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:

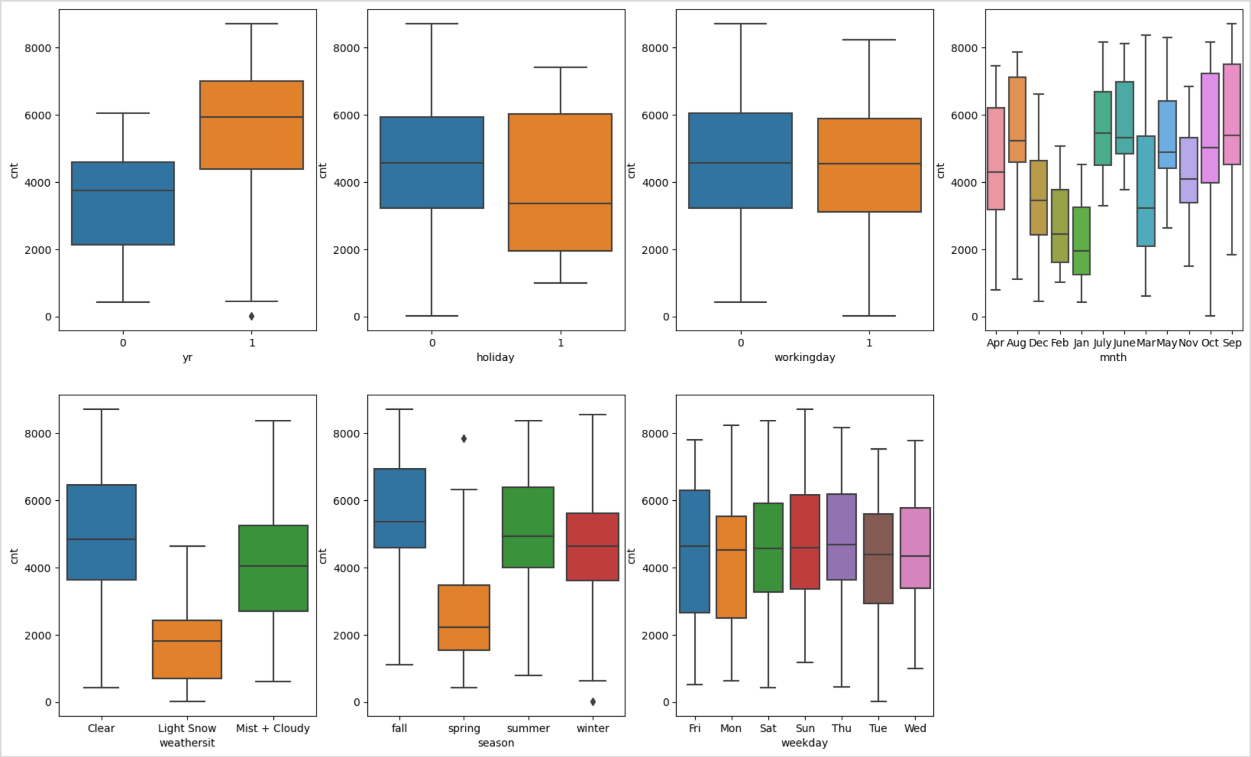


Fig: Box plots of Categorical Variables with Target Variable (Total Rental Counts of Bikes)

From the box plot of categorical variables with target variable (dependent variable) following inferences can be maid:

- It is pretty clear that demand in year **2019** (shown by 1 in plot) is greater than **2018**.
- Bike rentals booked **more** on **non-holidays** (shown by 0 in plot) as compared to holidays. Median of non-holidays is significantly high the holidays.
- On **working** or **non-working day** both we have almost similar results, as the median value is near 4500. Although non-working day has slight edge over working days, which we can also corelate with holidays as non-working days will be either holiday or weekend or both.
- From **Months** plot, we can see that maximum bikes were rented in the month of September, then October and then August.
- From **weather sit** we can see that initially we were give 4 categories out of only 3 are plotted on graph means no bikes were booked in 4th category i.e., Heavy Rain + Thunderstorm. Also, max. bikes were booked on clear or few cloudy weather then misty clouds and then light snow.
- From **Season** we can see that max. bikes were rented in fall season and then summer season and least in spring.
- From **Weekday** graph, we can see that most bikes were ranted on Friday, Sunday and Thursday.

