Assignment-based Subjective Questions

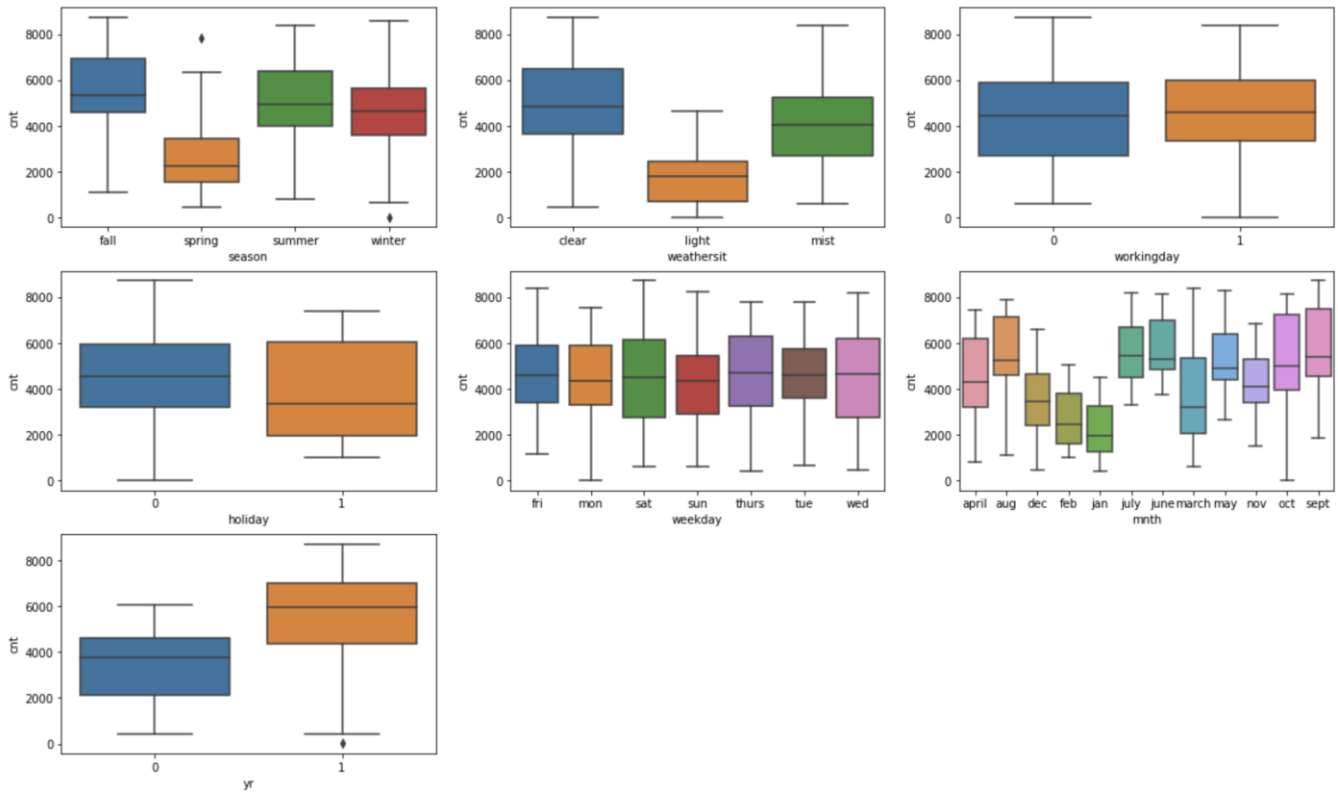
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans:

Categorical variables for the given data set are:

- Season
- Weathersit
- Workingday
- Holiday
- Weekday
- Mnth
- Yr.

Below is the visualization of impact of categorical variable on the dependent variable “cnt”



- Spring season is having the inverse impact on the “cnt” dependent variable. As per the correlation matrix its coefficient is –0.56 (matrix can be found in submitted notebook). So it means lesser number of bikes are booked during spring season.
- 50% of the bike booked during the fall season is 5000.
- Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, weather has inverse impact on cnt variable. Also, its coefficient is –0.26. It means lesser bikes are booked during light weather.
- 50% of bikes booked on each day of weekday are nearly equal.

- Jan months has least number of bikes booked.
- 2019 year has more no. Of bikes booked. 50% bikes booked in year 2019 is 6000. And 50% bikes booked in 2018 are 4000.

