Introduction

Health insurance and medical bills are integral aspects of our everyday lives, with medical charges being influenced by various factors such as age, gender, lifestyle, and pre-existing health conditions. Understanding these factors is crucial for optimizing healthcare costs and improving policy decisions.

This study leverages an intriguing dataset from an insurance company, which includes medical charges for individuals with diverse backgrounds. The dataset consists of seven variables: Medical Charges, Age, Body Mass Index (BMI), Number of Children, Gender, Smoking Status, and Region.

The goal of this comprehensive analysis is to uncover patterns, relationships, and anomalies within the data to gain a deeper understanding of the factors associated with medical charges. This involves a multi-step approach:

1. **Exploratory Data Analysis (EDA)**: To visualize data distributions, identify potential correlations, and detect outliers.
2. **Correlation Analysis**: To assess the strength and direction of relationships between medical charges and the other variables.
3. **Segmentation and Statistical Analysis**: To examine differences in medical charges across various demographic and lifestyle segments.
4. **Classification Analysis**: To predict the groupings of medical charges and identify important features using classification models.
5. **Regression Analysis**: To predict actual medical charges and quantify the impact of each significant feature.

By employing these methods, we aim to provide valuable insights into the factors driving medical charges, which can inform healthcare providers and policymakers in their efforts to enhance resource allocation, patient care strategies, and cost management.

Data Validation

We began by importing the necessary libraries and loading the dataset. To understand the structure of the dataset and validate the data, we performed several checks.

Step 1: Variable Definitions and Checking for Missing Values and Data Types

We verified the presence of missing values and confirmed the data types of each variable. The dataset consists of 1,338 records with no missing values. The data types for all columns are correctly identified: age (int64), sex (object), bmi (float64), children (int64), smoker (object), region (object), and charges (float64).

Step 2: Summary Statistics for Numerical Variables

We examined the summary statistics of the numerical variables to check for any numerical values that do not make sense and to verify data ranges:

- **Age**: 18 to 64 years (mean: 39).
- **BMI**: 15.96 to 53.13 kg/m² (mean: 30.66).
- **Number of Children**: 0 to 5 (mean: 1.1).
- **Medical Charges**: $1,121.87 to $63,770.43 (mean: $13,270.42).

All numerical variables fall within plausible ranges, with no values appearing as outliers or errors.

Step 3: Checking for Data Consistency in Categorical Variables

We examined the unique values for each categorical variable to ensure data consistency:

- **Gender**: 'female' and 'male'.
- **Smoking Status**: 'yes' and 'no'.
- **Region**: 'southwest', 'southeast', 'northwest', and ‘northeast'.

The unique values for each categorical variable are as expected, and there is no need for transformations due to inconsistencies.

Step 4: Duplicate Handling

We checked for duplicate rows to ensure data integrity. We identified and compared two duplicated rows. We removed the second duplicated row, keeping only the first occurrence. The number of records after removing duplicates is 1,337.

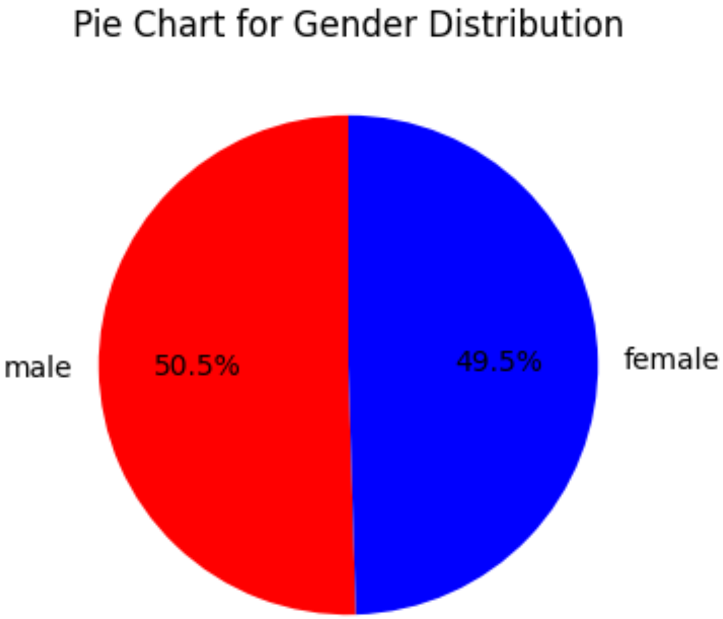
Exploratory Data Analysis

In this exploratory data analysis (EDA), we visually examine the distribution of each variable to gain an initial understanding of the data's structure and key characteristics.

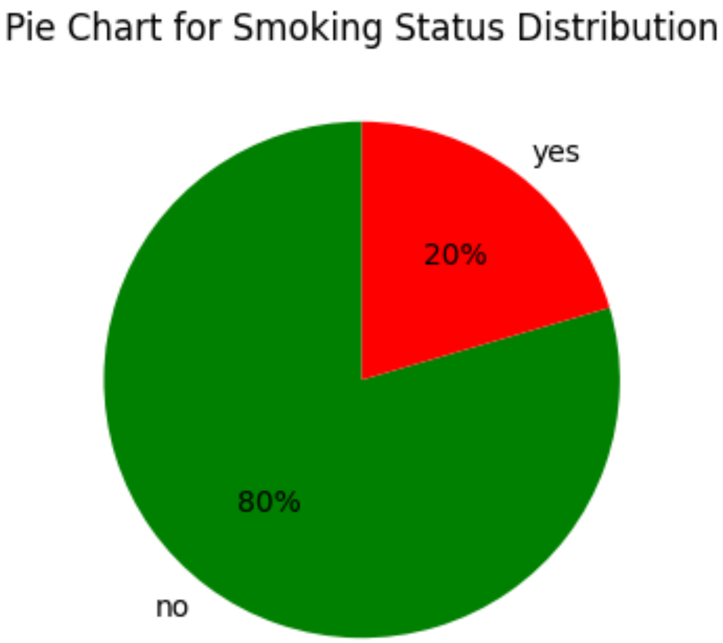
Step 1: Visualizing Categorical Variables

To explore the distribution of categorical variables, we create pie charts to illustrate the part-to-whole relationships for Gender, Smoking Status, Region, and Number of Children.

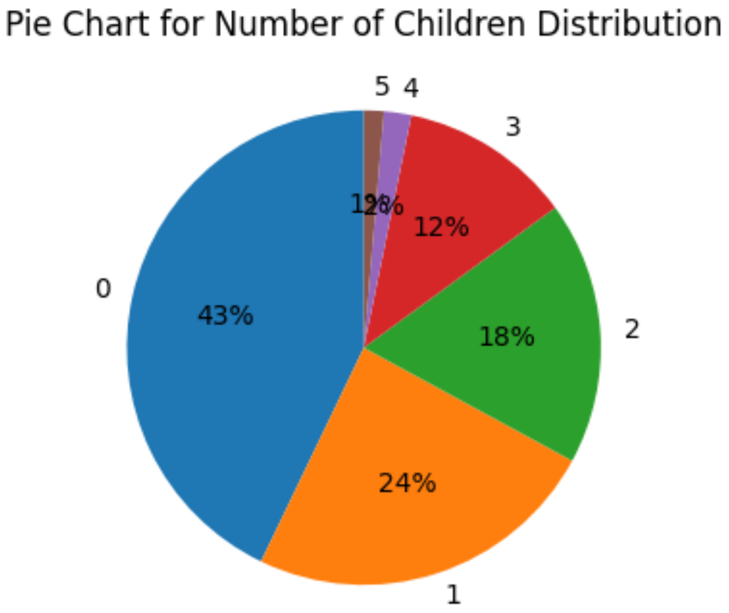
- **Gender:** The pie chart below shows the gender composition of the dataset. The distribution is relatively balanced with 50.5% male and 49.5% female.



