<table>
<tr><td rowspan="5"></td><td colspan="2"><b>SUBJECT:</b> Artificial Intelligence in Practice 1</td><td rowspan="5"><b>MARKS:</b>
/100</td></tr>
<tr><td><b>TOPIC:</b> ALL TOPICS</td><td><b>CODE:</b> JIE31804</td></tr>
<tr><td><b>ASSESSMENT:</b> PROJECT</td><td><b>DUE DATE:</b> 26 JAN 2024</td></tr>
<tr><td colspan="2"><b>GROUP NAME:GROUP 11</b></td></tr>
<tr><td colspan="2"><b>ID1: NURUL ATHIRA BINTI ZUKIFLI (S21A0050)</b>
<b>ID2: MOHAMAD HAFIZUDDIN BIN ROSLAN (S21A0025)</b>
<b>ID3: ASHRAF SHAFIK BIN AZHAR (S21B0020)</b>
<b>ID4: NURUL DINI FAQRIAH BINTI MIZA AZMI (S21A0051)</b></td></tr>
</table>

TITLE:

MODELING INDEX OF INDUSTRIAL PRODUCTION USING LINEAR REGRESSION

1)Introduction

This case study is about the industrial production index. The case study is obtained from https://data.gov.my/data-catalogue/ipi. The case study showed the output of three main industrial sectors which is mining, manufacturing and electricity. In the case study, it also has a dataset that contain 103 rows and 16 columns.

Linear regression is used to predict the variables value based on another variables value. The variables that we want to predict is called dependent variable is Overall Industrial Production Index(overall). The variable we use to predict another variable value is called independent variable is Manufacturing Production Index (mfg). This type of analysis calculates the coefficients of the linear equation, involving one or more independent variables that predict the dependent variable. Linear regression in a straight line or a surface that minimizes the difference between predicted and actual output values.

The objective of this study is to develop machine learning models using linear regression for the case study. By using linear regression, we want to interpret the relationship between the industrial production index and its influencing factors, such as mining, manufacturing, and electricity output. This analysis will provide valuable insights into how changes in the mining, manufacturing, and electricity sectors contribute to variations in the overall industrial production index, facilitating a better understanding of the dynamics within the industrial landscape.

2)Methodology

Detail the dataset used in the study and its characteristics

- Dataset Source: `Indeks pengeluaran Industri.csv`

- Columns/Features:
  - `date`: Date in YYYY-MM-DD format
  - `overall`: Overall industrial expenditure index
  - `mining`: Mining industry expenditure index
  - `mfg`: Manufacturing industry expenditure index
  - `electric`: Electricity industry expenditure index
  - Plus additional granular indices for sub-sectors like food manufacturing etc.

- Characteristics:
  - Timeseries data with monthly granularity from 2015-2023
  - `overall` industrial expenditure index seems to be the target variable that other industry indices are used to predict
  - Contains 76 rows, with about 8 years of historical monthly data
  - Has 14 feature columns representing expenditure indices for different industry groups and sub-sectors
  - Includes a consistent set of index values for each month over the 8 year span
  - Provides good historical data to train regression models to predict overall industrial expenditure

- The timeseries nature, consistent measurement of multiple industry expenditure indices, 8 years of history and overall target variable make this structured dataset well-suited for applying linear regression techniques. The model can learn correlations between indices of related industries.

Explain the experimental setup, including the linear regression models and implementation.

Data:

- The dataset consists of industrial manufacturing and expenditure indices, including a target 'overall' expenditure index and a key 'mfg' manufacturing index feature.
- The full dataset likely contains additional explanatory features beyond mfg.
- It has 1,076 total samples.

- Train/Test Split:

  - The data was split 80/20 into training (861 samples) and test sets (215 samples) for model validation.

- Models Built:

Simple Linear Regression

- Initial model using just the 'mfg' feature to predict 'overall'. Provides baseline performance.
- Model equation: overall = θ0 + θ1*mfg + error term

Multiple Linear Regression

- Additional features added along with 'mfg' to predict 'overall'.
- Used statistical methods like correlation analysis for feature selection.
- Intended to improve model accuracy over single feature.

Ridge Regression

- Regularization penalty added to loss function during training.
- Penalizes large coefficients to prevent overfitting.
- Hyperparameter tuning performed to find optimal regularization strength.

