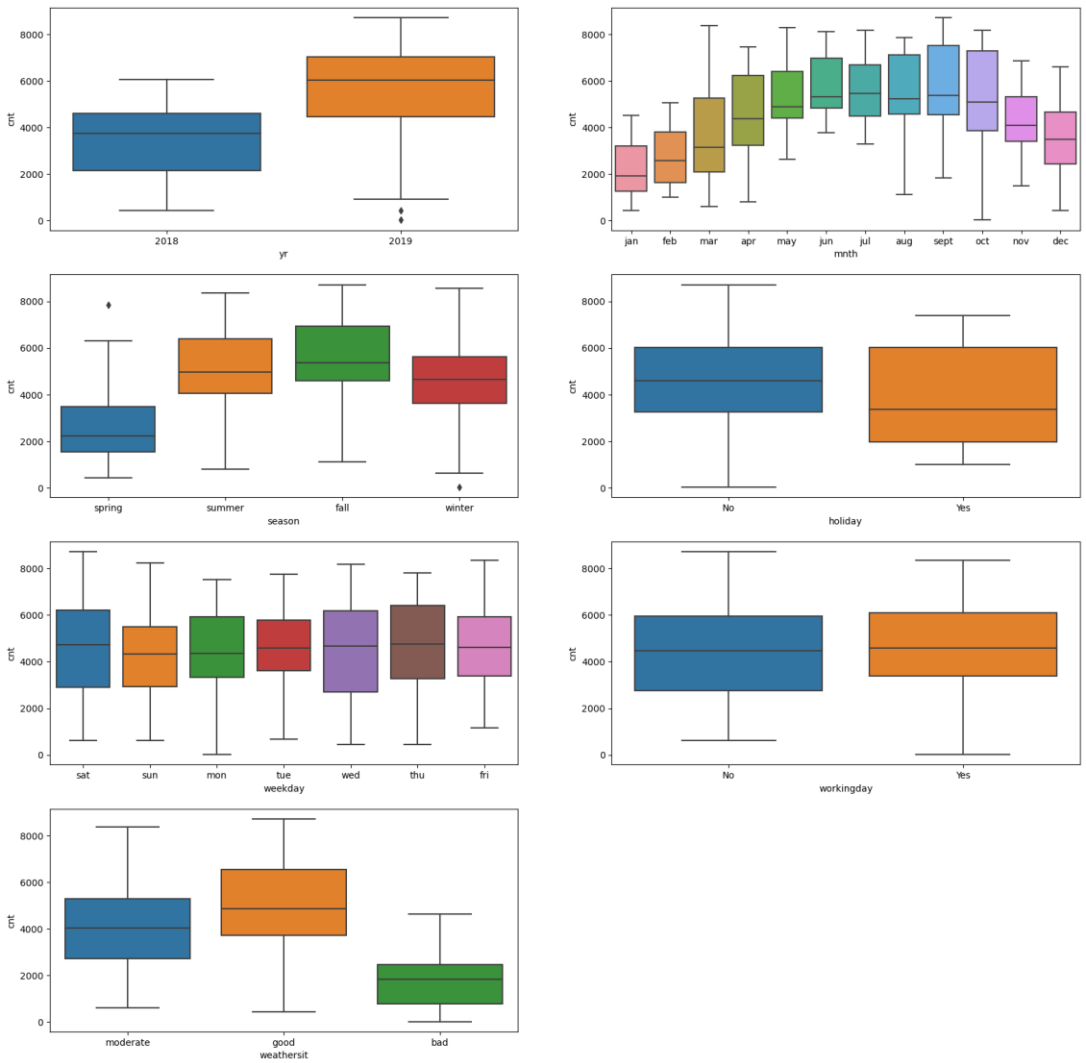
BIKE SHARING ASSIGNMENT

SUBJECTIVE QUESTIONS

SHRUTHIP VENKATESH
ML-C54

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)



Answer:

Based on the plots above, below are the inferences,

1.There is increase in demand for the rental bikes every year.

2.There is an incremental rise in demand till September month and then it declines till December. This might be due to the weather conditions in winter.

3.The demand is high during fall season which might be due to good weather conditions.

4.During holidays, there is a wide spread in demand and the median is less, which means that there is less demand.

5.There is no significant observation about week day, though Wednesday has a little bit more spread in demand.

6.There is no significant observation about working day, though non working day has a little bit more spread in demand.

7.When the weather is good, there is a significant rise in demand and when the weather is bad, there is significant decline in demand.