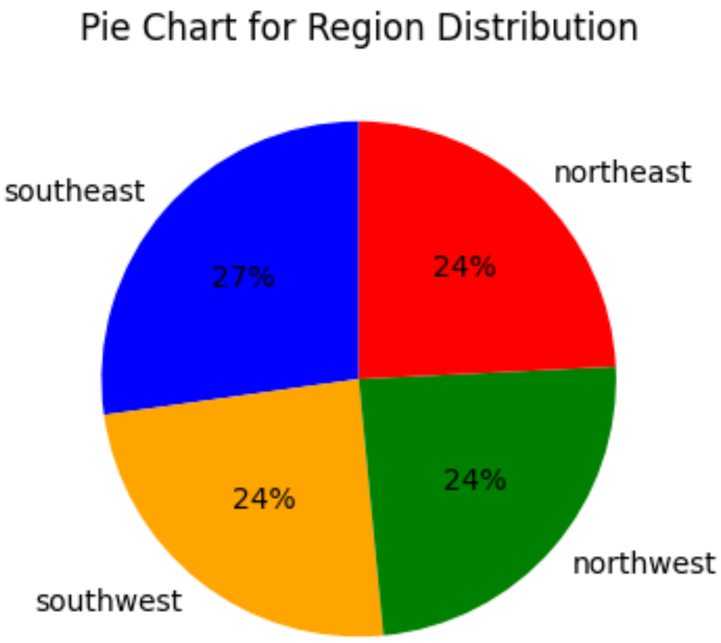
- **Smoking Status:** The pie chart below shows the smoking status composition of the dataset. A significant majority of the sample are non-smokers (80%) compared to smokers (20%).



- **Number of Children:** The pie chart below shows the number of children composition of the dataset. The distribution is as follows:
  - 43% of the sample has no children.
  - 24% has one child.
  - 18% has two children.
  - 12% has three children.
  - 2% has four children.
  - 1% has five children.



- **Region:** The pie chart below shows the region composition of the dataset. The distribution is relatively balanced across the regions with 27% from the southeast, and 24% from each of the southwest, northwest, and northeast regions.



Step 2: Visualizing Numerical Variables

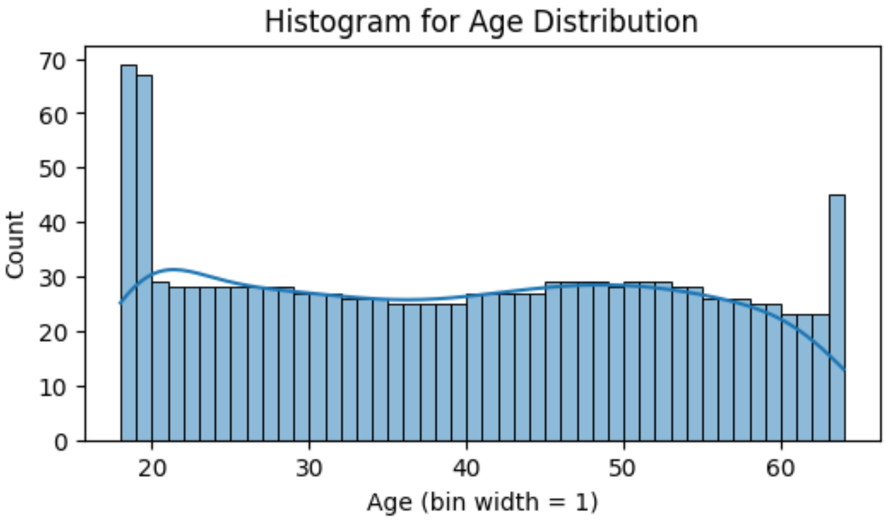
Summary statistics are provided below, followed by histograms to visualize the numerical variables, including Age, BMI, and Medical Charges.

Summary Statistics for Numerical Variables:

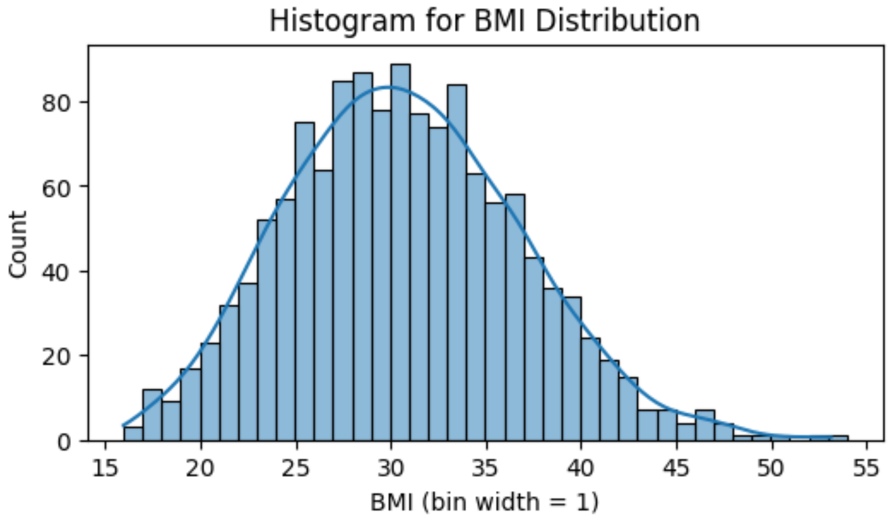
Statistic	Age	BMI	Charges
Count	1,338	1,338	1,338
Mean	39.21	30.66	13,270.42
Std Dev	14.05	6.10	12,110.01
Min	18	15.96	1,121.87
25th Pctl	27	26.30	4,740.29
Median	39	30.40	9,382.03
75th Pctl	51	34.69	16,639.91
Max	64	53.13	63,770.43

