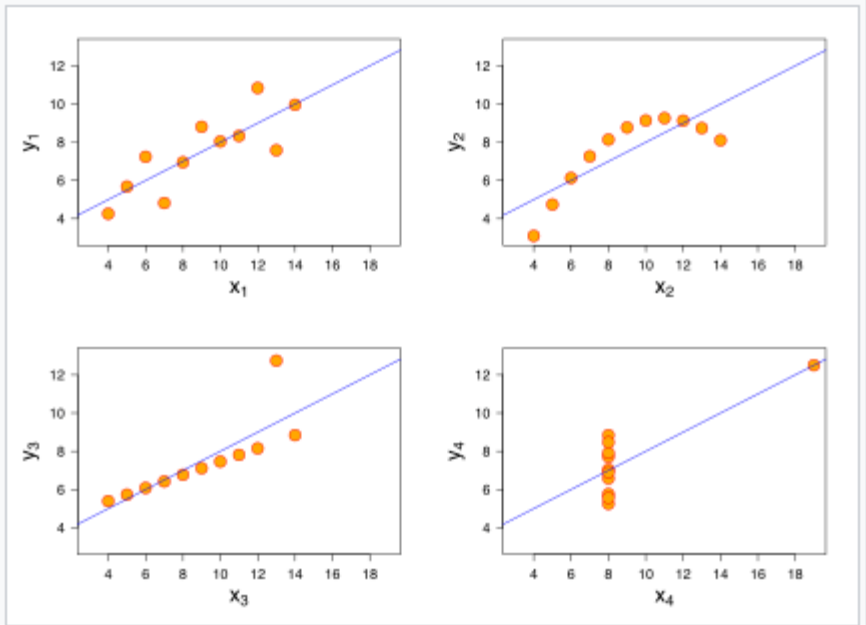
Steps performed in Linear Regression(Multiple)

- We analyze various categorical and continuous variables with respect to the target variable to see the association between them. To achieve this we use visualization plots.
- Create dummy variables for all categorical features
- Divide the data into train and test
- Perform scaling to standardize the continuous variables in a fixed range
- Perform Linear regression and check for significance of features using p value and VIF
- Eliminate the features which are not significant and fit the model
- Find the best fit model where p value <0.05 and VIF<5 with a good R2 and adjusted R2 value.
- Using the best fit model we check the normality of error using difference between y-train and y_train_pred
- Perform scaling on test data and drop all the variables as per X_train and then predict the y_test_pred based on best fit linear model
- Check the R2 between y_test and y_test_pred

  The best fit model aims to predict y value such that the error difference between predicted value and true value is minimum i.e. to have minimum Cost function(RMSE)

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

   Anscombe's quartet as the name suggests it comprises of 4 data set that have identical descriptive statistics yet have very different distributions and appear differently when graphed. Each data set consists of 11 (x,y) points. It was constructed to demonstrate both the importance of graphing the data before analyzing them and the effect of outliers and other influential observations on different statistical properties



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Dataset I (top left) consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II (top right) fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III (bottom left) looks like a tight linear relationship between *x* and *y*, except for one large outlier. Dataset IV (Bottom right) looks like *x* remains constant, except for one outlier as well.

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

It's the Pearson correlation coefficient also denoted as R. It measures the linear correlation between two variables X and Y. It ranges between -1 to +1 where +1 indicates positive linear correlation, 0 indicates no linear correlation and -1 indicates negative linear correlation.

The pearson correlation formula when applied for the population is

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where Cov(X,Y) is the covariance between X and Y

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

Where Cov(X,Y) is

$$\operatorname{cov}(X,Y) = \mathcal{E}[(X-\mu_X)(Y-\mu_Y)],$$

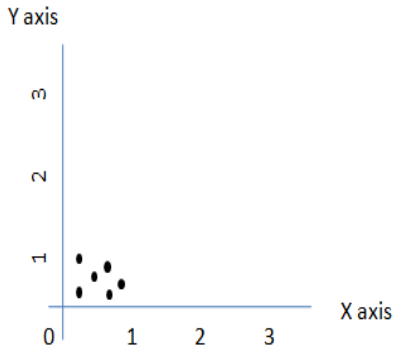
The pearson correlation formula when applied for the sample is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