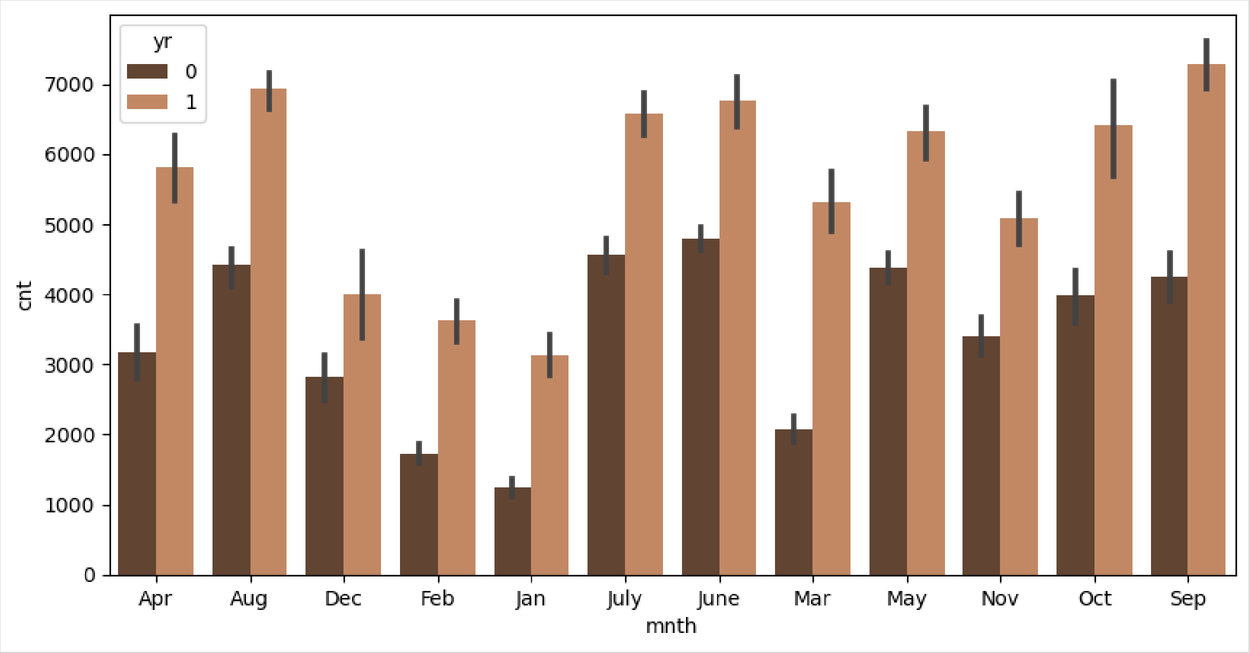


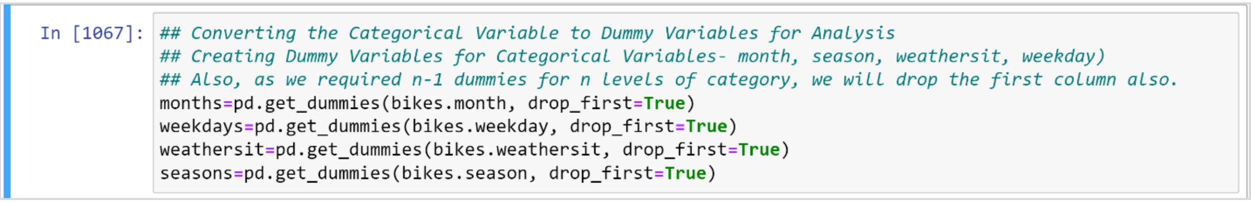
Fig: Month wise rental counts with year

- As we inferred from 1st plot that more bikes were rented in 2019 than 2018. This can also be seen from the above graph which is plotted month wise rental books in year 2018(0) & 2019(1).