- **Age:**

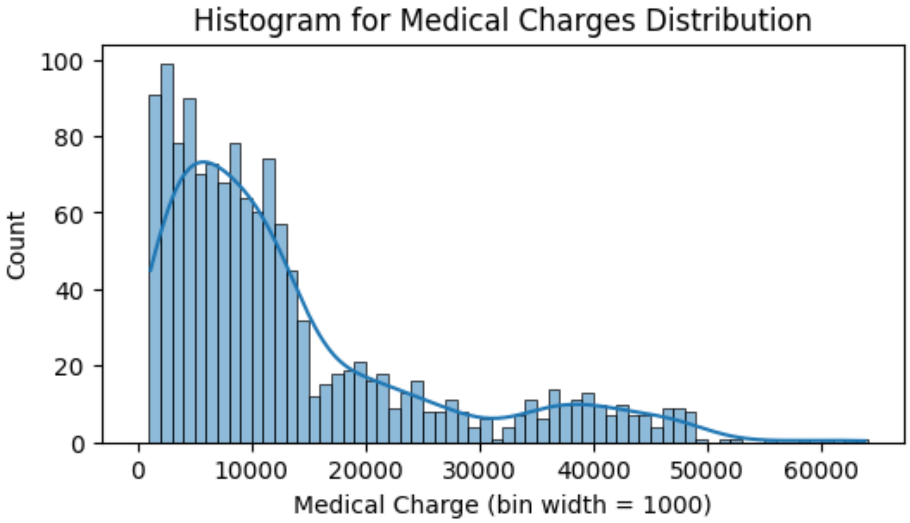
The age distribution is relatively uniform across most age ranges, with noticeable peaks at ages 18 and 64. The mean age is 39.21 years, with a standard deviation of 14.05 years. The interquartile range (IQR) spans from 27 to 51 years.



- **BMI:** The BMI distribution exhibits a bell-shaped pattern with most values concentrated around the mean of 30.66. The standard deviation is 6.10, indicating moderate variation in BMI values. The IQR for BMI ranges from 26.30 to 34.69.



- **Medical Charges:** The distribution of Medical Charges is right-skewed, with most charges below $16,640. The mean charge is $13,270, with a high standard deviation of $12,110, reflecting significant variability in medical expenses. The IQR spans from $4,740 to $16,640.



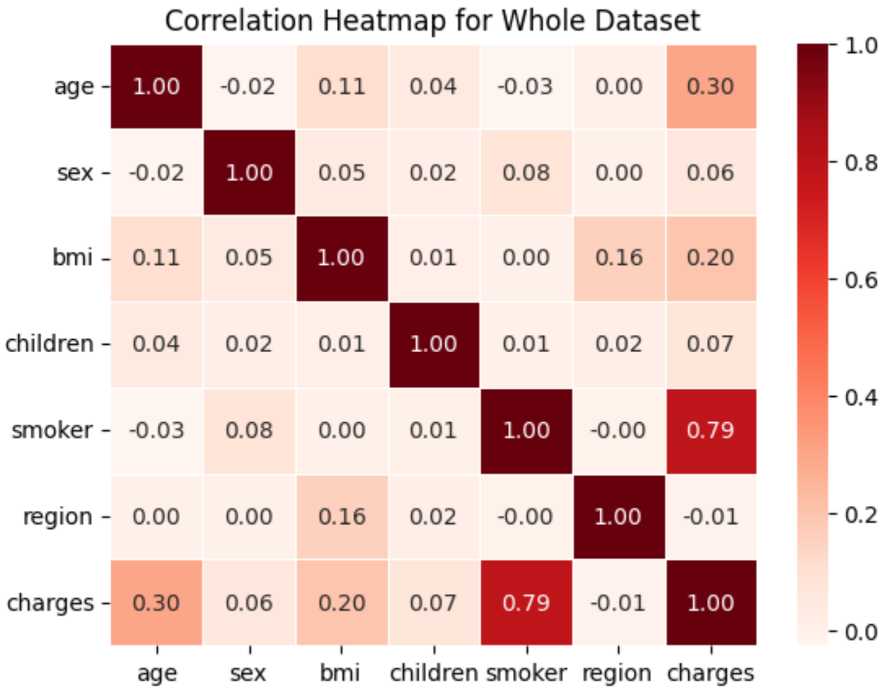
Correlation Analysis

In this section, we aim to uncover the relationships between the variables in our dataset through a detailed correlation analysis. By examining how variables such as Age, BMI, Smoking Status, and other factors correlate with Medical Charges, we can gain insights into the factors associated with medical costs.

We will begin by transforming categorical variables into numerical values to facilitate the correlation analysis. Next, we will analyze the entire dataset to identify overall trends and patterns. Subsequently, we will delve deeper by creating Correlation Matrices and Heatmaps for Smokers and Non-Smokers separately, allowing us to compare and contrast the driving factors for each group. This approach will help us build a robust understanding of the influences on Medical Charges.

Correlation Analysis for Whole Dataset

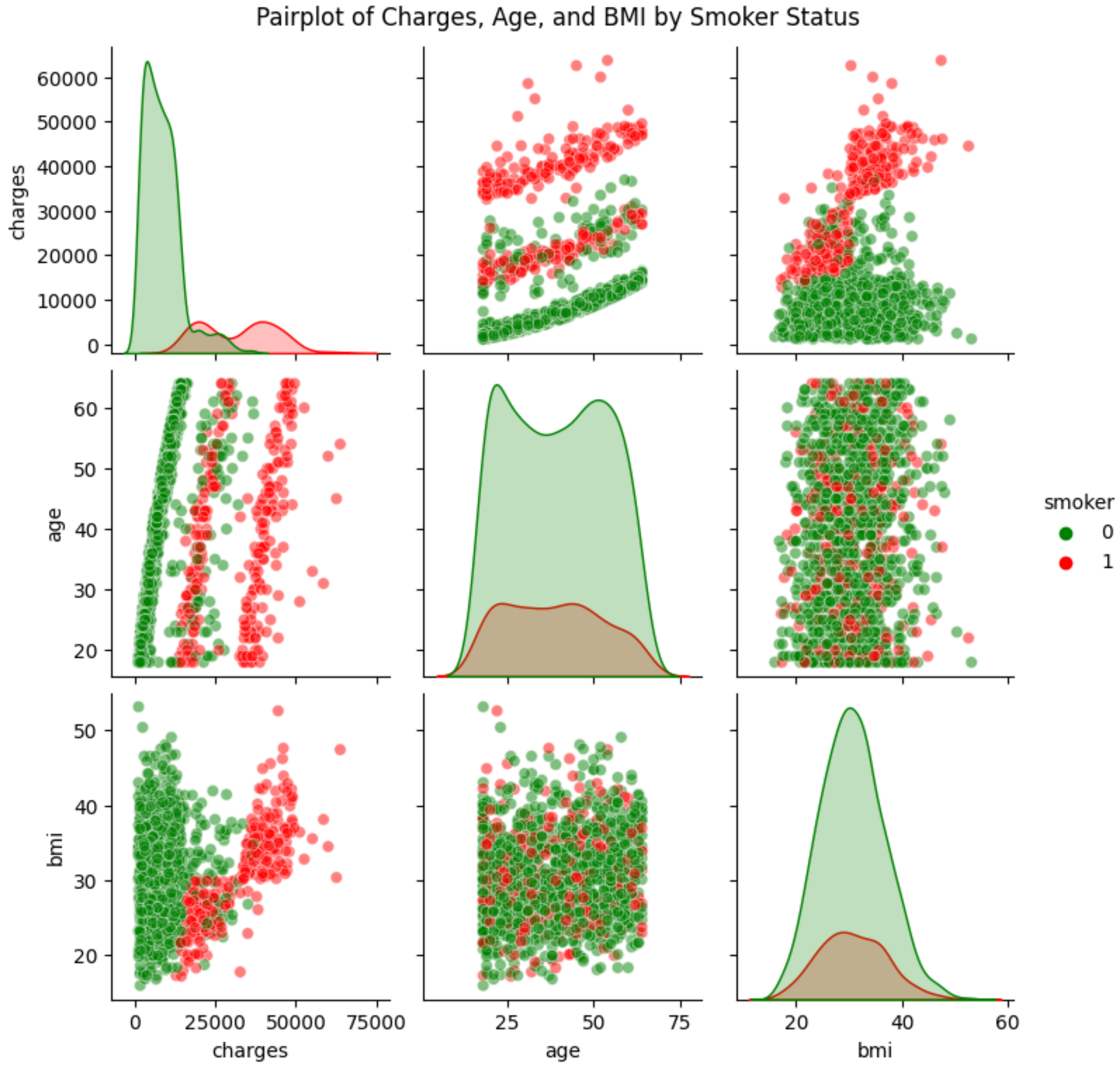
To start, we transform the categorical variables (Sex, Smoker, and Region) into numerical values. Next, we construct a Correlation Matrix for the entire dataset to understand the relationships between the variables. To visualize these relationships, we create a Correlation Heatmap for the entire dataset.



