Medical Insurance Cost Analysis and Prediction Using Statistical Learning Methods

Course: Data Analytics

Dataset: Kaggle “Insurance” https://www.kaggle.com/datasets/mirichoi0218/insurance

Instructor: Dr. Ruxian Wang

Group 8: Fiona Ma, Yicheng Gu, Juan Sun, Andrea Zhang, Chen Nuo

Date: 12.16.2025

- Fiona Ma: Led data preparation and EDA, defined research questions, and created the shared train–test split.
- Yicheng Gu: Built baseline and extended linear regression models and interpreted key coefficients and diagnostics.
- Juan Sun: Conducted model selection using adjusted R^2/AIC/BIC and finalized the parsimonious regression model.
- Andrea Zhang: Developed and evaluated logistic regression and decision tree models for high-cost classification.
- Chen Nuo: Performed K-means clustering and integrated analyses into the final report and presentation.

1. Questions and Hypotheses

Our project investigates how demographic and lifestyle factors drive annual medical insurance charges, and how these variables can be used to identify high-cost customers and risk segments. Based on economic intuition and the initial exploratory analysis, we focus on the following research questions and hypotheses.

Q1. Does smoking significantly increase annual medical charges after controlling for other individual characteristics?

H1. Smokers have significantly higher expected annual charges than non-smokers, even after controlling for age, BMI, number of children, sex, and region. In a multiple linear regression framework, we expect the coefficient on the smoker indicator to be positive and statistically significant.

Q2. Is the relationship between BMI and medical charges approximately linear, or do charges increase more than proportionally at higher levels of BMI?

H2. BMI is positively associated with annual charges, and a nonlinear specification (for example including a BMI^2 term) improves model fit. This would indicate that individuals with very high BMI experience a more-than-proportional increase in expected charges relative to individuals with normal BMI.

Q3. After controlling for age, BMI, and smoking status, do sex and region still have a meaningful impact on annual charges?

H3. Once age, BMI, smoking status and family size are taken into account, sex and region have weaker and possibly insignificant effects on charges. We expect model selection procedures to drop at least some of these variables from the final regression model, suggesting that they add little incremental explanatory power.

Q4. Can we predict which individuals fall into a "high-cost" group using observable characteristics such as age, BMI, smoking status, number of children, sex, and region?

H4. By defining a high-cost indicator based on the upper quantiles of the charges distribution, logistic regression and classification trees can achieve substantially higher classification accuracy than random guessing. We expect smoking status, BMI and age to be among the most important predictors of high-cost status.

Q5. Are there distinct groups of customers with different risk and cost profiles that can be uncovered through clustering?

H5. K-means clustering on variables such as age, BMI, number of children and annual charges reveals several clusters that correspond to meaningful customer segments. For example, we expect to find a low-risk segment of younger, non-smoking individuals with relatively low charges, and a high-risk segment of older or smoking individuals with higher BMI and significantly higher charges.

2. Data Description

We use the publicly available "Insurance" dataset from Kaggle, which simulates annual medical insurance charges for individuals in the United States. The dataset contains 1,338 observations and seven variables:

- age (numeric): age of the primary beneficiary, measured in years.
- sex (factor): gender of the insured individual ("female" or "male").
- bmi (numeric): body mass index, a standard measure of weight relative to height.
- children (numeric): number of dependents covered by the insurance plan.
- smoker (factor): smoking status ("yes" or "no").

- region (factor): residential region (“northeast”, “northwest”, “southeast”, or “southwest”).
- charges (numeric): annual medical insurance charges in U.S. dollars.

The dataset contains no missing values. Age ranges from 18 to 64 years (mean 39.2), BMI from 15.96 to 53.13 (mean 30.7), and the number of dependents from 0 to 5 (mean 1.1).

Annual charges are highly right-skewed: the mean is $13,270, but the median is only $9,382. While the interquartile range spans $4,740 to $16,640, the distribution has a long right tail, with the 90th and 95th percentiles reaching $34,800 and $41,200, and a maximum of $63,770.

Smoking status is the strongest driver of costs. Although smokers represent only 20% of the sample, their average annual charge ($32,050) is nearly four times that of non-smokers ($8,434). Regional variation exists but is modest by comparison, with mean charges ranging from $12,347 in the southwest to $14,735 in the southeast.

