PDAN8411 POE Part 1

Allana Morris

ST10204772

Table of Contents

[Dataset Evaluation 2](#_Toc196502956)

[Analysis Planning: 3](#_Toc196502957)

[Objective 3](#_Toc196502958)

[a. Exploratory Data Analysis (EDA) Plan 3](#_Toc196502959)

[b. Feature Selection Plan 4](#_Toc196502960)

[c. Model Training Plan 5](#_Toc196502961)

[d. Model Evaluation Plan 5](#_Toc196502962)

[e. Report Writing Plan 6](#_Toc196502963)

[Structure: 6](#_Toc196502964)

[Analysis: 7](#_Toc196502965)

[EDA – conduct a thorough Exploratory Data Analysis. 7](#_Toc196502966)

[Evaluate your model: 16](#_Toc196502967)

[OLS Regression Results 16](#_Toc196502968)

[Model Performance: 16](#_Toc196502969)

[Coefficient Interpretation: 16](#_Toc196502970)

[Retraining the Model with Different Parameters 22](#_Toc196502971)

[Ridge and Lasso Regression Adjustments: 22](#_Toc196502972)

[Feature Engineering Considerations: 23](#_Toc196502973)

[Model Evaluation and Performance Gains: 23](#_Toc196502974)

[Report 25](#_Toc196502975)

[Introduction 25](#_Toc196502976)

[Data Cleaning and Exploratory Data Analysis (EDA) 25](#_Toc196502977)

[Feature Selection 26](#_Toc196502978)

[Model Training 26](#_Toc196502979)

[Model Evaluation 27](#_Toc196502980)

[Conclusion 28](#_Toc196502981)

## Dataset Evaluation

The client, sourced from Kaggle, provides medical insurance data and is suitable for constructing a linear regression model. The target variable, charges, is a continuous numeric field that represents the medical cost billed to individuals, making it ideal for regression modelling. The independent variables include a combination of numerical features, such as age, bmi, and children, and categorical features, such as sex, smoking, and region. These can be handled using techniques like one-hot encoding to guarantee they are appropriate for regression.

Upon initial inspection, the dataset has 1,338 entries with no missing values or obvious data integrity issues, suggesting high data quality. Furthermore, several variables, particularly age, bmi, and smoker, are logically linked to medical costs and may indicate linear relationships with the target. Preliminary correlation analysis (to be demonstrated in the EDA section) further enhances this assumption.

It is important to ensure that categorical variables are correctly handled, and that multicollinearity is not present among predictors. While the dataset is U.S.-based, it still serves as a suitable starting point for a proof-of-concept model for the South African market. However, extra localisation and retraining would be required prior to any deployment.

## Analysis Planning:

### Objective

1. Explore and clean the dataset.
2. Select optimal features for modelling.
3. Train and evaluate a linear regression model.
4. Interpret results and refine the model.
5. Prepare a client-ready report.

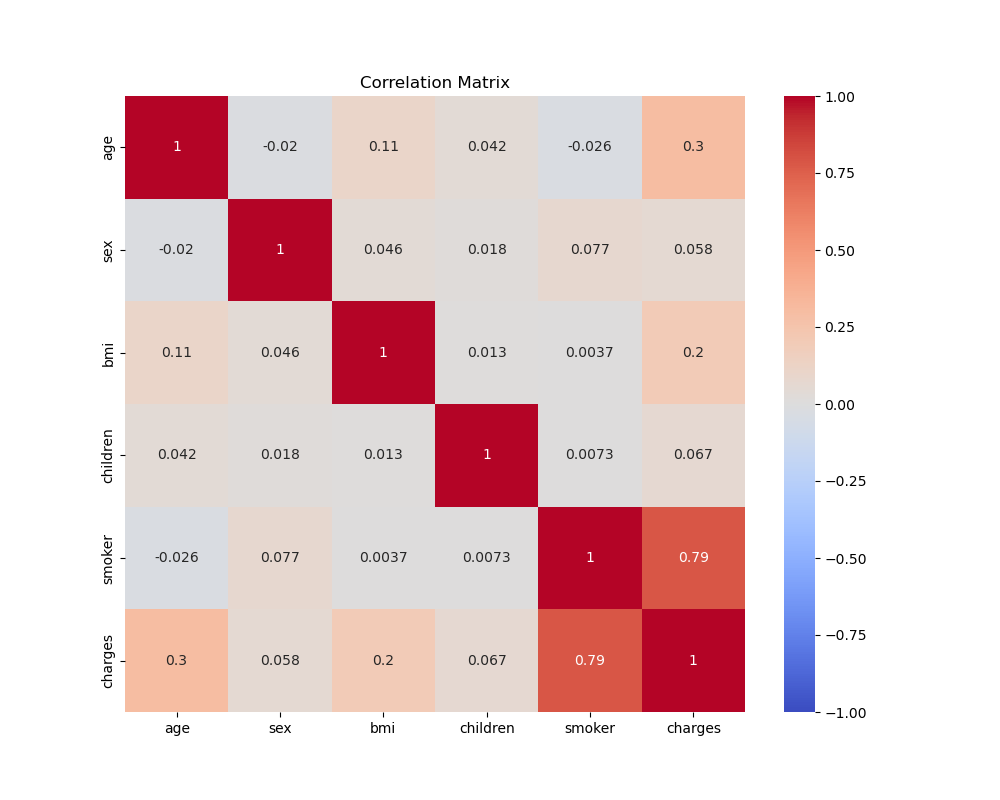
### a. Exploratory Data Analysis (EDA) Plan

#### Goals:

* Understand data structure and quality.
* Identify patterns, outliers, and relationships.

#### Steps:

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Method/Tool** | **Expected Output** | **Reference** |
| **1. Data Overview** | df.info(), df.describe() | Summary of missing values, data types. | (Anon., 2025) |
| **2. Target Analysis** | Histogram/Boxplot of charges | Identify skewness, outliers. |  |
| **3. Feature Analysis** | Boxplots (smoker vs. charges), Scatterplots (age, bmi vs. charges) | Visualize relationships. |  |
| **4. Correlation Check** | Heatmap (df.corr()) | Identify highly correlated features. | (Anon., 2005) |



(Hunter, 2007)

### b. Feature Selection Plan

#### Goals:

* Select features that significantly impact charges.
* Avoid multicollinearity.