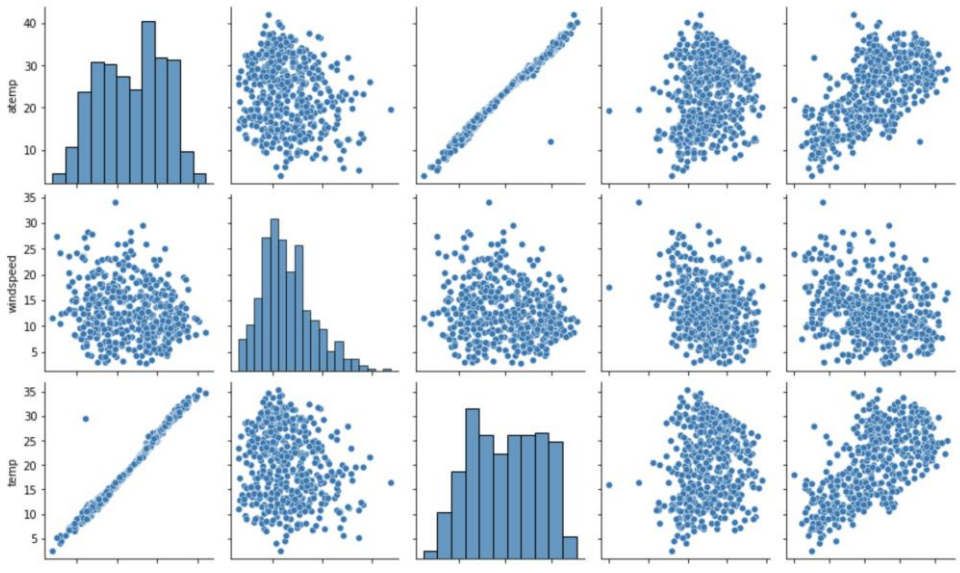
2. **Why is it important to use drop_first=True during dummy variable creation?**
    a. For example, we have 3 categorical variables : Single, In Relationship and Married. If we create. If we create a dummy table of these variables it will look like:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

We can drop single variable and re construct the table as below, still we will be able to convey the value of all the three categorical variable:

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

Hence, drop_first=True, drops the first variable while creating the dummy variable.

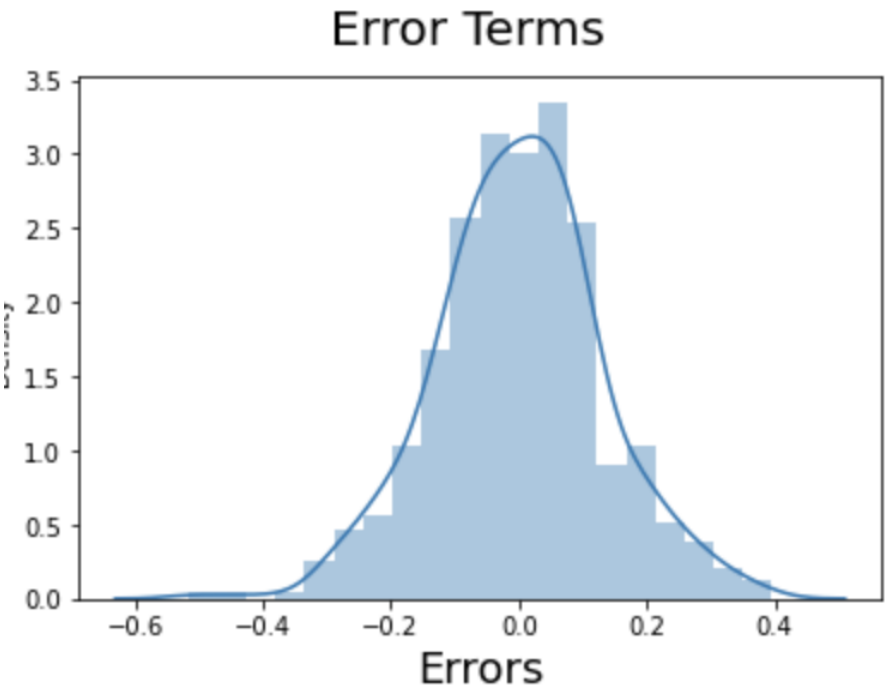
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    a. Atemp and temp variable has the highest correlation with the cnt (target) variable.
    b. 

    Last column is of “cnt” variable.

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

Ans: Before trying out the model for prediction we have to be sure if model is reliable or not. For this we perform residual analysis.

In residual analysis, we create the y-pred using the created model using the X-train. And then calculate the residual using : y-train – y-pred. And then plot the distribution plot using. And error curve should be normally distributed as shown below.



