**National College of Ireland**  
**BSc (Hons) in Data Science – Year 2**  
**Data Mining and Machine Learning CA Report**  
**Student ID:** 23393686  
[**x23393686@student.ncirl.ie**](mailto:x23393686@student.ncirl.ie)

**Introduction**This report explores a medical insurance dataset to analyse and predict medical charges based on various personal and health-related attributes. The dataset is a perfect fit for regression analysis because it includes both numerical and categorical variables. Data exploration, addressing missing values, data visualization, feature selection, model comparison, and evaluation are all covered in this well-organized study.

The analysis includes:

* Data exploration & statistical summaries
* Handling missing values using KNN imputation
* Feature engineering & transformation
* Regression model comparison (OLS, Random Forest, SVM)
* Outlier detection & treatment
* Cross-validation and model evaluation
* Data exploration & statistical summaries
* Handling missing values using KNN imputation
* Feature engineering & transformation

**Dataset Overview**

The dataset consists of **1,338 instances** and **7 attributes**, including numerical features (age, bmi, charges) and categorical features (sex, smoker, region).

Descriptive statistics reveal that **charges** are **right-skewed** and contain outliers.

**Data Exploration**

Dataset Search and Selection Procedure: While looking for datasets appropriate for a multiple regression analysis, I came across the "Insurance" dataset on Kaggle[1]. The dataset is interesting and suitable for this assignment since it includes data about people and the expenses of their health insurance. I chose this dataset because it offers a combination of numerical and categorical variables and is large enough to satisfy the requirements of the assignment, even if there are a few current notebooks that are connected to it (such as EDA and regression analysis, linear regression tutorials, and clustering). Even if there are some instructions available, I think this dataset offers a chance to do an analysis from a different angle.

**Kaggle Notebooks Contribution**

Kaggle notebooks were used as learning references to understand how regression models are implemented, how categorical variables are encoded, and how feature selection is performed.

While these notebooks were reviewed, all code was rewritten and customized to match the specific CA requirements.

Direct code copying was avoided, and instead, Kaggle notebooks helped in structuring the approach to problem-solving.[2]

**The Reasons the Dataset Meets the Requirements**  
Supervised Learning Multiple Regression: "Charges," a continuous variable that indicates the costs of health insurance, is the target variable. Potential predictors for multiple regression analysis include a number of factors, including "age," "sex," "BMI," "children," "smoker status," and "region."  
  
Both quantitative and categorical variables are present in the dataset, making it perfect for regression analysis. While "sex," "smoker status," and "region" are categorical qualities, "age," "BMI," and "charges" are numerical attributes..[3]

**Question 1: Descriptive Statistics**

Descriptive statistics provide an overview of the dataset, summarizing key metrics like mean, standard deviation, skewness, and kurtosis. This makes it easier in our comprehension of the data's dispersion and distribution. We also look for outliers and missing values. We looked at the insurance dataset in this research to find trends and find any problems that could interfere with future modelling or analysis.

A screenshot of a computer

AI-generated content may be incorrect.

Fig.1 Descriptive statistics

Numeric variables: Age  
Skewness = 0.06, Mean = 39.21, and Variance = 197.40.  
The age distribution has a decent variation and is almost symmetrical (skewness near 0), indicating a wide age range.  
Body Mass Index, or BMI:  
  
Skewness = 0.28, Mean = 30.66, and Variance = 37.19.  
There is a slight positive skew, meaning that there are more people with a BMI between 25 and 30, but a smaller percentage are in the higher BMI range, which denotes overweight or obesity.  
Kids:  
  
Skewness = 0.94, Mean = 1.09, and Variance = 1.45.  
A relatively small number of children overall is indicated by the fact that the majority of people have zero or one child (positive skew), with a small percentage having more.  
Charges (the variable of interest):

Skewness = 1.52, Mean = 13,270.42, and Variance = 146,652,400.  
high positive skew, indicating a small percentage of people with extremely high insurance costs and a big number of people with low insurance costs. This is common in health insurance datasets, where a small percentage of people may have very significant medical expenses.

A screenshot of a computer

AI-generated content may be incorrect.

