Project Title: Predicting Medical Insurance Costs — A Comparative Regression Analysis

Otajon Yuldashev 473457

Motivation & Problem Definition

Insurance profitability hinges on **accurate risk assessment**. Underestimating medical costs exposes the insurer to **underwriting losses**, while overestimating drives customers to competitors with lower premiums. The challenge is to **price policies competitively yet profitably**, balancing precision and fairness.

Technical Challenge: Medical cost data is **inherently non-linear and skewed**:

- **Skewness:** The majority of clients incur low costs, but a **small, high-risk minority** (e.g., smokers, obese individuals, or those with chronic conditions) generates disproportionately high expenses. This right-skewed distribution complicates modeling, as traditional linear methods struggle to capture extreme values.

- **Interaction Effects:** Risk factors **amplify each other multiplicatively**. For example, smoking and high BMI individually increase costs, but their combination leads to **exponentially higher expenses**—a pattern linear models fail to address.

Project Objective

To **systematically compare three regression architectures**—Ridge Regression (linear baseline), Random Forest (bagging), and XGBoost (boosting)—and identify which best captures the **non-linear, interactive nature** of healthcare costs while retaining interpretability for stakeholders.

Dataset

- The dataset comprises **1,337 entries** (after removing 1 duplicate) with **7 features**: age, sex, BMI, number of children, smoker status, region, and medical charges.
- **Age Distribution:** Clients range from **18 to 64 years** (median = 39), representing a **diverse adult population** with a mix of younger and older individuals.
- **BMI Distribution:** The median BMI is **30.4 (overweight)**, with a significant proportion of clients likely falling into the obese category (BMI ≥ 30). This suggests a **high prevalence of lifestyle-related health risks** (e.g., diabetes, cardiovascular diseases).
- **Smoking Status:** A binary variable (yes/no), expected to be a **primary driver** of cost variability due to its strong association with chronic diseases (e.g., lung cancer, heart disease).

- **Family Size:** The median number of children is **1**, with a maximum of 5. Family size alone is unlikely to be a major cost driver unless interacting with other factors (e.g., pediatric care for younger parents).

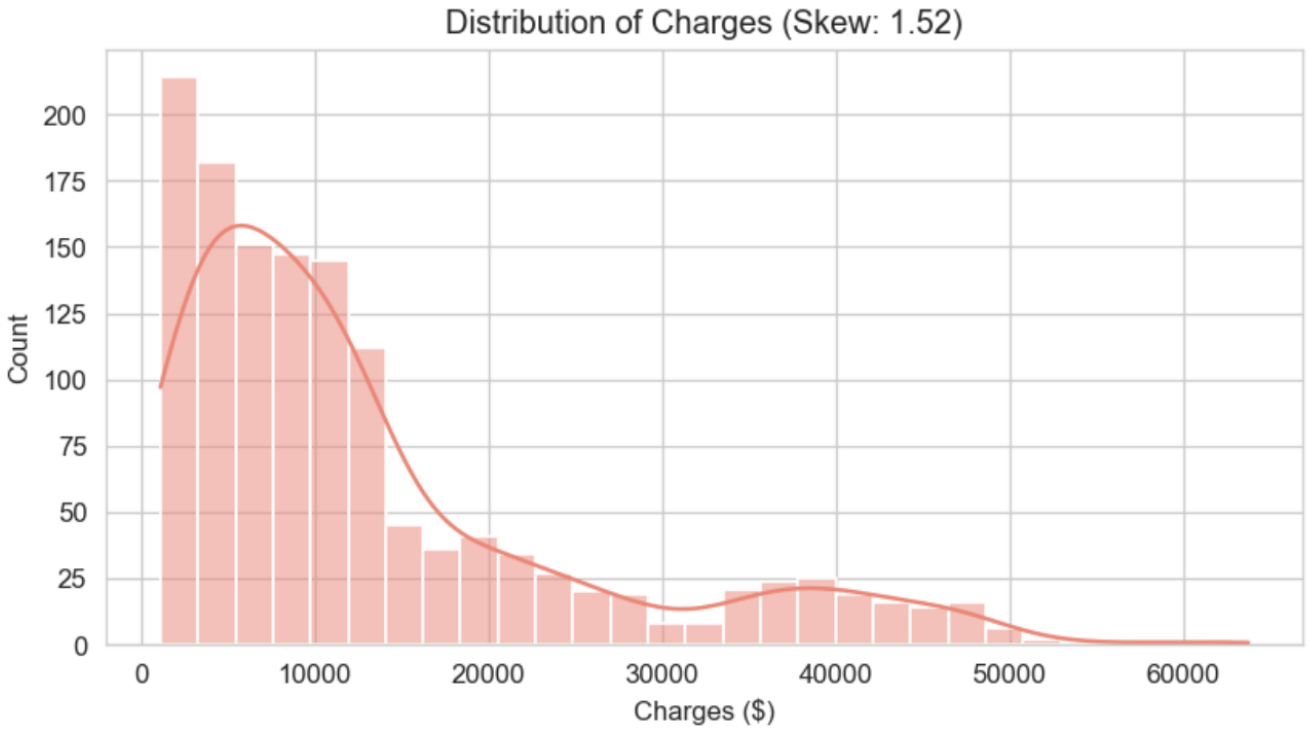
Medical Charges Distribution

- The distribution of medical charges is **highly right-skewed** (mean = $13,279, median = $9,386), indicating that most clients incur relatively low costs, but a **small subset drives the majority of expenses**.
- **25% of clients exceed $16,658** in charges, and the **maximum charge reaches $63,770** (over 5× the median). This reflects the **Pareto principle** in healthcare, where a minority of high-risk clients account for a disproportionate share of costs.
- The **gap between the 75th percentile ($16,658) and the maximum ($63,770)** underscores the presence of **catastrophic or chronic cases**, likely tied to **smoking, obesity, and age interactions**.

Key Implications for Modeling

- **Non-linearity and Interactions:** The data suggests that **risk factors interact multiplicatively** (e.g., smoking × BMI, age × smoking). Linear models may fail to capture these complexities.
- **Tree-Based Models:** Methods like **Random Forest and XGBoost** are theoretically superior for this dataset, as they **automatically detect and leverage interactions** without requiring manual feature engineering.

3. Exploratory Data Analysis (EDA)



Histogram (Skew: 1.52)

- Most clients have charges **below $15,000**, with a peak concentration around **$5,000–$10,000**, aligning with the median charge of **$9,386**.