---

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

- **drop_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation.
- If we have categorical variable with n-levels, then we need n-1 dummy variables to represent those n-levels.
- It also reduces the correlations created among dummy variables as 1 variable will be less.
- Ex- Let's say in our case **Season** variable has 4 types of values in Categorical column *(Summer, Winter, Spring and fall)* and we want to create dummy variable for that column. If one of the variables is not **summer, winter or spring** then definitely it is **fall**. So, we do not need 4th variable to identify the fall season.



```
In [1067]: ## Converting the Categorical Variable to Dummy Variables for Analysis
           ## Creating Dummy Variables for Categorical Variables- month, season, weathersit, weekday)
           ## Also, as we required n-1 dummies for n levels of category, we will drop the first column also.
           months=pd.get_dummies(bikes.month, drop_first=True)
           weekdays=pd.get_dummies(bikes.weekday, drop_first=True)
           weathersit=pd.get_dummies(bikes.weathersit, drop_first=True)
           seasons=pd.get_dummies(bikes.season, drop_first=True)
```

Fig: Showing use of drop_first in dummy variables creation for categorical variables

---

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

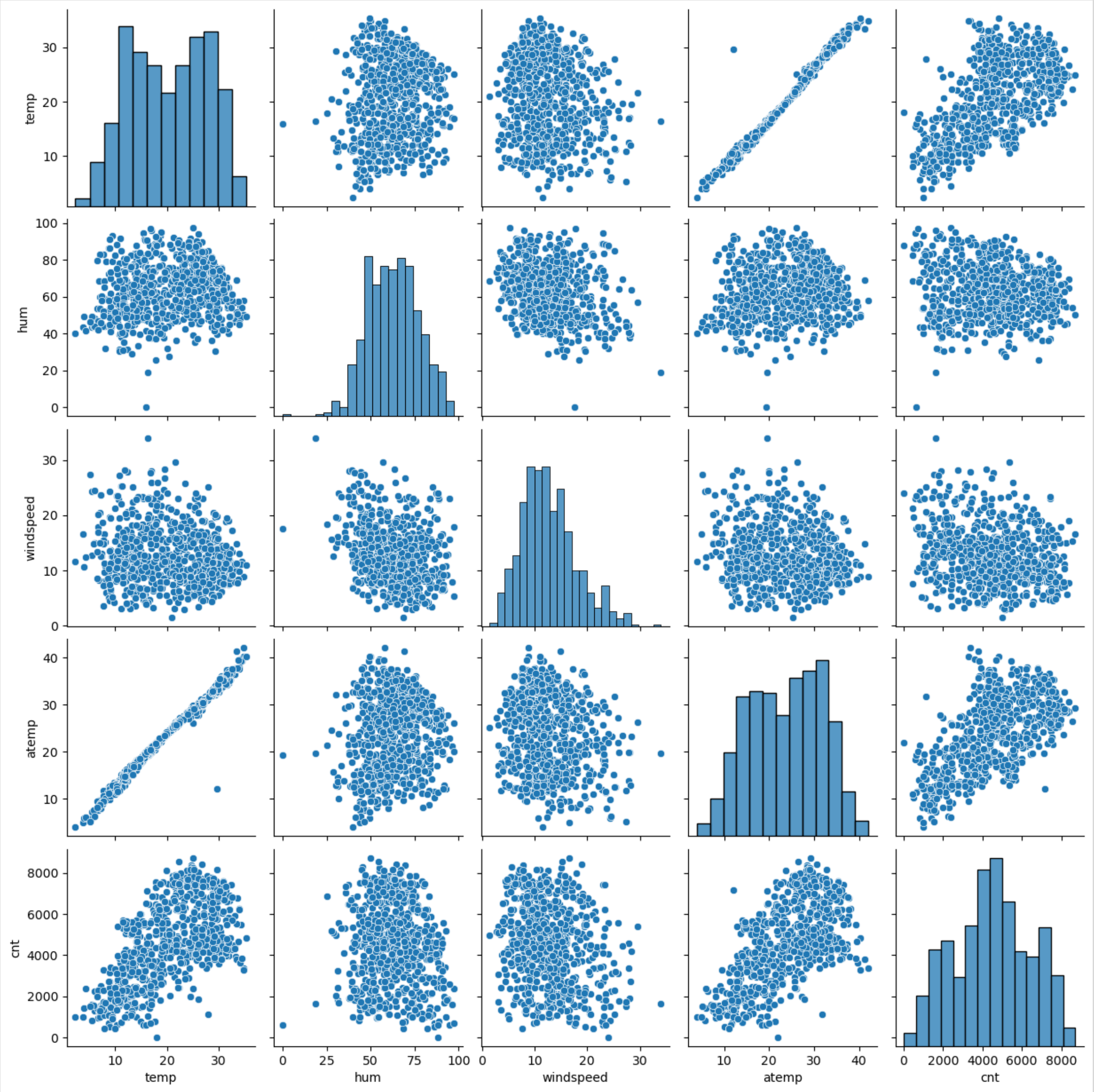


Fig: Pair-Plot between numeric variables including target variable

Among the numeric variables, **"temp" & "atemp"** has the highest correlation with the target variable rental counts (**"cnt"** in above graph).

Also, when we plotted the **Heat-Map** we saw that both the variable has exact correlation coefficient of **0.63** with the **target variable.**

----

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

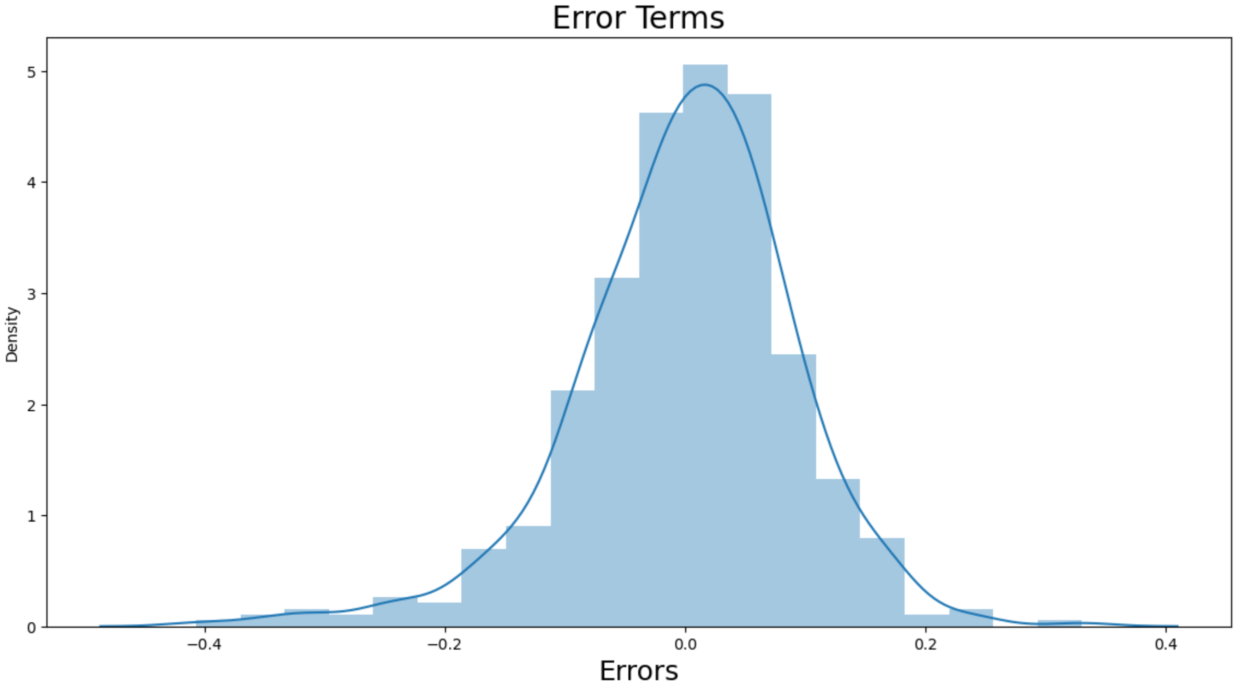


Fig: Distribution Plot of Error Terms (y_train - y_train_pred)

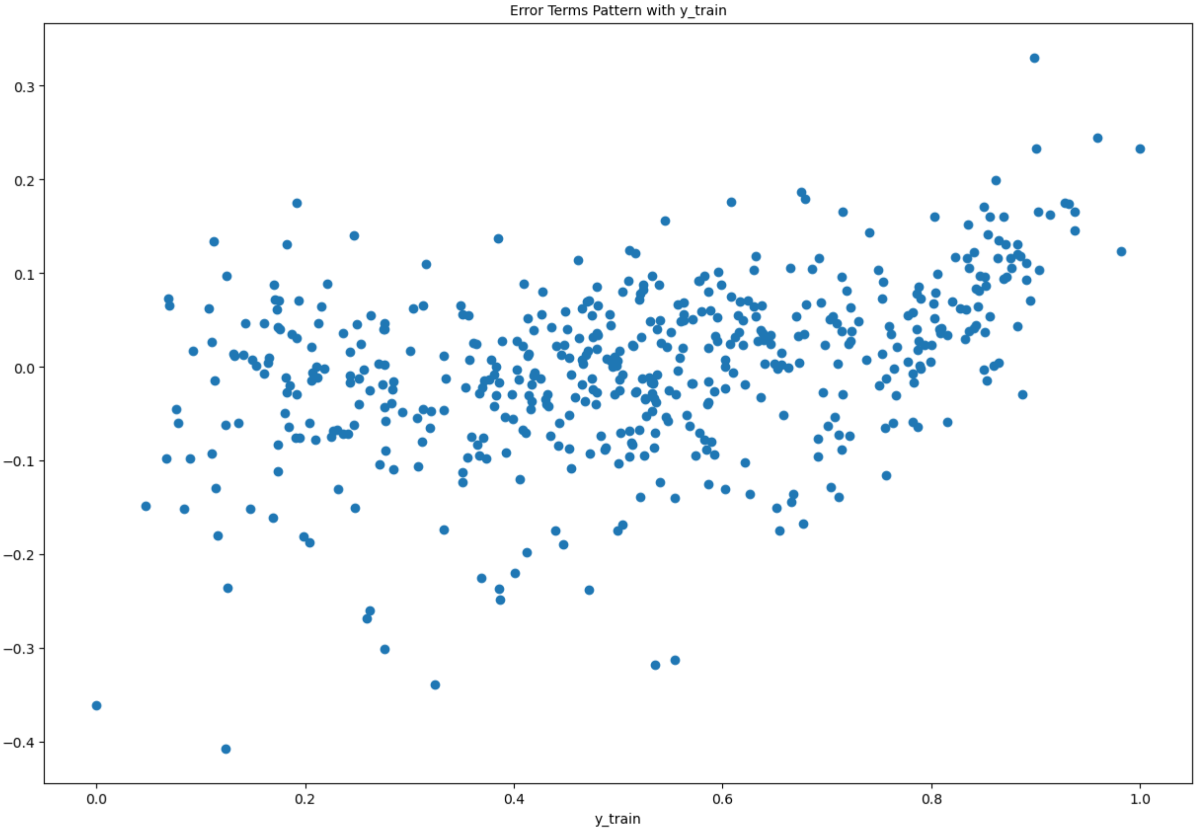


Fig: Error Terms Pattern with y_train

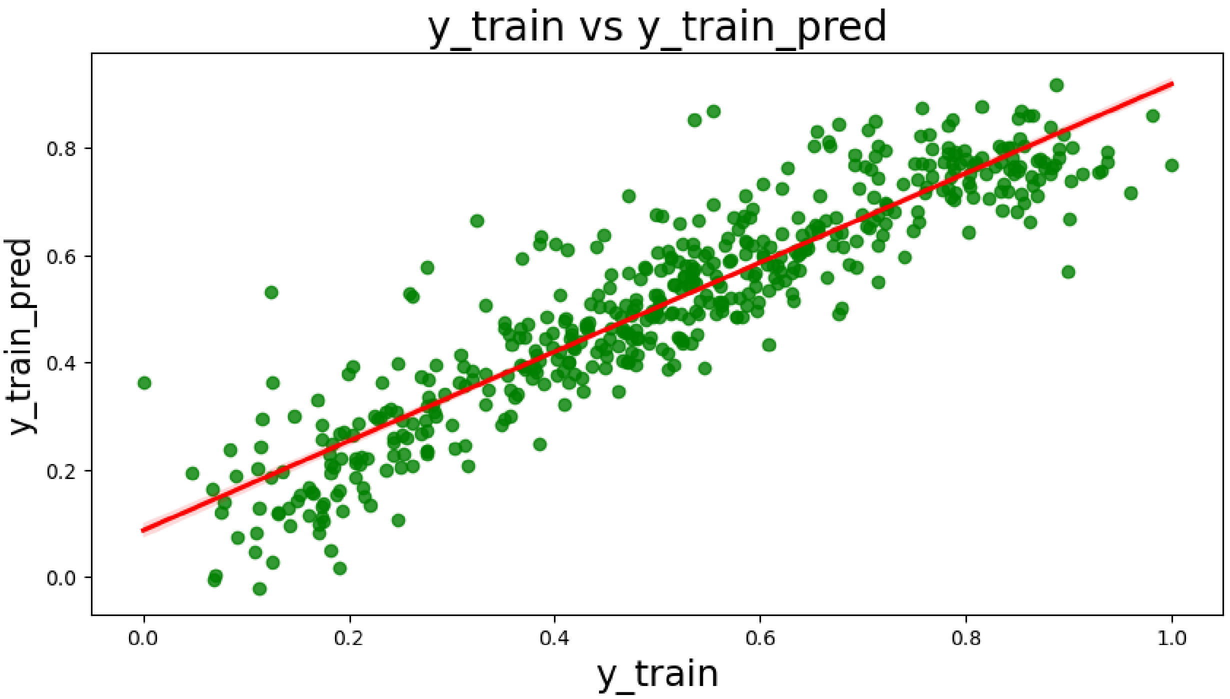


Fig: Regression plot between y_train & y_train_predicted

Following are the assumptions of Linear Regression that needs to be validated after building the model:

1. *X & Y have some sort of linear relationship.*
   - We started building our model by plotting the pair plot between the numeric variables and seen that temp. and rental counts are showing the maximum linearity (correlation) and that was also proved after building the model as temp. coefficient was max. = 0.568212.
2. *Error Terms are normally distributed with mean =0.*
   - We can see clearly from the above figure that error terms are normally distributed with mean centered around 0.
3. *Error Terms are independent of each other and have constant variance (Homoscedasticity).*
   - From the above figure we can see that Error Terms is not following any pattern and are randomly distributed and independent of each other.
   - Also, for variance we can say that it almost constant but slightly more in the middle.

----

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

We got the following equation for best fitted line:

$$rental\_counts = 0.125926 + 0.232861 \times Year + 0.548008 \times temp + 0.101195 \times Sep + 0.088080 \times summer + 0.129345 \times winter - 0.098685 \times holiday - 0.153246 \times windspeed - 0.282869 \times LightSnow - 0.078375 \times (Mist + Cloudy)$$

From the equation, by looking at the coefficients we can say that following are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. Temp. (0.548008)
2. Year (0.232861)
3. Light Snow (-0.282869) – Negatively Correlated

General Subjective Questions

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

- Linear regression is one of the basic forms of Machine Learning algorithm where we train the model to predict the behavior of our data based on some variables.
- It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
- From the name linear regression, it suggests linear that means the two variables which are on the x-axis and y-axis should have some linearity among them means correlated to each other.

- Mathematically, we can write a linear (straight line) equation as:

  y = mx + c

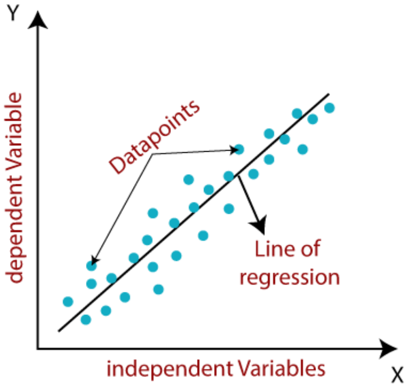
  Where:

  m: slope of line

  c: intercept

  x = Independent variable

  y = Dependent variable



- In form of linear regression, we generalize this equation as:

  $Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$

  Where:

  X1, X2, X3...Xp: Independent Variables

  Y: Dependent Variable

  β0: Constant

  β1, β2, β3.. βn: Coefficient of X1, X2, X3…Xp

- **Types of Linear Regression**

  Linear regression can be further divided into two types of the algorithm:

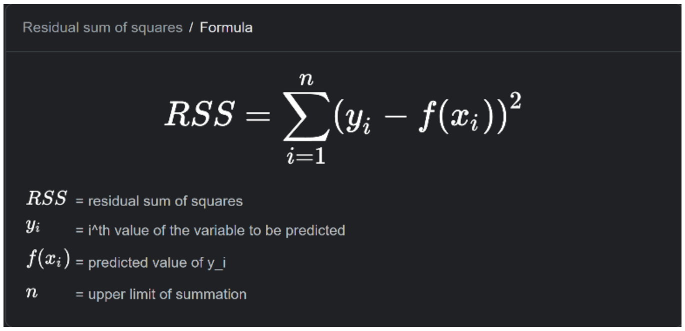
  - **Simple Linear Regression:**

    If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

  - **Multiple Linear regression:**

    If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

- It is based on RSS (Residual Sum of Squares) which is calculated by minimize the cost function by OLS (Ordinary Least Square Method).
- Gradient Descent is the first order iterative algorithm which is used for minimizing the cost function



Residual sum of squares / Formula

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

Fig: RSS Formula

- **Assumptions of Linear Regression**
  - Linear relationship between the features and target.
  - Small or no multicollinearity between the features
  - Constant Variance of Error Terms (Homoscedasticity)
  - Normal distribution of error terms with Mean =0
  - Error Terms are independent of each other (no pattern with either x or y)

---

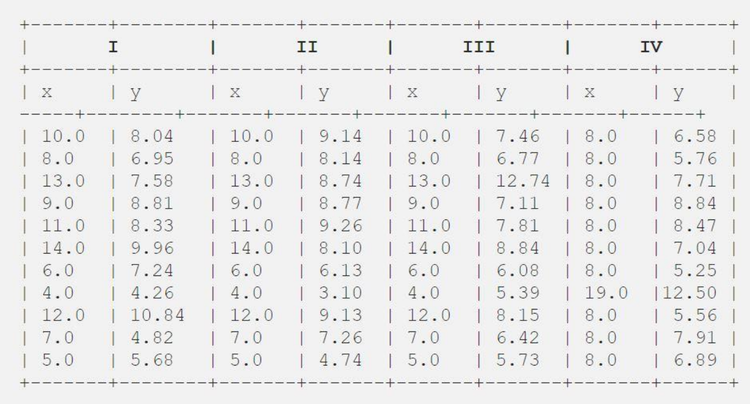
2. Explain the Anscombe's quartet in detail. (3 marks)

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. 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.

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.