For model development, the data were randomly split 70/30 using a fixed seed, producing a training set of 937 observations and a test set of 401 observations for out-of-sample evaluation.

3. Methodology

3.1 Linear Regression: Baseline and Extended Models

We use multiple linear regression to model annual medical insurance charges as a function of individual characteristics. Linear regression is appropriate because the response variable (charges) is continuous and we seek to quantify the marginal effect of each predictor while controlling for other factors.

Baseline Model Specification

The baseline model includes all available predictors as main effects:

$$\text{charges}=\beta0+\beta1\cdot\text{age}+\beta2\cdot\text{bmi}+\beta3\cdot\text{children}+\beta4\cdot\text{smoker}+\beta5\cdot\text{sex}+\beta6\cdot\text{region}+\varepsilon\text{charges}=\beta0+\beta1\cdot\text{age}+\beta2\cdot\text{bmi}+\beta3\cdot\text{children}+\beta4\cdot\text{smoker}+\beta5\cdot\text{sex}+\beta6\cdot\text{region}+\varepsilon$$

This model addresses Hypotheses H1 and H3 by estimating the smoking effect while controlling for demographics and region. The model is fitted using ordinary least squares on the training set (n = 937).

Diagnostic Assessment

The residuals-versus-fitted plot reveals two distinct bands corresponding to smokers and non-smokers, suggesting that a single linear slope for BMI may be inadequate. The Q-Q plot shows right-skew deviation, and the Shapiro-Wilk test confirms non-normality ($W = 0.905$, $p < 0.001$). Approximately 45 observations (4.8%) have standardized residuals exceeding ±2, indicating high-cost outliers. These findings motivate exploring nonlinear terms and interactions.

Extended Model Specifications

To address Hypothesis H2 and improve model fit, we consider four extensions: (1) a BMI^2 term to test nonlinearity, (2) a smoker × BMI interaction allowing the BMI effect to differ by smoking status, (3) a smoker × age interaction, and (4) a full extended model combining BMI^2 and smoker × BMI. Each model is evaluated using adjusted R^2, residual standard error, AIC, and BIC.

3.2 Model Selection & Final Regression Model Lead

While the extended model improves predictive accuracy, it includes many predictors and relies on judgment-driven feature expansion. To obtain a more objective and parsimonious specification, we applied best subset selection using the leaps package. The candidate set contained all baseline predictors, a nonlinear BMI^2 term, and the smoker × BMI interaction. Models were compared using Adjusted R^2, Mallows' Cp, and BIC. While Adjusted R^2 and Cp favored larger models (≈10 predictors), BIC selected a 4-variable model: age, children, smoker, and smoker × BMI.

On the test set, the extended model achieved the lowest RMSE, but the BIC-selected model delivered nearly identical accuracy with less than half the predictors—offering the best balance between interpretability and performance. The inclusion of smoker × BMI confirms that BMI meaningfully affects costs only for smokers, consistent with earlier findings. Based

on these considerations, the BIC-selected specification is adopted as our final regression model.

3.3 Classification – Logistic Regression

To extend our analysis beyond continuous prediction of medical charges, we reframed the problem as a binary classification task aimed at identifying "high-cost" customers. Using the training sample (insurance_train), we defined a policyholder as high-cost if their annual medical charges were at or above the 75th percentile of the training-set distribution. This focuses attention on customers who fall into the highest cost tier.

We estimated a logistic regression model using the predictors age, BMI, number of children, smoking status, sex, and region. The model was trained on insurance_train, and predicted probabilities were generated for insurance_test. We evaluated several probability cutoffs (0.4, 0.5, and 0.6) and computed accuracy, sensitivity, and specificity for each.

The logistic regression model is specified as:

$$\textit{Log-odds of being high-cost} = -6.82 + 0.0276 \cdot Age + 0.0859 \cdot BMI + 0.2599 \cdot Children + 5.4602 \cdot Smoker + 0.1253 \cdot Male + Region\ effects$$

This formulation shows that smoking status has a disproportionately large effect on the log-odds of being high-cost, while age, BMI, and number of children also contribute positively to cost risk.

