Insurance Charges Prediction

5. Analysis Report

Jishen Harilal

4/25/25

Linear Regression

Table of Contents

Introduction

This report explores the use of a linear regression model to predict individual medical insurance charges based on demographic and lifestyle factors. While the dataset is from the United States, it serves as a proof of concept for potential implementation in a South African medical aid context. The goal is to understand which features are most predictive of medical costs and how this insight can inform data-driven pricing strategies.

Linear regression was selected due to the continuous nature of the target variable (charges). The task aligns well with supervised learning and is appropriate given the structure and types of features in the dataset.

Dataset Overview

The dataset (insurance.csv) contains 1338 records and 7 features:

- **Numerical**: age, bmi, children
- **Categorical**: sex, smoker, region
- **Target**: charges

From a domain perspective, the independent variables make logical sense as cost predictors. For example:

- age can correlate with increased health risks.
- bmi can suggest obesity-related conditions.
- smoker is a known health cost risk factor.

Initial Findings:

- No missing values were found.
- Data was structured and clean upon loading.
- The charges column exhibited right-skew due to individuals with very high charges.

Exploratory Data Analysis (EDA)

Statistical Analysis

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Figure 1: Statistical summary of numeric variables

Based on the statistical summary, several observations were made regarding the numeric variables in the dataset:

- **Age** ranged from 18 to 64 years, with an average of approximately 39 years, indicating a broad and balanced representation of the adult population.
- **BMI**, which reflects body weight relative to height, averaged around 30 — slightly above the healthy range. Several individuals exhibited high BMI values, suggesting a diverse mix of body types.
- **Children**, representing the number of dependents, ranged from 0 to 5, with most individuals having 1 or 2 children.
- **Charges**, denoting medical insurance costs, varied significantly — from just over $1,000 to more than $63,000. This wide range suggests substantial differences in risk and cost among individuals.

Overall, the data appeared reliable and well-suited for modelling, with good variability across features and no signs of irregular or missing values.