The Correlation Heatmap indicates that Medical Charges are strongly correlated with Smoking Status (0.79), followed by Age (0.30) and BMI (0.20), indicating weak to moderate correlations.

Analysis of Charges, Age, and BMI by Smoking Status

We further visualize the relationships by creating a Pairplot of Charges, Age, and BMI, differentiated by Smoking Status.

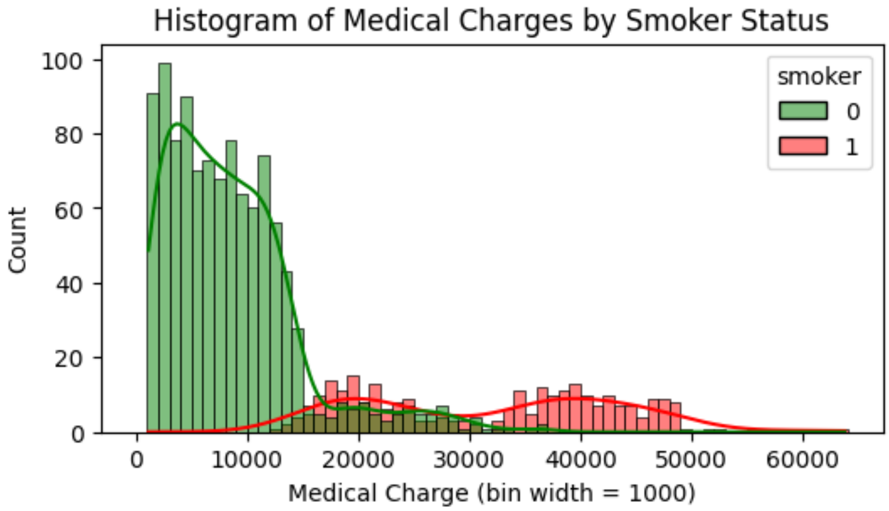


- **Distribution of Medical Charges, Age, and BMI**

The plots on the diagonal show the distributions of Medical Charges, Age, and BMI for Smokers (in red) and Non-Smokers (in green).

Among these distributions, the distributions of Age and BMI appear similar for both groups, except for the frequency of counts, as 80% of the sample are Non-Smokers.

In contrast, Medical Charges show the most significant differences in distribution and range between Smokers and Non-Smokers, indicating that Smokers tend to have higher Medical Charges and a wider range of Charges. An additional Histogram of Charges by Smoking Status further clarifies this distribution.

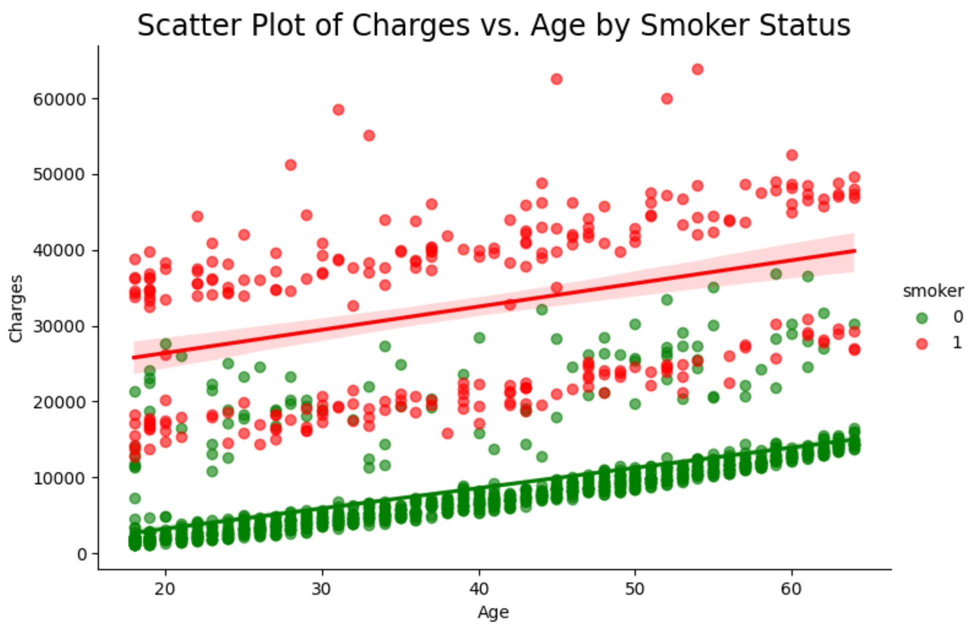


It is interesting to note that Smokers tend to have higher Medical Charges while Non-Smokers tend to have lower Medical Charges. However, some Smokers and Non-Smokers fall in between these extremes, suggesting the influence of other factors driving such distribution and grouping.

- ## Pairwise Relationships

The Pairplots reveal no clear pattern between Age and BMI for either group, suggesting that these two variables do not directly influence each other within this dataset. However, the Charges vs. Age and Charges vs. BMI Pairplots show interesting patterns. To further investigate these relationships, we created additional scatter plots with trend lines for both Smoking and Non-Smoking groups.