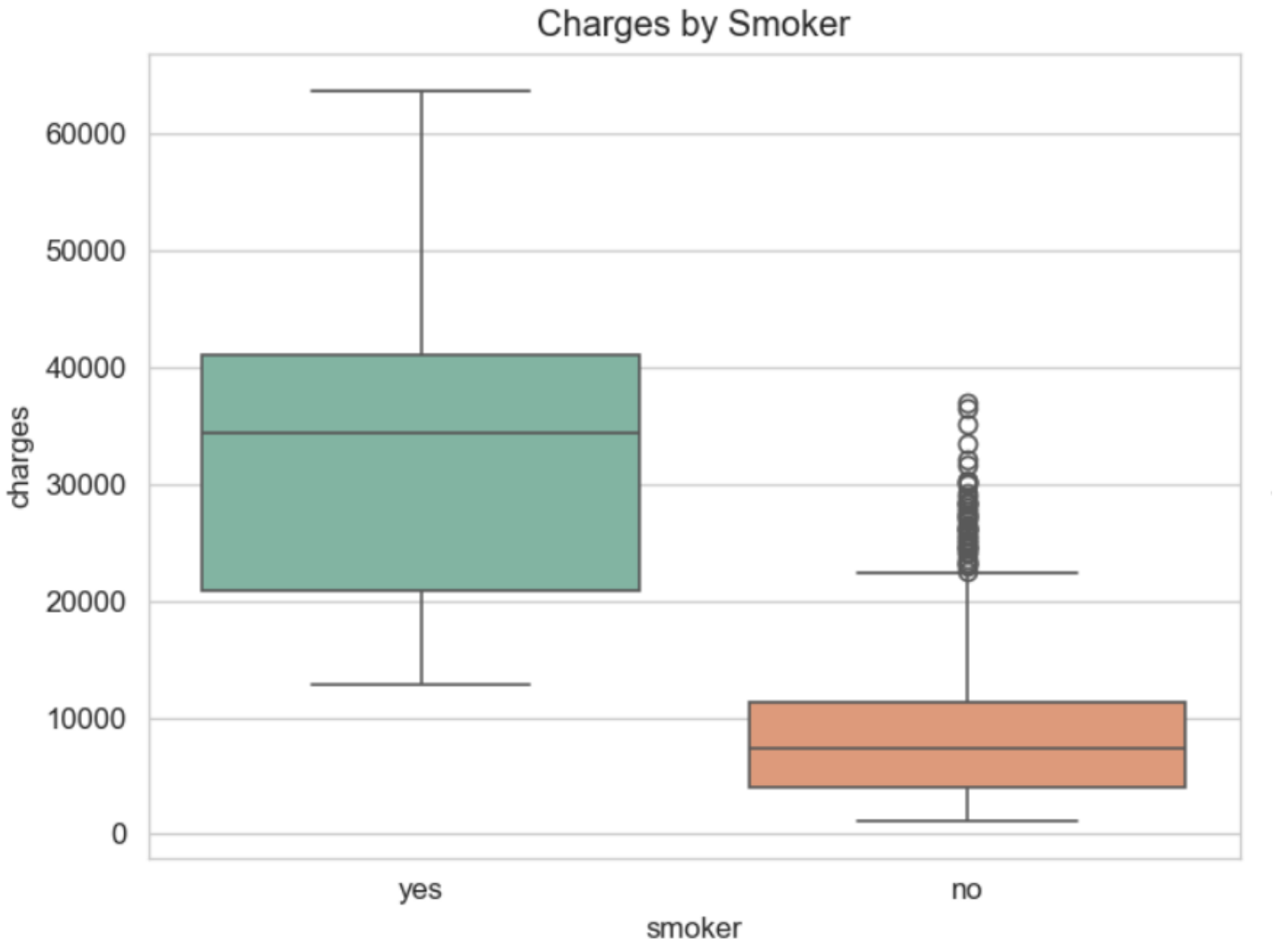
- A **long tail extends to $60,000+**, indicating the presence of **high-cost outliers** (e.g., chronic or catastrophic cases).
- A **secondary hump around $30,000–$40,000** suggests a subgroup of clients with moderately high charges, likely tied to specific risk factors (e.g., smokers, obese individuals).

Implications for Modeling and Business:

- **Model Selection:** The right skew and outliers suggest that **non-linear models** (e.g., Random Forest, XGBoost) are better suited than linear models. Alternatively, a **log transformation** of the target variable could improve linear model performance.

- **Risk Assessment:** Accurately predicting high-cost outliers is **critical to avoid underwriting losses**. These cases likely represent **high-risk clients** (e.g., smokers, obese individuals, or those with chronic conditions).

- **Business Strategy:** Insurers should **stratify clients by risk factors** (e.g., smoking, BMI, age) to price policies competitively while ensuring adequate reserves for high-cost cases. Ensuring adequate reserves for high-cost cases. The secondary hump may indicate a need for **specialized underwriting or wellness programs** for specific subgroups.



Smokers:

- The median charge for smokers is **approximately $30,000**, significantly higher than for non-smokers.
- The interquartile range (IQR) for smokers spans from **about $25,000 to $40,000**, indicating a wide spread of high medical costs.
- Smokers exhibit **multiple high-cost outliers**, with charges extending up to **$60,000+**, reflecting catastrophic or chronic health conditions.

Non-Smokers:

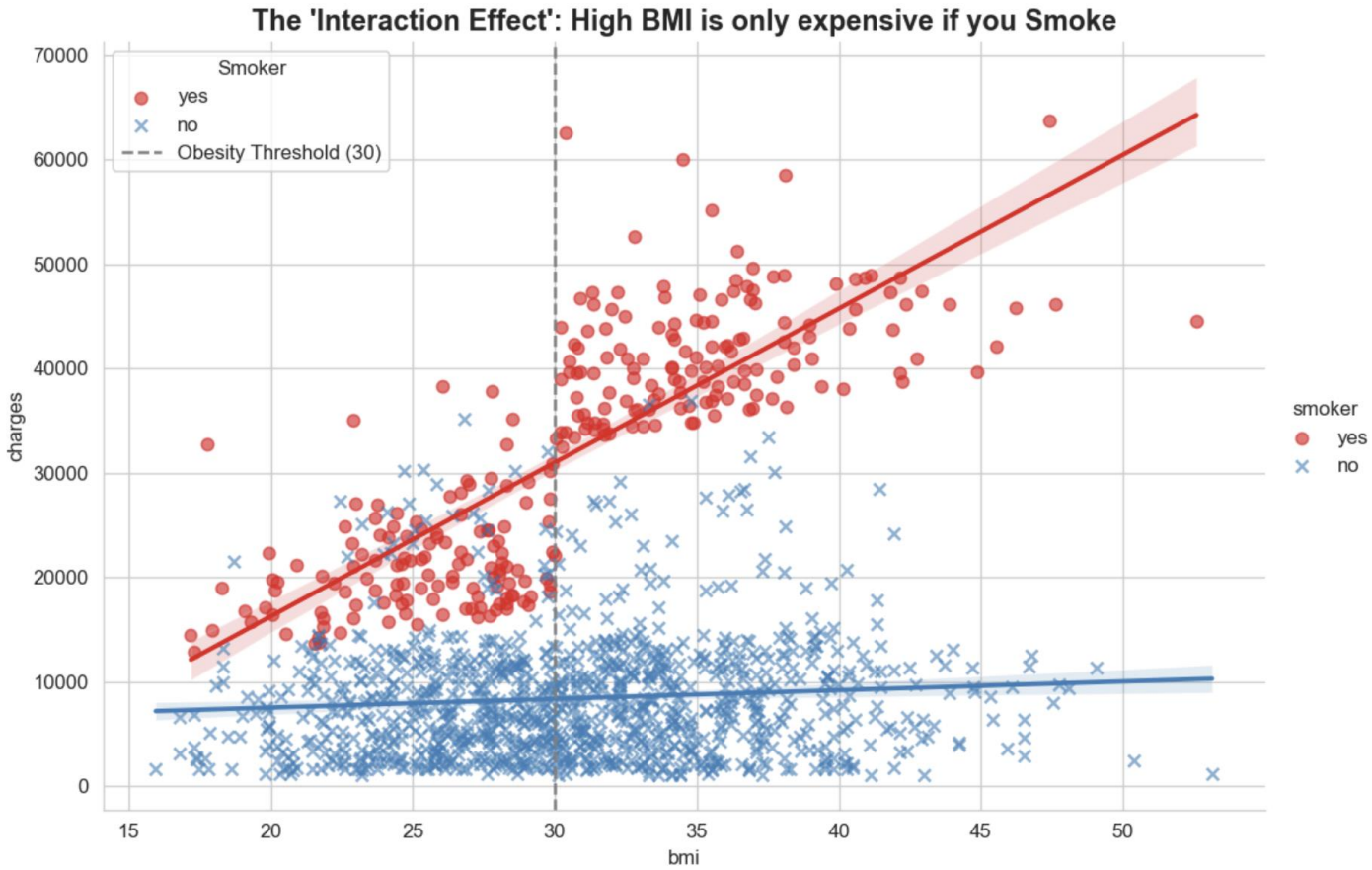
- The median charge for non-smokers is **around $7,000**, roughly **4 times lower** than that of smokers.
- The IQR for non-smokers ranges from **approximately $4,000 to $12,000**, indicating relatively lower and less variable medical costs.

Implications for Modeling and Business:

- **Risk Stratification:** Smoking status is a **critical factor** in determining medical costs, with smokers incurring **substantially higher charges** than non-smokers.

- **Modeling Approach:** Models must **effectively differentiate between smokers and non-smokers** to predict medical costs accurately. **Non-linear models** (e.g., Random Forest, XGBoost) are likely to perform well in capturing the distinct cost patterns associated with smoking.