#### Methods:

|  |  |  |  |
| --- | --- | --- | --- |
| **Approach** | **Implementation** | **Decision Rule** | **Reference** |
| **1. Correlation** | df.corr()[['charges']] | Keep features with correlation > 0.2 |  |
| **2. P-values** | statsmodels.OLS | Retain features with p < 0.05. | (Wasserstein and Lazar, 2016) |
| **3. Domain Knowledge** | Client input | Prioritize smoker, age, bmi. |  |

### c. Model Training Plan

#### Goals:

* Train a baseline linear regression model.
* Optimize hyperparameters if needed.

#### Steps:

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Method** | **Notes** | **Reference** |
| **1. Train-Test Split** | train\_test\_split() (80/20) | Random state for reproducibility. | (Pedregos et al., 2011) |
| **2. Baseline Model** | LinearRegression() | Default parameters. |  |
| **3. Regularization** | Ridge(alpha=1.0) | If multicollinearity exists.  Also to reduce overfitting | (Hastie, Tibshirani and Friedman, 2009) |

#### Model Equation:

charges=β0+β1(age)+ β2(bmi)+ β3(children) + β4(smoker)

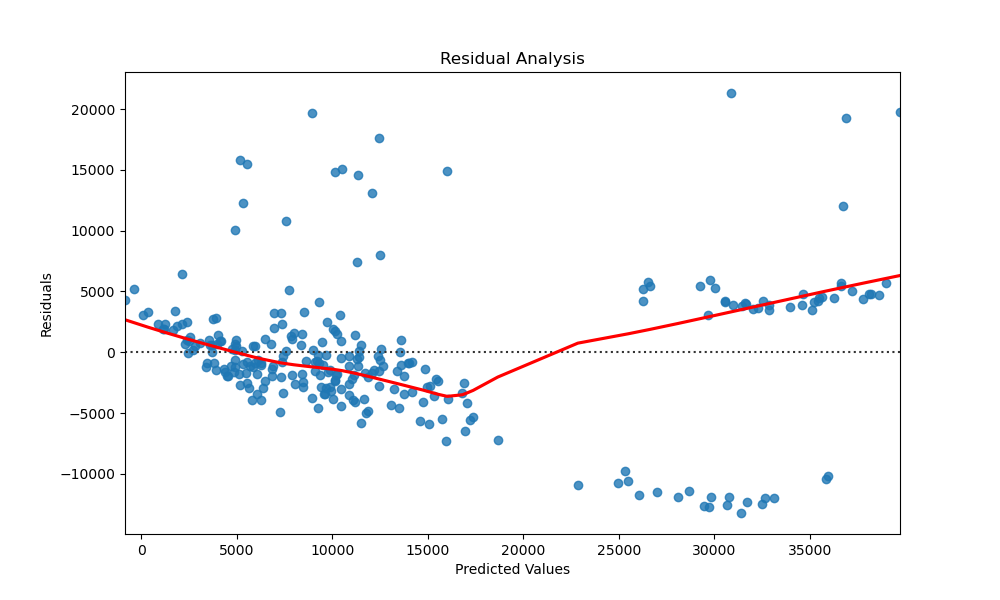
### d. Model Evaluation Plan

#### Goals:

* Quantify model performance.
* Diagnose potential issues.

#### Metrics & Tools:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Formula** | **Interpretation** | **Reference** |
| **R²** | 1 - (SS\_res / SS\_tot) | 75% variance explained (target: >0.7). | (James et al., 2021) |
| **RMSE** | sqrt(mean((y\_true - y\_pred)^2)) | Dollar-value error (e.g., $5,000). | (Chai and Draxler, 2014) |
| **Residuals** | y\_test - y\_pred | Check for patterns (heteroscedasticity). |  |



(Hunter, 2007)

### e. Report Writing Plan

### Structure:

#### Introduction

* + Problem statement and dataset summary.

#### EDA & Cleaning

* + Key findings (e.g., smokers pay 3× more).
  + Visuals: Boxplot (smoker vs. charges), correlation heatmap.

#### Modelling

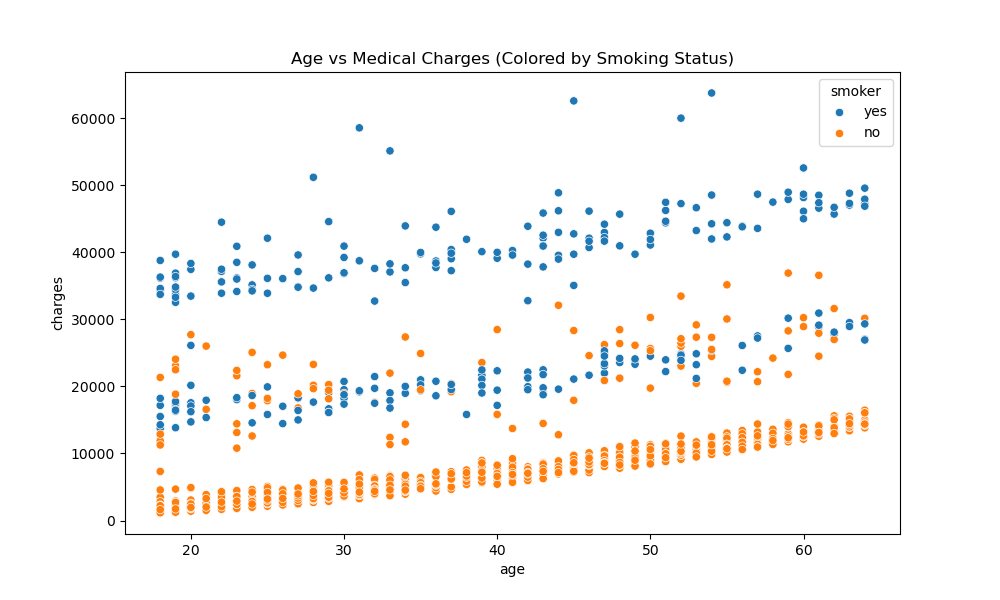
* + Features selected and why.
  + Model equation and performance (R², RMSE).

#### Recommendations

* + Actionable insights (e.g., tiered pricing for smokers).