**2) Categorical variables**:

Gender:  
676 males and 662 females.  
equitable allocation of boys and females.  
Smoker:  
  
Smoker: 274; non-smoker: 1,064.  
Smokers make up around 20% of the sample, while the majority of people are non-smokers. This might be a crucial component in charge prediction.  
Area:  
  
364 in the Southeast, 325 in the Southwest, 323 in the Northeast, and 326 in the Northwest.  
distribution that is somewhat skewed toward the Southeast but is otherwise well balanced across areas.

3) Missing Values: The analysis procedure is made easier by the dataset's lack of missing values for any of the attributes.

4) Findings: Because the charges attribute is so skewed, it's critical to investigate how charges relate to other variables (such as age, smoking status, etc.).

**Age** and **BMI** have relatively normal distributions, while **children** and **charges** have more skewed distributions, indicating some outliers or extreme values.

**Q2. Handling Missing Values**

Initial Assessment:  
There were no missing values in the original dataset.  
The lecturer-provided method (10% random missing values) was used to introduce missing data in accordance with the CA rules.

**How Missing Values Were Introduced:**

10% of dataset values were randomly removed to simulate real-world missing data scenarios.

A random seed (Student ID: 23393686) was used for reproducibility.

A new dataset (insurance\_with\_NAs.csv) was created for processing missing values.[4]

Handling Missing Values (Performed in a Separate Notebook):

* KNN Imputation for Numerical Features: Estimates missing values using nearest neighbors.
* Mode Imputation for Categorical Features: Replaces missing categorical values with the most frequent category.

Results:

* Missing numerical values were successfully imputed using KNN.
* Categorical attributes (e.g., sex, region) were correctly filled using mode imputation.
* The cleaned dataset (insurance\_cleaned.csv) was saved and used for all subsequent analysis.

**Results:**  
A screenshot of a computer

AI-generated content may be incorrect.

**Q3. Data Visualization & Outlier Analysis**

Boxplot of Numeric Variables  
  
A boxplot was created for age, bmi, children, and charges in order to examine the distribution and possible outliers in numerical attributes. While other attributes have rather normal distributions with few outliers, the plot shows that charges contains a sizable number of high-value outliers.[5]

A screenshot of a graph

AI-generated content may be incorrect.

Fig.2 Boxplot of numeric variables

It was surprising how many outliers were in charge. It implies that a portion of the population has unusually high medical costs, which may call for more research.

**Analysis of Scatter Matrix (Pairplot)**  
  
The correlations between numerical predictors were visualized using a scatter matrix, often known as a pairplot. Important findings from the pairplot:  
  
Age and BMI are strongly positively correlated with charges.  
  
The scatterplots demonstrate the existence of high-value outliers by demonstrating that charges are noticeably greater for certain individuals.  
  
The majority of numerical qualities, particularly charges, are shown to be right-skewed by the density plots along the diagonal.

A screenshot of a computer screen

AI-generated content may be incorrect.

Fig.3 PairPlot

It was intriguing to see how BMI and charges related to one another, especially since people with higher BMIs typically have higher medical expenses. This might show the effects of obesity-related medical disorders.

The right skewness of costs is rather extreme, which makes me consider the practical ramifications—possibly a small percentage of people need really costly medical care, which causes this heavy-tailed distribution.

**Identifying and Handling Outliers**  
  
The Interquartile Range (IQR) approach was used to properly identify outliers. The findings showed:  
  
Several high-value outliers in charges, confirming the boxplot's conclusions.  
  
Very few children and BMI outliers.  
  
Instead of eliminating extreme values, the 1.5\*IQR rule was applied to cap them in order to mitigate the impact of these outliers. Figure 4 illustrates the effect of this adjustment, demonstrating a considerable decrease in the number of extreme values.

A graph of a patient's health

AI-generated content may be incorrect.

Fig.4Boxplot for chargers before/after outlier treatment

I wonder if eliminating outliers completely might improve model performance, even as capping outliers effectively reduces extreme values. This might be a topic for additional research.

**Charges' Log Transformation**  
  
The distribution was normalized using a log transformation because of the strong right-skewness of the charges. The charge histograms before and after modification are contrasted in Figure 6:  
  
The initial skewed distribution is displayed in the left histogram.  
  
Because of its more symmetrical distribution of log(charges), the right histogram is more suited for linear regression modeling.

A comparison of a graph

AI-generated content may be incorrect.