- **Business Strategy: Targeted interventions**, such as smoking cessation programs, could help reduce medical costs for high-risk clients. Pricing strategies should reflect the **higher risk and costs** associated with smokers.



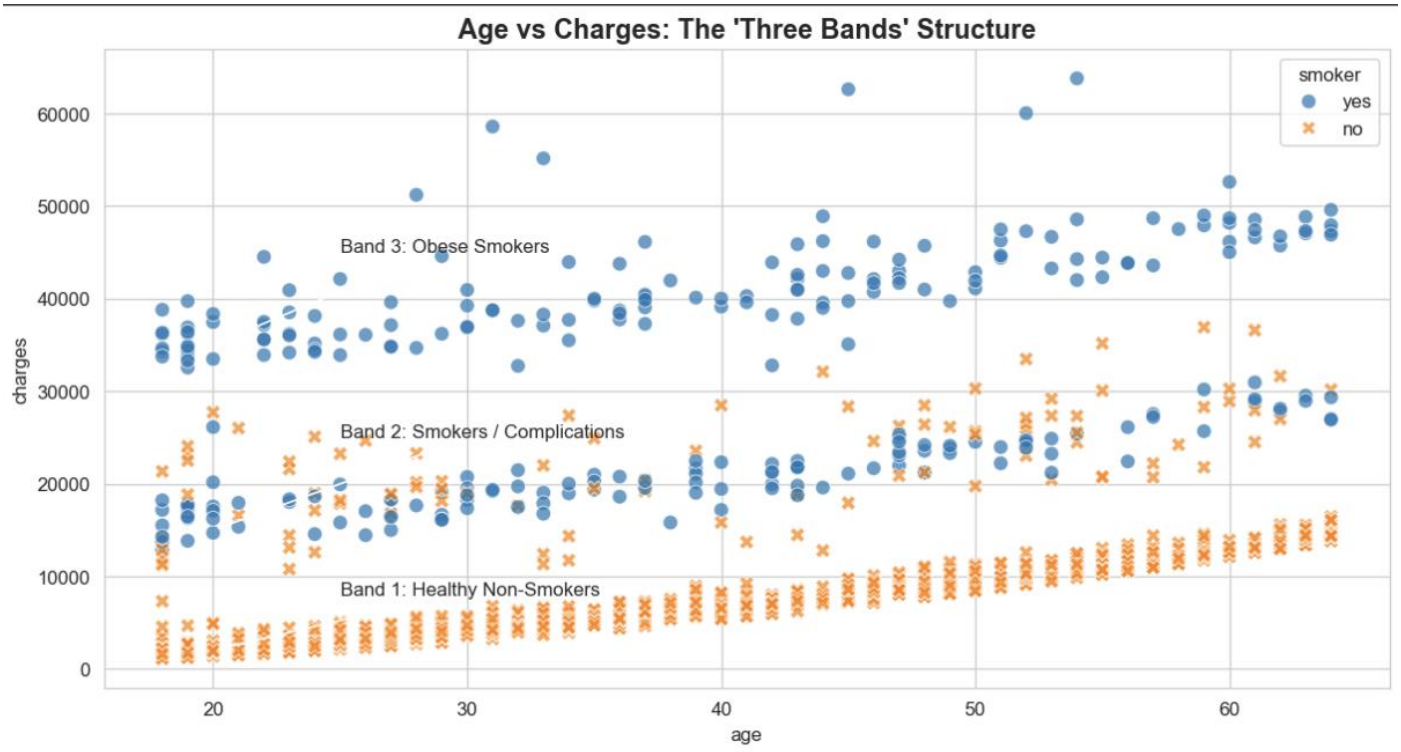
Smokers (Red Dots):

- There is a **strong positive relationship** between BMI and charges for smokers.
- As BMI increases, medical charges **rise steeply**, particularly for BMI values **above the obesity threshold (30)**.
- Smokers with high BMI (above 30) frequently incur charges **exceeding $40,000**, with some reaching **$60,000 or more**.
- **Non-Smokers (Blue Crosses):**
    - The relationship between BMI and charges is **weak and almost flat**.
    - Even at high BMI levels, non-smokers generally incur charges **below $20,000**, with minimal increase as BMI rises.

Implications for Modeling and Business:

- **Interaction Effect:** The plot clearly demonstrates a **multiplicative interaction** between smoking and BMI. High BMI is **only associated with high charges if the individual smokes**.
- **Modeling Approach:** Models must incorporate **interaction terms** (e.g., smoker × BMI) to capture this effect. **Non-linear models** (e.g., Random Forest, XGBoost) are well-suited to automatically detect and leverage such interactions.
- **Risk Stratification:** Insurers should **prioritize clients who are both smokers and have high BMI** for risk management and underwriting. **Wellness programs** targeting smoking cessation could be particularly effective in reducing medical costs for high-BMI individuals.

2. Age vs. Charges by Smoker Status



Band 1: Healthy Non-Smokers (Orange Crosses):

- Non-smokers consistently exhibit **low charges** ($0 to $15,000), regardless of age.

Band 2: Smokers / Complications (Lower Blue Dots):

- Smokers generally incur **moderate charges** ($15,000 to $30,000), with charges increasing slightly with age.

Band 3: Obese Smokers (Upper Blue Dots):

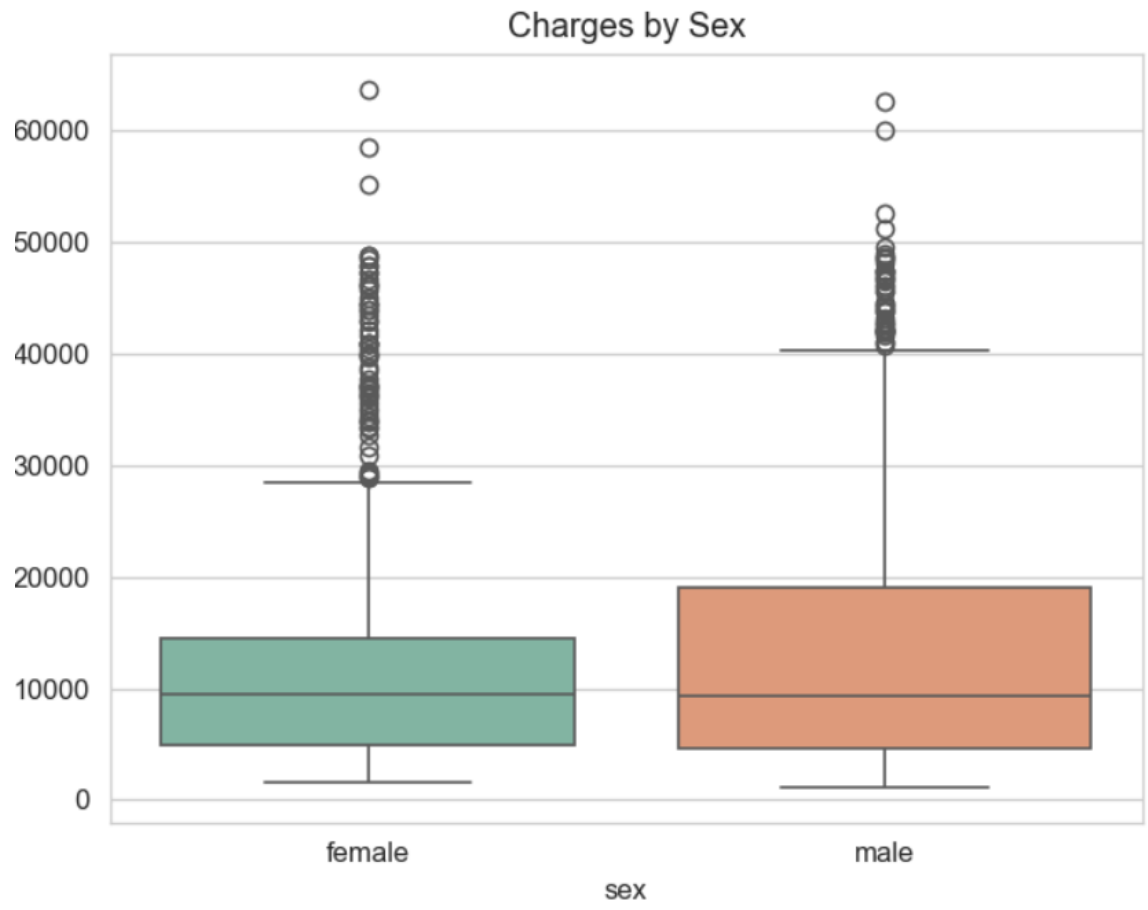
- A subset of smokers, likely those with additional risk factors such as obesity, incur **high charges** ($30,000 to $60,000), increasing with age.

Implications for Modeling and Business:

- **Risk Stratification:** The plot identifies **three distinct risk groups**, necessitating accurate risk assessment to distinguish between them.

- **Modeling Approach:** Models must capture the **non-linear and interactive effects** of age and smoking status on medical charges. **Non-linear models** (e.g., Random Forest, XGBoost) are well-suited to detect and leverage these complex patterns.

- **Business Strategy:** Pricing strategies should account for the **distinct risk profiles**, with higher premiums for smokers, particularly those in the high-charge band. **Wellness programs** targeting smoking cessation and weight management could help transition high-risk clients to lower-risk bands.



- The median charge for both females and males is approximately $9,000–$10,000, indicating similar central tendencies across sexes.

- The interquartile range (IQR), represented by the box, is comparable for both groups, suggesting similar variability in charges within the central 50% of the data.

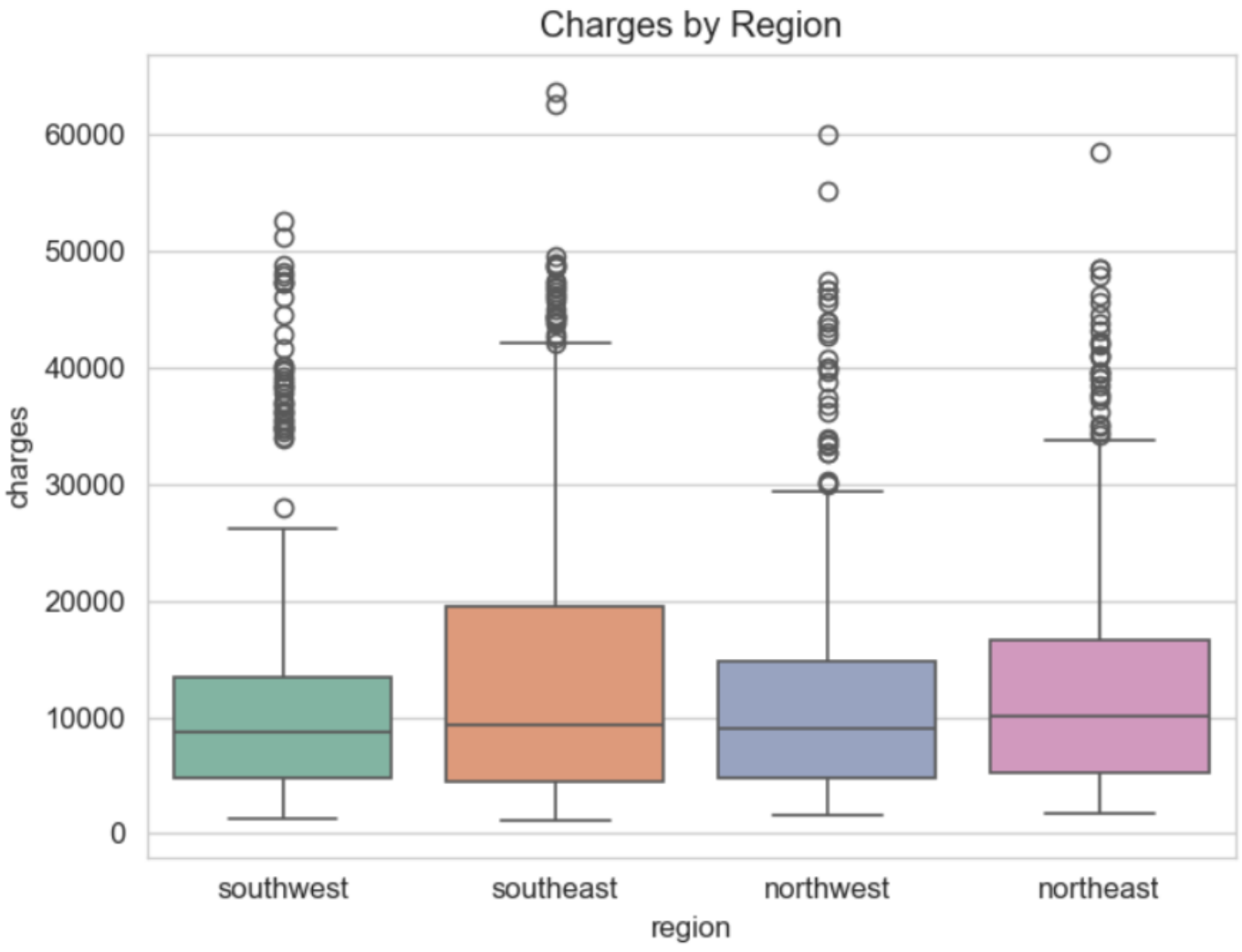
- The whiskers extend from approximately $0 to $30,000, capturing the typical range of charges for most individuals.

- Both distributions exhibit high-cost outliers, with charges exceeding $60,000. These outliers may indicate high-risk cases or extreme events, though further investigation is required to identify specific contributing factors.

Implications for Modeling and Business:

- **Minimal Impact of Sex:** Sex is **not a significant predictor** of medical charges in this dataset.

- **Focus on Other Risk Factors:** The analysis should prioritize other variables, such as **smoking status, BMI, and age**, which have a stronger impact on medical charges.

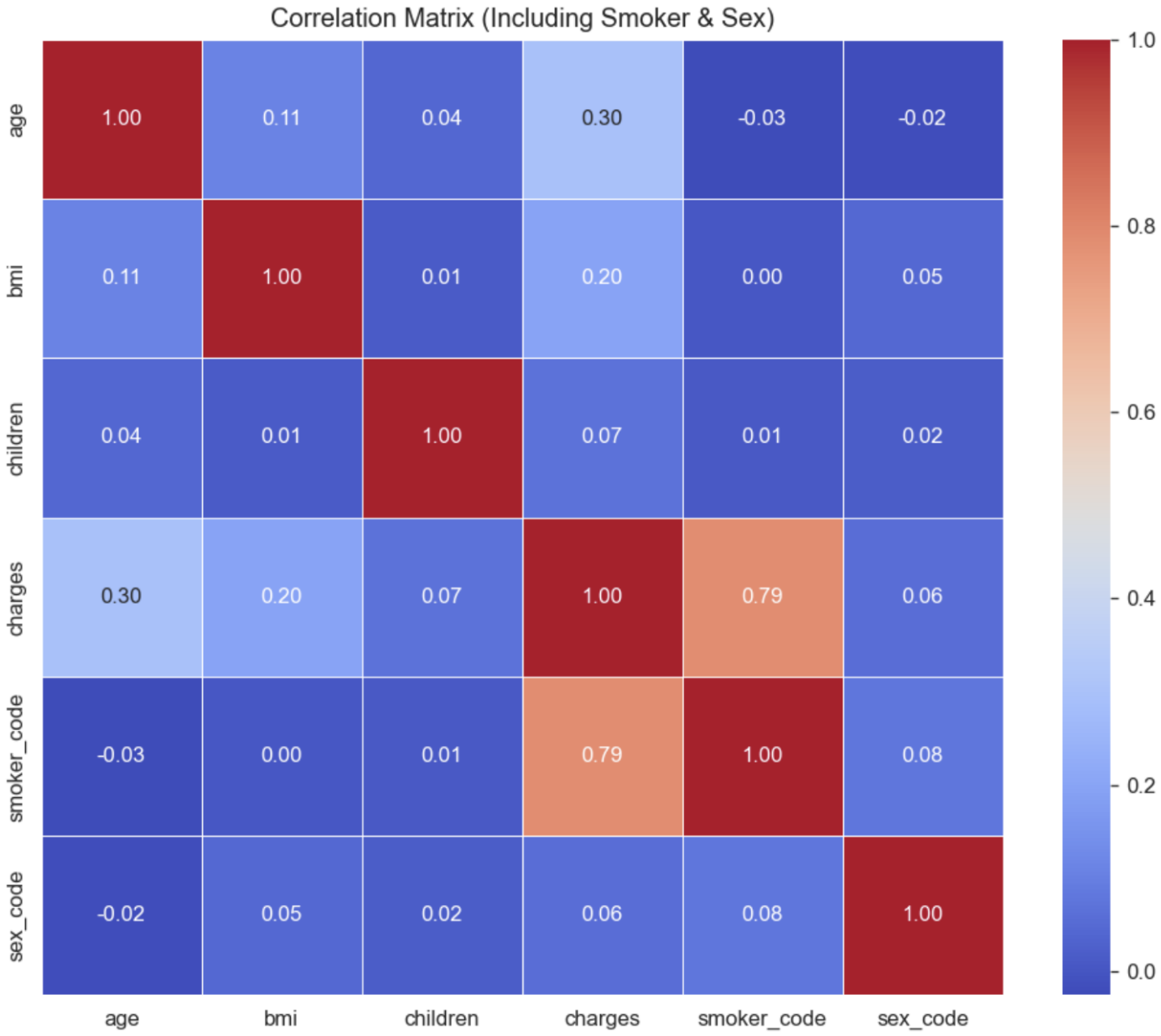


- The median charges for all regions are **similar**, ranging from approximately **$9,000 to $11,000**.
- The IQRs for all regions are **comparable**, indicating similar variability in charges within each region.
- All regions have **high-cost outliers** (up to $60,000+), likely representing high-risk individuals (e.g., smokers, obese individuals).

Implications for Modeling and Business:

- Region is **not a significant predictor** of medical charges in this dataset.

- The analysis should prioritize other variables, such as **smoking status, BMI, and age**, which have a stronger impact on medical charges.



- **Smoker Status (smoker_code):** Shows the **strongest positive correlation with charges (0.79)**, indicating that smoking is the **dominant risk factor** for higher medical costs.

- **Age:** Has a **moderate positive correlation with charges (0.30)**, suggesting that older clients tend to incur higher medical expenses.

- **BMI:** Exhibits a **positive correlation with charges (0.20)**, indicating that higher BMI is associated with increased medical costs.

- **Children/Sex:** Display **negligible correlations with charges (0.07 and 0.06, respectively)**.

Implications for Modeling and Business:

- **Feature Importance:** Smoker status should be prioritized as the **most influential predictor** in modeling medical charges. Age and BMI are also relevant but less impactful.
- **Interaction Effects:** While the correlation matrix highlights linear relationships, **interaction effects** (e.g., smoking × BMI, smoking × age) are likely to be significant and should be explored further.

Key EDA Findings

1. **Smoking is the dominant risk factor**, driving **4× higher median charges** compared to non-smokers.
2. **BMI and age amplify costs for smokers**, creating **explosive interaction effects** that must be captured in modeling.
3. **Sex and region** have **minimal impact** on charges; children are negligible.
4. **Skewness and outliers** necessitate **non-linear models** (XGBoost/Random Forest) or log transformation for linear models.

Interpretation of Preprocessing, Model Tuning, and Final Evaluation

1. Feature Engineering: Capturing Key Insights from EDA

What We Did:

- Created an **interaction term** (bmi_smoker = bmi × smoker_status) to capture the **multiplicative effect** of BMI and smoking on medical charges.
- Applied a **log transformation** (np.log1p) to the target variable (charges) to address skewness.
- Split the data into **80% training and 20% test sets**, using **stratified sampling** by smoker status.

Why We Did It:

- The EDA revealed a **strong interaction effect** between BMI and smoking. The interaction term allows the model to capture this relationship explicitly.
- The **right-skewed distribution** of charges necessitated a log transformation to normalize the target variable and improve model performance.

- Stratified sampling ensures that the **minority class (smokers)**, which drives a significant portion of the costs, is adequately represented in both training and test sets.

3. Model Training & Hyperparameter Tuning: Finding the Best Parameters

A. Ridge Regression: Tuning Regularization Strength

What We Did:

- Tuned the **regularization parameter (alpha)** to control model complexity.
- Used **GridSearchCV** with **5-fold cross-validation** and **negative RMSE** as the scoring metric.

Best Parameters:

- **Alpha: 10.0**

Why This Matters:

- A higher alpha reduces overfitting but may increase bias. The optimal alpha of 10.0 balances these trade-offs.
- The **tuning curve** (Negative RMSE vs. Alpha) shows how model performance varies with different alpha values, flattening out at higher values.

B. Random Forest: Tuning Tree Complexity and Ensemble Size

What We Did:

- Tuned the **number of trees (n_estimators)**, **tree depth (max_depth)**, and **minimum samples per leaf (min_samples_leaf)**.
- Used **GridSearchCV** with **3-fold cross-validation** and **negative RMSE** as the scoring metric.

Best Parameters:

- **n_estimators: 100**
- **max_depth: 5**
- **min_samples_leaf: 4**

Why This Matters:

- **100 trees** are sufficient to capture the data patterns without excessive computational cost.
- A **maximum depth of 5** prevents overfitting while allowing the model to capture non-linear relationships.
- **4 minimum samples per leaf** further reduces overfitting by ensuring that each leaf contains a reasonable number of samples.
- The **OOB (Out-of-Bag) score vs. number of trees** plot shows how the model's performance stabilizes as the number of trees increases, demonstrating the robustness of Random Forest.

C. XGBoost: Tuning Learning Rate and Tree Complexity

What We Did:

- Tuned the **number of trees (n_estimators)**, **learning rate (learning_rate)**, and **tree depth (max_depth)**.
- Used **GridSearchCV** with **3-fold cross-validation** and **negative RMSE** as the scoring metric.

Best Parameters:

- **n_estimators: 100**
- **learning_rate: 0.05**
- **max_depth: 3**

Why This Matters:

- **100 trees** are sufficient for this dataset, similar to Random Forest.
- A **lower learning rate (0.05)** allows the model to learn more gradually, often leading to better generalization.
- A **shallower tree depth (3)** prevents overfitting while capturing the essential patterns in the data.

4. Intermediate Evaluation: Log-Scale Performance

Evaluated the models on the **log-transformed test set** using RMSE and R^2.

Results:

Model	RMSE (log scale)	R^2 (log scale)
Ridge	0.3480	0.8547
Random Forest	0.2529	0.9233
XGBoost	0.2579	0.9202

- **Random Forest** outperformed the other models on the log scale, achieving the **lowest RMSE (0.2529)** and **highest R^2 (0.9233)**.

- **XGBoost** performed comparably to Random Forest but with a slightly higher RMSE and lower R^2.

- **Ridge Regression** lagged behind, confirming that **non-linear models are better suited** for this
- dataset.

5. Final Evaluation: Real-Dollar Performance

- **Converted predictions** back to the original dollar scale using np.expm1.

- **Evaluated models** using RMSE, MAE, and R^2 on the original scale to understand the actual monetary impact of prediction errors.

Final Leaderboard:

Model	RMSE ($)	MAE ($)	R^2
Random Forest	3280.54	1601.93	0.9253
XGBoost	3332.75	1591.77	0.9230
Ridge	7710.34	3656.92	0.5876

Random Forest: The Optimal Model

- **RMSE ($3,280.54):** On average, predictions are within $3,280 of actual medical charges, indicating **high precision** for most cases. This level of accuracy is crucial for setting competitive yet profitable insurance premiums.
- **MAE ($1,601.93):** The average prediction error of $1,602 provides a practical measure of accuracy, suggesting that the model is **reliable for day-to-day underwriting decisions**.
- **R^2 (0.9253):** Explains **92.53% of the variance** in medical charges, demonstrating a **strong fit** and the model's ability to capture the underlying patterns in the data.

- Random Forest is the **top-performing model**, offering a balance of accuracy, reliability, and interpretability. Its ability to capture non-linear relationships and interaction effects makes it well-suited for predicting medical charges in a real-world insurance context.

XGBoost: A Strong Contender

- **RMSE ($3,332.75):** Slightly higher than Random Forest, indicating **marginally less accurate predictions**, but still within an acceptable range for practical applications.
- **MAE ($1,591.77):** Comparable to Random Forest, reinforcing that **average prediction errors are similar** between the two models.
- **R^2 (0.9230):** Explains **92.30% of the variance**, nearly as strong as Random Forest, indicating a **highly effective model**.

- XGBoost is a **close second**, offering nearly equivalent performance to Random Forest. It could be considered as an alternative if specific use cases or computational constraints favor its implementation.

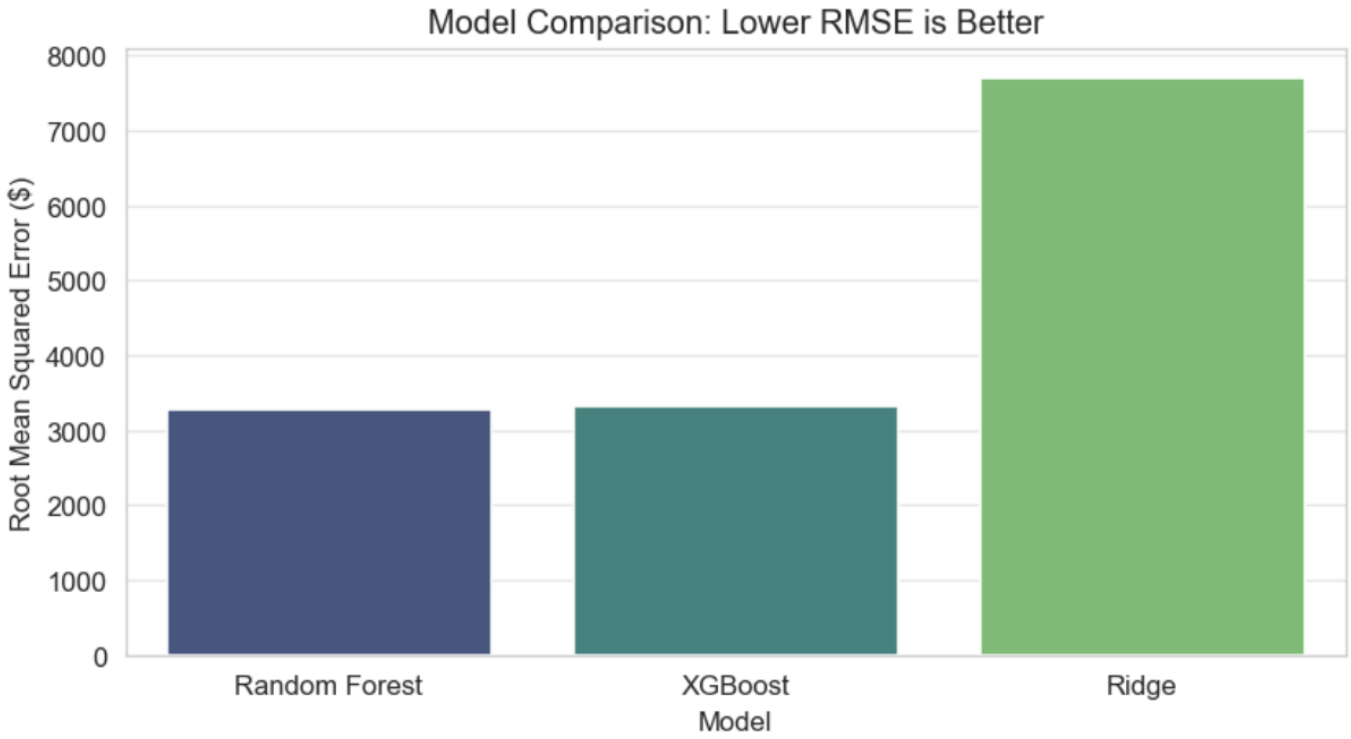
Ridge Regression: Inadequate for Non-Linear Data

- **RMSE ($7,710.34):** Significantly higher, indicating **much larger prediction errors** and a lack of precision, particularly for high-cost cases.
- **MAE ($3,656.92):** More than double that of Random Forest and XGBoost, highlighting **substantial average errors** that could lead to inaccurate premium pricing.
- **R^2 (0.5876):** Explains only **58.76% of the variance**, indicating a **weak fit** and inability to capture the complexities of the data.

- Ridge Regression performs poorly, confirming that **linear models are inadequate** for this dataset. The non-linear relationships and interaction effects present in the data require more sophisticated models like Random Forest or XGBoost.

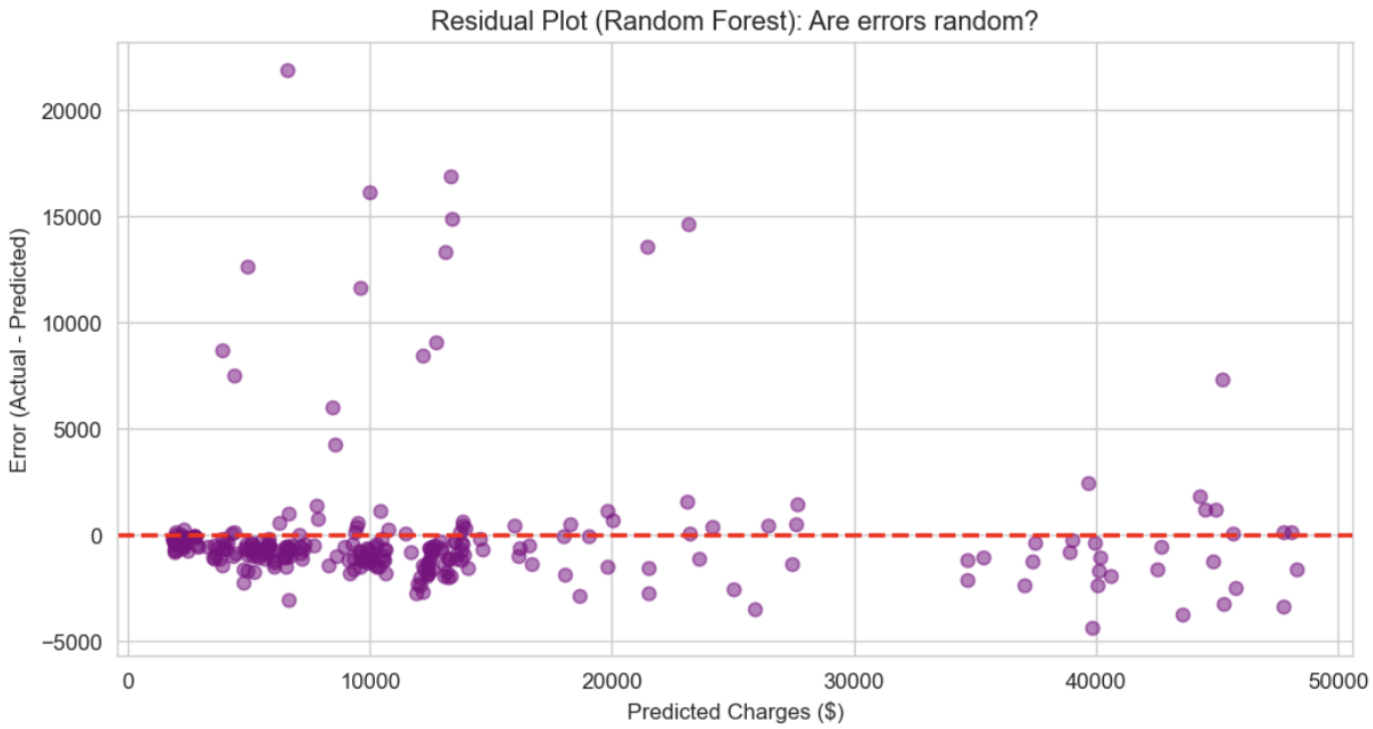
6. Performance Visuals: Understanding Model Behavior

A. Model Comparison (Bar Chart)



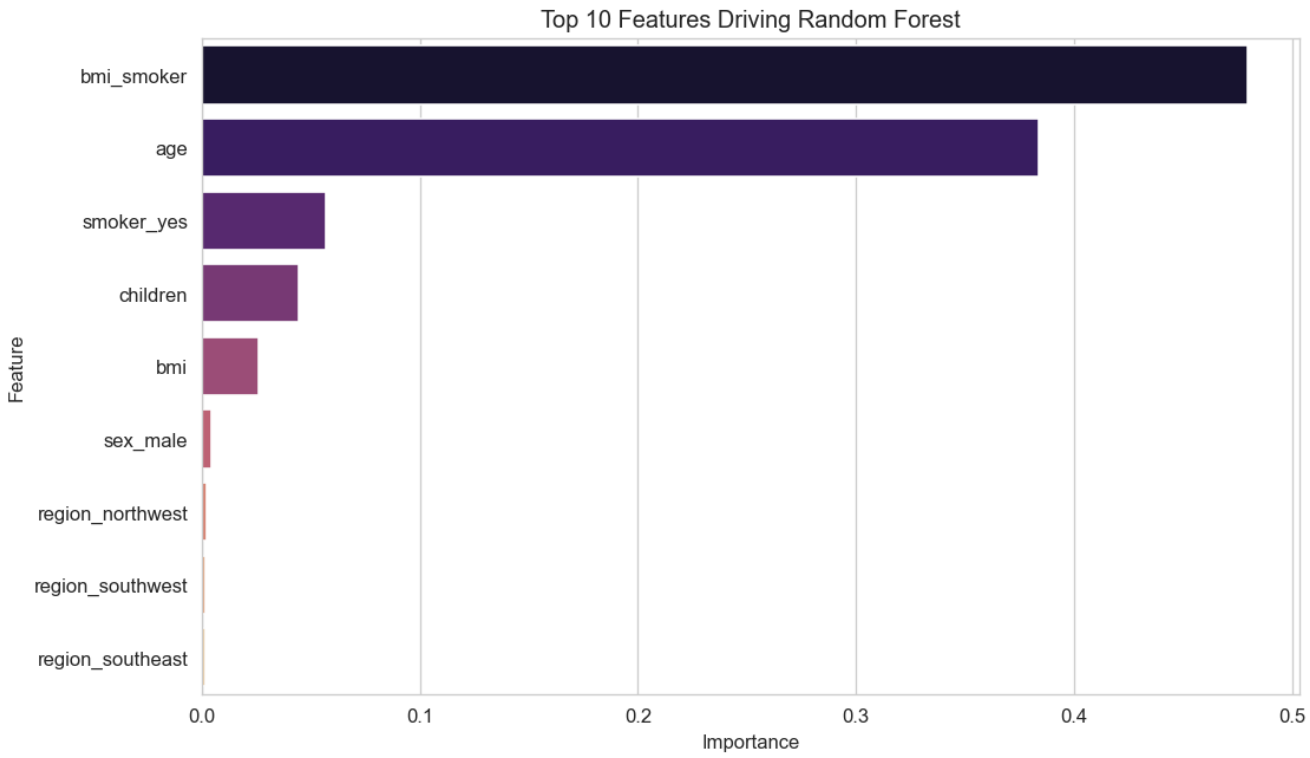
- The bar chart comparing the Root Mean Squared Error (RMSE) in dollars for Random Forest, XGBoost, and Ridge models shows that Random Forest has the lowest RMSE, followed closely by XGBoost, while Ridge has a significantly higher RMSE. This indicates that Random Forest is the most accurate model for predicting medical charges, with XGBoost performing nearly as well, while Ridge Regression, being a linear model, struggles to capture the non-linear relationships and interaction effects present in the data. The substantial difference in RMSE values highlights the importance of using non-linear models for this dataset, as they significantly outperform the linear baseline, ensuring more accurate predictions of medical charges.

B. Residual Plot (Random Forest)



- The residual plot for the Random Forest model shows that residuals (actual − predicted charges) are mostly centered around zero, indicating generally accurate predictions and no strong systematic bias in the central range of the data. The errors appear largely randomly distributed; however, their variance increases with higher predicted charges, suggesting heteroscedasticity.
- Notably, there are larger positive residuals for high-cost cases (above approximately $30,000), indicating that the model occasionally underestimates insurance charges for high-risk individuals. This behavior is common in right-skewed datasets with extreme outliers and reflects increased uncertainty in predicting rare, high-cost events.
- From a business perspective, this underestimation is particularly relevant for insurance underwriting, as it may lead to underpricing of high-risk clients. Consequently, additional risk management measures—such as safety margins or supplementary risk factors—may be required.
- For further model improvement, additional feature engineering or alternative modeling approaches could be considered to better capture variability in high-cost cases. Continuous validation is also important to ensure stable performance over time.

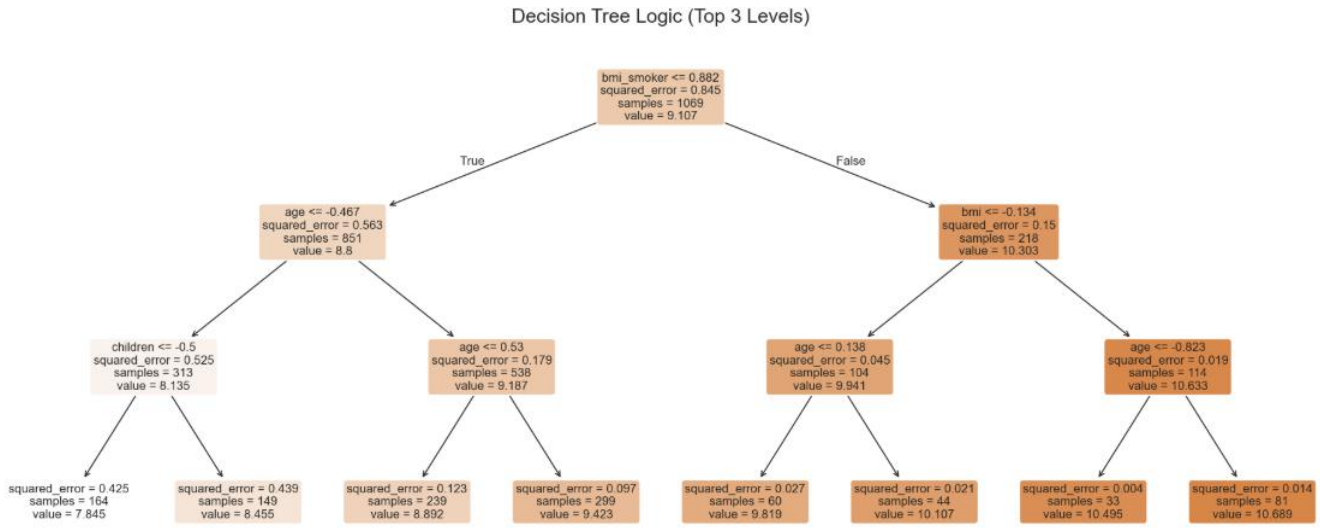
C. Feature Importance (Random Forest)



- This bar chart illustrates the top 10 features driving the Random Forest model's predictions for medical charges. The feature "bmi_smoker" is by far the most important, indicating that the interaction between BMI and smoking status is the strongest predictor of medical charges. Age is the second most important feature, followed by smoking status and the number of children. Features related to sex and region have minimal importance. This aligns with the EDA findings, which highlighted the significant impact of smoking and BMI on medical costs, and supports the decision to focus on these variables for modeling. The dominance of "bmi_smoker" underscores the importance of capturing interaction effects in the model.