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

Answer:

- drop_first = True helps in reducing one extra column generated during dummy variables creation.
- If we do not drop the extra column, it will increase the multicollinearity since there is redundant data.
- Setting it true will help in reducing the multicollinearity created during these dummy variables creation by dropping one extra column.

- Example, let's say the salary of a person can be categorized as 'high', 'medium' and 'low. When drop_first is set to false, then the dummy variables will be like 'is_high', 'is_medium' and 'is_low'.

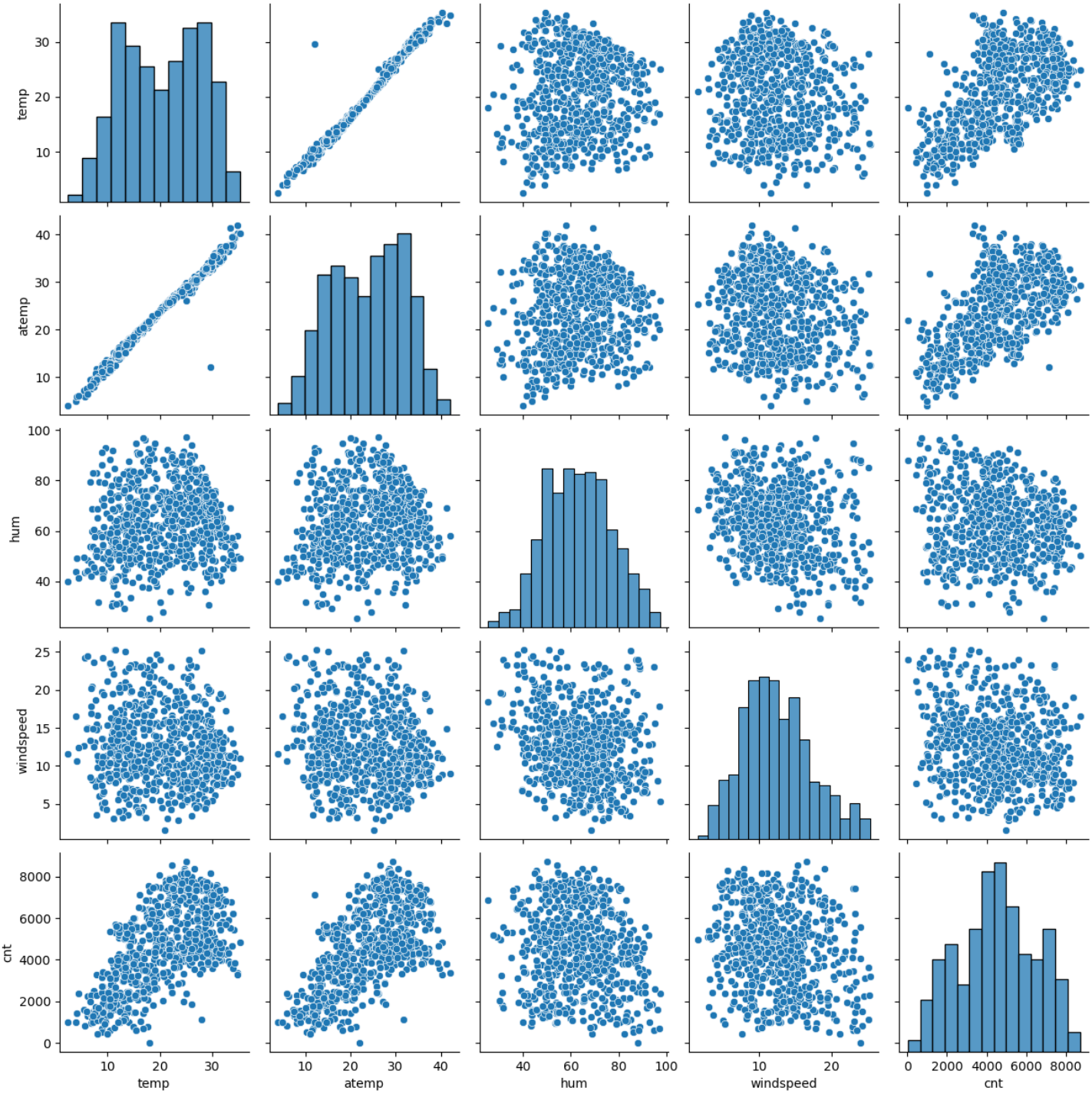
is_high	is_medium	is_low	Category
1	0	0	high
0	1	0	medium
0	0	1	low

- We can achieve the above same result by reducing one column by setting drop_first=True like below.

is_high	is_medium	Category
1	0	high
0	1	medium
0	0	low since we know this is the only remaining category

- Thus, it is important to use drop_first = True like explained before.
- In general, k levels can be explained using k-1 variables.