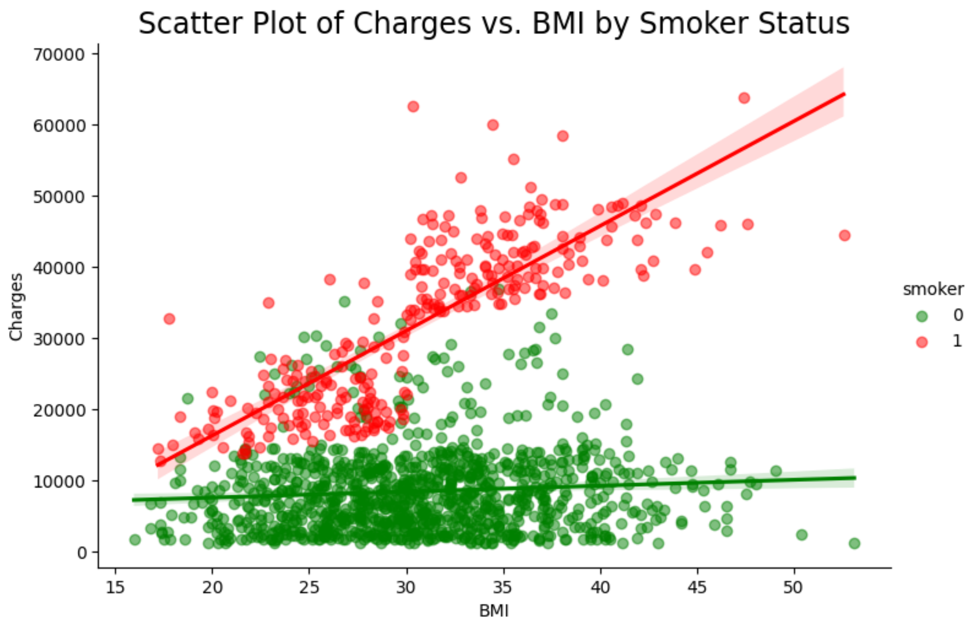
- Charges vs. Age



The Scatter Plot of Charges vs. Age with trend lines for both Smoking and Non-Smoking groups further confirmed that Medical Charges tend to increase with Age.

While most Non-Smokers' data points align closely with their associated trend line, some are noticeably above it. In contrast, Smokers' data points are more dispersed from their trend line. Two distinct groups of Smokers' data points are observed: one above and one below their trend line. These observed groups suggest the possible influence of other important factors on Medical Charges.

- Charges vs. BMI



The Scatter Plot of Charges vs. BMI by Smoking Status highlights two findings:

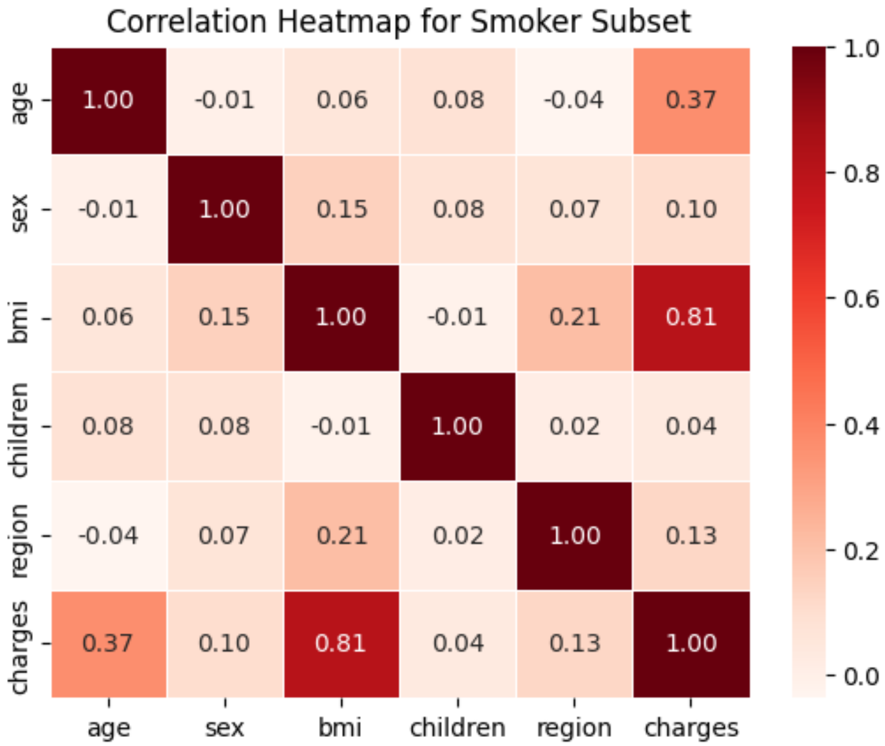
First, Medical Charges show a much stronger correlation with BMI for Smokers than for Non-Smokers. This indicates that BMI is more closely associated with Medical Charges for Smokers, suggesting an interaction between BMI and Smoking Status.

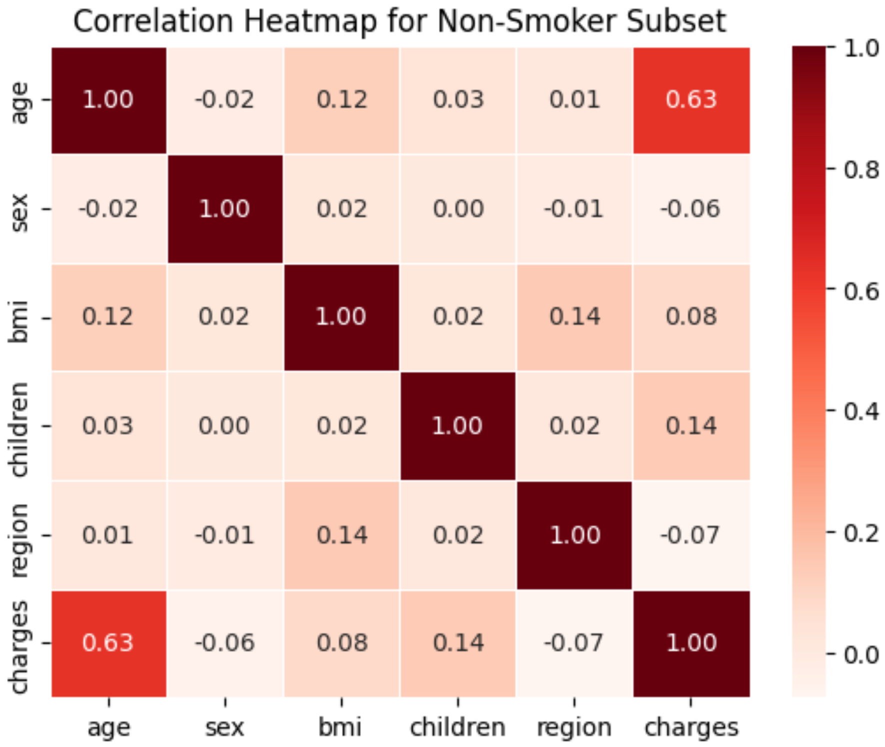
Second, higher BMI Smokers tend to have much higher Medical Charges than lower BMI Smokers. From the plot, it appears that Medical Charges enter a significantly higher range once BMI exceeds 30. This observed discrepancy in Medical Charges between high and low BMI Smokers could account for the grouping seen in the Charges vs. Age Pairplot, but further analysis is required to investigate the reason.

Correlation Analysis for Smokers and Non-Smokers Subsets

To further investigate the roles of Age and BMI in Medical Charges for Smokers and Non-Smokers, we partition the dataset into two subsets and create the Correlation Matrices and Heatmaps for Smokers and Non-Smokers accordingly.

To further investigate the roles of Age and BMI in Medical Charges for Smokers and Non-Smokers, we partition the dataset into two subsets and create the Correlation Matrices and Heatmaps for Smokers and Non-Smokers accordingly.