Distribution Analysis

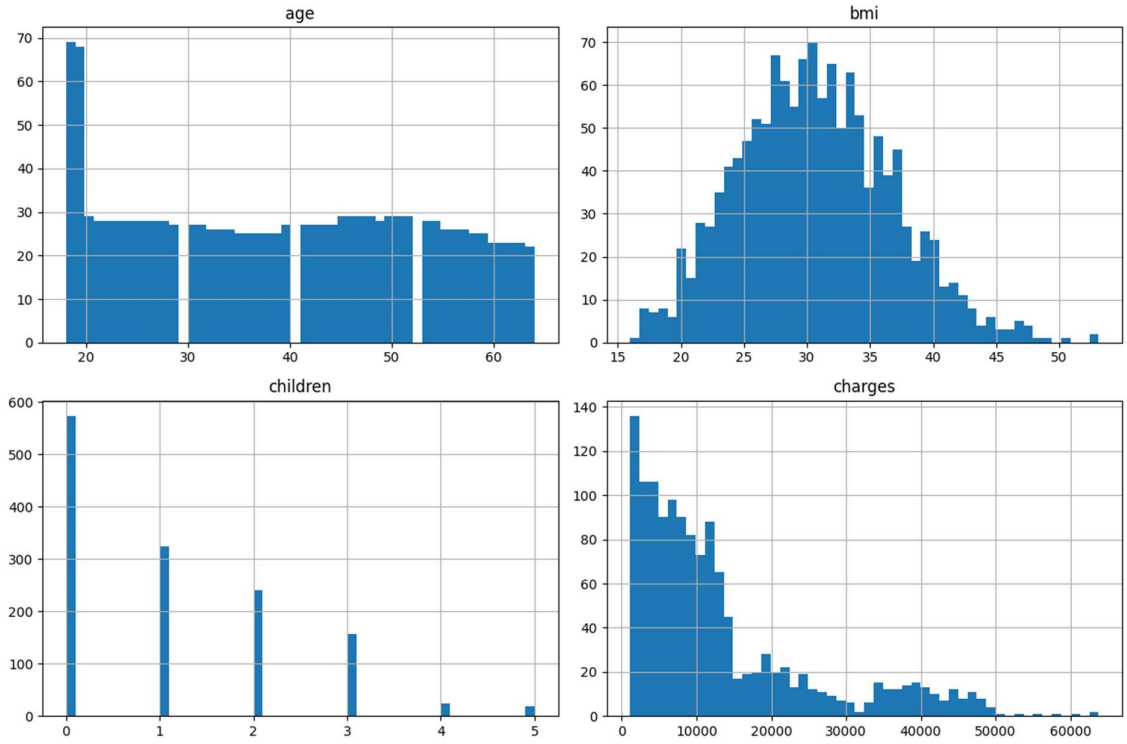

Figure 2: Distribution of numerical features.

Figure 2 illustrates the distribution of the numerical features in the dataset. Several key patterns were observed:

- Individuals aged between 18 and 20 appeared more than twice as frequently as those in other age groups, indicating a potential overrepresentation of younger adults.
- The distribution of **BMI** was relatively uniform across the sample.
- The number of children per individual followed a generally expected linear pattern; however, the high proportion of younger adults likely contributed to a bias toward individuals with no children.
- **Charges** exhibited significant variability, largely influenced by a small number of outliers with substantially higher costs. This right-skewed distribution may pose challenges for the model in accurately predicting charges for these extreme cases.

Correlation Analysis

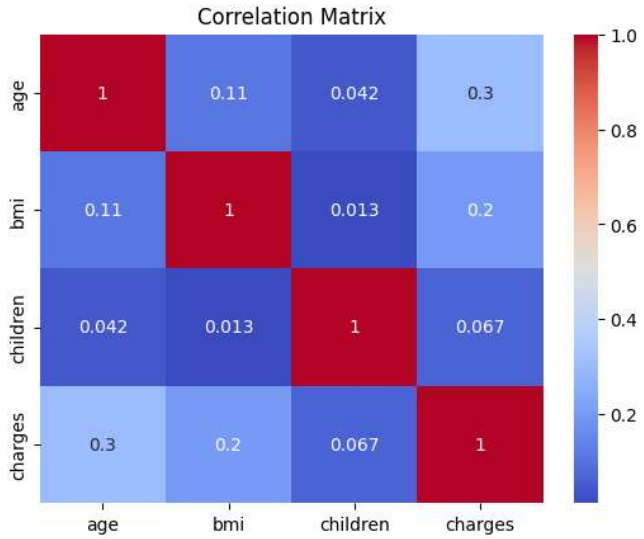

Figure 3: Correlation matrix of numerical features

Figure 3 presents the correlation coefficients between the numerical features, offering insight into the linear relationships within the dataset:

- **Age and Charges** (correlation = 0.30):
  A moderate positive correlation was observed, indicating that insurance charges tend to increase with age. This is consistent with expectations, as older individuals typically require more healthcare services.
- **BMI and Charges** (correlation = 0.20):
  A weak to moderate positive correlation suggests that individuals with higher BMI may incur higher charges, possibly due to obesity-related health risks.
- **Children and Charges** (correlation = 0.067):
  A very weak correlation was found, suggesting that the number of children does not significantly impact medical charges in this dataset.
- **Other Feature Pairs**:
  All other correlations—such as between **age and BMI**, or **children and any other variable**—were low, mostly below 0.1, indicating minimal linear relationships.

Key Insights:

- **Age** and **BMI** emerged as the most relevant numerical predictors of charges.
- The number of **children** is unlikely to contribute significantly to the linear regression model.
- The absence of strong correlations among the independent variables suggests that **multicollinearity is not a concern**, supporting the stability of the regression coefficients.