Example :

```
y_train_pred = lr6.predict(X_train_lm6)

res = y_train-y_train_pred

fig = plt.figure()

sns.distplot((res), bins = 20)

fig.suptitle('Error Terms', fontsize = 20)

plt.xlabel('Errors', fontsize = 18)
```

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

Ans: Below is the correlation coefficient of 9 features items on dependent variables “cnt”

```
lr6.params
```

```
const                  0.587019
yr                     0.247088
holiday               -0.108669
windspeed             -0.170448
season_spring         -0.295092
season_summer         -0.041506
season_winter         -0.071465
mnth_sept              0.074532
weekday_sun           -0.049155
weathersit_light      -0.302387
weathersit_mist       -0.092621
dtype: float64
```

Out of them list in sequence of impact.:

- light weather has positive impact on cnt variable
- Season spring has negative impact on cnt variable
- Yr has positive impact on cnt variable.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    a. Linear regression is a machine learning algorithm based on supervised learning. It provides a linear relationship between independent and dependent variable to predict the future events.
    b. Supervised learning system is a kind of system that takes some input and known output and tries to create a model that can generate a response to unknown provided input.
    c. Linear regression, is a stastical method. Where the independent variable is consider to be present on x axis of 2-d graph and dependent on y-axis of the graph. Then it tries to find the linear relationship b/w bother the variables that best fits.

As we know the linear equation of a line is y = mx + c.

So, in linear regression technique we try to find the value of m (slope) and c (y intercept) that will create a best fit equation. By this way we will be able to predict the value of y for the given value of x.

A best fit line will also be a line, that has min difference between the y-predicted value and actual value. So, final motive is to reduce the difference. There are different ways of finding the minimum difference

1. **Root Mean Square Method (RMSM).**
2. **Ordinary Least Square methos (OLSM):**

In OLSM, we find the eL = yL – ypred. Here L be the different values: 1, 2,3,4,…

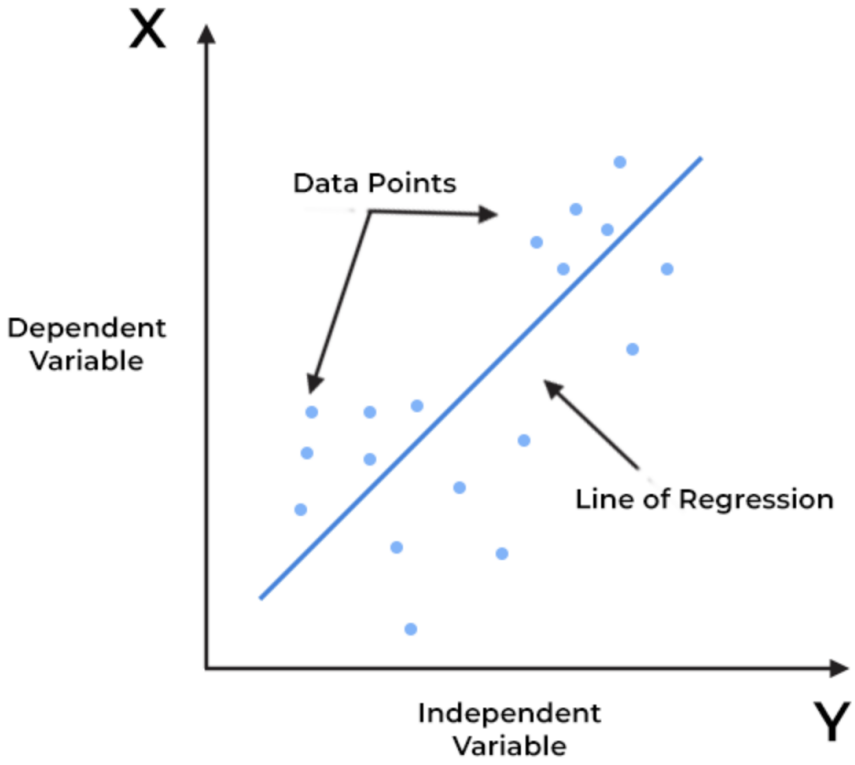
And OLSM calculates e1^2 + e2^2 + e3^2 + … en^2

Less will be the value of OLSM more linearly related will be the variables.

First we start with drawing the scatter plot between independent and dependent variable. To check do they have any kind of linear relationship. If yes, then we proceed to find the best with line b/w them using OLSM or RMSM or any other method.

Apart from scatter plot we can also find the coefficients correlation between the variable Coeff correlation's value is between –1 and 1. If its value is 0, then it means there is no linear regression between the variables.

1.