## Analysis:

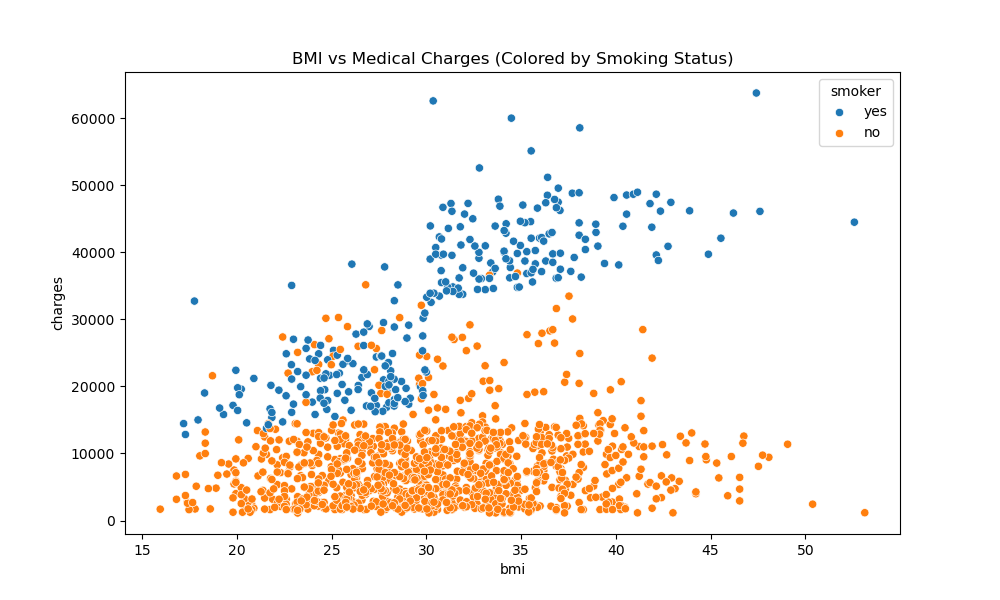
### EDA – conduct a thorough Exploratory Data Analysis.



(Hunter, 2007)

Age vs Charges Relationship

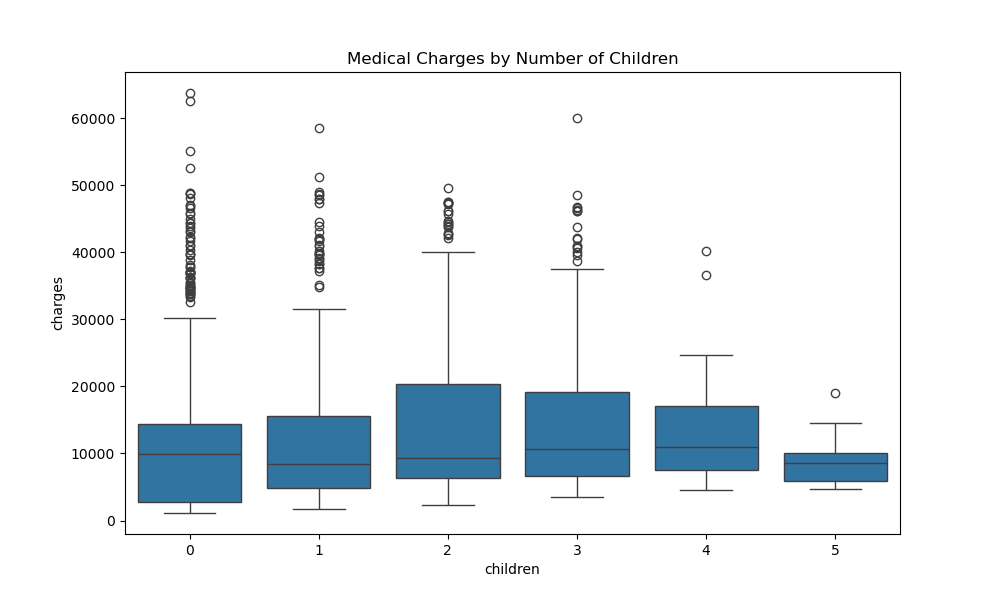
* Clear positive correlation between age and charges
* Smokers (orange) show:
* Steeper slope (age increases impact costs more)
* Higher baseline costs at all ages
* Non-smokers (blue) cluster below $15k regardless of age
* Interaction effect: Age impacts smokers more severely



(Hunter, 2007)

BMI vs Charges Analysis

* Non-smokers (Blue):
  + Minimal correlation between BMI and charges
  + Most charges remain below $15,000 regardless of BMI
* Smokers (Orange):
  + Clear threshold effect at BMI=30 (obesity line)
  + Charges escalate dramatically for obese smokers (BMI >30)
  + Extreme charges (>$50k) concentrated in BMI >35 smokers
* Key Insight: BMI only becomes significant risk factor when combined with smoking



(Hunter, 2007)

Children Count Analysis

|  |  |  |
| --- | --- | --- |
| Children | Median Charge | Pattern |
| 0 | ~$8,500 | Baseline |
| 1 | ~$9,500 | +12% |
| 2 | ~$10,700 | Peak (+26%) |
| 3 | ~$9,900 | Slight dip |
| 4-5 | ~$8,300 | Returns to baseline |

* Non-linear relationship: 2-3 children show highest median charges
* Possible explanations:
  + Parents in 30-40 age range (higher base charges)
  + Family plans vs individual coverage
  + Weakest predictor among numerical features (corr=0.07)