Categorical Feature Analysis

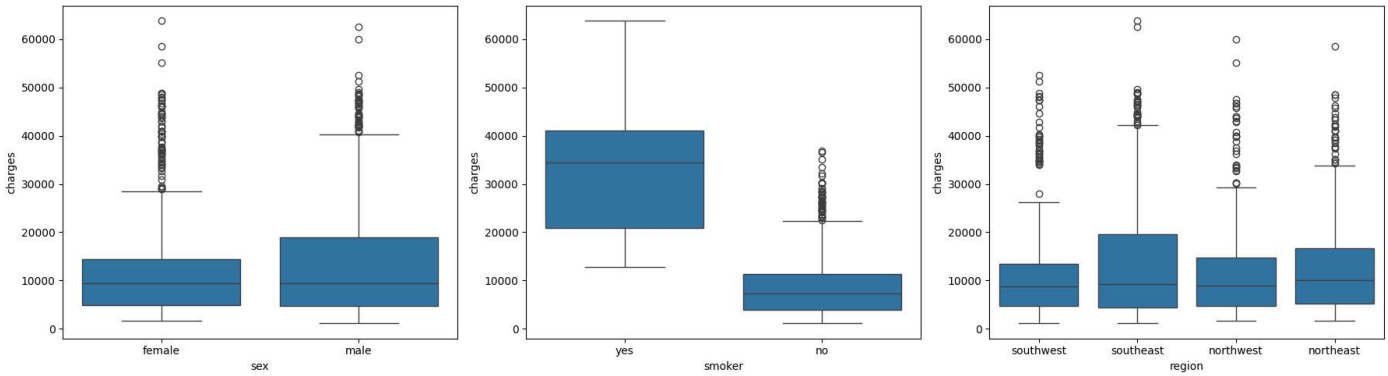

Figure 4: Boxplots for categorical feature analysis

A series of boxplots was used to examine the distribution of charges across different categories of the categorical variables:

1. **Sex vs Charges**:
   Male and female individuals exhibited nearly identical distributions in medical charges. The proximity of the medians and the presence of outliers in both groups suggest that gender does not play a significant role in determining charges.

2. **Smoker vs Charges**:
   A substantial difference was observed between smokers and non-smokers. Smokers had noticeably higher median charges, with the upper range extending significantly beyond that of non-smokers. This indicates that **smoking status is a strong predictor of higher medical costs**, reinforcing its importance as a key variable in the modelling process.

3. **Region vs Charges**:
   The impact of region on charges appeared minimal. All four regions showed similar medians and ranges, with the **southeast** region showing slightly greater variability and the **southwest** slightly less. Overall, region did not demonstrate a meaningful influence on medical charges.

Based on these observations and subsequent statistical analysis, **'sex' and 'region' were excluded** from the final model due to their lack of significance, as indicated by high p-values during model training.