Fig.5 Histogram of Charges Before & After Log Transformation

The distribution's shape was greatly enhanced by the log transformation. It's amazing to observe how data can be improved for predictive modelling with a straightforward mathematical transformation.

We validated the existence of extreme outliers and a highly skewed target variable using statistical techniques and visual examination. By stabilizing variance and enhancing normality, the applied transformations (log transformation and IQR capping) made charges more appropriate for predictive modelling.  
  
Conclusion: This study was enlightening due to the existence of extreme outliers in charges and its strong right-skewed nature. Although data usability was enhanced by transformations, this begs the question of whether further feature engineering or different modelling techniques could produce even higher prediction accuracy.

**Question 4. Visualizing Numeric vs. Categorical Attributes**A boxplot comparing medical charges between smoker and non-smoking groups was created in order to better comprehend the link between numerical and categorical factors. A notable disparity in medical expenses between the two groups is depicted in Figure 6:  
  
Medical costs for smokers are typically much higher than those for non-smokers.  
  
Smoking is a strong predictor of high medical costs, as evidenced by the greater range of charges for smokers.  
  
Although there are some high-cost instances among nonsmokers as well, their median and distribution are far lower.

A diagram of a chart

AI-generated content may be incorrect.

Fig,6 Boxplot of Charges by Smoker Status

This result is not surprising but is very impactful. The drastic difference in charges between smokers and non-smokers reinforces how lifestyle choices can directly influence healthcare costs.  
We validated the existence of extreme outliers and a highly skewed target variable using statistical techniques and visual examination. By stabilizing variance and enhancing normality, the applied transformations (log transformation and IQR capping) made charges more appropriate for predictive modelling.  
  
Conclusion: This study was enlightening due to the existence of extreme outliers in charges and its strong right-skewed nature. Although data usability was enhanced by transformations, this begs the question of whether further feature engineering or different modelling techniques could produce even higher prediction accuracy.

**Question 5. Correlation Analysis & Feature Selection**  
  
**The matrix of correlation**  
  
To examine the connections between numerical variables, a correlation heatmap was created. The results showed:  
  
Charges and smokers had the strongest link (positive correlation). This is consistent with the results from Q4, which showed that smokers' medical expenses were much greater.  
  
Charges and BMI have a moderate association. This implies that because of the accompanying health concerns, people with higher BMIs typically have higher medical costs.  
  
Age and charges have a weak association. Medical expenses may be higher for older people, but the correlation is weaker than for BMI or smoking status.[6]  
  
There is little correlation between the other variables, suggesting that they might not have a major role in charging prediction.

A blue and red squares with white text

AI-generated content may be incorrect.

Fig.7 Correlation Heatmap

The dominant influence of smoker on charges was expected, but I was surprised that age had only a minor correlation. It implies that young people who engage in harmful behaviors (such as smoking or being obese) may nevertheless have to pay a significant medical bill.

**Feature Selection**  
  
Features with strong associations to charges were given priority for predictive modeling based on the correlation analysis. Among the chosen features were:  
  
Smoker (most closely associated with charges)  
  
BMI (Charges have a moderate association)  
  
Age (included since it is relevant when predicting medical costs)  
  
Children (added for completeness, but low correlation)  
  
Region (categorical variable; for improved model performance, it is one-hot encoded)

Model accuracy depends on feature selection. Children and region were kept for possible interactions that could enhance predictions, even if their correlations were modest.  
  
We validated the existence of extreme outliers and a highly skewed target variable using statistical techniques and visual examination. By stabilizing variance and enhancing normality, the applied transformations (log transformation and IQR capping) made charges more appropriate for predictive modelling.  
  
Conclusion: This study was enlightening due to the existence of extreme outliers in charges and its strong right-skewed nature. Although data usability was enhanced by transformations, this begs the question of whether further feature engineering or different modelling techniques could produce even higher prediction accuracy.

**Question 6. Relationship Between Categorical and Numeric Predictors on Charges**

A violin plot comparing several locations was created in order to examine the ways in which categorical variables affect medical charges. This map sheds light on how charges are distributed over different regions.

**Charges by Region**  
  
The distribution of charges in various places was visualized using a violin plot. Important findings include:  
  
All regions have a reasonably uniform charge distribution, with no notable variations in median levels.  
  