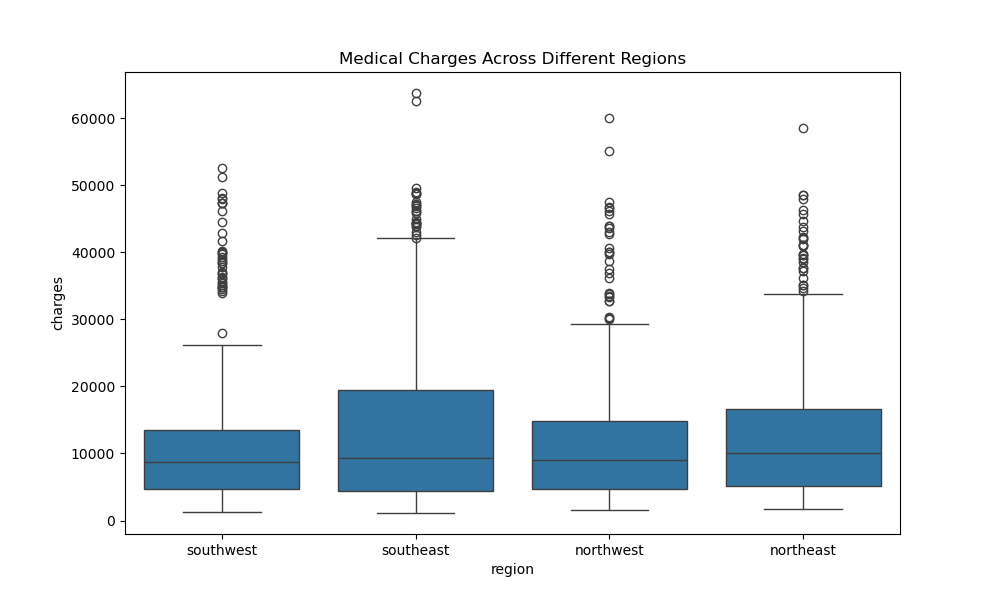
A graph of a distribution of medical charges

AI-generated content may be incorrect.

(Hunter, 2007)

Distribution of Medical Charges

* The distribution is right skewed with most charges below $20,000
* A small number of extreme cases exceed $50,000
* Bimodal pattern suggests two distinct groups (likely smokers vs non-smokers)
* Primary peak around $4000, secondary around $20000

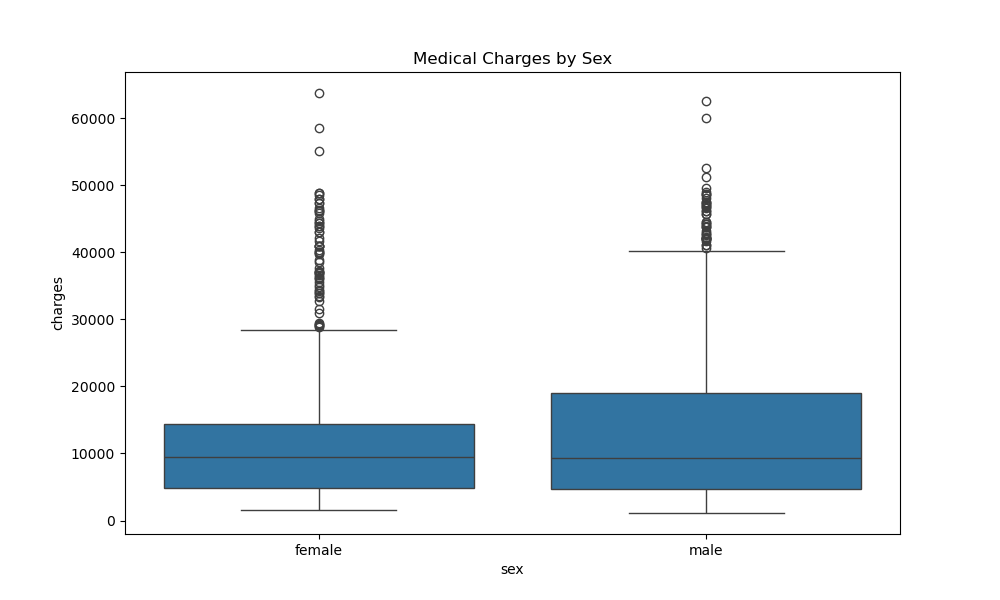


(Hunter, 2007)

Regional Cost Variations

1. Southeast has:
   * Highest median charges (~$9,500)
   * Most extreme outliers (>$50,000)
   * Wider IQR ($6,000-$17,000)
2. Southwest shows:
   * Lowest median charges (~$8,000)
   * Most compact distribution
   * Fewest high-cost outliers
3. Northeast/Northwest:
   * Similar median charges (~$8,500-$9,000)
   * Moderate outlier presence

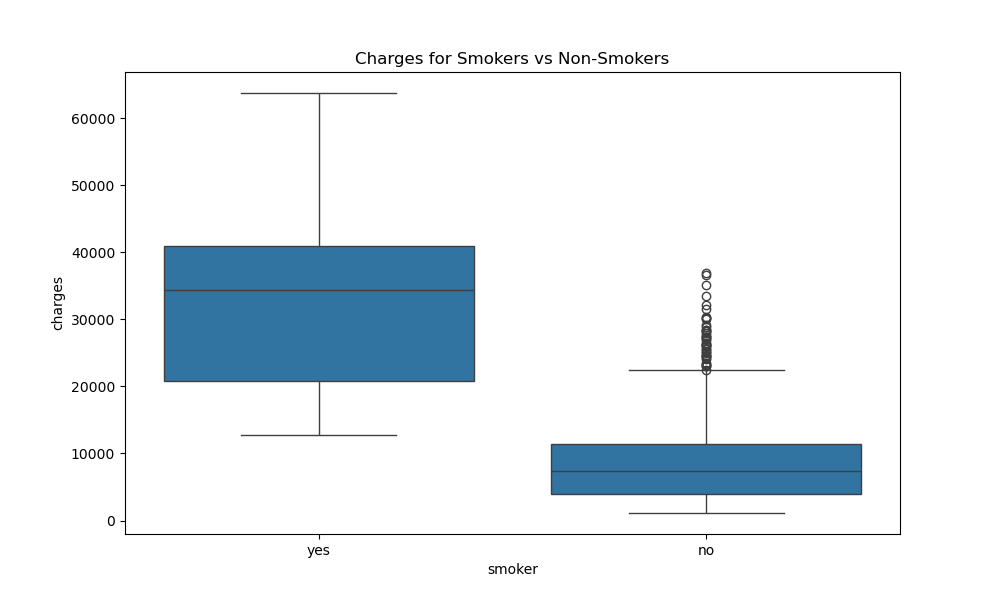
* Possible Explanations:
  + Regional cost-of-living differences
  + Varying healthcare provider networks
  + Population health demographics



(Hunter, 2007)