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



Answer:

Based on the above pair plot, cnt (dependent variable) has highest correlation with temp and atemp variables (independent variable).

Also,

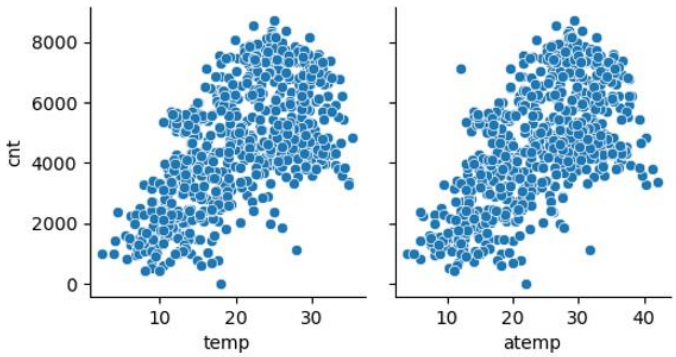
- temp and atemp has very high linear correlation. So, these two could be considered as single variable.

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

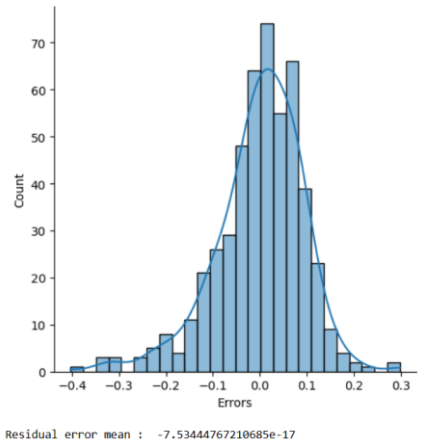
Answer:

The assumptions are validated as explained below,

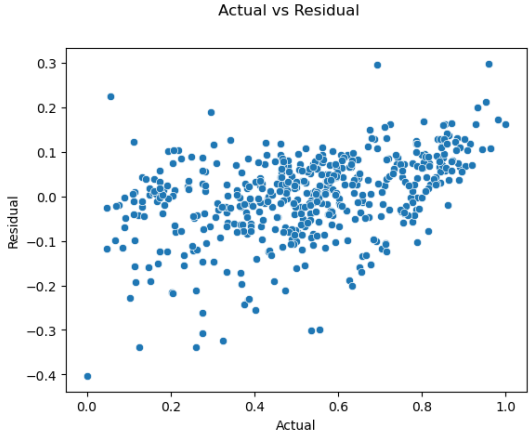
1) *There is a linear correlation between dependent and independent variable* - cnt (dependent variable) has a very good linear correlation with temp and atemp variables (independent variable).



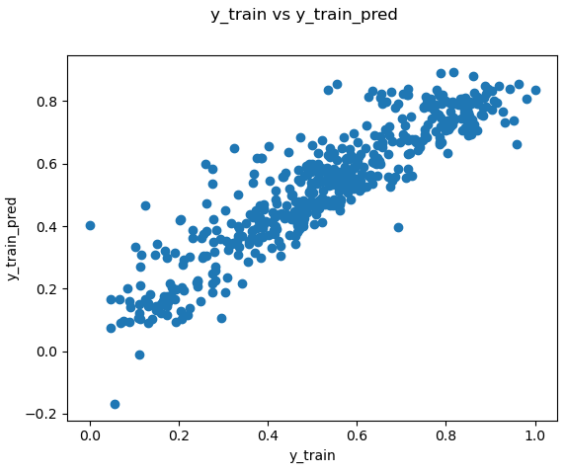
2) *Error terms are normally distributed with mean 0* – the below error distribution plot proves the assumption.



3) *Error terms do not follow any pattern* – the below plot between true value and the error proves the assumption. Also the Durbin Watson test value of the final model is **1.973** which is close to 2 and we can safely claim that there is *no auto correlation*.



4) *Error terms have constant variance* – the below plot between true value and predicted value shows there is no significant deviation in spread of prediction.