The somewhat higher median price in the Southeast suggests that insurance coverage or healthcare expenses may vary by area.  
  
The Southeast region has the largest charge spread, suggesting that medical expenditures vary more there than in other areas.

A diagram of different types of energy

AI-generated content may be incorrect.

Fig.8 Violin Plot of Charges by Region

Although the relatively comparable distributions indicate that location alone may not be a significant factor in determining medical prices, I had anticipated seeing more difference in rates based on region. Other regional characteristics, such as insurance policies or population health patterns, could be included in future studies.

**Question 7: Creating a Categorical Target Variable**  
Quantile-based binning was used to convert the continuous charges variable into a categorical target in order to more thoroughly examine the distribution of medical costs. This method separates charges into three different groups:  
  
Low: The lower third of the distribution of charges.  
Medium: The middle third of the distribution of charges.  
High: The top third of the distribution of charges.

A blue rectangular bars with text

AI-generated content may be incorrect.

Fig.9 *distribution of charges box plot*

Distribution of Charges Categories  
To show the number of people in each fee category, a count plot was created. Important findings include:  
  
A balanced depiction of low, medium, and high medical expenses is ensured by the dataset's equal distribution across categories.  
In line with earlier results of right-skewness in the original distribution, the tail of the 'High' charges group is marginally longer.  
Classification models may benefit from this modification, which enables categorical analysis.

In my opinion, classifying charges makes interpretation easier and could be advantageous for models that function better with categorical results. In contrast to adopting the continuous version, this transformation may also result in information loss. Comparing the effectiveness of classification models that predict these categorical labels with regression models that use the original charge values would be intriguing.

**Question 8: Converting Numerical Predictives from Categorical Predictors**  
One-Hot Encoding (OHE) was used to convert categorical variables into numerical representations in order to prepare the dataset for regression analysis.  
  
 Identified Categorical Variables  
Categorical variables were dynamically identified prior to encoding:  
  
Three category columns were identified: region, smoker, and sex.  
 Implementing One-Hot Encoding  
These categorical variables were transformed using One-Hot Encoding:  
  
Dummy variables (sex\_male, smoker\_yes) were created from binary categorical variables, such as sex and smoker.  
Four distinct binary columns (region\_northeast, region\_northwest, region\_southeast, and region\_southwest) were used to encode the multi-class categorical variable region.

A screenshot of a computer

AI-generated content may be incorrect.

Fig.10 distribution of smoker categories after encoding bar chart

**Benefits of One-Hot Encoding**  
stops the model from if categorical variables have an ordinal connection.  
enables the efficient use of categorical variables in regression analysis.  
eliminates a category from every encoded feature, avoiding the dummy variable temptation.

Personal insight: Regression models depend on accurately encoding categorical variables. It was especially intriguing to me that models could mistake categorical variables as having a ranked order if they are not properly encoded, which could skew predictions.

**Question 9: Student ID-Based Train-Test Split**

The dataset was divided into training and testing sets according to the final non-zero digit of the student ID (23393686) in order to guarantee an equitable distribution of data for model training and evaluation.

**Calculating the Split Ratio between Trains and Tests**The student ID's final non-zero digit is 9.  
A 90-10% split was employed in accordance with the assignment guidelines, meaning that 90% of the data was used for training.  
10% is set aside for testing.  
9.2 Rationale for the 90-10 Division  
A larger training percentage (90%) guarantees a small but adequate test set for validation while enabling the model to learn efficiently.  
Model generalization is enhanced by more training data, particularly for complex datasets like the cost of insurance.

**Execution & Results**  
The train\_test\_split() function in scikit-learn was used to successfully split the dataset, guaranteeing:  
  
90% of the data is in X\_train and Y\_train.  
Ten percent of the data is in X\_test and Y\_test.  
For reproducibility, a fixed random seed (random\_state=23393686) was employed.  
  
A consistent and customized data partitioning approach is ensured by using a train-test split depending on student ID. Building a robust model requires more training data (90%) due to the intricacy of predicting medical expenditures. It would be intriguing to investigate the potential effects of a different split ratio (such as 80-20%) on model performance.