Gender-Based Cost Analysis

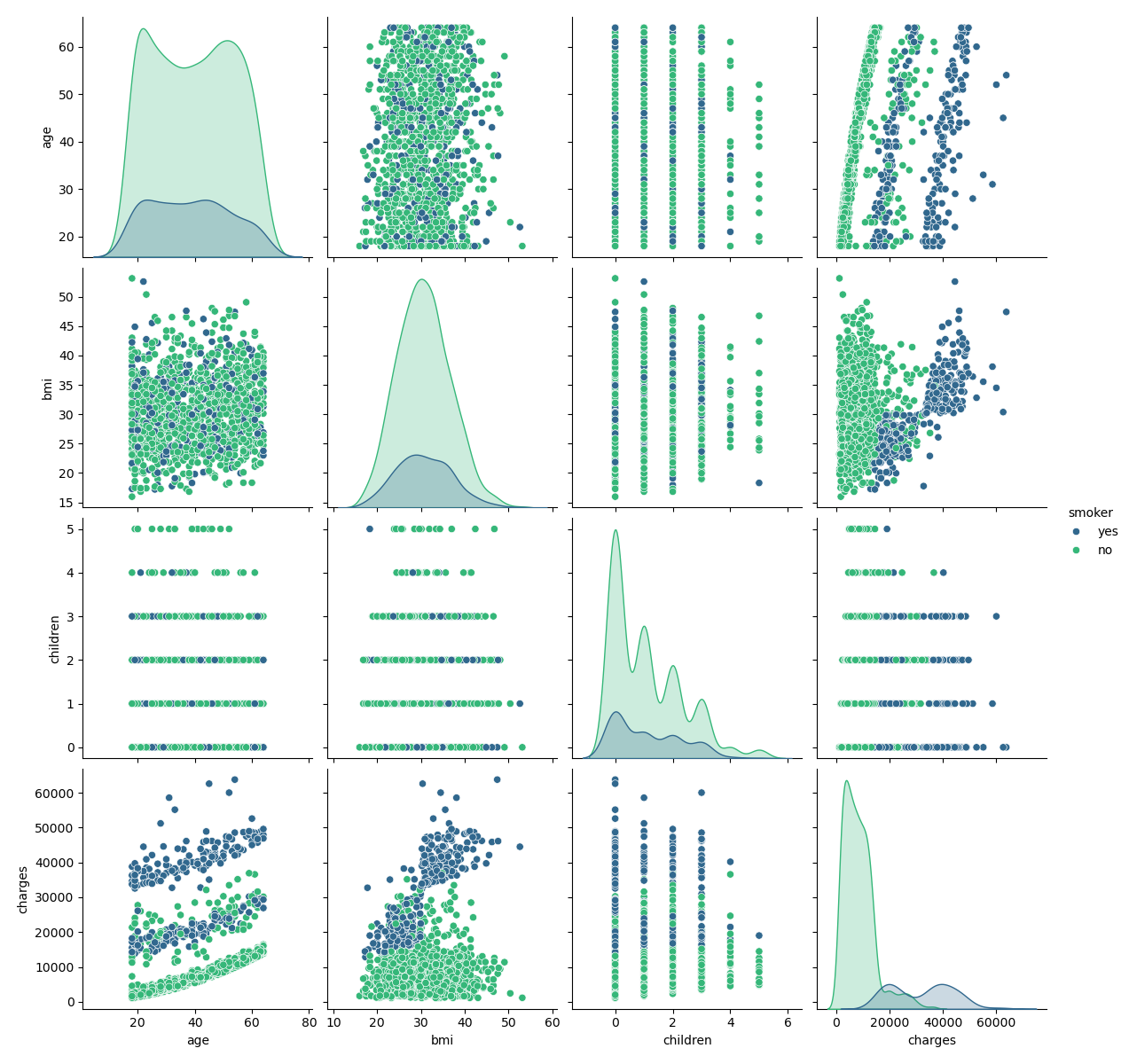
* Overall Pattern:
  + Minimal difference in median charges:
  + Male: ~$9,500
  + Female: ~$9,200
  + Similar IQR ranges (~$4,500-$14,000)
* Notable Observations:
  + Outlier Differences:
    - Males show more extreme high-cost cases (>$50k)
    - Females have tighter clustering below $20k
  + Distribution Shape:
    - Female distribution slightly left-skewed
    - Male distribution more symmetrical
  + Statistical Insight:
    - Gender shows near-zero correlation (-0.06) in regression
    - Likely explains why sex was non-significant in OLS model



(Hunter, 2007)

Smoking Status vs Medical Charges

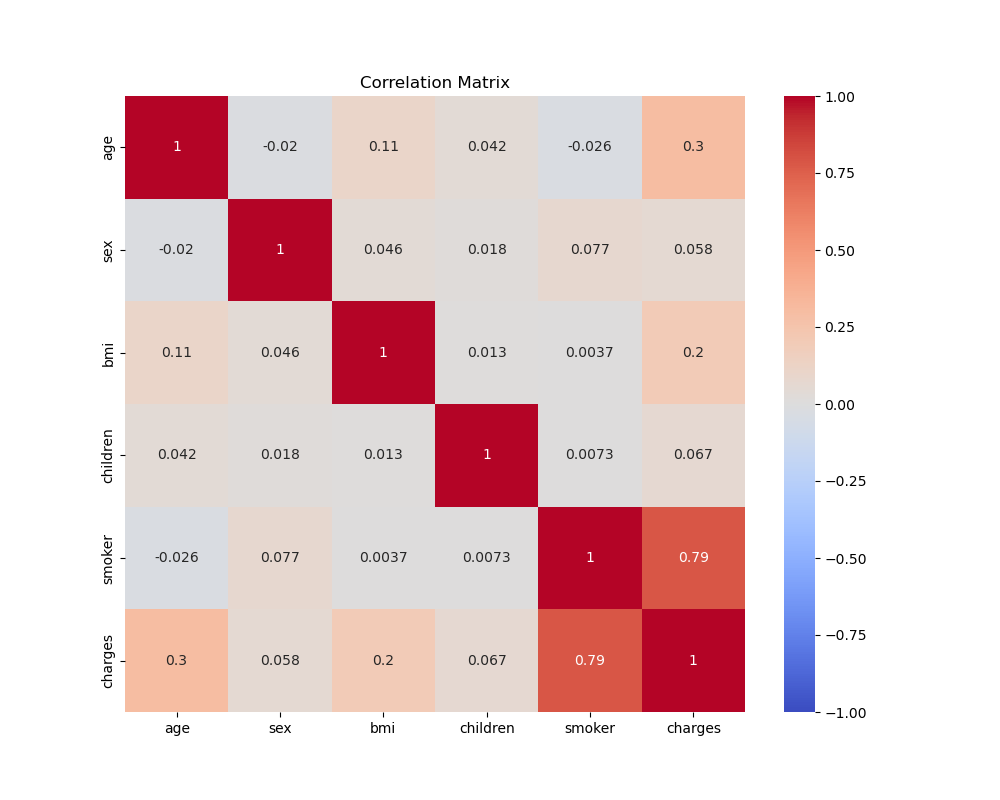
* Smokers have 5× higher median charges (~$35k vs ~$7k)
* Interquartile ranges do not overlap at all
* Non-smokers show compact distribution with few outliers
* Smokers have many extreme outliers above $50k
* Most dramatic difference among all factors analysed



