Predictive Analysis of Health Insurance Charges Using Data Science Techniques

**Executive Summary**

This project applies data science techniques to analyse and predict health insurance charges, using a publicly available dataset of 1,338 individuals from Kaggle (2025). The dataset includes demographic and lifestyle variables such as age, sex, BMI, number of children, smoking status, and region, alongside the dependent variable of annual insurance charges. The purpose of the project is twofold: first, to build and evaluate a predictive model for charges, and second, to generate insights that can support better decision-making in health insurance pricing and risk management.

The project was carried out using Microsoft Excel 365, chosen for accessibility and transparency, ensuring the workflow can be easily reproduced by stakeholders without coding expertise. The analysis followed a structured process: exploratory data analysis, feature engineering, regression modelling, classification of high-cost cases, and clustering to identify population subgroups. Results were summarised in a dedicated metrics tab and visualised through charts and dashboards for clarity.

The baseline model, which predicted the mean training charge for all individuals, produced an RMSE of £12,466 and MAE of £9,593 on the test set. The multiple linear regression model, incorporating age, BMI, smoker status, and interaction terms, significantly improved accuracy, achieving an RMSE of £4,538 and MAE of £2,755 (reductions of 64% and 71% respectively). Classification analysis correctly identified most high-cost individuals (≥£20,000), with precision of 84.6% and recall of 77.2%. Clustering separated the population into three interpretable groups, with smokers forming the highest-cost cluster.

The findings clearly show that smoking, BMI, and age are the strongest cost drivers. The results provide actionable insights for insurers, such as identifying high-risk individuals for targeted health programmes. While the analysis highlights limitations of linear models, it demonstrates the value of Excel-based workflows for producing reproducible, interpretable, and impactful results.

**Data Infrastructure & Tools**

This project was completed entirely in Microsoft Excel 365. The decision to use Excel was deliberate: although more advanced tools such as Python or R offer greater flexibility, Excel provides transparency, accessibility, and familiarity for most business users. For this project, these qualities were prioritised, as the workflow needed to be reproducible and easy to follow for both technical and non-technical stakeholders (James et al., 2021).

The infrastructure followed a structured organisation of the Excel workbook, supported by clear documentation in a GitHub repository. The dataset (*insurance.csv*) was imported into Excel and stored on a dedicated Data tab. Data cleaning and preparation were handled with Power Query, which allowed for standardising column names, checking data types, and ensuring consistency. Engineered variables and additional transformations were implemented using formula-based columns, which made every step auditable and transparent (Kuhn and Johnson, 2019).

Excel’s Analysis ToolPak provided the regression modelling functionality. While limited compared to programming environments, the ToolPak offers ordinary least squares regression with outputs including coefficients, R², and residuals. PivotTables and conditional formatting were employed to summarise data and explore key relationships. Charts such as histograms, scatterplots, and bar graphs were created to visualise distributions and trends. A dedicated Metrics tab consolidated performance measures (RMSE, MAE, R²), ensuring results were easy to track and compare.

The GitHub repository served as the project’s version control system, hosting the raw dataset, Excel workbook, exported figures, and final report. This ensured full reproducibility, as all components could be accessed and re-run by others. Using GitHub also adds transparency and professional polish, aligning the project with open data science practices (Han, Kamber and Pei, 2011).

Overall, the infrastructure combined the strengths of Excel for straightforward analysis with GitHub for reproducibility and sharing. This dual approach delivered a workflow that was not only effective in producing insights, but also practical for business use. The GitHub repository served as the project’s version control system, hosting the raw dataset and final report. This ensured full reproducibility, as all components could be accessed and re-run by others. Using GitHub also adds transparency and professional polish, aligning the project with open data science practices.

**Data Engineering**

The dataset contained seven main columns: age, sex, BMI, number of children, smoking status, region, and charges. Initial checks confirmed that all variables were present in the expected format, with no missing values. To improve model performance and capture known patterns in health costs, several engineered features were created:  
• Dummy variables were created for categorical variables including sex, smoker, and region. This ensured that regression models could treat these variables appropriately by converting categories into binary indicators.  
• Age squared was added as a variable to capture the accelerating costs associated with ageing, which are not strictly linear in nature.  
• Smoker × BMI interaction was introduced to reflect the combined impact of smoking and higher BMI on charges. This aligns with health research that shows smoking and obesity together increase risks significantly (Han, Kamber and Pei, 2011).  
• Standardised values (z-scores) of age and BMI were created to support clustering analysis, ensuring that variables on different scales contributed equally to distance calculations.

The data was split into training (80%) and testing (20%) sets to allow for objective evaluation of predictive performance. This split was fixed and stored as values to ensure reproducibility.