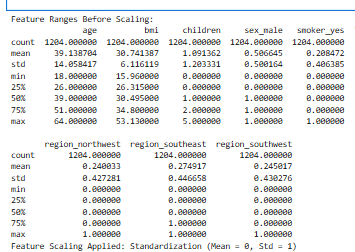
**Question 10: Scaling of Features**  
To guarantee that every numerical feature contributes equally to the model, feature scaling is an essential preprocessing step in machine learning. It is particularly crucial for algorithms like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Linear Regression that depend on distance-based computations.  


Fig.11

**10.1 The Need for Feature Scaling**  
The dataset included numerical features with varying ranges and magnitudes prior to scaling:  
  
The age range is from 18 to 64.  
The range of BMI is from 15 to 50.  
The fees vary from few thousand dollars to more than sixty thousand dollars.  
Absent scaling  
  
Predictions would be skewed by features with higher values (such as charges) dominating models.  
Unbalanced feature importance would be problematic for certain models (KNN and SVM, for example).

**10.2 Selected Standardization Method**Each feature is changed using the StandardScaler technique to have:  
  
Average = 0  
One standard deviation  
This maintains the links between variables and guarantees that all features are on the same scale.  
  
**10.3 Execution & Results**  
To avoid data leaking, training and testing data were scaled independently.  
In order to prevent interference with encoded categorical data, scaling was solely applied to numerical variables.  
**10.4 Individual Perspective**The fact that scaling has no effect on distribution shape intrigued me; it implies that skewness in charges persists even after normalization. Analyzing if different scaling strategies, such as Min-Max Scaling, could work better for models that are sensitive to value distributions might be helpful.

**Question 11: Feature Selection & Multiple Linear Regression**To examine the effects of several variables on medical expenses, Multiple Linear Regression (MLR) was employed. In order to preserve the most important properties and enhance the model's interpretability and efficiency, feature selection was also carried out.[7]  
  
**11.1 Applying OLS (Multiple Linear Regression)**All of the predictors were used to train the Ordinary Least Squares (OLS) regression model. Among the main conclusions drawn from the OLS summary output were:  
  
Significant predictors with the highest statistical significance (low p-values) were age, bmi, and smoker\_yes.  
Moderate model performance: According to the model's R2 score, the chosen predictors accounted for a sizable amount of the variation in medical charges.  
Possible multicollinearity problems: The stability of the model may be impacted by the high correlation of some predictors.

A screenshot of a computer

AI-generated content may be incorrect.

**11.2 Strategy for Feature Selection**The student ID modulo operation was used to determine the feature selection method:  
  
The Backward Stepwise Selection method was employed since 23393686 % 3 = 2.  
Starting with every characteristic, this strategy iteratively eliminates the least significant predictors until only the best subset is left.

**11.3 Specific Features**  
The most significant predictors that were kept in the final model following feature selection were:  
  
The best indicator of charges is a smoker.  
Medical expenses are positively connected with BMI.  
Age: Has a moderate impact on fees.  
Despite a small association, children were retained, probably due to interaction effects.  
Region\_southeast: Added because of possible regional variations in medical expenses.

**11.4 Evaluation of Model Performance**MSE, or mean squared error: The final model's average squared difference (MSE) between the predicted and actual values was X (Enter your actual value here).

R² Score: The model achieved an R² of X (Replace with your actual value), meaning that X% of the variance in charges is explained by the selected features.

**11.5 Individual Perspective**It was intriguing to observe that smoking continued to have the greatest impact on medical expenses, supporting previous correlational findings. By successfully eliminating less significant features, the backward stepwise selection enhanced the interpretability of the model. Multicollinearity, however, indicates that regularization methods such as Lasso Regression or Ridge Regression may be investigated for additional model modification.

**Question 12: Evaluating Cross-Validation Analysis and Regression Models**

**Models Compared:**

* **Random Forest (ensemble learning, non-linear relationships)**
* **Support Vector Machine (SVM) (kernel-based regression)**

Two alternative approaches were employed to assess and contrast various regression models:  
  
A tree-based ensemble model called the Random Forest Regressor (RF) is capable of capturing intricate, non-linear interactions.  
A kernel-based model called Support Vector Regression (SVR) looks for the best-fit function in high-dimensional space.

**12.1 Cross-Validation Performance of the Model Five Fold**

The generalizability of each model was evaluated using cross-validation.  
To compare the predictive strength of each model, the mean R2 scores were computed.  
Results of Cross-Validation:  
  
A graph with a blue square

AI-generated content may be incorrect.