Further Analysis

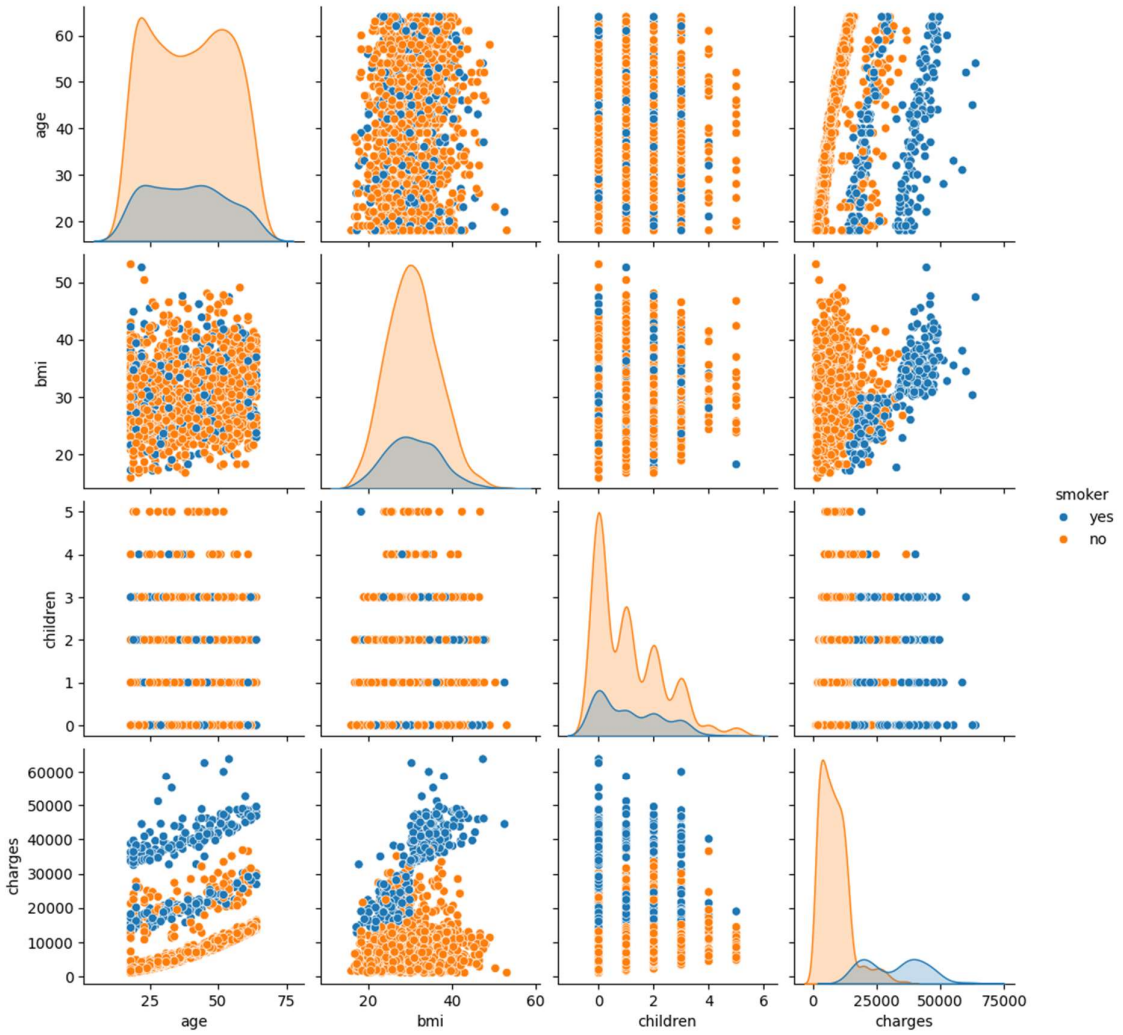

Figure 5: Pair plot showing potential linear relationships

Variable Pair	Relationship Type
Smoker vs Charges	Strong positive (especially smokers)
Age vs Charges	Moderate positive
BMI vs Charges	Mild to moderate (stronger if smoker)
Children vs Charges	Very weak or none

Figure 6: Relationships between variable pairs

A pair plot distinguishing smokers from non-smokers was used to further explore potential linear relationships within the dataset. The plot revealed a balanced distribution of smokers across different age groups and BMI levels, suggesting that **smoking status is not biased toward specific age or BMI ranges**.

Additionally, individuals with varying numbers of children were observed across all ages, indicating that the **'children' variable is not strongly influenced by age**. This supports the conclusion that there is no notable multicollinearity between these two variables, which enhances the reliability of the model's feature set.

Model Training Process

Feature Engineering

- Categorical variables (sex, smoker, region) were one-hot encoded.
- Care was taken to avoid the dummy variable trap by excluding one category per variable.

Model Setup

- Data was split into training (80%) and testing (20%) sets.
- Both Scikit-learn and Statsmodels were used:
    - Scikit-learn for performance metrics.
    - Statsmodels for detailed coefficient analysis and p-values.

Feature Selection

Using backward elimination:

- Variables with high p-values were iteratively removed.
- 'Sex' and 'region' variables were among the least significant and were dropped.

Model Evaluation

Metrics Used

- R^2 Score: Measured model explanatory power.
- MAE: Easy to interpret average error.
- RMSE: Penalized larger errors more harshly.

The R^2 Score of 0.805 means about 80.5% of the variation in insurance charges is explained by this model. It suggests that this model has good explanatory power.

The Root Mean Squared Error (RMSE) of 5992.88, means that on average, the predictions are off by about $5993. Bigger mistakes carry more weight with this metric, which means that the errors with outliers in the data get magnified.

Mean Absolute Error (MAE) of 4198.59 suggests that on average the predictions are off by $4199, this metric is calculated by treating all errors equally regardless of the size of the error unlike RMSE. This is a more realistic metric since we know our dataset contains some large outliers.