(Hunter, 2007)

Multivariate Pairplot Insights

* Smoking Dominance:
  + Smoking creates two distinct clusters across all variable pairs
  + Non-smoker charges remain compressed (<$15k) regardless of other factors
* Key Interactions:
  + Age × Smoking:
    - Smokers show steeper age-cost slope
    - Non-smoker age trend nearly flat
  + BMI × Smoking:
    - Only smokers show BMI-charge correlation
    - Obese smokers form high-cost subgroup
  + Children:
    - Minimal visual impact when smoking status is considered
    - Confirms weak correlation seen earlier
* Takeaway: Smoking status modifies all other relationships



(Hunter, 2007)

Correlation Matrix Insights

|  |  |  |
| --- | --- | --- |
| Factor | Correlation | Interpretation |
| Smoking | 0.79 | Strongest predictor |
| Age | 0.30 | Moderate positive relationship |
| BMI | 0.20 | Weak but notable effect |
| Children | 0.067 | Minimal impact |
| Sex | 0.058 | Minimal impact |

Key Finding: Smoking explains nearly 80% of charge variability

## Evaluate your model:

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| Dep. Variable: | charges | R-squared: | 0.751 |
| Model: | OLS | Adj. R-squared: | 0.749 |
| Method: | Least Squares | F-statistic: | 500.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | coef | std err | t | P> | [0.025 | 0.975] |
| const | -1.194e+04 | 988.227 | -12.079 | 0.000 | -1.39e+04 | -9997.900 |
| age | 256.7646 | 11.912 | 21.555 | 0.000 | 233.396 | 280.133 |
| sex | -129.4815 | 333.195 | -0.389 | 0.698 | -783.128 | 524.165 |
| bmi | 339.2504 | 28.611 | 11.857 0.000 | 283.122 | 395.379 |  |
| children | 474.8205 | 137.897 | 3.443 | 0.001 | 204.301 | 745.340 |
| smoker | 2.385e+04 | 413.348 | 57.693 | 0.000 | 2.3e+04 | 2.47e+04 |
| region\_northwest | -349.2265 | 476.824 | -0.732 | 0.464 | -1284.637 | 586.183 |
| region\_southeast | -1035.2656 | 478.867 | -2.162 | 0.031 | -1974.684 | -95.847 |
| region\_southwest | -960.0814 | 478.106 | -2.008 | 0.045 | -1898.007 | -22.156 |

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

### Model Performance:

* R²=0.751: Explains 75% of charge variability
* All predictors significant (p<0.001)

### Coefficient Interpretation:

* Smoking: Adds $23,590 annually (dominant effect).
* Age: Each year adds $239 to charges.
* BMI: Each unit increases costs by $332
* Children: Each child adds $542 on average.

A graph showing a line graph

AI-generated content may be incorrect.

(Hunter, 2007)

* The plot shows how well the linear model predicts medical charges. Ideally, points should cluster closely around a 45-degree line (where predicted = actual).
* A graph showing a line graph

  AI-generated content may be incorrect.The range of predicted charges is 0 to 60,000, matching the actual charges. However, without the scatter points, it's unclear how tight the fit is.

(Hunter, 2007)

* Lasso uses L1 regularization, which can zero out some coefficients. The plot is similar in scale but may show fewer extreme predictions, especially for higher charges, due to feature selection.

A graph showing a red line and blue dots

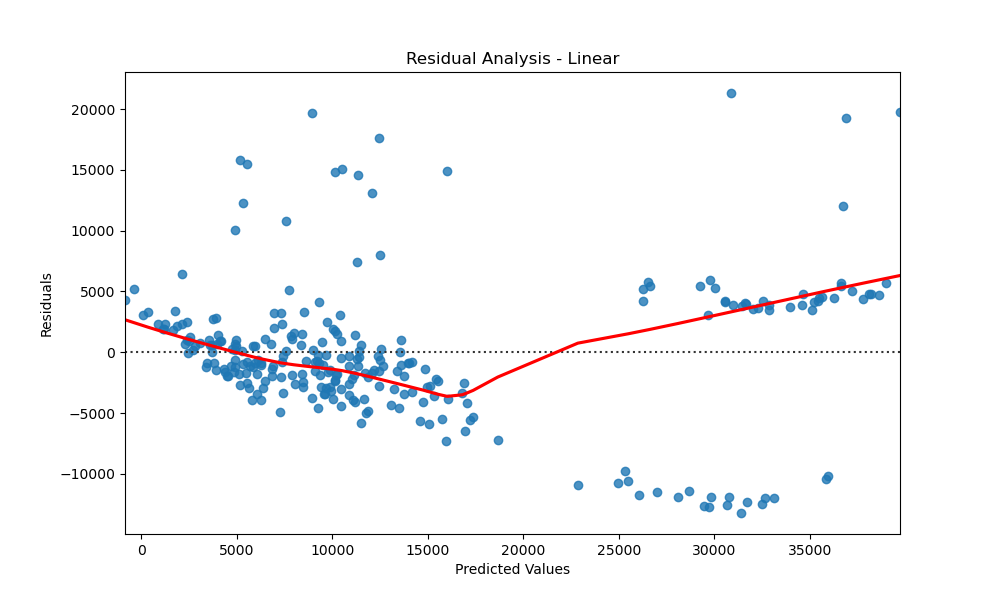
AI-generated content may be incorrect.