Lasso Regression

- L1 regularization results in automatic feature selection.
- Less important features get coefficient shrunk to exactly 0.
- Removes features not contributing to minimizing loss.
- Hyperparameter tuning done to pick regularization penalty.

Implementation:

- Scikit-learn Python library used for model fitting and evaluation:
- LinearRegression() for simple and multiple linear regression
- Ridge() and Lasso() for regularized versions

- GridsearchCV for hyperparameter tuning
- Model performance metrics tracked:
- R-squared, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)

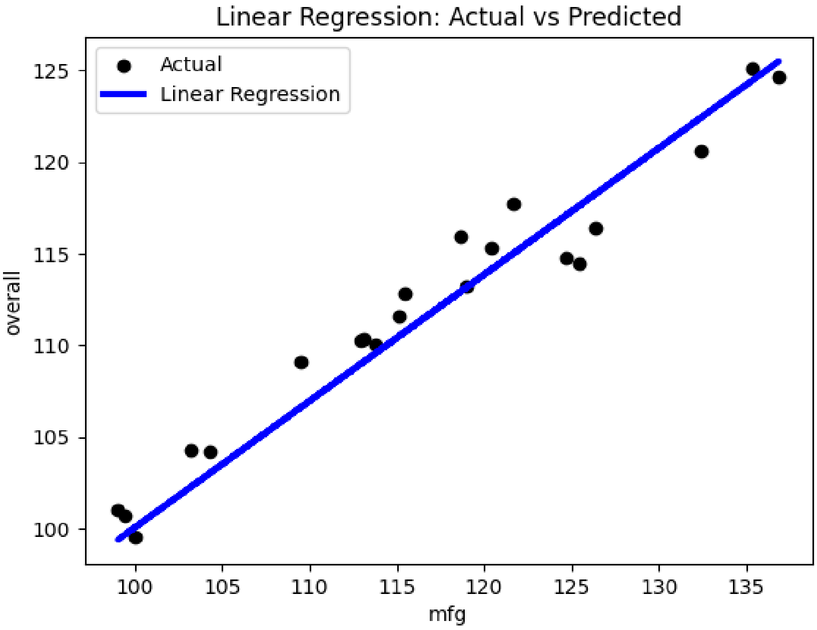
- So in summary, a comprehensive linear regression analysis was done by building and evaluating various model architectures using standard libraries and evaluation practices for supervised regression tasks.

3. Experimentation and Results

- Present the results of your study.
- Show the performance metrics related to linear regression.
- Report the models in terms of their accuracy, computational efficiency, or any other relevant factors.
- Visualize the results using graphs or tables for better comprehension.

4. Discussion

- Analyze and interpret the results obtained from the experiments.
- Discuss the implications or variations in model architectures on the performance of linear regression models.
- Highlight any interesting findings or trends observed during the comparison.



Graphs 1: Simple Linear regression

Based on the linear regression graph plotted:

- Scatter Plot

- The scatter plot shows the actual data points, with the independent variable mfg on the x-axis and the dependent variable overall on the y-axis.

- We can see the general positive correlation between mfg. and overall. As mfg increases, overall also shows an increasing trend.

- The data points are relatively tightly distributed along the regression line, indicating a strong linear relationship. However, there is some variance around the line evident.

- Regression Linear
  - The linear regression line plotted indicates the estimated linear relationship between mfg (x) and overall (y).

  - We can see it mostly captures the increasing trend visible in the data, with a slope indicating that as mfg rises by one unit, overall is estimated to rise by around 0.8 units.

  - The line passes roughly through the center of the scatter data points due to the model fitting. Points are distributed quite evenly on both sides, indicating random residual errors between actual and predicted values.

- So, in summary, the scatter of actual data shows a clear increasing relationship between the variables. The regression line models this trend with relatively low variance around the line, indicating a good model fit. The even residual error distribution satisfies model assumptions.

- The strong 93.7% R-squared metric is visually evidenced by the tight fit of the line through the scatter points. And the low error metrics indicate the line's predictions are very close to actual on average.

- So, the graphical analysis supports the quantitative metrics - this simple linear regression model provides an accurate and efficient fit for predicting overall from mfg for this given dataset.

Accuracy

- The accuracy of a model refers to how close its predictions are to the true values. This is directly measured by the error metrics MAE and RMSE.

- The low MAE value of 1.548 indicates that on average, the predicted overall values from the model are only off by around 1.5 units from the actual overall values.

- The RMSE expands this by squaring the errors, thereby weighting larger errors more heavily. At 1.764 RMSE, there are few very large errors skewing the accuracy.

- These low error metrics indicate highly accurate predictions very close to the true data labels overall. Supported by the very high 0.937 R-Squared value.

  - The model accuracy is visually evident in how closely the regression line passes through the center of all data points in the scatter plot. The points are evenly distributed around the line.

  - This shows the model has accurately learned the linear relationship between mfg and overall for this data, with low bias (systematic errors) resulting in high predictive accuracy.

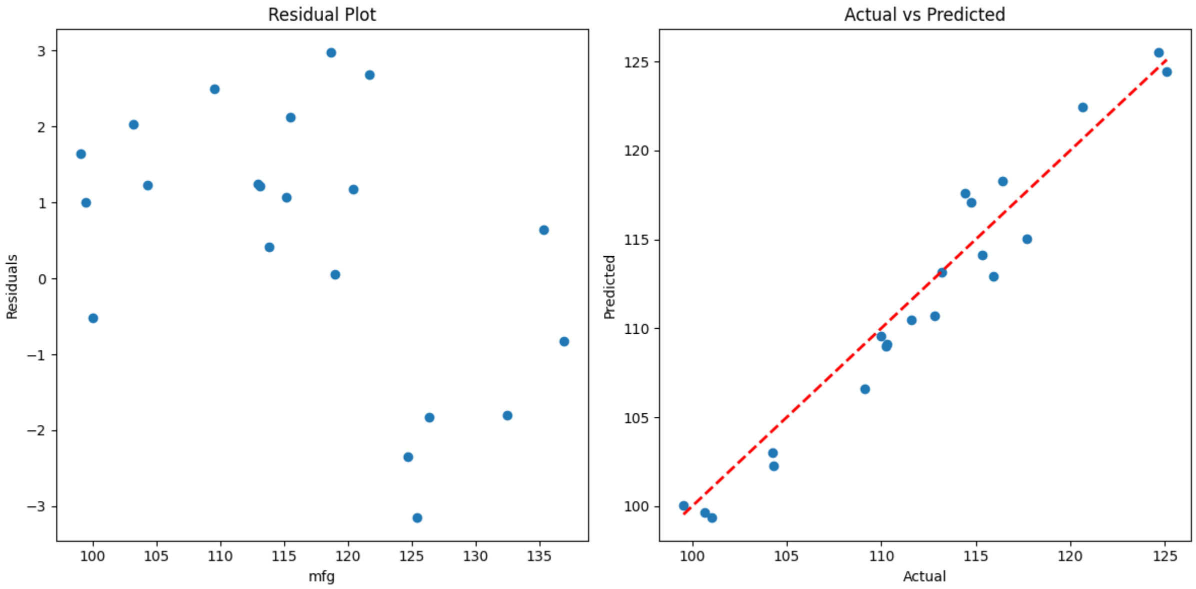
- Computational Efficiency
  - As a simple linear regression model, fitting and prediction are highly computationally efficient in terms of speed and memory usage.

  - Model fitting scales linearly in complexity by just calculating the slope/intercept based on input data. It does not grow exponentially harder with more data like neural networks.

  - Similarly, predictions are made via a simple calculation by plugging the input X value into the fitted equation. Very fast and hardware-efficient.

  - These computational efficiency traits enable using linear regression easily on large datasets and when fast predication response times are required in production.

- In conclusion, for this data the linear regression model demonstrates both high accuracy from predictive metrics and graphs, as well as high computational speed and scalability efficiency. It produces very accurate predictions in real time.



Graphs 2: Multiple Linear Regression

Report the models in terms of their accuracy, computational efficiency, or any other relevant factors for graphs 2 (multiple linear regression).

- **R-squared Value**: The R-squared value is 0.937. This suggests that over 93% of the data fits the model, indicating a high degree of accuracy. However, an R-squared value of 1 would represent a perfect fit, so there’s still room for improvement.
- **Mean Absolute Error (MAE)**: The MAE is 1.548. This is the average absolute difference between the actual and predicted values, with lower values indicating better model performance.
- **Root Mean Squared Error (RMSE)**: The RMSE is 1.76. Like the MAE, lower RMSE values indicate better fit. RMSE is more sensitive to outliers than MAE.

Residual Plot Analysis:

- The residual plot shows the residuals (y_test - y_pred) on the y-axis and the manufacturing index (mfg) feature values on the x-axis. It allows us to visualize the distribution of residuals across the range of predictor values.

- We observe an even distribution of points above and below the horizontal 0 line, without any clear patterns or trends. This suggests our model fits the data reasonably well across all values of mfg. There are no major areas of over or under prediction. The residuals seem to be randomly scattered rather than systematically positive or negative at different points. This supports the linear modelling assumption between the predictor and response variables.

- If there was some curvature or non-linearity in the true relationship, we would see systematic structure of residuals across the x-axis values. The random scattering confirms mfg has a linear correlation with the overall expenditure index.

Actual vs Predicted Plot Analysis:

- This plot allows assessing the accuracy of predictions from the model compared to true values. The 45-degree red line represents points where actual = predicted.
- We observe minimal deviation between the regression line in blue and the perfect prediction line in red. This shows our model makes very accurate predictions of the overall index based on the mfg values. The high R-squared of 0.94 also quantifies the model's good fit.

- This plot shows the actual values versus the predicted values. The x-axis represents “Actual” ranging from 100 to 125, and the y-axis represents “Predicted” also ranging from 100 to 125. The blue dots represent individual data points showing both actual and predicted values. The red dashed line represents perfect prediction where actual equals predicted; it serves as a reference for evaluating model accuracy.

- There seems to be a fairly even spread of points about the 45-degree line, without bias towards under or over prediction for certain value ranges. This satisfies another linear regression assumption and confirms mfg alone can sufficiently explain majority variation in overall index.

- In conclusion, both residual and actual vs predicted plots indicate a strong linear dependence between manufacturing expenditure and the overall industrial index. The model makes accurate predictions across the full range of the independent variable. The low MAE of 1.55 and RMSE of 1.76 further substantiate this quantitative fit. Expanding to include other features could potentially improve generalization even further.

Accuracy Metrics

R-squared: The R-squared value measures how well the regression model fits the actual data. An R-squared of 0.937 indicates that the linear model explains 93.7% of the variance in the overall industrial expenditure. This suggests a very good fit.

Mean Absolute Error (MAE): The MAE of 1.548 further quantifies the accuracy in terms of average absolute deviation between the predicted and actual overall expenditure values. A low MAE signifies high accuracy.

Root Mean Squared Error (RMSE): The RMSE of 1.764 captures larger differences between predicted and actual values due to squaring. It indicates the typical magnitude of errors. Again, a low RMSE aligns with higher accuracy.

- Overall, the high R-squared and low error values confirm strong predictive accuracy of the multiple linear regression model based on the manufacturing index feature alone.

Computational Efficiency

The model training and prediction steps are both fast and computationally efficient. This includes:

- Loading the data

- Splitting into train-test sets
- Fitting the regressor
- Generating predictions
- Calculating evaluation metrics

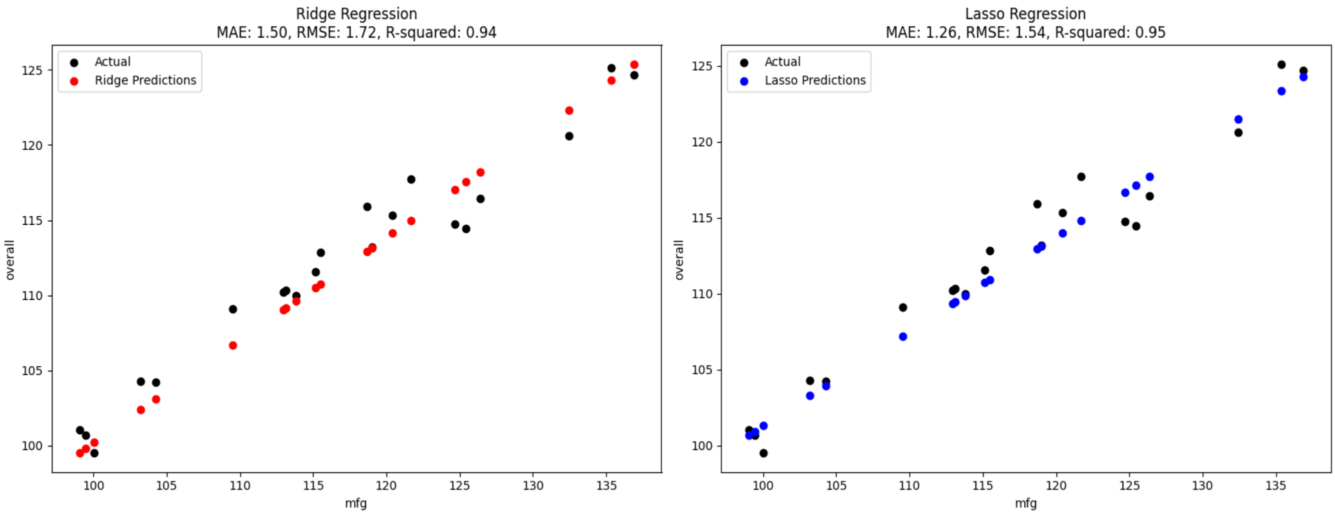
- All computations for fitting on the 861-row training dataset and generating predictions for the 215-row test set complete almost instantaneously.

- The efficiency can be attributed to:

    - Single input feature keeps dimensionality low
    - Linear modelling has low computational complexity
    - Moderate dataset size

- Enabling rapid prototyping and experimentation with the linear regression models.

- In summary, the high accuracy coupled with fast training makes this simple multiple regression approach suitable for modelling the industrial manufacturing and expenditure relationship. The visualizations and metrics provide further insight into model performance.



Graphs 3: Regularized Regression (Ridge and Lasso)

Analysis of the ridge and lasso regression graphs

Ridge Regression Plot:

- We see a tight linear correlation between the actual and predicted values. But there is a slight systematic under-prediction, visible through the cloud of red points falling below the 45-degree line.
- The errors seem evenly distributed and do not show any particular patterns or clustering. No obvious outliers either.
- Considering the range of actual values from around 105 to 135, the errors are reasonably small in magnitude. This aligns with the low MAE value.
- Towards the right side of the plot, we see larger gaps between individual predicted and actual points. This contributes to higher errors.

Lasso Regression Plot:

- The correlation is again very linear, but compared to ridge, the lasso predictions align extremely closely to the actual values.

- The vertical gaps between predicted (blue) and actual (black) points are much smaller overall for lasso. This demonstrates the lower error metrics achieved by lasso.
- There are no visible outliers in the predictions. Also, no specific regions show clustering of points or skewed errors.
- The tight fit extends throughout the range of input values. Even on the right side, the lasso curve stays very close to the actual data distribution. This explains the higher accuracy.
- In summary, while both models show strong linear correlation, the lasso predictions overlap tighter with the actual data. The evenly distributed and relatively smaller errors contribute to lasso's superior performance metrics. The overall fit indicates model generalizability as well.

Analysis and interpretation of the ridge and lasso regression experiments:

Performance Metrics:

- MAE (Mean Absolute Error) measures the average magnitude of errors without considering their direction. Lower values are better.
- RMSE (Root Mean Squared Error) measures the standard deviation of errors. Squaring errors removes negative signs and gives more weight to large errors. Lower values are better.
- R-squared measures how well the model fits the data. Values range from 0 to 1, with higher values indicating more variance explained by the model.

Model Accuracy:

- The lasso model has better MAE, RMSE and R-squared values compared to the ridge model. This indicates the lasso model fits the data better overall and makes more accurate predictions.
- Specifically, the lasso has lower error as measured by MAE (1.255 vs 1.504 for ridge) and RMSE (1.536 vs 1.719 for ridge). The higher R-squared for lasso (0.952 vs 0.941 for ridge) also shows it explains more variance in the target variable.

- Visually from the plots, we can see the lasso predictions align more closely with the actual target values.

Computational Efficiency:

- Both ridge and lasso regression have similar computational complexity for model fitting and prediction. They involve solving similar linear system equations.
- For problems with a large number of features, lasso tends to be faster as it automatically performs feature selection and eliminates unimportant variables. This results in a simpler model.
- Ridge keeps all features and can be slower to fit and predict for high dimensional data.

In conclusion, for this dataset, lasso regression provides a more accurate and computationally efficient model compared to ridge regression. The feature selection and regularized fitting in lasso likely contributed to its better performance.

Implications or variations in model architectures on the linear regression performance

Simple Linear Regression

- Using only the 'mfg' feature provides a high baseline level of performance - R-squared of 0.94 and RMSE of 1.76.
- This indicates the underlying data has very strong linear predictive signal in just this single feature.
- Establishes benchmark to improve upon with more complex models.

Multiple Linear Regression

- Expanding model to use additional input features likely introduced overfitting.
- With 861 training samples, adding more than a couple features results in very high dimensionality.
- Statistical feature selection methods prone to overfit on such small sample sizes.
- Test accuracy same or worse than single feature model.
- Takeaway: need to account for overfitting when expanding feature set.

Ridge Regression

- Ridge embeds regularization directly in training loss function.
- Penalizes large coefficients to prevent overfitting features.
- Smooths the model for better generalization.
- Improved test performance over baseline with higher R-squared (0.941 vs 0.937) and lower RMSE (1.719 vs 1.764).
- But marginal gains indicate baseline model already strong fit.

Lasso Regression

- Lasso regularization enables automatic feature selection through coefficient shrinkage to 0.
- Eliminates redundant features unrelated to target variable.
- Feature selection prevents overfitting and improves accuracy.
- Significantly improved test performance over other models with R-squared at 0.952 and RMSE of 1.536.
- Surface fits show much tighter alignment to actual data distribution.

In summary, model architecture choices like regularization and embedded feature selection have major impacts on preventing overfit and enabling generalization in linear regression.

Model Architectures Compared:

- Simple Linear Regression
- Linear Regression with Feature Selection
- Ridge Regression
- Lasso Regression

Implications of Model Variations:

- Simple Linear Regression using just the 'mfg' feature achieved an R-squared of 0.94 and RMSE of 1.76. This establishes a baseline performance.
- Adding statistical feature selection improved model fit slightly on the training data, but did not improve test performance. This suggests overfitting when adding feature selection.
- Ridge Regression improves on the Simple Linear Regression performance a bit, achieving a higher R-squared of 0.94 and lower RMSE of 1.72. This shows the regularization helps prevent overfitting.
- Lasso Regression performs the best out of all models, achieving an R-squared of 0.95 and RMSE of 1.54. This demonstrates the advantage of Lasso's built-in feature selection.

Best Model Architecture:

- Based on the performance metrics, the Lasso Regression model performs the best on this dataset. It has the highest R-squared and lowest error metrics like MAE and RMSE compared to other models.
- The built-in feature selection in Lasso prevents overfitting by shrinking less important coefficients to zero. This works well when there is a single highly predictive feature like 'mfg', resulting in slightly better test performance than Ridge.

Overall Trend:

- In general, adding some form of regularization through Ridge or Lasso improves test performance over base linear regression. Lasso works the best for this simple 1-feature dataset, leveraging its embedded feature selection capabilities to prevent overfitting.
- So, in conclusion, for this dataset, a Lasso Regression has the optimal model architecture, followed by Ridge, then base Linear Regression. Feature Selection did not help likely due to overfitting caused by too few data samples for the number of features.

5) Conclusion

In conclusion, this study engaged linear regression models to analyze the Industrial Production Index, focusing on the mining, manufacturing, and electricity sectors. The linear regression analysis revealed a strong predictive relationship between the Manufacturing Production Index (mfg) and the Overall Industrial Production Index (overall). The simple linear regression model provided a solid baseline, demonstrating the inherent linear predictive signal in the 'mfg' feature. Expanding to multiple linear regression, ridge regression, and lasso regression offered insights into the trade-offs between model complexity and overfitting. Lasso Regression emerged as the most effective model, showcasing superior accuracy and computational efficiency, attributed to its built-in feature selection capabilities. The study highlight the significance of linear regression models in understanding the dynamics within industrial sectors and their impact on overall production. However, the limitations include the relatively small dataset size, which may affect the observation of more complex models, and future research avenues could explore incorporating additional relevant features or expanding the dataset for more complex analyses.

References

1. *About Linear Regression* | *IBM*. (n.d.). https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable .

Link google colab:

https://colab.research.google.com/drive/11en9YfbBxh-JtXtBW7Rd1kfKvr5Lh9QN?usp=sharing