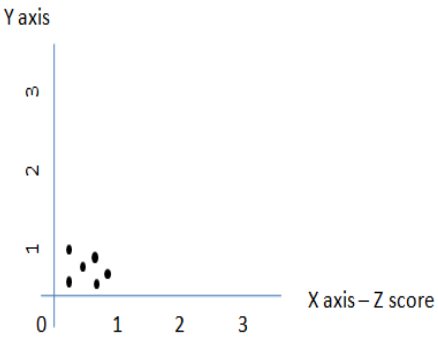
Where n is the sample size
X bar is the sample mean of xi
Xi and yi are individual sample points for I ranging from 1 to n

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

Scaling is a technique to standardize the independent variables present in the data in a fixed range. It helps to normalize the data with a particular range.
It is performed during data pre processing to handle highly varying magnitudes or values or units. It we don't normalize then when predicting the model it will yields in wrong predictions because of varying units.
Normalization usually means to scale a variable to have values between 0 and 1 whereas standardized scaling transforms data to have mean of 0 and standard deviation as 1.



Normalization scaling



Standardization where Z axis represents Z score

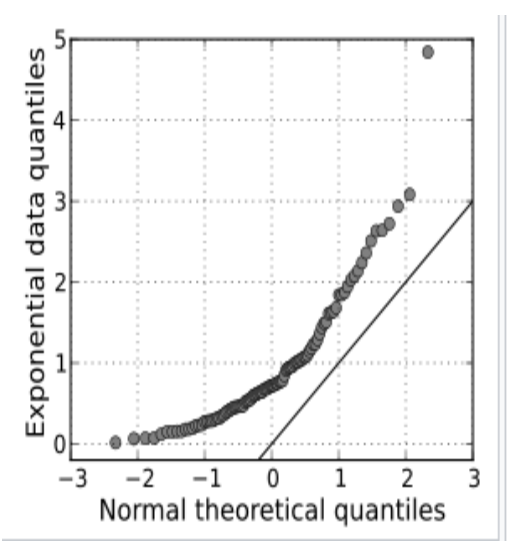
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

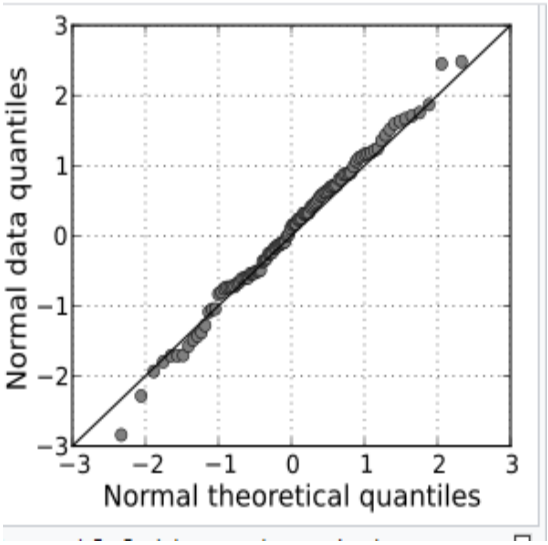
VIF refers to variance inflation factor. It is used for identifying multicollinearity i.e. to check if there is any dependence between the independent variables.
Higher the VIF the greater the correlation of that particular variable with other independent variables.
When VIF=infinite it represents perfect correlation i.e. 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. (3 marks)

Q-Q(quantile-quantile) plot is a probability plot which is a graphical method for comparing two probability distribution by plotting their quantiles against each other.
If the two distributions are similar, the points in Q-Q plot will approximately lie on the line y=x.If the distribution are linearly related the points in the Q-Q plot will lie on the line but not necessarily on line y=x.
Q-Q plot helps to assess the assumption of normality. Q-Q plot helps us to assess if a set of data has come from some theoretical distribution such as Normal, exponential or Uniform. It also helps to determine if two data sets come from populations with common distribution.

In linear regression when we divide the data set into training and test, use of Q-Q plot can confirm if the data sets are from populations with same distribution.





Graph 1 indicates non-linear pattern. It indicates that they don't follow standard normal with mean 0 and standard deviation as 1

Graph2 indicates its linearly related and the data are normally distributed.