The results show that, for Smokers, Medical Charges have a strong correlation with BMI (0.81) and a moderate correlation with Age (0.37). For Non-Smokers, Medical Charges have a strong correlation with Age (0.63) but not with BMI (0.08).

These findings suggest that higher BMI and older age Smokers tend to have higher Medical Charges, while older age Non-Smokers tend to have higher Medical Charges with relatively minimal influence of BMI. The interaction between BMI and Smoking Status is more prominent among Smokers, indicating that BMI plays a significant role in determining Medical Charges for Smokers.

Summary

In summary, the correlation analysis indicates that Smoking Status, BMI, and Age are associated with Medical Charges, with Smoking Status being the most significant factor. Smokers tend to have higher Medical Charges, while Non-Smokers generally have lower Medical Charges. However, some Smokers and Non-Smokers fall into an intermediate range of Medical Charges, suggesting that other factors may contribute to these distributions.

Both Pairplots and correlation analyses for Smokers and Non-Smokers indicate similar findings visually and quantitatively: for Smokers, BMI is strongly associated with Medical Charges, followed by Age. For Non-Smokers, Age is the primary factor related to Medical Charges. The interaction between BMI and Smoking Status suggests that BMI significantly influences Medical Charges for Smokers.

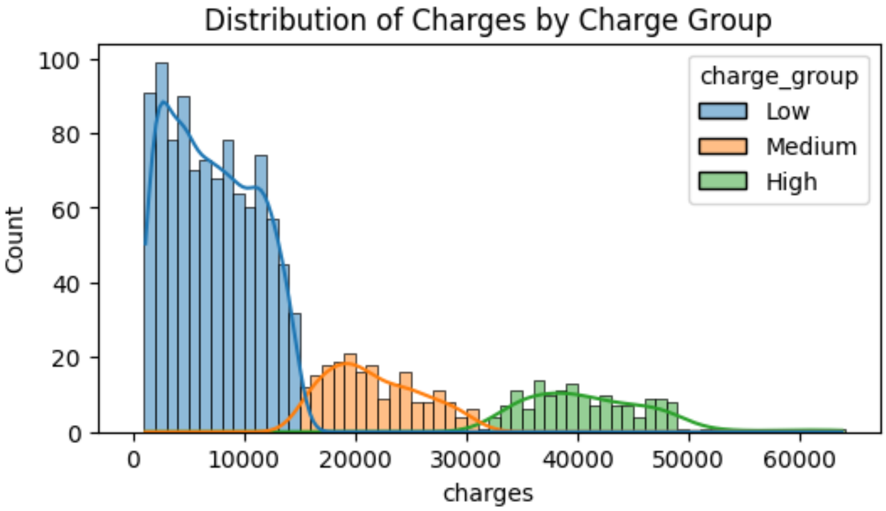
Segmentation and Statistical Analysis

This comprehensive analysis aims to understand the factors driving Medical Charges and identify significant variables that differentiate individuals across different cost groups. By examining the distribution of Medical Charges, especially in relation to Smoking Status, we gain insights into why some individuals fall into Low-, Medium-, or High-Charge categories. The analysis explores the characteristics of each group and delves into the statistical differences between them to identify the key driving factors.

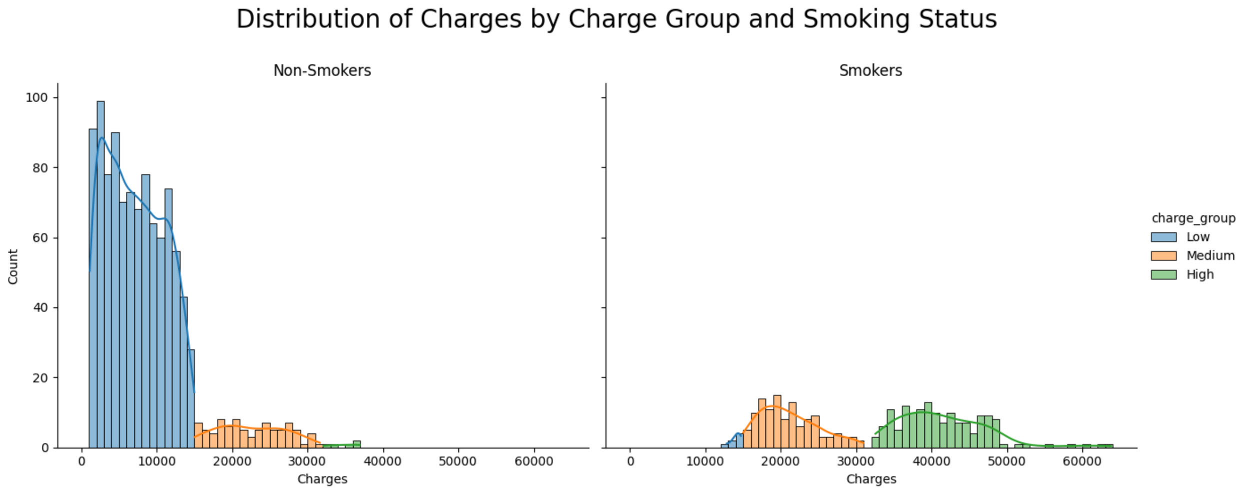
Defining Medical Charge Groups

In previous analysis, we observed distinct groupings of Medical Charges: higher charges for Smokers, lower charges for Non-Smokers, and medium charges for a mix of both Smokers and Non-Smokers.

To understand these observed groupings, we divided the distribution into three groups by defining the cutoffs ($15,000 and $32,000) for Low-, Medium-, and High-Charges and assigned each record to a "Charge Group" based on the defined ranges. The Distribution of Charges by Charge Groups plot below shows the distinct, segmented cost groupings.



We also created separate Charge Distributions for Non-Smokers and Smokers, highlighted by Charge Group, to visualize how they are distributed across these three Charge Groups. This shows that while the majority are in either the Low- or High-Charge Groups, some Non-Smokers and Smokers fall into the Medium-Charge Group.

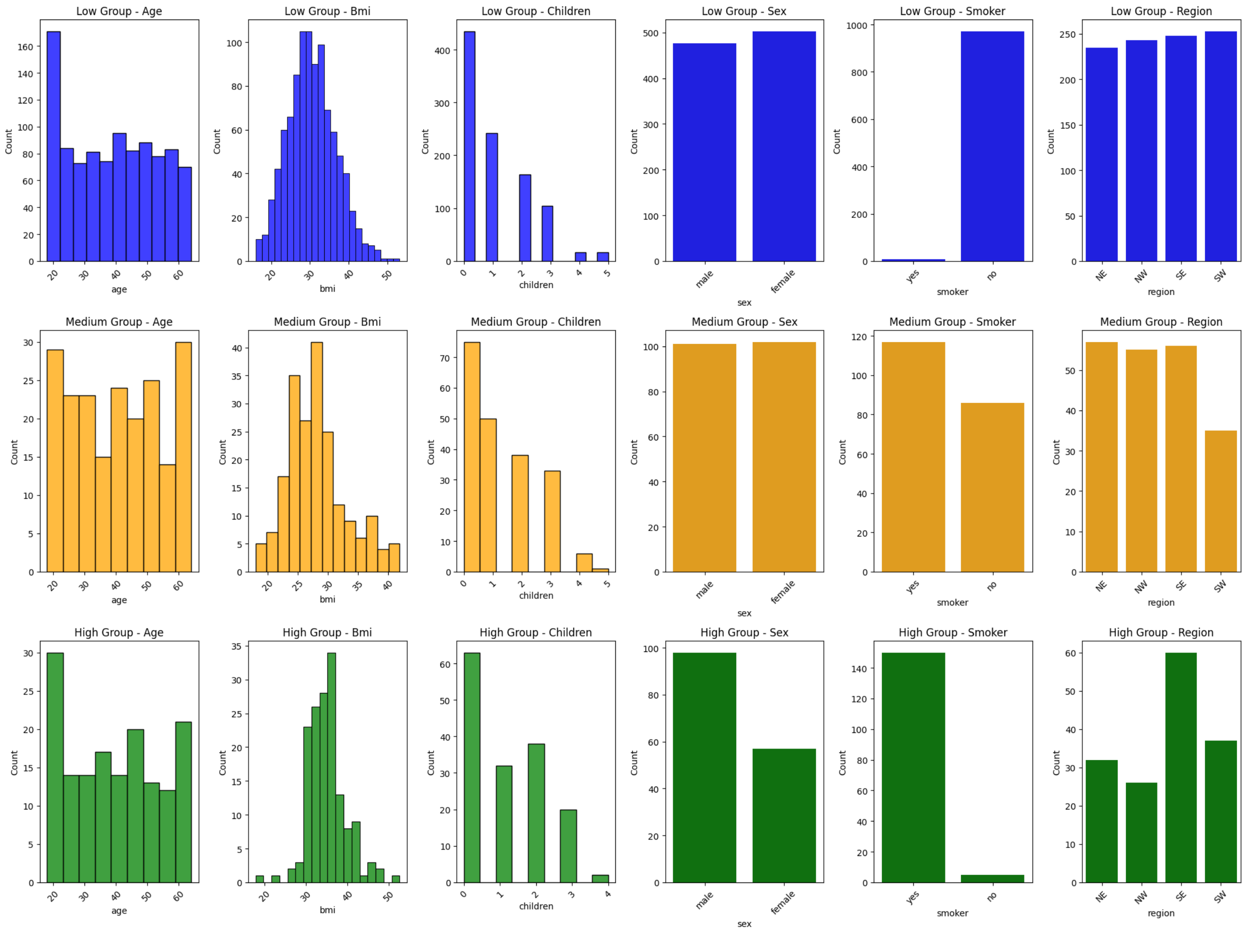


The three Charge Groups are summarized below, detailing the charge range and the composition of Smokers and Non-Smokers:

- **Low Charges:** $0 to $15,000, covering the majority of Non-Smokers and a smaller proportion of Smokers.
- **Medium Charges:** $15,000 to $32,000, including a mix of both Smokers and Non-Smokers.
- **High Charges:** Above $32,000, primarily observed among Smokers, with very few Non-Smokers in this range.

Feature Distribution Analysis

To understand the characteristics of each Charge Group, we visualized the distribution of key features. These plots below offer a visual representation of various features across the Low-, Medium-, and High-Charge Groups. The analyzed features include Age, BMI, Number of Children, Sex, Smoking Status, and Region.



Initial findings are summarized as follows:

- **Age:** There is a peak of younger individuals (20s) in the Low-Charge Group, while older individuals (60s) are slightly more prevalent in the Medium- and High-Charge Groups.
- **BMI:** Higher BMI values are more prevalent in the Medium- and High-Charge Groups.

- **Number of Children:** There is no visually significant difference in the number of children between the groups.
- **Sex:** There is no drastic difference, but there is a slight male predominance in the higher Charge Groups.
- **Smoking Status:** Smoking is a significant factor, with a high concentration of Smokers in the High-Charge Group and Non-Smokers in the Low-Charge Group.
- **Region:** Regional variations might affect Medical Charges, with certain regions showing higher concentrations in different Charge Groups.

Statistical Analysis

To further investigate the observed groupings in Medical Charges, we conducted statistical analysis to compare variables such as Age, BMI, Number of Children, Sex, and Region across two sets of groups: Low- vs. Medium-Charge Non-Smokers and Medium- vs. High-Charge Smokers. The objective was to identify features that show statistically significant differences within these comparisons.

For continuous variables (Age, BMI, Number of Children), we first performed normality and variance tests to ensure the validity of further testing. The Shapiro-Wilk test checked if the data followed a normal distribution, and Levene's test assessed the homogeneity of variances across groups. Depending on the results, we chose between a t-test and a Mann-Whitney U test. The t-test was used if the assumptions of normality and equal variances were met; otherwise, the Mann-Whitney U test was employed.

For categorical variables (Sex and Region), we ensured that each group had at least three samples before performing the chi-square test. This test determined if there were significant associations between these variables and the Charge Groups.

- Statistical Results:

The results revealed significant differences in Age and Number of Children between Low- and Medium-Charge Non-Smokers, suggesting that older Non-Smokers and those with more children are more likely to incur higher Medical Charges. For Smokers, significant differences were found in BMI, Sex, and Region between Medium- and High-Charge Groups. Higher BMI in Smokers was strongly associated with higher Medical Charges. Gender and regional differences also impacted Medical Expenses among Smokers.

Significant variables and their p-values are summarized as follows:

- **Non-Smokers:**
  - **Age:** Mann-Whitney U test, p-value = 0.0084. Moderately significant. Older Non-Smokers tend to have higher Medical Charges.
  - **Number of Children:** Mann-Whitney U test, p-value = 0.0116. Moderately significant. Non-Smokers with more children tend to have higher Medical Charges.
- **Smokers:**
  - **BMI:** Mann-Whitney U test, p-value ≈ 0. Highly significant. Strong difference in BMI between Medium and High Charge Groups. Higher BMI is strongly associated with higher Medical Charges.
  - **Sex:** Chi-square test, p-value = 0.0376. Borderline significant. Significant difference in sex distribution, with male Smokers tending to have higher Medical Charges.
  - **Region:** Chi-square test, p-value = 0.0486. Borderline significant. Significant regional differences in Medical Charges.

Summary of Segmentation and Statistical Analysis

In summary, this Segmentation and Statistical Analysis provides key insights into the factors influencing Medical Charges. The three-peak distribution of Medical Charges can be explained by the interplay of Age, BMI, Number of Children, Sex, and Region. This comprehensive analysis aimed to explore the role of these factors and identify significant variables that differentiate individuals across different cost groups.

For Non-Smokers, Age is a key factor in Medical Charges, with a moderate influence of the Number of Children. Older Non-Smokers and those with more children are more likely to incur higher Medical Charges.

For Smokers, BMI is the most significant factor in Medical Charges, with additional influences of Sex and Region. Higher BMI is strongly associated with higher Medical Charges, with male Smokers and regional differences also impacting costs.