Benefits of Linear Regression:

1. It is easy to implement.
2. It doesn't require lot of computational power.
3. It is straight forward.

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

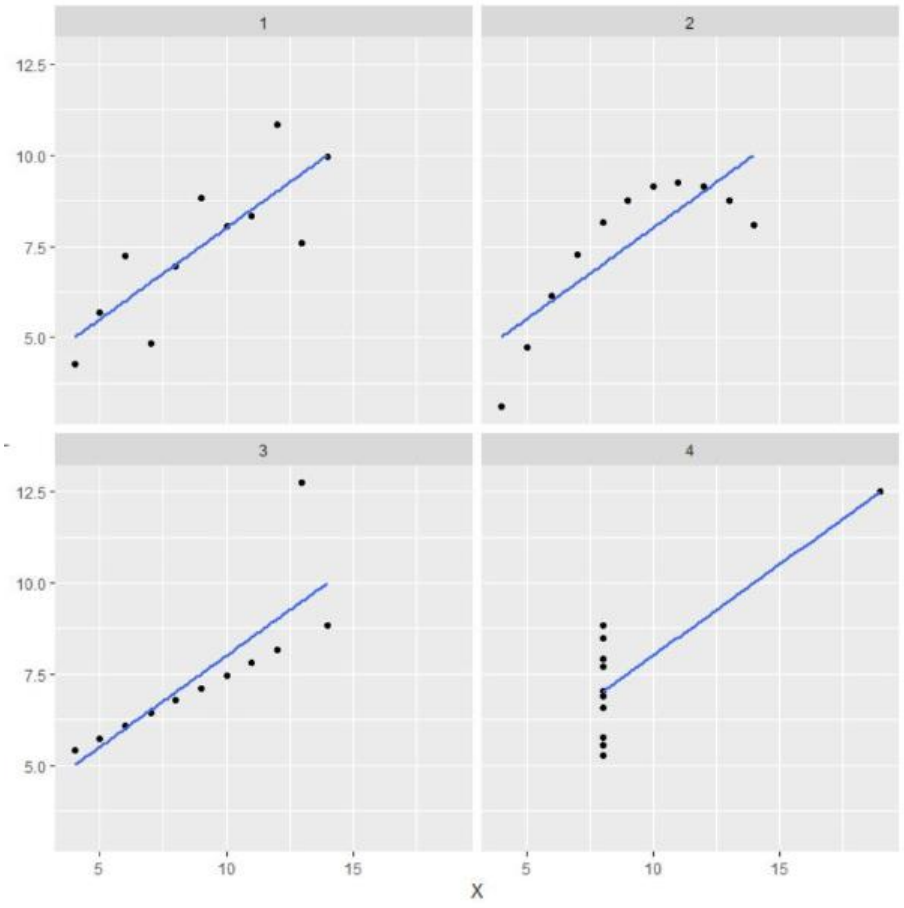
   Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

   Those 4 datasets are:

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

IF we calculate the mean, standard deviation and correlation for each of the set between x and y, the values will be same.

And when we draw them on scatter plot, each plot of diff data set will appear to be different.



The main objective of this is to explain the importance of plotting the data graphically before doing any particular type of analysis.

3. **What is Pearson's R?**

Ans:

Pearson's R is the numerical summary of the strength of linear association between the variables.

It's correlation value lies between –1 and +1. If two variable increases with increases in there values then they are said to be having positive correlation. IF with the increase in value of one variable other decreases then they are said to be having negative correlation.

And if value is 0, then there is no linear relationship b/w variables.

Or in other words if we increase one variable then other variable increases then there is a positive effect or if changing one variable other variable decreases then there is a negative effect.

It's values is calculated as:

For the given values of variables X and Y. Calculate the following in steps:

- For each row calculate XY = X * Y . And store it in column XY
- Each row calculate X^2 and Y^2 and store it in column X^2 and Y^2
- Take sum of all the columns : Σ X , ΣY, ΣXY, ΣX^2, ΣY^2
- Now calculate R using above results as:
    i. R= ((N * ΣXY) - (ΣX * ΣY)) / Sqrt(((N * ΣX^2) - ((Σ X)^2))((N * ΣY^2) - ((ΣY)^2)))

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

Ans: Scaling is a data pre-processing method to generalize the different independent variables present in data set. If the difference between the variables are very far apart, then scaling is method to bring them together.

Scaling is performed, let say we have a feature those are different in magnitude, units and range. If modelling is performed it will only take magnitude in account, that will result in incorrect modelling.

Scaling only impacts the correlation coefficient, not the F-stastics, p-value, r sqaure values.

Normalized scaling: Also known as min-max scaling. It brings all the data points between 0 and 1.

from sklearn.preprocessing import MinMaxScaler : helps to implement scalling in python.

Scaling fn , x = (x – mean(x))/(max(x) - min(x))

Standardization Scalling: It replaces the values by their z-scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

sklearn.preprocessing.scale helps to implement standardization in python.

Disadvantage: It doesn't take outliers in the account. Outliers get lost.

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

Ans: VIF formula is

VIF = 1/(1 – R^2). For perfect correlation R squared value is 1. It means VIF = 1/(1-1) = infinity.

Hence, vif havin a value infinity means there is a perfect correlation between independent and dependent variables. In other words, there will be a perfect linear line.

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

Ans: 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 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.