5) *Low multicollinearity* – the below table shows the VIF values for the predictors used in the final model. It is evident that the VIF values are well below **5** indicating less redundancy among the predictors.

```
-----
                Features   VIF
1                   temp  2.30
2                     yr  1.95
3  weathersit_moderate  1.41
4        season_spring  1.18
5            mnth_sept  1.17
6             mnth_oct  1.13
7       weathersit_bad  1.06
0              holiday  1.03
```

6) We saw that the test data R2 score is **0.81** whereas for training data it is **0.82**. The difference of 0.01 is acceptable. Hence the model did not overfit and has **generalized** the learning.

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

Ranking	Predictor	Coefficient
1	temp	0.351334
2	weathersit_bad	-0.342787
3	yr	0.230702
4	season_spring	-0.166929
5	holiday	-0.088471
6	mnth_sept	0.094459
7	weathersit_moderate	-0.078991
8	mnth_oct	0.076377

Answer:

The above ranking table shows the predictors used in the final model in the ranked order.

Below are the top three inferences from this table, the demand for bikes

1) Increase when temperature increase => The company can expect high demand during high temperature.
2) Decrease when weather is bad => It might be because of difficulty for the user during these bad conditions which is not good for driving. Hence, company can focus on preparing for meeting higher demands later like servicing the bikes, repairing the bike docks etc.
3) Increase every year => The company should focus on staying in business.

Based on the above table and inferences, below are the **top three features** contributing significantly towards explaining the demand of the shared bikes

1) Temperature ('temp')
2) Weather ('weathersit')
3) Year ('yr')

General Subjective Questions

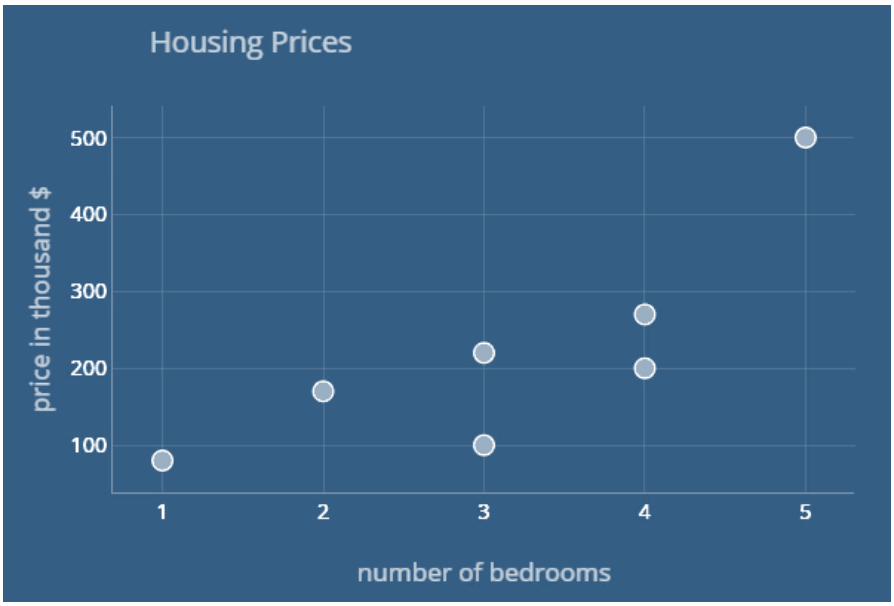
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.
- When the number of the independent feature is one, it is called as Univariate Linear regression. When the case of more than one feature, it is known as Multivariate linear regression.
- The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line or plane that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.
- Here y is called a dependent or target variable and X is called independent variable also known as the predictor of y. In regression, we have to predict the value of y using a learnt linear function given X as independent features.

Example :

- Imagine we have a dataset of houses for a specific city, where we are given the number of bedrooms for each house as well as the price of each house. The dataset might look like this:



- We can try to fit a line that explains the price of houses based on the number of bedrooms, which might look like below