These engineering steps are consistent with recommended practices in predictive modelling. As Kuhn and Johnson (2019) highlight, well-chosen features often improve model performance more than changes in algorithms. The emphasis in this project was on simple, interpretable transformations that reflect real-world health risk dynamics.

**Data Visualisation & Dashboards**

Visualisations were used to explore data, validate assumptions, and communicate results effectively. These were produced directly in Excel using standard chart types, ensuring accessibility for business stakeholders.

Key exploratory visuals included:  
• Histogram of charges, showing a right-skewed distribution with many moderate cases and a smaller group of high-cost individuals.  
• Bar chart of average charges by smoking status, which clearly highlighted that smokers face much higher costs.  
• Scatterplot of BMI versus charges split by smoking, which illustrated that the cost slope with BMI is steeper for smokers.  
• Boxplots of charges by region and sex, confirming that these variables had relatively smaller effects.

For model evaluation, a residual versus predicted plot was created for the test set, showing that residuals centred around zero but with wider spread at higher charges, indicating heteroscedasticity. A calibration plot by prediction decile confirmed that predicted averages aligned closely with actual averages, suggesting good overall fit.

All visuals were consolidated into a Dashboard tab alongside summary tables of metrics, classification results, and cluster profiles. This dashboard allowed stakeholders to quickly grasp the key findings without reviewing all details. The focus was on clarity and relevance, as data visualisation should support decision-making rather than overwhelm users (Kuhn and Johnson, 2019).

**Data Analytics**

*Baseline Model*

A baseline model that predicted the mean training charge for all individuals achieved:  
• RMSE = £12,466 (Test)  
• MAE = £9,593 (Test)

This established a reference point against which more sophisticated models could be judged.

*Regression Model*

The multiple linear regression model improved significantly on the baseline:  
• RMSE = £4,538 (Test) → 64% improvement  
• MAE = £2,755 (Test) → 71% improvement  
• R² = 0.87 (Test)

These results show strong explanatory power, consistent with regression’s ability to model additive relationships (James et al., 2021). The coefficients confirmed that smoking had the largest positive impact on charges, followed by BMI and the smoker × BMI interaction. Age and age² also contributed, capturing rising costs with age. Region and sex effects were minor.

Residual analysis indicated heteroscedasticity, with larger errors at higher charges. This is common in cost modelling. A log transform of charges or use of tree-based models could mitigate this issue (James et al., 2021).

*Classification of High-Cost Cases*

Using a £20,000 threshold, classification results were:  
• Precision = 84.6%  
• Recall = 77.2%  
• F1 score = 0.81  
• Confusion matrix: TP = 44, FP = 8, FN = 13, TN = 203.

This shows the model correctly identified most high-cost individuals while keeping false positives relatively low. In practice, such a system could help insurers flag customers for preventative support.

*Clustering*

Clustering produced three interpretable groups:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Size** | **Mean Age** | **Mean BMI** | **% Smokers** | **Mean Charges** |
| 1 | 594 | 49.8 | 32.0 | 0% | £11,235 |
| 2 | 274 | 38.5 | 30.7 | 100% | £32,050 |
| 3 | 470 | 26.2 | 28.9 | 0% | £4,895 |

Cluster 2, containing all smokers, had the highest costs, while Cluster 3 consisted of young non-smokers with low costs. Cluster 1 represented older non-smokers with moderate costs. These profiles are simple yet actionable, demonstrating the value of segmentation in understanding populations (Han, Kamber and Pei, 2011).

**Conclusion**

This project demonstrated an end-to-end data science workflow using Excel. It produced a regression model that improved substantially on a baseline, confirmed smoking, BMI, and age as the strongest cost drivers, and provided practical tools for classification and clustering. The workflow was fully documented and reproducible, supported by GitHub.

Ethical and practical considerations include the importance of fairness in insurance modelling. While smoking and BMI are legitimate health risk factors, care must be taken that models are not used in ways that unfairly penalise individuals. Instead, insights should guide supportive interventions, such as wellness programmes. Data privacy must also be upheld, ensuring that sensitive health information is handled responsibly (Kuhn and Johnson, 2019).

Overall, the project achieved its aims and demonstrated how accessible tools like Excel can generate valuable business insights when paired with structured data science practices.Cluster 2, containing all smokers, had the highest costs, while Cluster 3 consisted of young non-smokers with low costs. Cluster 1 represented older non-smokers with moderate costs. These profiles are simple yet actionable, demonstrating the value of segmentation in understanding populations (Han, Kamber and Pei, 2011).

**References**  
• Han, J., Kamber, M. and Pei, J. (2011) Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann.  
• James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning with Applications in R. 2nd ed. Springer.  
• Kaggle (2025) Medical Cost Personal Dataset. Available at: <https://www.kaggle.com/mirichoi0218/insurance> (Accessed: 20 July 2025).  
• Kuhn, M. and Johnson, K. (2019) Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press.