Ans of Que 2:- Salary and years experience

**Objective** - Delivery\_time -> Predict delivery time using sorting time

**Data pre-processing and Inferences from the data Set**

Data set talks about the years of experience with respect to salary with 30 Observations

**Columns:**

Delivery Time

Sorting Time

Dataset Size: 30

Data given is found to be a continuous data for which a simple linear regression can be performed getting deeper in to the data analysis and its behavior.

**years of experience**

Delivery Time Sorting Time

Min. : 8.00 Min. : 2.00

1st Qu.:13.50 1st Qu.: 4.00

Median :17.83 Median : 6.00

Mean :16.79 Mean : 6.19

3rd Qu.:19.75 3rd Qu.: 8.00

Max. :29.00 Max. :10.00

Ranges Between :- 1.1 – 10.5

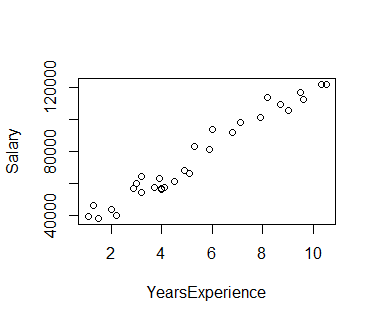
* For this years of experience, the mean is 5.3, it is just the average of the years of experience
* The median for the given data is 4.7, it speaks about the center of the data
* A comparison between mean and median tell us that data is skewed, if data was not skewed we would have considered mean but here it is positive skewed so we take median to talk about the data.
* The data is right skewed, skewness=0.36
* The data is positive kurtosis, it means it has thin peak and wider tails.
* No Outliers and missing values
* Data is normally distributed

**Salary**

Ranges Between :- 37731-122391

* For this Salary, the mean is 76003, which is more the median=65237
* The data is right skewed, skewness=0.34
* The data is positive kurtosis, it means it has thin peak and wider tails.
* No Outliers and missing values
* Data is normally distributed

**plot(YearsExperience,Salary)**

****

The above scatter diagram infer that the salary and years of experience are positively related

**Correlation coefficient:**

cor(`Delivery Time`,`Sorting Time’)

0.8259973

Based on the on the correlation value obtained which is 0.82 tells that it is moderately correlated

**Model Building**:

We use **lm() function from Base Package in R-studio** to estimate the years of experience using the other variable Salary whereas in python **LineraRegression() is used from the sklearn package**

Call:

lm(formula = `Delivery Time` ~ `Sorting Time`)

Residuals:

Min 1Q Median 3Q Max

-5.1729 -2.0298 -0.0298 0.8741 6.6722

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.5827 1.7217 3.823 0.00115 \*\*

`Sorting Time` 1.6490 0.2582 6.387 3.98e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.935 on 19 degrees of freedom

Multiple R-squared: 0.6823, Adjusted R-squared: 0.6655

F-statistic: 40.8 on 1 and 19 DF, p-value: 3.983e-06

Call:

lm(formula = Salary ~ YearsExperience)

Residuals:

Min 1Q Median 3Q Max

-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*

YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5788 on 28 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

p-values:

values are significant as it is less than 0.5

also we observed Multiple R squared value is 94.50, which is greater than 0.8 ((in general)

The Probability Value for F-Statistic is 2.2e-16(Overall Probability Model is also less than 0.05

Evaluation:

RMSE value = 0.55

We may have to do transformation of variables for better R v-squared value

Lets Apply some transformations there are different types of transformation techniques like log, exponential, Quadratic

Let’s also look in to the plot how they are behaving

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Input | Output | R-Squared value | Plot |
| Simple Regression | Sorting Time | Delivery Time | 0.6823 |  |
| Logarithmic  Transformation | Sorting Time | Log(Delivery Time) | 0.6954 |  |
| Exponential  Transformation | Log(Sorting Time) | Delivery Time | 0.7109 |  |
| Quadratic Transformation | Sorting Time | Delivery Time + I(Delivery Time ^2) | 0.6934 |  |
| Polynomial  Transformation | Sorting Time | Delivery Time + I(Delivery Time^2)+  I(Delivery Time ^3) | 0.7034 |  |

Based on obtained r-squared values and the plot the best transformation technique is polynomial with 0.96 R- squared value

**Packages used**

R studio

* Readr
* Ggplot2
* Moments

Python

* import numpy as np
* import pandas as pd
* import matplotlib.pylab as plt
* import statsmodels.formula.api as smf

Please refer the attached R and Python file for codes.