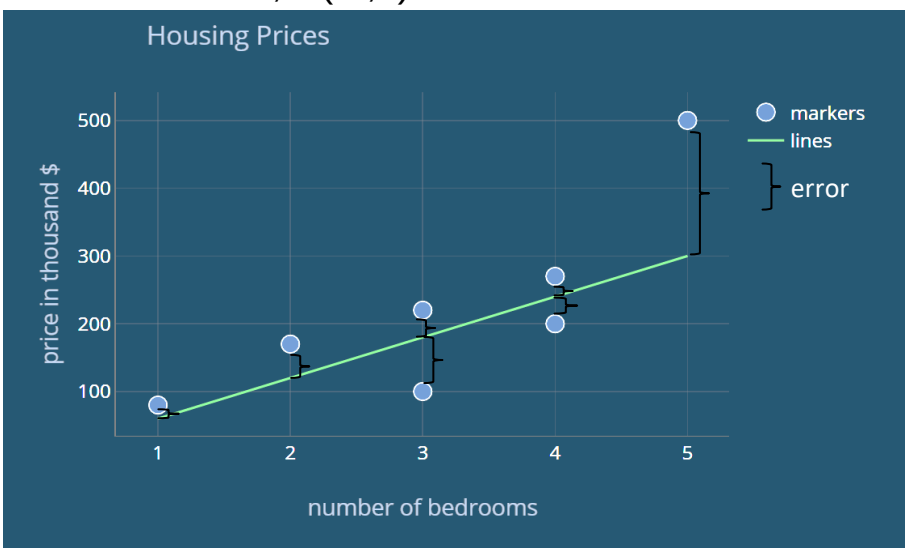
- We can see there are many lines possible, which can be expressed in general with a straight line equation such as,

$$y = \mathbf{m}*X + \mathbf{c}$$

- Here X is independant variable (number of bedrooms) and y is independent variable (price of house) and are known. **m** (slope) of the line and **c** (y-intercept) of the line are unknown.
- Linear regression modelling is to find the best values of **m** and **c** such that we can predict y as close to reality for X.

- We introduce a term called, cost function, which is the error between the predicted y and the actual y in simple terms.

Cost function, J (**m**,**c**) = sum of error terms



- The goal is to minimize this cost function to find the best fitting line for this dataset.
- The best fitting linear line equation is called the linear regression model.

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

Answer:

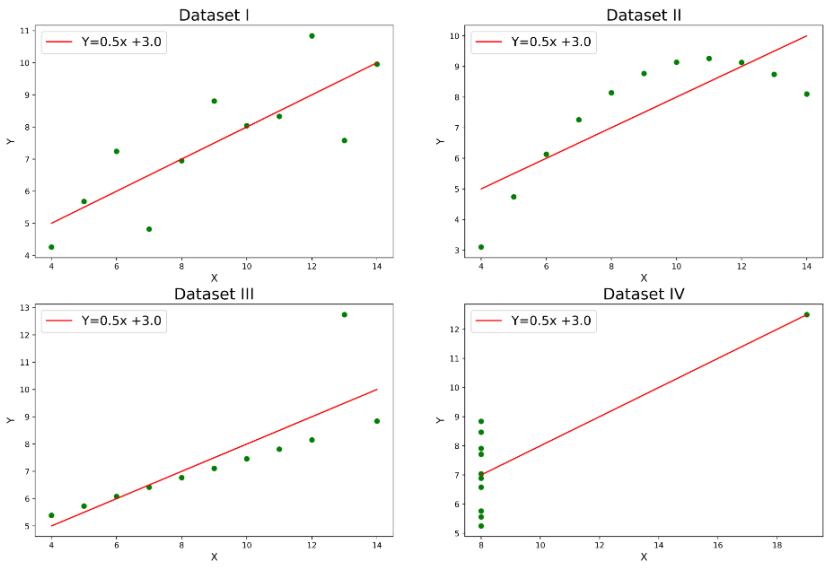
- Anscombe's quartet is used to demonstrate the importance of visualizing dataset. Summary statistics alone is not sufficient.
- It comprises a set of four dataset which have identical descriptive statistical properties but having different representations when we plot on graph.
- Each dataset consists of 11 (x,y) points as shown below,

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

- These datasets have the identical descriptive statistical summary as below,

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

- But, when we plot these datasets, we get a different perspective,



- We could see that all the datasets have the same linear regression line whereas the reality is different as we can see from the distribution plot.

We could infer the following from visualizing the data,

- Dataset I: There is a linear relationship between x and y and hence the line equation could be considered as valid.
- Dataset II: There is a non-linear relationship between x and y and hence line equation cannot be considered.
- Dataset III: There is an outlier present in the data which cannot be explained by the line equation.
- Dataset IV: Because of one big outlier present in the data, it cannot be explained by the line equation.

- To summarize, the quartet explains the importance of explanatory data analysis before starting to analyze relationship, and the drawback of basic statistic properties for describing realistic datasets.