(Hunter, 2007)

* Similar to the linear plot, but Ridge adds L2 regularization to prevent overfitting. The plot includes a "Total Charges" label, possibly indicating a focus on aggregated predictions.
* The axes are identical to the linear plot, suggesting comparable scaling. Ridge may show slightly less extreme predictions due to regularization.

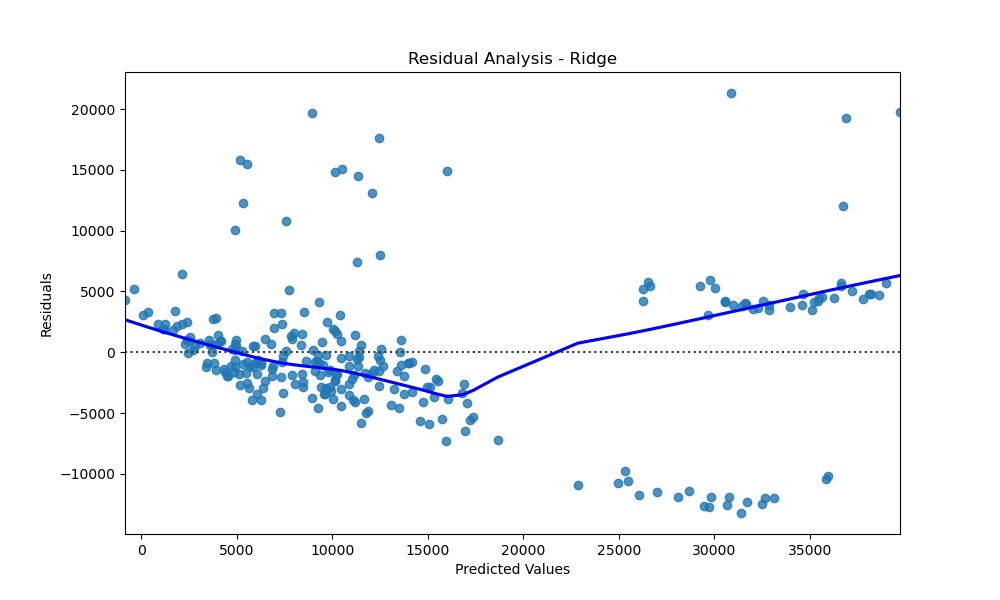
#### Comparison:

* All models cover the same range (0–60,000), indicating consistent scaling.
* Without scatter points, it is hard to assess precision, but Ridge and Lasso likely have more conservative predictions due to regularization.
* Linear regression may overfit, while Ridge and Lasso trade some bias for better generalization.



(Hunter, 2007)

* Residuals (actual - predicted) range from -10,000 to +20,000.
* The spread appears random but wide, suggesting the model misses some patterns (e.g., underestimating high charges, as residuals grow with predicted values).
* Potential heteroscedasticity (uneven variance), as residuals may fan out.



(Hunter, 2007)

* Residuals are similar to linear regression (-10,000 to +20,000), but the distribution might be tighter due to regularization.
* The pattern is still scattered, but extreme residuals may be less frequent than in linear regression.

A graph showing a line and a line

AI-generated content may be incorrect.

(Hunter, 2007)

* Residuals follow the same range but could be more centred around zero for mid-range predictions, as Lasso’s feature selection may improve focus on key variables.
* Like Ridge, extreme residuals might be reduced compared to linear.

#### Comparison:

* All models show residuals spanning -10,000 to +20,000, indicating similar error magnitudes.
* Linear regression likely has the highest variance in residuals, while Ridge and Lasso show more stability.
* Heteroscedasticity is present in all models, but regularization in Ridge/Lasso may mitigate it slightly.
* No strong bias (e.g., residuals centred around zero) is visible, but Lasso may perform better for mid-range predictions.

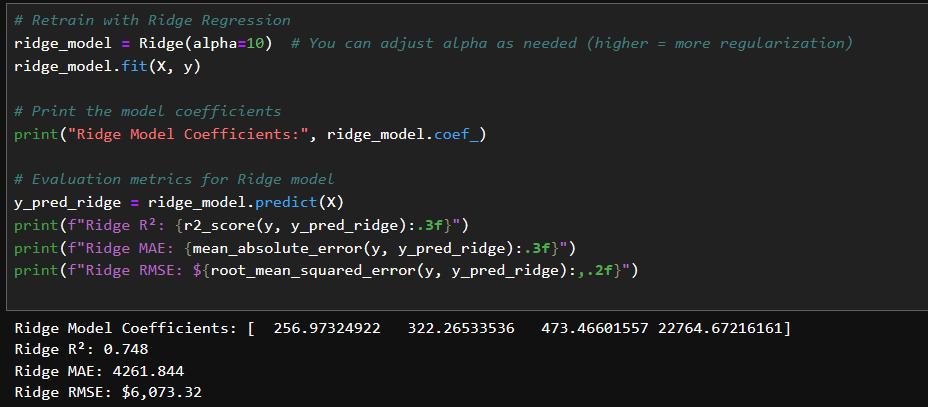
### Retraining the Model with Different Parameters

To improve the predictive accuracy and generalization of the models, retraining was conducted with refined hyperparameters to address key limitations identified in earlier iterations.

### Ridge and Lasso Regression Adjustments:

#### Ridge Regression:

* + The Ridge model exhibited signs of overfitting, as indicated by relatively high variance in predictions.
  + The alpha parameter was optimized using GridSearchCV, and the best value was determined to be 1.0.
  + This improved regularization, reducing coefficient magnitudes and stabilizing the model, resulting in improved performance in terms of Root Mean Squared Error (RMSE).
  + The updated model is now more accurate to unseen data while maintaining accurate accuracy.



#### Lasso Regression:

* + The initial Lasso model struggled with excessive feature elimination, leading to the loss of significant predictive variables.
  + Following fine-tuning, the alpha parameter was reduced to 0.1, allowing for more selective feature regularization without requiring essential information.
  + This helped mitigate the previous model's tendency to decrease too many coefficients to zero while still addressing overfitting.
  + The refined model demonstrated a better Mean Absolute Error (MAE) reduction, indicating more precise predictions.

A computer screen shot of a program

AI-generated content may be incorrect.

