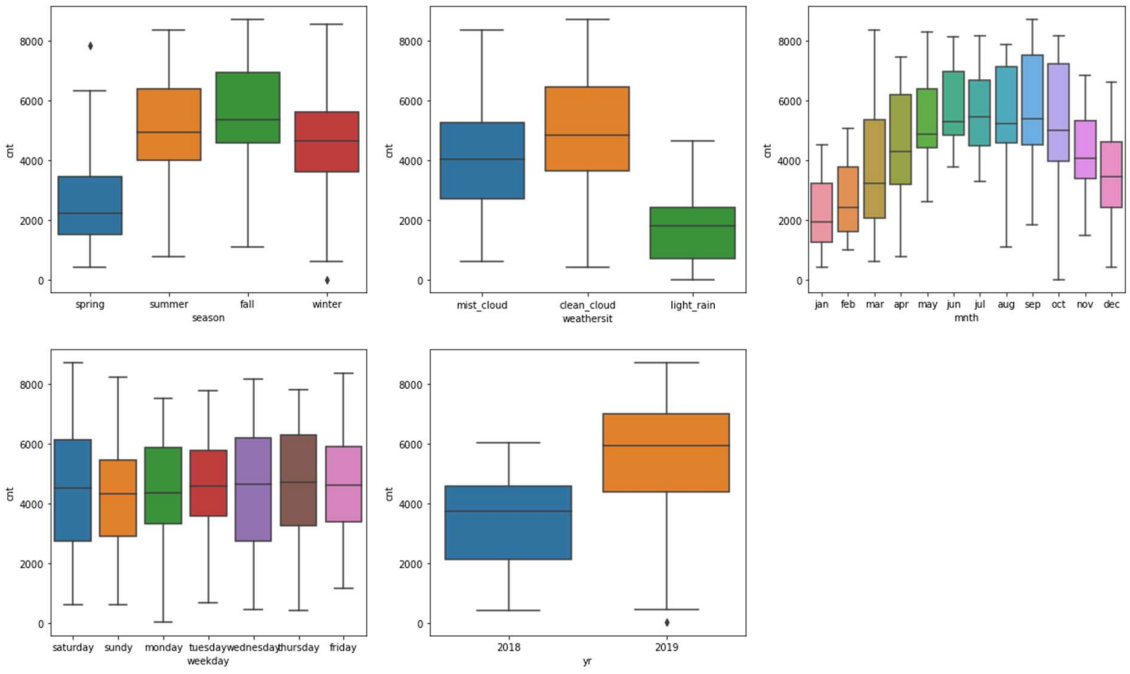
Assignment-based Subjective Questions

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

->From observing the data set and dictionary we came to conclusion and categorical variable are as fallows:

1. Season : Target variable changes as season changes from spring to snow fall increasing manner and again decreases in winter
2. Weathersit : target variable dependence with comparison to weathersit is as fallows: Light_rain < mist_cloud < clean_cloud
3. Mnth : target variable changes(observe median) as we go from jan to mid of the year it increases then again after the sep we saw drastically decrease.
4. Weekday : there is no change in median its nearly constant but we can observe the difference in range and density which is more for Saturday and Wednesday.
5. Yr : due to corona company poor progress in 2018 and 2019 it got better.



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

->After creation of dummy variable for n number of values we create n column, If we know n-1 column as value 0 then nth variable is the 1 and hence we can able to consider it by reducing the column.

Ex: created head and tail both dummy variable, if we remove fst column that is head we can recognize if not tail its definitely head. By this we able to reduce the column also without removing value to the model.

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

-> temp and a temp both are having highest correlation with target variable.