The gap between the MAE and RMSE indicates that **some larger errors exist in the predictions, but not to an extreme degree.**

Residual Analysis

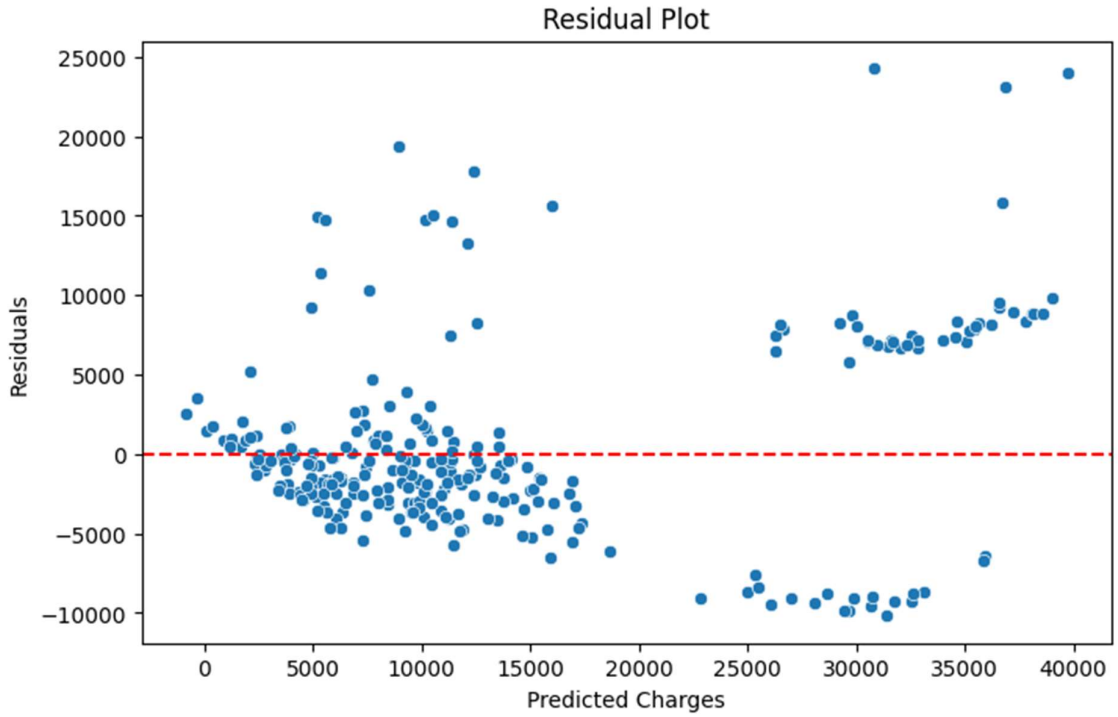

Figure 7: Residual plot

Ideally, residuals should appear randomly scattered around the zero line, indicating that the model's errors are evenly distributed. However, in this case, the residual plot in Figure 7 displayed distinct horizontal banding, with residuals clustered in separate ranges. This pattern strongly suggests the presence of underlying subgroups within the data. It is likely that the distinction between smokers and non-smokers is contributing to this effect, as earlier exploratory analysis — particularly the pair plot — also revealed visibly distinct groupings based on smoking status. This indicates that modelling smokers and non-smokers separately could potentially improve prediction accuracy.