3.4 Classification – Decision Tree

A classification tree was also estimated to capture nonlinear relationships and natural threshold effects among predictors. Using insurance_train, we fit an initial tree model with the same predictor set used in logistic regression.

To avoid overfitting, cost-complexity pruning was performed using cross-validation. The optimal tree size was selected based on the misclassification deviance curve, and the model was pruned accordingly. The pruned tree was then evaluated on insurance_test, and we examined both its predictive performance and its interpretability.

4. Results and Conclusion

4.1 Baseline Model Results

The baseline model explains approximately 74.6% of the variance in annual charges (adjusted $R^2 = 0.746$). Table 1 summarizes the coefficient estimates.

Table 1: Baseline Model Coefficients

Variable	Estimate	95% CI	p-value
Intercept	−$12,481	[−14,866, −10,095]	< 0.001
Age	$240	[211, 269]	< 0.001
BMI	$370	[302, 438]	< 0.001
Children	$674	[350, 998]	< 0.001
Smoker (yes)	$23,940	[22,971, 24,908]	< 0.001
Sex (male)	−$247	[−1,048, 554]	0.546
Region (NW)	−$595	[−1,734, 544]	0.305
Region (SE)	−$953	[−2,101, 196]	0.104
Region (SW)	−$1,117	[−2,258, 24]	0.055

Interpretation of Key Findings

Hypothesis H1 (Smoking Effect): Strongly supported. Controlling for all other variables, smokers incur approximately $23,940 more in annual charges than non-smokers ($p < 0.001$). This represents the largest effect in the model.

Hypothesis H3 (Sex and Region Effects): Supported. Sex is not statistically significant ($p = 0.546$), and most regional indicators fail to reach significance. Once we account for age, BMI, smoking, and family size, gender and geography contribute little additional explanatory power.

Other Effects: Age and BMI are both positive and highly significant. Each additional year of age increases expected charges by $240, each unit increase in BMI raises charges by $370, and each additional child adds approximately $674 to annual costs.

4.2 Extended Model Results

Table 2: Model Comparison

Model	Adj R^2	RSE	AIC	BIC
Baseline	**0.746**	**6,182**	**19,029**	**19,077**

+ BMI²	**0.747**	**6,166**	**19,025**	**19,078**
+ Smoker:BMI	**0.840**	**4,912**	**18,599**	**18,652**
+ Smoker:Age	**0.746**	**6,184**	**19,031**	**19,084**
Extended (BMI² + Smoker:BMI)	**0.841**	**4,891**	**18,592**	**18,650**

Hypothesis H2 (BMI Nonlinearity): Partially supported. The BMI² coefficient is negative and significant ($\beta = -9.24$, $p = 0.016$), indicating a concave relationship where the marginal effect of BMI diminishes at very high values. However, BMI² alone yields only modest improvement (adjusted R^2 from 0.746 to 0.747).

Smoker × BMI Interaction: The most substantial improvement comes from this interaction term. The coefficient of \$1,441 ($p < 0.001$) indicates that each unit increase in BMI raises charges by \$1,441 more for smokers than for non-smokers. With this interaction included, the main BMI effect becomes non-significant for non-smokers ($\beta = \$27$, $p = 0.390$), while for smokers the total BMI effect is \$1,468 per unit.

This finding has practical implications: among non-smokers, BMI has minimal impact on charges, but among smokers, higher BMI dramatically amplifies costs. This synergistic effect is consistent with medical evidence that smoking and obesity compound cardiovascular and metabolic disease risk. Figure 1 visualizes this interaction.

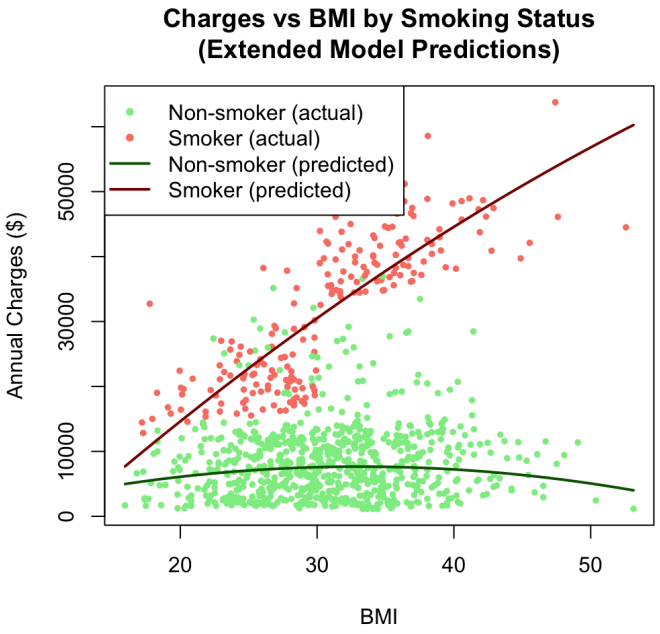


Figure 1: Predicted annual charges by BMI and smoking status.

The smoker × age interaction is not significant ($p = 0.625$), suggesting that the age effect does not differ by smoking status.

Final Extended Model

The extended model achieves adjusted $R^2 = 0.841$, explaining 84.1% of variance in charges. An F-test confirms significant improvement over the baseline ($F = 278.15$, $p < 0.001$). On the test set ($n = 401$), the extended model achieves RMSE = $4,707, compared to $5,829 for the baseline, a 19% reduction in prediction error.

Residual Diagnostics

The extended model diagnostics show improvement. The residuals-versus-fitted plot no longer displays the two-band pattern, indicating that the interaction term captures the differential BMI effect. Some right-skew remains in the Q-Q plot due to high-cost outliers that linear models struggle to predict precisely.

4.3 Model Selection Results and Final Model

To balance predictive performance and interpretability, we compared the baseline and extended models from Section 4.1–4.2 with a model selected using best subset selection. Table 3 summarizes the training adjusted R^2 and the test RMSE for the three candidates.

Table 3: Comparison of Baseline, Extended, and Final Models

Model	Adjusted R^2	Test RMSE
Baseline	0.764	5,829
Extended	0.841	4,707
Final(BIC-selected)	0.838	4,739

The extended model achieves the best predictive accuracy (RMSE = $4,707), but the BIC-selected model achieves nearly identical test performance with substantially fewer predictors. The final model includes only:

- age
- children

- smoker
- smoker × BMI

The presence of the smoker × BMI interaction again highlights the strong compounding effect between smoking and body weight. Importantly, BMI itself does not enter the final model independently, reinforcing the finding that BMI has little predictive value for non-smokers.

Because the BIC-selected model offers a more compact structure with minimal loss in predictive accuracy, we identify it as the final regression model used in our overall findings.

4.4 Classification Results

Logistic Regression Results

Using the 75th percentile of training-set charges as the high-cost threshold, the logistic regression model identifies smoking status, BMI, age, and number of children as statistically significant predictors of high-cost outcomes ($p < 0.01$). Smoking is by far the strongest driver: the estimated log-odds coefficient of **5.46** implies that, holding other variables constant, smokers have dramatically higher odds of falling into the high-cost segment.

BMI also plays a substantial role. Each one-unit increase in BMI raises the odds of being high-cost by approximately **8.6%**, and age contributes positively as well, with older customers more likely to incur higher medical expenses. Sex and geographic region, in contrast, do not appear to have significant explanatory power within this classification framework.

When evaluated on the test set, the logistic regression model exhibits strong predictive performance:

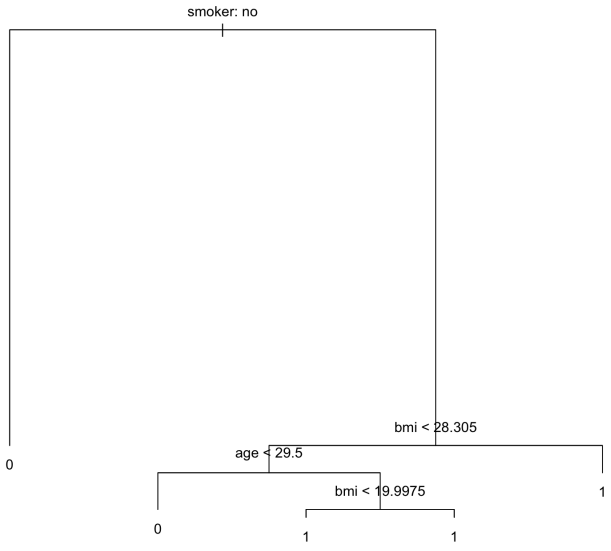
Cutoff	Accuracy	Sensitivity	Specificity
0.4	0.910	0.691	0.978
0.5	0.913	0.691	0.980
0.6	0.915	0.691	0.984

Increasing the cutoff slightly improves overall accuracy and specificity while keeping sensitivity constant at 69%. All cutoffs produce accuracy above 91%, indicating consistent predictive strength with relatively few false positives.

Classification Tree Results

The pruned decision tree achieves a test accuracy of **0.913**, comparable to the logistic regression model. The tree produces a clear, hierarchical segmentation rule that is straightforward for managers to interpret.

Figure 2 below illustrates the pruned classification tree used to identify high-cost customers, highlighting smoking status as the primary split, followed by BMI and age thresholds among non-smokers.



The first and most important split is **smoking status**:

- **If smoker = no → low-cost (predict 0)**
- **If smoker = yes → additional splits on BMI and age**

For smokers, the tree identifies BMI and age thresholds that distinguish higher- vs lower-risk profiles:

1. **smoker = yes**
    - If **BMI < 28.3**, then:
        - If **age ≥ 29.5**, predict *high-cost*
        - If **age < 29.5**, predict *low-cost*

- If **BMI ≥ 28.3** → predict *high-cost*

These rules highlight intuitive relationships: high BMI among non-smokers still substantially elevates medical cost risk, and younger, low-BMI non-smokers represent the lowest-risk segment.

Model Comparison and Business Implications

Both models have strong performance, with test-set accuracies between 0.91 and 0.915. Logistic regression provides smooth probability estimates and clear coefficient-based interpretations, while the tree model translates the same patterns into simple, actionable decision rules.

Overall, smoking is the main driver of high-cost medical payout, while higher BMI and older age further increase the risk even among non-smokers, and the resulting tree-based decision rules provide insurers with a transparent and actionable framework for customer segmentation, pricing, and targeted wellness interventions.

The classification analysis confirms the core drivers identified in earlier regression models and provides operational tools for identifying customers most likely to generate high medical expenses.

5.1 Clustering – K-means

To explore whether distinct groups of policyholders with different risk and cost profiles exist (Hypothesis H5), we apply K-means clustering as an unsupervised approach to customer segmentation. This method identifies natural groupings based on demographic and cost-related characteristics without imposing a predefined outcome.

The clustering variables include age, BMI, number of children, annual charges, and a binary smoking indicator—the dominant risk factor identified in earlier regression and classification analyses. All variables are standardized to ensure that differences in scale, particularly the large magnitude of charges, do not dominate the distance calculations.

We determine the optimal number of clusters using the elbow plot of within-cluster sum of squares and the average silhouette width. The elbow plot shows clear improvement up to K =

3, after which gains become marginal. Balancing statistical fit and interpretability, we select K = 3 as the final solution.

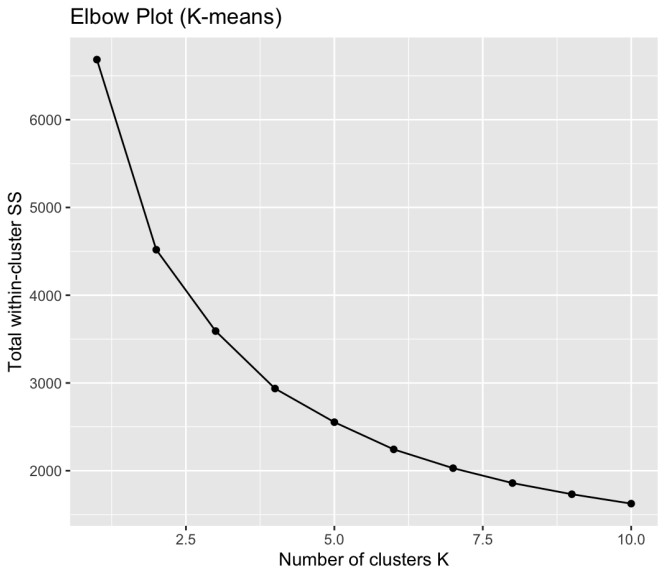


Figure 3.Elbow plot

The K = 3 clustering yields three clearly interpretable customer segments:

- **Cluster 1: Low-Risk Young Non-Smokers (n = 531)**
  This group consists primarily of younger individuals with relatively low medical expenditures. The average annual charge in this cluster is approximately $5,029, with a median of $3,982. Members of this cluster are almost exclusively non-smokers and represent the lowest-risk segment in the dataset.
- **Cluster 2: Older Non-Smokers with Moderate Costs (n = 533)**
  Individuals in this cluster are older on average and exhibit slightly higher BMI levels. Despite these risk factors, they are largely non-smokers, which keeps their medical costs moderate. The average annual charge in this group is approximately $11,827, substantially higher than Cluster 1 but far below the highest-cost group.
- **Cluster 3: High-Risk Smokers with Extremely High Costs (n = 274)**
  This cluster is characterized almost entirely by smokers and exhibits dramatically higher medical expenditures. The average annual charge exceeds $32,000, with a median of approximately $34,456. This segment corresponds closely to the high-cost group identified in both the linear regression and classification analyses.

Cluster	n	Avg.Age	Avg.BMI	Avg. Children	Avg. Charges (USD)	Median Charges (USD)	Smoker Share
1	531	27.3	29.3	1.02	5,029	3,982	~0
2	533	51.4	32.0	1.17	11,827	10,737	~0
3	274	38.5	30.7	1.11	32050	34,456	~1

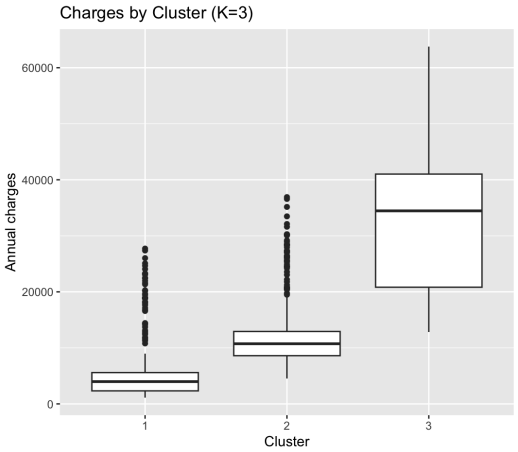


Figure 4.Cluster boxplot

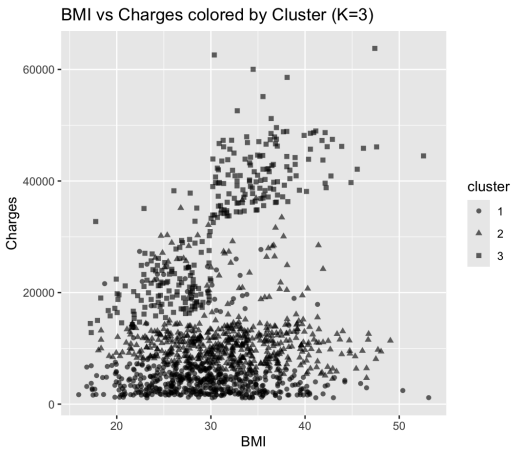


Figure 5.BMI vs Charges colored by Cluster

Overall, the clustering results strongly support Hypothesis H5. The identified clusters align closely with the key drivers of medical costs uncovered in earlier sections, particularly smoking status, age, and BMI. The high-risk smoker cluster mirrors the high-cost segment detected through logistic regression and decision tree models, reinforcing the conclusion that smoking is the dominant determinant of extreme medical expenditures. From a managerial perspective, this segmentation provides insurers with a transparent framework for risk-based pricing, targeted wellness interventions, and customer management strategies.

6 Integration and Conclusion

This project analyzes the determinants of medical insurance charges using regression, classification, and clustering methods. Across all approaches, smoking status consistently emerges as the most important driver of costs, with age and BMI providing additional explanatory power. Linear regression quantifies these effects and reveals a strong smoking-BMI interaction; classification models translate them into predictive decision rules;

and K-means clustering identifies distinct customer segments, including a high-risk group composed primarily of smokers.

These methods complement rather than duplicate each other, forming a coherent analytical framework that highlights the central role of lifestyle-related risk factors in insurance pricing and risk management.