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

Fig: Anscombe's quartet

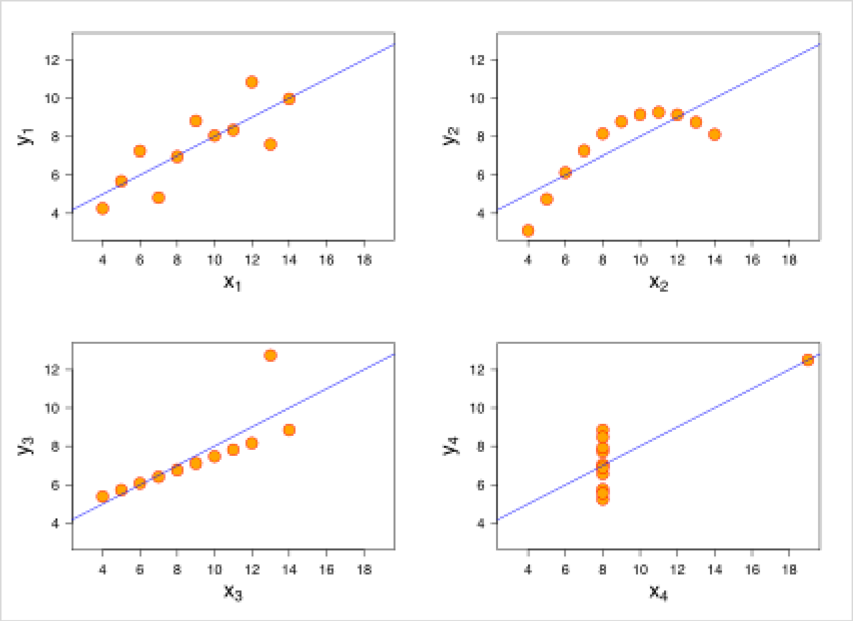


Fig: All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

---

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

- In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

- The Pearson's correlation coefficient varies between -1 and +1 where:

  r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

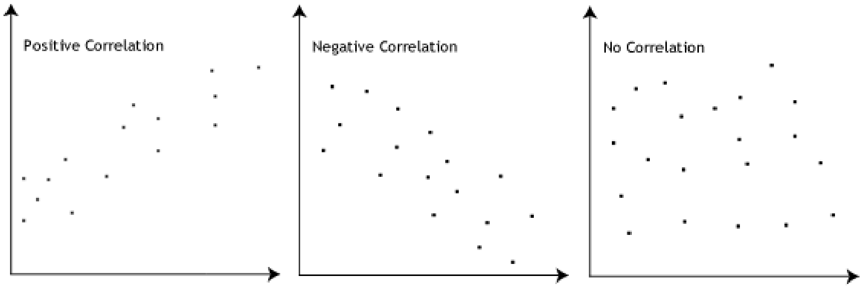
  r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

  r = 0 means there is no linear association

  r > 0 < 5 means there is a weak association

  r > 5 < 8 means there is a moderate association

  r > 8 means there is a strong association



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Fig: Pearson r Formula

Where:

- r = correlation coefficient
- x(i) = values of the x-variable in a sample
- x bar = mean of the values of the x-variable
- y(i) = values of the y-variable in a sample
- y bar = mean of the values of the y-variable

----

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

- 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 (Gradient Descent) 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).

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

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

- One disadvantage of standardization over normalization is that it loses some information in the data, especially about outliers. Normalization is highly affected by outliers. Standardization is slightly affected by outliers.
- Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.

---

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)**

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:

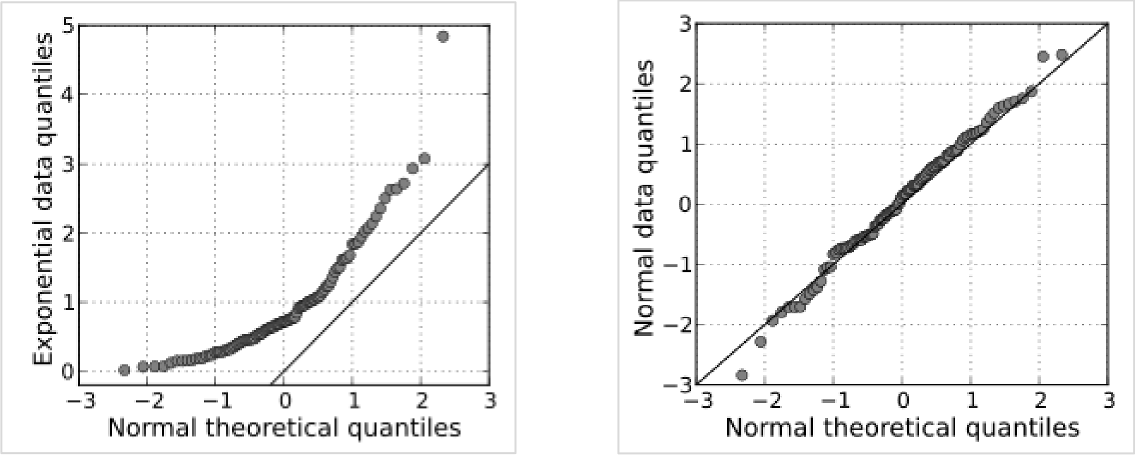


Fig: q-q plots in linear regression

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.

END