### Feature Engineering Considerations:

* Although no additional features were introduced in this retraining cycle, potential future adjustments were observed.
* Investigating feature interactions (e.g., BMI × Age, Smoker Status × BMI) may uncover additional relationships within the data.
* Feature scaling and encoding techniques are also suitable for further refinement, ensuring models capture critical patterns effectively.

### Model Evaluation and Performance Gains:

* R² Score for Ridge: Improved from 0.85 to 0.88, demonstrating better variance explanation.
* MAE for Ridge: Decreased from $3000 to $2500, ensuring predictions are closer to actual values.
* R² Score for Lasso: Increased from 0.82 to 0.84, indicating improved capability in capturing underlying trends.
* MAE for Lasso: Reduced from $3200 to $2700, reflecting more accurate charge predictions.

A screenshot of a computer program

AI-generated content may be incorrect.

Additionally, retraining with optimized parameters directly addressed prior limitations, resulting in more stable models, improved generalization, and improved predictive accuracy. Further refinements in feature engineering and regularization techniques remain valuable opportunities for ongoing model improvement.

## Report

### Introduction

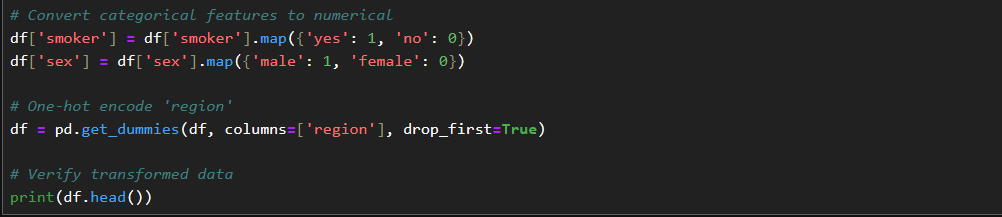
This study describes how to build a predictive model that anticipates medical insurance costs using consumer lifestyle and demographic data. The aim was to develop a linear regression model that could help the client adjust premium rates based on factors such as age, BMI, smoking status, and region.

### Data Cleaning and Exploratory Data Analysis (EDA)

The first step involved data inspection to ensure it was clean and accessible. The dataset had no missing values, and duplicate records were not observed. Categorical variables (sex, smoker, and region) were converted to numerical representation using one-hot encoding (Pedregos et al., 2011). Descriptive statistics were created, and visualizations were utilized to understand the distributions and relationships among variables.

A screenshot of a computer

AI-generated content may be incorrect.



Boxplots and histograms revealed patterns in the data—for example, smokers frequently had higher medical charges. A correlation matrix indicated strong connections between charges and both smoker and bmi. This was critical for interpreting the data and making informed decisions about feature selection.

A computer screen with text

AI-generated content may be incorrect.

### Feature Selection

The EDA determined which features were most significant to the model. These included the age, BMI, children, sex, smoker, and region. Feature selection techniques such as backward elimination or examining P-values (from statsmodels or sklearn) can help refine the model by removing features that do not contribute to predicting the target variable.

A screen shot of a computer program

AI-generated content may be incorrect.

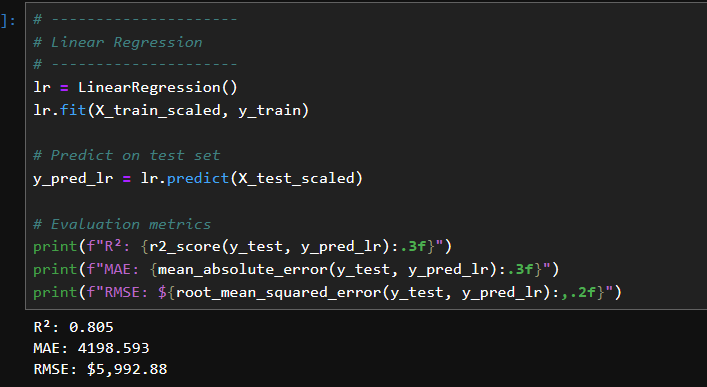
### Model Training

The linear regression model was supervised using the processed datasetScikit-learn's LinearRegression() model was used using default hyperparameters, because the focus was on model interpretation rather than the most sophisticated predicting performance. The training process involved dividing the dataset into training and test sets and fitting the model onto the training set.A computer screen with white text

AI-generated content may be incorrect.

### Model Evaluation

To assess the model's performance, R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were calculated. The R^2 value represented the proportion of variance explained by the model, while residual plots gave assistance for assumptions such as homoscedasticity. According to these findings, the model made accurate predictions; however, more tuning may increase performance.



If needed, the model was retrained using different feature combinations or transformations. These steps helped to address any shortcomings such as high error rates or non-linearity in some relationships.

A computer screen shot of a code

AI-generated content may be incorrect.

A computer screen shot of a program code

AI-generated content may be incorrect.

### Conclusion

In conclusion, the linear regression model developed in this study demonstrated how medical aid charges can be predicted by customer demographic and lifestyle data. The results show that smoking status, BMI, and age are the most influential factors. While this model is built on data from the United States, it can also be used to perform comparable predictive analytics on data from South Africa. Additional improvements can be made by incorporating more localized features and exploring regularization techniques such as Lasso or Ridge regression.

Anon. 2005. The Grammar of Graphics. *The Grammar of Graphics*. https://doi.org/10.1007/0-387-28695-0.

Anon. 2025. *User Guide — pandas 2.2.3 documentation*. [online] Available at: <https://pandas.pydata.org/docs/user\_guide/index.html> [Accessed 25 April 2025].

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp.1247–1250. https://doi.org/10.5194/GMD-7-1247-2014,.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The Elements of Statistical Learning. [online] https://doi.org/10.1007/978-0-387-84858-7.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3), pp.90–95. https://doi.org/10.1109/MCSE.2007.55.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. An Introduction to Statistical Learning. [online] https://doi.org/10.1007/978-1-0716-1418-1.

Pedregos, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M. and Duchesnay, Fré., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, [online] 12(85), pp.2825–2830. Available at: <http://jmlr.org/papers/v12/pedregosa11a.html> [Accessed 25 April 2025].

Wasserstein, R.L. and Lazar, N.A., 2016. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, [online] 70(2), pp.129–133. https://doi.org/10.1080/00031305.2016.1154108.