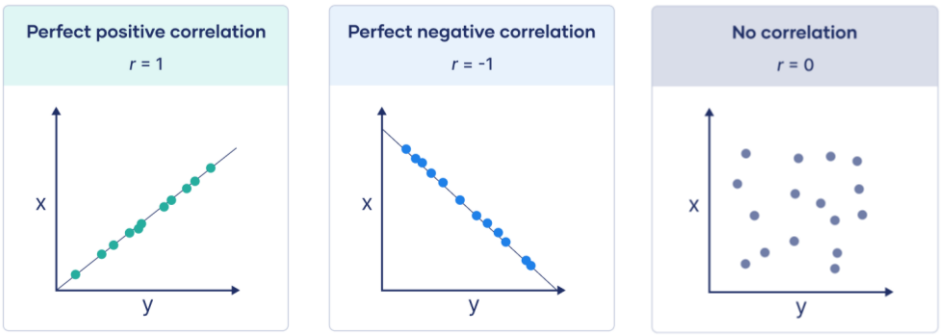
3. What is Pearson's R? (3 marks)

Answer:

- Pearson correlation coefficient (R) is the popular way of measuring a linear correlation between two variables.
- It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (R)	Correlation type	Interpretation
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.

- Graphically, we can see the correlation types like below,



- Pearson's R equation is,

$$r = [n(\Sigma xy) - \Sigma x \Sigma y] \,/\, \text{Square root of} \sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}$$

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

Answer:

- Scaling is a technique to standardize all the independent variables in the data set to a fixed range.
- It is done during data preparation step.
- If scaling is not done, then the algorithm will weigh greater values as high and smaller values as low since the algorithm works on numbers and not on units.
- Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc
- Example, in the bike sharing assignment, holiday variable has 0 or 1 values whereas windspeed is having higher values up to 25. When these values are given to algorithm without scaling, windspeed will be weighed higher due to the bigger values. But in reality, from pair plot, it is clear that windspeed has no pattern with the demand for bikes. When these values are scaled, it is seen that holiday comes as one of the best variables whereas windspeed is not.

Reasons for scaling:

- Scaling guarantees that all the independent variables are in comparable scales and have comparable ranges.
- The algorithm's performance improves by converging quickly.
- Numerical instability can be prevented by avoiding significant scale disparities between features.
- Scaling ensures that each independent variable is given the same consideration during the learning process.

Normalized scaling:

- This is also called Min-Max scaling. This uses maximum and minimum of the data.
- This method scales the data between 0 and 1. (0 being minimum and 1 being maximum.
- Formula, $$X_{\text{scaled}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

Standardized scaling:

- This method of scaling is basically based on the central tendencies and variance of the data.
- This method scales the data to achieve a normal distribution with mean 0 and standard deviation 1.
- Formula, $$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

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

Answer:

- Let us look into the Variance Inflation Factor (VIF) formula,

$$VIF_i = \frac{1}{1 - R_i^2}$$

- We know that VIF tells about the multicollinearity of ith independent variable with respect to the other independent variables.
- R2 score is known as the coefficient of determination. We can consider it as correlation between two variables.

- When a variable can be perfectly expressed by a linear combination of other variables, then R2 score will be 1.
- When R2 score is 1, denominator in VIF becomes 0 and value of VIF will become infinite.

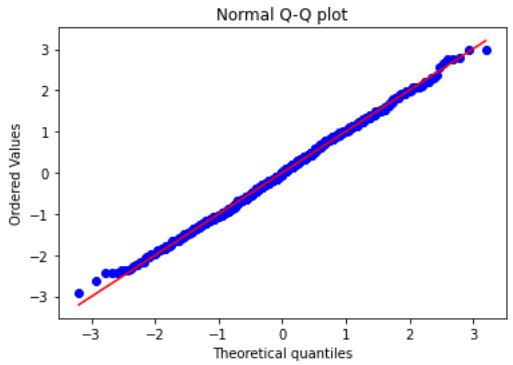
- Thus a VIF value of infinite means that a variable is perfectly expressed by linear combination of other variables.

- Note,
  - When VIF is above 10, then there is very high correlation.
  - When VIF is above 5, then we might need to check the significance of that variable.
  - When VIF is less than 5, we can consider there is less correlation.

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

Answer:

- In statistics, a Q–Q plot aka Quantile-Quantile plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.
- Below is an example Q-Q plot of a normally distributed data,



Interpretation:

- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity 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

Usage:

To determine

- If two samples are from the same population.
- If two samples have the same tail.
- If two samples have the same distribution shape.
- If two samples have common location behavior.
- If scale and skewness are similar or different in the two distributions.

Advantages:

- Since Q-Q plot is like probability plot, when comparing two datasets, the sample size need not to be equal.
- Since we need to normalize the dataset, we don’t need to consider about the dimensions of values.
- Q–Q plot is generally more diagnostic than comparing the samples histograms
- This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic.