Ans of Que 1:- **Weight Gained and Calories Consumed**

**Objective** - calories\_consumed -> predict weight gained using calories consumed

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

Data set talks about the weight gained with respect to calories\_consumed with 14 Observations

**Columns:**

Weight Gained

calories Consumed

Dataset Size: 14

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.

**Weight Gained**

Weight gained (grams) Calories Consumed

Min. : 62.0 Min. :1400

1st Qu.: 114.5 1st Qu.:1728

Median : 200.0 Median :2250

Mean : 357.7 Mean :2341

3rd Qu.: 537.5 3rd Qu.:2775

Max. :1100.0 Max. :3900

Ranges Between :- 62 – 1100

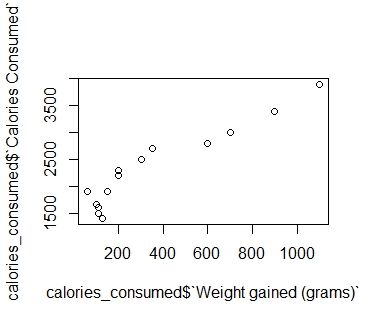
* For this weight gained, the mean is 200, it is just the average of the weight gained
* The median for the given data is 200, it speaks about the center of the data
* A comparison between mean and median tell us that data is skewed, because mean is greater than median.
* The data is right skewed, skewness=1.11
* The data is positive kurtosis, it means it has thin peak and wider tails.
* No Outliers and missing values and NA values
* Data is normally distributed

**Calories Consumed**

Ranges Between :- 1400-3900

* For this calories consumed, the mean is 2341, which is more the median=2250 so it states that it is skewed
* The data is right skewed, skewness=0.58
* The data is positive kurtosis, it means it has thin peak and wider tails.
* No Outliers and missing values and NA values
* Data is not normally distributed we applied log transformation to make it normally distributed

**plot(weight gained, weigh gained)**

****

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

**Correlation coefficient:**

cor(`Weight gained (grams)`,`Calories Consumed`)

0.946991

Based on the correlation value obtained which is 0.94 tells that it is strongly 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 = `Weight gained (grams)` ~ `Calories Consumed`)

Residuals:

Min 1Q Median 3Q Max

-158.67 -107.56 36.70 81.68 165.53

Coefficients:

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

(Intercept) -625.75236 100.82293 -6.206 4.54e-05 \*\*\*

`Calories Consumed` 0.42016 0.04115 10.211 2.86e-07 \*\*\*

---

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

Residual standard error: 111.6 on 12 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8882

F-statistic: 104.3 on 1 and 12 DF, p-value: 2.856e-07

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 0.89, 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 = 103.30

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 | Weight Gained | Calories Consumed | 0.89 |  |
| Logarithmic  Transformation | Weight Gained | Log(Calories Consumed) | 0.807 |  |
| Exponential  Transformation | Log(Weight Gained) | Calories Consumed | 0.877 |  |
| Quadratic Transformation | Weight Gained | Calories Consumed + I(Calories Consumed ^2) | 0.9521 |  |
| Polynomial  Transformation | Weight Gained | Calories Consumed + I(Calories Consumed ^2)+  I(Calories Consumed ^3) | 0.9811 |  |

Based on obtained r-squared values and the plot the best transformation technique is polynomial with 0.98 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.