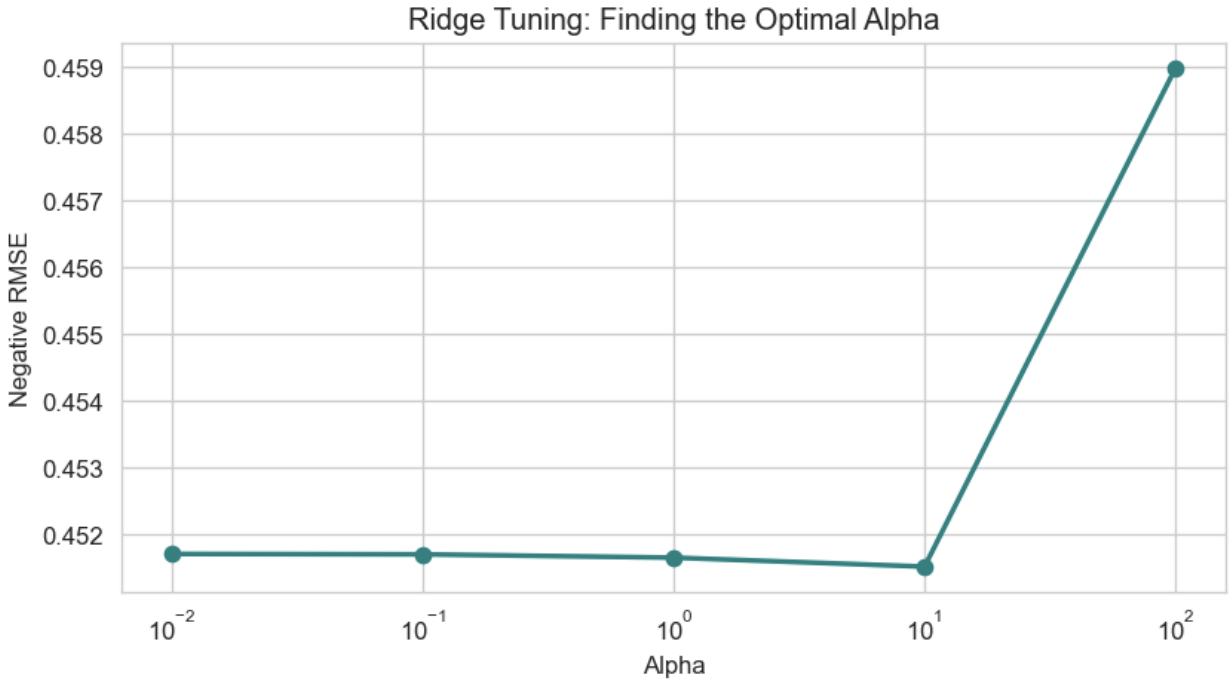
7. Process Visuals: Demonstrating Model Logic

A. Decision Tree Diagram



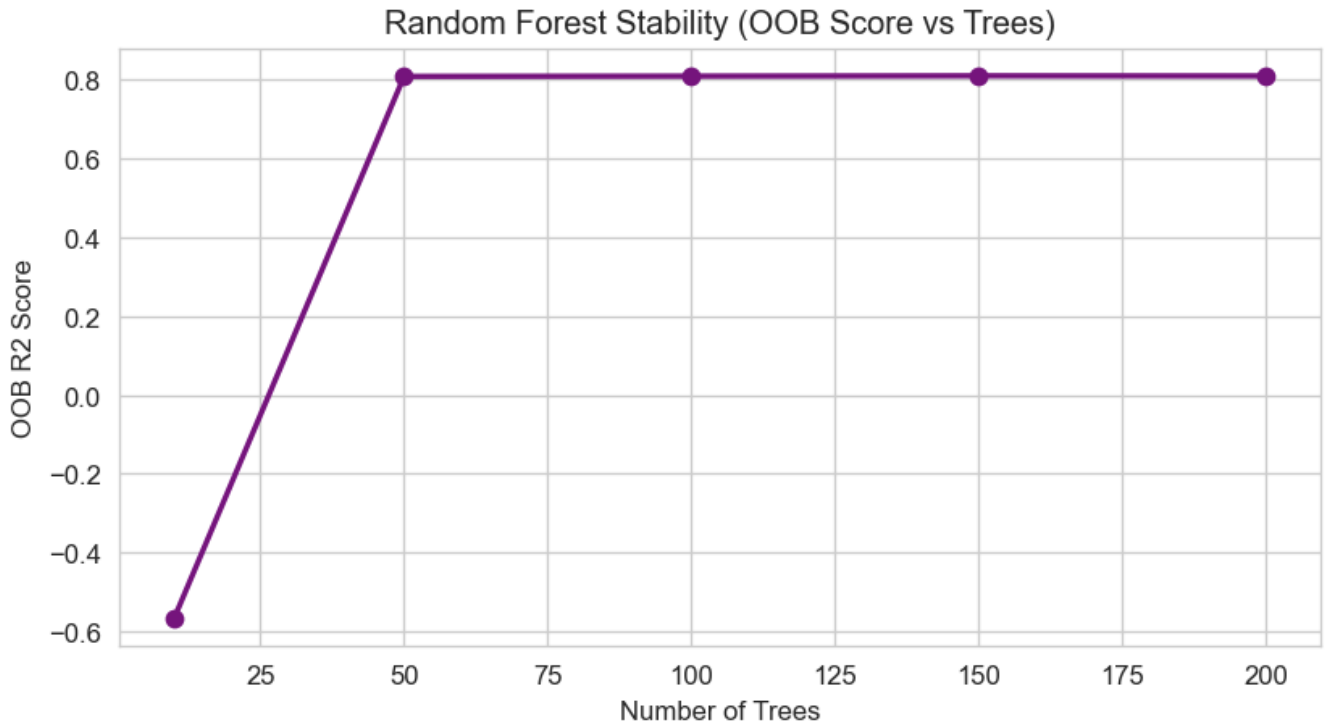
- This decision tree diagram illustrates the logic of the top 3 levels of a decision tree used in the Random Forest model for predicting medical charges. The first split is based on the "bmi_smoker" feature, confirming its importance as the primary driver of predictions, which aligns with the feature importance chart. If "bmi_smoker" is less than or equal to 0.882, the tree further splits on age, and then on the number of children. If "bmi_smoker" is greater than 0.882, the tree splits on BMI and then on age. Each node shows the feature and threshold used for splitting, the squared error, the number of samples, and the predicted value for that node. This visualization demonstrates how the model uses features to make predictions and confirms that "bmi_smoker" and age are critical factors, supporting the EDA findings and the feature importance analysis.

B. Ridge Tuning Curve



This plot illustrates the tuning process for Ridge Regression by showing the relationship between different Alpha values and the Negative Root Mean Squared Error (RMSE). The x-axis represents Alpha values on a logarithmic scale, and the y-axis represents the Negative RMSE. For lower Alpha values (0.01 to 1), the Negative RMSE remains relatively constant, indicating that the model performance does not change significantly. However, at Alpha = 100, there is a sharp increase in Negative RMSE, indicating poorer model performance. The optimal Alpha value, found to be 10, balances model complexity and performance, providing the best trade-off between bias and variance. This tuning process ensures that the Ridge Regression model is not overfitting while maintaining predictive power.

C. Random Forest Stability (OOB Score)



- This plot illustrates the stability of the Random Forest model by showing the Out-of-Bag (OOB) R^2 Score as a function of the number of trees. The x-axis represents the number of trees, and the y-axis represents the OOB R^2 Score. The plot shows a steep increase in the OOB R^2 Score from 25 to 50 trees, after which the score stabilizes around 0.8. This indicates that the model's performance significantly improves as the number of trees increases up to 50, and then plateaus, suggesting that adding more trees beyond this point has diminishing returns on performance. The stability of the OOB R^2 Score after 50 trees demonstrates the robustness of the Random Forest model and confirms that 100 trees, as chosen during hyperparameter tuning, are sufficient for capturing the data patterns without excessive computational cost.

Final Conclusion

This analysis successfully identified **Random Forest** as the optimal model for predicting medical insurance charges, outperforming both XGBoost and Ridge Regression in accuracy and robustness. The model's strength lies in its ability to capture **non-linear relationships and interaction effects**, particularly the critical interaction between BMI and smoking status, which emerged as the most influential predictor.

While the model demonstrates strong overall performance, it occasionally **underestimates high-cost cases**, suggesting a need for enhanced risk management strategies for high-risk clients. The transparency and interpretability of the Random Forest model make it well-suited for practical implementation in insurance underwriting, enabling **data-driven pricing and risk stratification**.

For future work, focusing on improving predictions for high-cost cases—through additional feature engineering or alternative modeling techniques—could further enhance the model's utility in real-world applications.