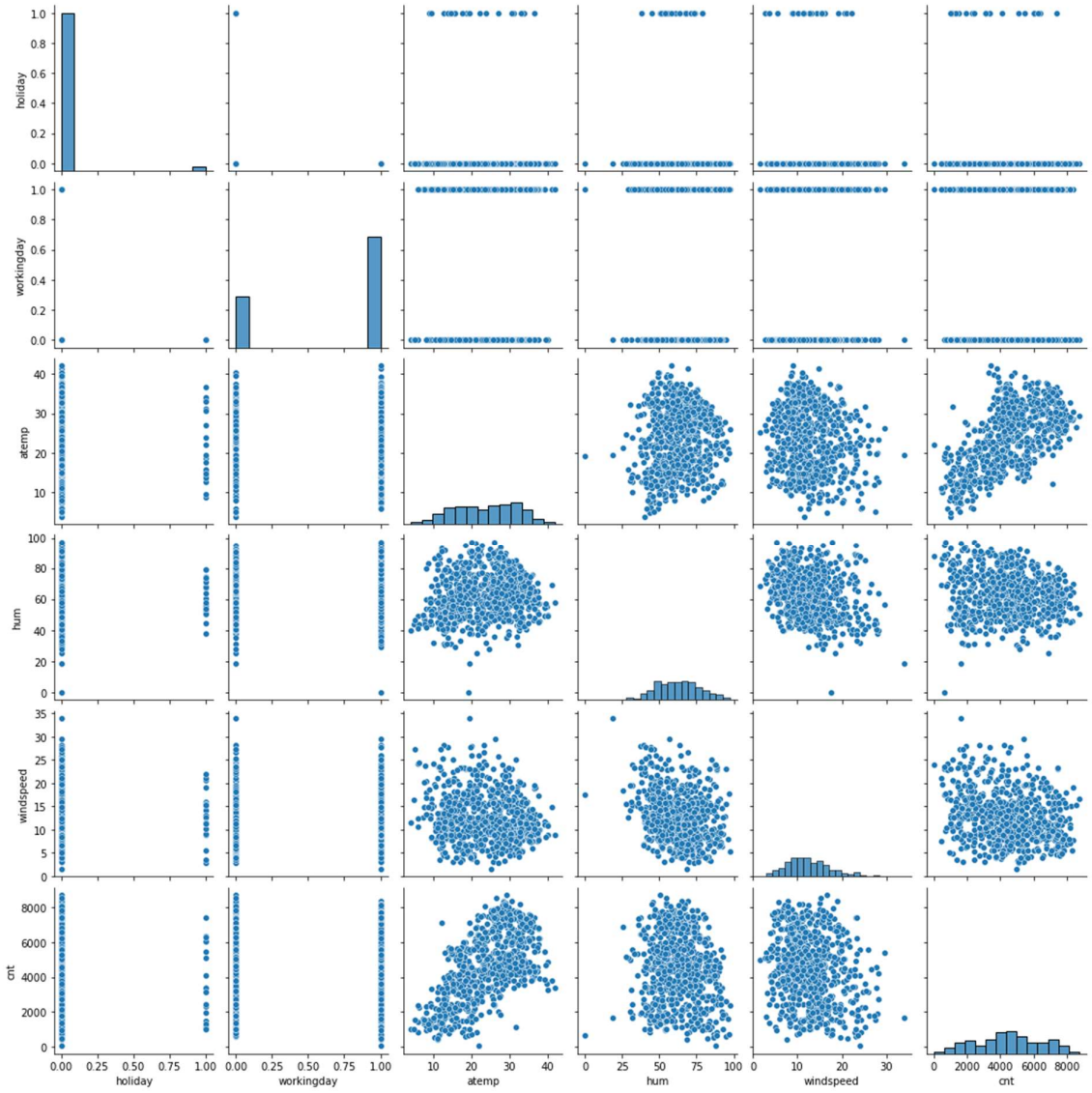


Figure 1 Scattered plot for numerical variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

-> **Linear Regression Assumption 1 — Independence of observations**

The first assumption of linear regression is the independence of observations. Independence means that there is no relation between the different examples. Atemp and temp had highly correlated hence dropped from train dependent variable, categorical variables deleted after creating dummy variables