Actual vs Predicted Charges Analysis

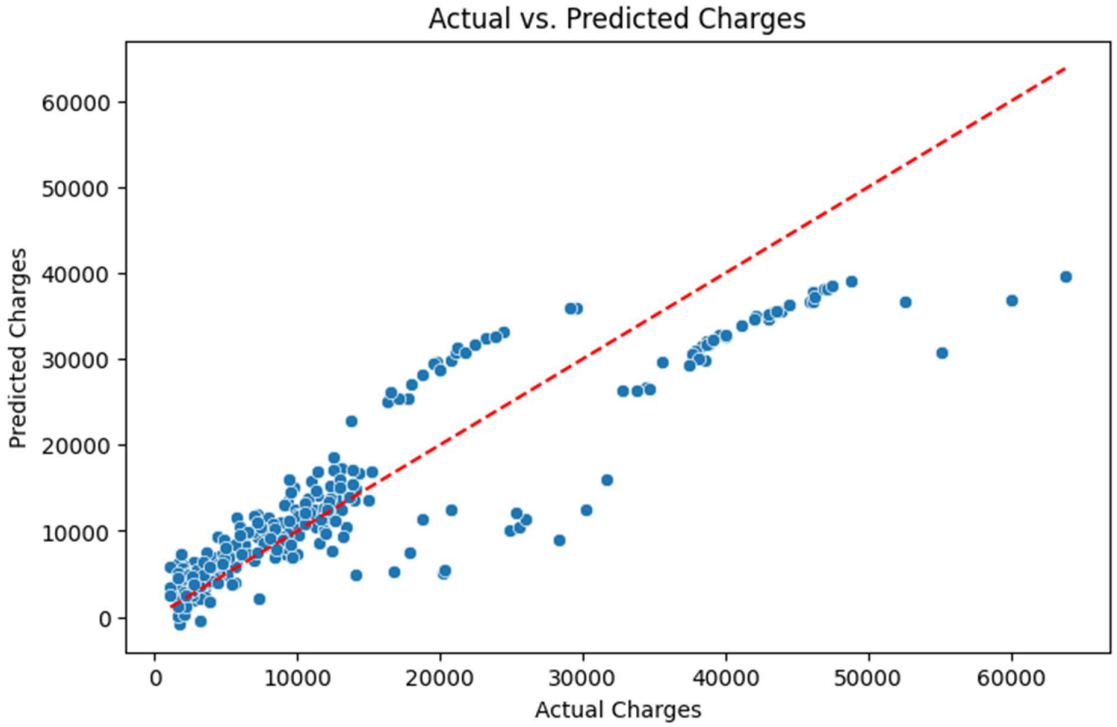

Figure 8: Actual vs Predicted Charges Plot

The Actual vs. Predicted Charges plot in Figure 8 revealed that predictions initially clustered closely around the regression line. However, distinct bands of points emerged just above and below the predicted values, following a similar gradient. This banding effect suggests that the model may be averaging between two clearly defined subgroups within the dataset. Based on prior exploratory findings, it is strongly suspected that these groups correspond to smokers and non-smokers, whose charge distributions differ significantly. This reinforces the consideration that separate modelling for these groups could yield improved performance.

Model Retraining

In response to concerns that the initial model was averaging between distinct subgroups - particularly between smokers and non-smoker - a revised modelling approach was implemented. Unlike the first model, which relied on backward elimination for feature selection, the retraining process involved standardising the dataset and applying Lasso regularisation. Standardisation was a necessary preprocessing step to ensure that all features contributed equally to the regularisation process. Lasso was chosen for its ability to perform both regularisation and automatic feature selection, reducing the risk of multicollinearity and simplifying the model without manual intervention.

Despite this refined approach, the new model did not resolve the previously observed issues. The evaluation metrics showed only marginal differences from the earlier model, and the same residual and prediction patterns persisted. Notably, the distinct banding of values - particularly those associated with smoker and non-smoker groups - remained visible. This outcome

indicates that, although Lasso regularisation provided a more constrained model, it was still insufficient to fully capture the complex subgroup dynamics present in the dataset.

Recommendations

The model's findings can support smarter, data-driven insurance pricing:

- **Risk-Based Pricing**: Key factors like smoking status, BMI, and age can be used to classify applicants into risk tiers, helping align premiums with expected medical costs.
- **Health Engagement Programs**: The cost gap between smokers and non-smokers highlights an opportunity for targeted wellness initiatives, such as programs to help people quit smoking, to reduce long-term claims.
- **Premium Estimation Tools**: The model's simplicity and interpretability make it suitable for integration into pricing tools or early underwriting processes.
- **Customer Segmentation**: Insights from features like age and BMI can guide the tailoring of insurance products to suit different risk profiles.

Limitations and Future Work

Several issues were identified that limit the model's performance and suggest areas for further improvement:

- **Outliers**: High-cost individuals skew predictions. Future work could involve outlier handling or segmentation.
- **Banding Effects**: Distinct prediction bands, likely tied to smoking status, suggest that separate models for smokers and non-smokers may yield better results.
- **Feature Gaps**: Important factors like chronic conditions or healthcare usage are missing, reducing predictive accuracy.
- **Demographic Relevance**: As the dataset is US-based, retraining on South African data is essential for practical application.

End