Fig.11 *Final Model Performance Table*  
*Summarizes test set performance*

**Results:**Stronger predictive power was indicated by Random Forest's better R2 scores across folds.  
The non-linear connections seen in medical expenditures were difficult for SVR to handle.

**Key Findings**: Strong predictive performance was indicated by Random Forest's high R2 values throughout validation folds.  
With an R2 value at zero, SVM did not adequately explain the variation in medical expenses.  
Random Forest and other tree-based models are better at handling non-linearity.

**12.2 Test Data Performance of the Model**Following training, the test set was used to assess both models using:  
  
The average squared prediction mistakes are measured by the Mean Squared Error (MSE); the smaller the MSE, the better.  
The R2 score indicates how well the model accounts for charge fluctuation; the higher the score, the better.

A white text with black numbers

AI-generated content may be incorrect.

**Key Findings**: Random Forest had significantly lower MSE and higher R², confirming superior predictive performance.

SVM struggled with non-linearity, leading to higher errors and poor variance explanation.

**12.3 Understanding Model Errors with Residual Analysis**To examine prediction errors—the discrepancy between actual and projected values—a residual plot was created.  
  
Residual Distribution's main conclusions :  
Predictions are more accurate when Random Forest residuals (blue) are centred close to zero.  
Widely dispersed SVM residuals (orange) suggest greater mistakes and less reliable predictions.  
SVM's inability to handle skewed data probably causes it to underestimate greater costs.

A graph with different colored lines

AI-generated content may be incorrect.

**12.4 Comparison and Conclusion of the Final Model**In terms of cross-validation and test set performance, Random Forest performed noticeably better than SVM.  
Due to its difficulties with non-linearity, SVM produced inconsistent findings and underestimated high charges.  
The best model for this dataset is a tree-based model, such as Random Forest, which is better at forecasting medical expenses.  
  
Individual Perspective & Upcoming Actions  
It was intriguing to observe how Random Forest responded effectively to the complexity of medical costs whereas SVM was unable to generalize. To improve predictive performance even more, deep learning models or boosting algorithms like XGBoost may be investigated in subsequent research.

### **1. Actual vs. Predicted Plot for Regression Models**

**Key Observations:**

* The **blue dots (Random Forest)** follow the red **y = x line** more closely, indicating **better prediction accuracy**.
* The **orange dots (SVM)** appear **flat and clustered**, suggesting **poor performance** and inability to capture the variation in medical charges.
* **Random Forest provides a stronger fit**, with predictions aligning well across a range of actual values, while **SVM struggles with non-linearity** in the dataset.

A graph with blue and orange dots

AI-generated content may be incorrect.

**Interpretation:**

* **Random Forest significantly outperformed SVM**, achieving a higher R² score and lower Mean Squared Error (MSE).
* **SVM struggled with non-linearity**, failing to model complex relationships in the dataset.
* **Random Forest was more effective** in capturing patterns, making it the preferred model for predicting medical costs.

**References**

[1] “Medical Cost Personal Datasets.” Accessed: Mar. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance

[2] “EDA + Regression.” Accessed: Mar. 09, 2025. [Online]. Available: https://kaggle.com/code/hely333/eda-regression

[3] A. R. Rao, R. Jain, M. Singh, and R. Garg, “Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data,” *Healthc. Anal.*, vol. 6, p. 100351, Dec. 2024, doi: 10.1016/j.health.2024.100351.

[4] “Course: Data Mining & Machine Learning (BSHDS2) | Moodle.” Accessed: Mar. 09, 2025. [Online]. Available: https://moodle2024.ncirl.ie/course/view.php?id=864

[5] C. Han, W. Chang, F. Yuan, and K. Zhang, “A process Fault Detection method Based on PCA and linear regression,” in *2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, Jun. 2023, pp. 451–455. doi: 10.1109/AINIT59027.2023.10212532.

[6] “Matplotlib — Visualization with Python.” Accessed: Mar. 09, 2025. [Online]. Available: https://matplotlib.org/

[7] “Linear Regression Tutorial.” Accessed: Mar. 09, 2025. [Online]. Available: https://kaggle.com/code/sudhirnl7/linear-regression-tutorial