Linear Regression Assumption 2 — No Hidden or Missing Variables

*By checking info we get to know no null value exist

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 730 entries, 1 to 730
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   dteday      730 non-null    int64
 1   season      730 non-null    int64
 2   yr          730 non-null    int64
 3   mnth        730 non-null    int64
 4   holiday     730 non-null    int64
 5   weekday     730 non-null    int64
 6   workingday  730 non-null    int64
 7   weathersit  730 non-null    int64
 8   temp        730 non-null    float64
 9   atemp       730 non-null    float64
 10  hum         730 non-null    float64
 11  windspeed   730 non-null    float64
 12  casual      730 non-null    int64
 13  registered  730 non-null    int64
 14  cnt         730 non-null    int64
dtypes: float64(4), int64(11)
memory usage: 91.2 KB
```

Linear Regression Assumption 3 — Linear relationship

*checked for linear relationships easily by making a scatter plot(Figure 1 refer 3rd answer) for each independent variable with the dependent variable.

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

-> July month, light rain and holidays are top 3 features contributing significantly towards demand of shared bikes

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 a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression : y = C + Bx

C -> intercept

B -> co-efficient of x

y -> dependent variable

x -> independent variable

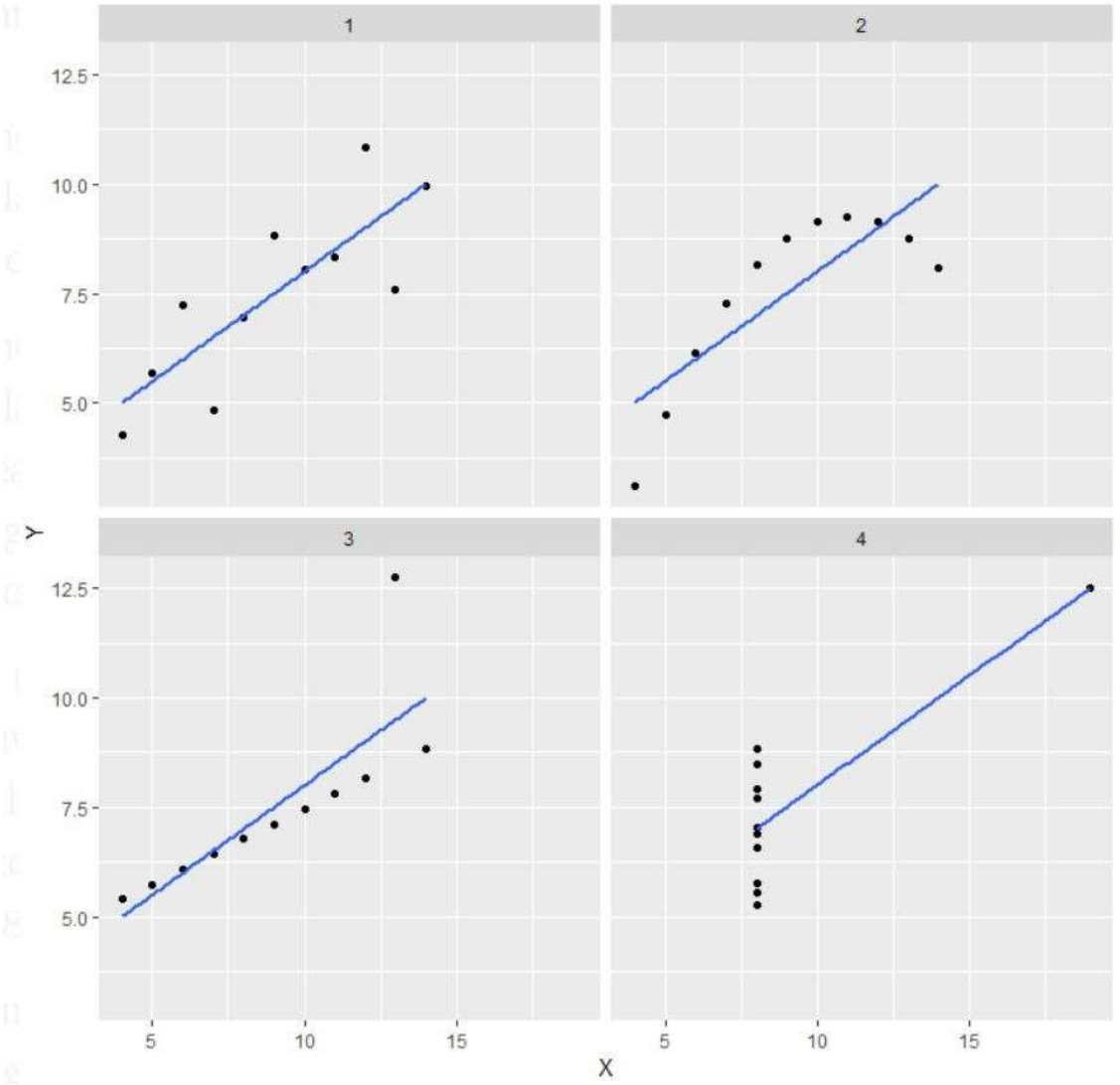
Best fitting model has to find best C and B then only model will be best fitted.

It can be find by RSS root mean squared Error = 1/n(y – y pred)^2 which is to be minimal.

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

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.

Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



Explanation of 4 datasets with similar statistical property with graph output:

- ➔ In the first one: the scatter plot you will see that there seems to be a linear relationship between x and y.
- ➔ In the second one: 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.

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

->Pearson's Correlation Coefficient ®

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

-> What?

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.

Why?

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. MinMaxScaling: x = 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).

Standardisation: x = x-mean(x)/sd(x)

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.

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

- ➔ 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).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

- ➔ If two populations are of the same distribution
- ➔ If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- ➔ Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.