The composition of individuals in each Charge Group is summarized as follows:

Low-Charge Group:

- Primarily consists of younger Non-Smokers.
- Includes a small number of very young Smokers.

Medium-Charge Group:

- Includes a mix of both Smokers and Non-Smokers:
  - **Non-Smokers:** Tend to be older on average compared to those in the Low-Charge Group (43.13 vs. 38.98). The number of children is slightly higher (1.37 vs. 1.06).
  - **Smokers:** Have lower BMI compared to those in the High-Charge Group (25.66 vs. 35.13).

High-Charge Group:

- Dominated by Smokers with significantly higher BMI, a higher proportion of males (64.67%), and regional variations (39.33% from Southeast as the dominant region).
- Includes a small number of Non-Smokers with exceptionally high BMI.

Classification Analysis

Introduction

The classification analysis was conducted to complement the findings from the exploratory data analysis (EDA), correlation analysis, and statistical analysis, aiming to understand the observed grouping of medical charges and identify the driving factors. By using classification models, we can quantify the importance of various features in predicting medical charges, providing a clearer understanding of how different factors contribute to the observed charge groups.

To achieve this, we explored four different classification models:

1. **Classification without Feature Engineering**: This model uses the original set of variables without introducing any additional interaction features.
2. **Classification Excluding Less Important Variables (Threshold 0.02)**: This model excludes the variables shown to have lower importance based on initial findings to assess the impact on classification performance.
3. **Classification with Feature Engineering (Threshold 0.01)**: This model incorporates additional interaction features to capture more complex relationships between the variables and medical charges, using a threshold of 0.01 for feature importance.
4. **Classification with Feature Engineering (Threshold 0.02)**: Similar to the previous model, but with a stricter threshold of 0.02 for feature importance to evaluate its impact on the model's performance.

Methodology

For our classification analysis, we developed four models with different sets of features using a Random Forest Classifier to compare their effectiveness in predicting medical charges. The methodology for these models is detailed below:

- Data Preparation

The data preparation process was consistent across all models:

1. **Encoding categorical variables**: Categorical variables (Sex, Region, Smoker) were encoded into binary variables using one-hot encoding.
2. **Normalization**: All variables were normalized to ensure they contributed equally to the model.

3. **Handling class imbalance**: Class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset.

- Models Developed

- **Model 1: Classification without Feature Engineering**
  - **Feature Set:** Original variables including Smoking Status, BMI, Age, Number of Children, Sex, and Region.

- **Model 2: Classification Excluding Less Important Variables (Threshold 0.02)**
  - **Feature Set:** Original variables excluding Region variables with importance scores less than 0.02, as identified in the classification analysis.

- **Model 3: Classification with Feature Engineering (Threshold 0.01)**
  - **Feature Set**: Original variables plus interaction features to capture more complex relationships. Interaction terms were created for Non-Smokers with Age and Number of Children, and for Smokers with BMI, Sex, and Region. The initial set of features included age_nonsmoker, children_nonsmoker, bmi_smoker, sex_male_smoker, region_northwest_smoker, region_southeast_smoker, and region_southwest_smoker.
  - **Feature Selection**: Iterative process to drop features with importance scores below 0.01.

- **Model 4: Classification with Feature Engineering (Threshold 0.02)**
  - **Feature Set:** Same as Model 3, but with a stricter threshold of 0.02 for feature importance.

- Model Training and Evaluation

1. **Classifier**: A Random Forest Classifier was used for all models due to its robustness, ability to handle a large number of features, and effectiveness in capturing complex interactions between variables.
2. **Cross-Validation**: Cross-validation was performed on the same split train/test dataset for all models to ensure fair comparison. We used a 3-fold cross-validation strategy.
3. **Evaluation Metrics**: Model performance was assessed using accuracy, confusion matrix, precision, recall, and F1-score. These metrics provided a comprehensive view of the model's predictive capabilities and highlighted areas of strength and improvement.

This structured approach allowed us to systematically compare the effectiveness of different feature sets and engineering strategies in predicting medical charges, ensuring robust and fair evaluation of each model's performance.

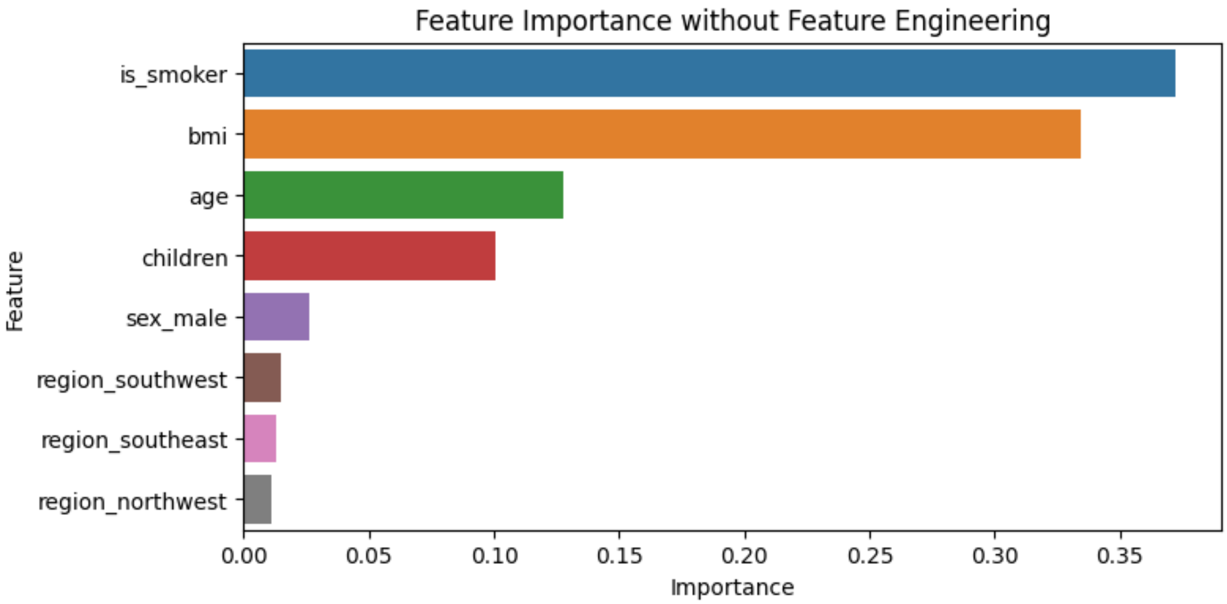
Comparison of Classification Model Results

Model 1: Classification without Feature Engineering

- Accuracy: 0.925
- Classification Report:

	precision	recall	f1-score	support
Low	0.85	0.96	0.90	196
Medium	0.94	0.84	0.89	196
High	1.00	0.97	0.98	196
accuracy			0.93	588
macro avg	0.93	0.93	0.93	588
weighted avg	0.93	0.93	0.93	588

- **Important Features:** is_smoker (0.3722), bmi (0.3343), age (0.1275), children (0.1006), sex_male (0.0261), region_southwest (0.0150), region_southeast (0.0129), region_northwest (0.0114).



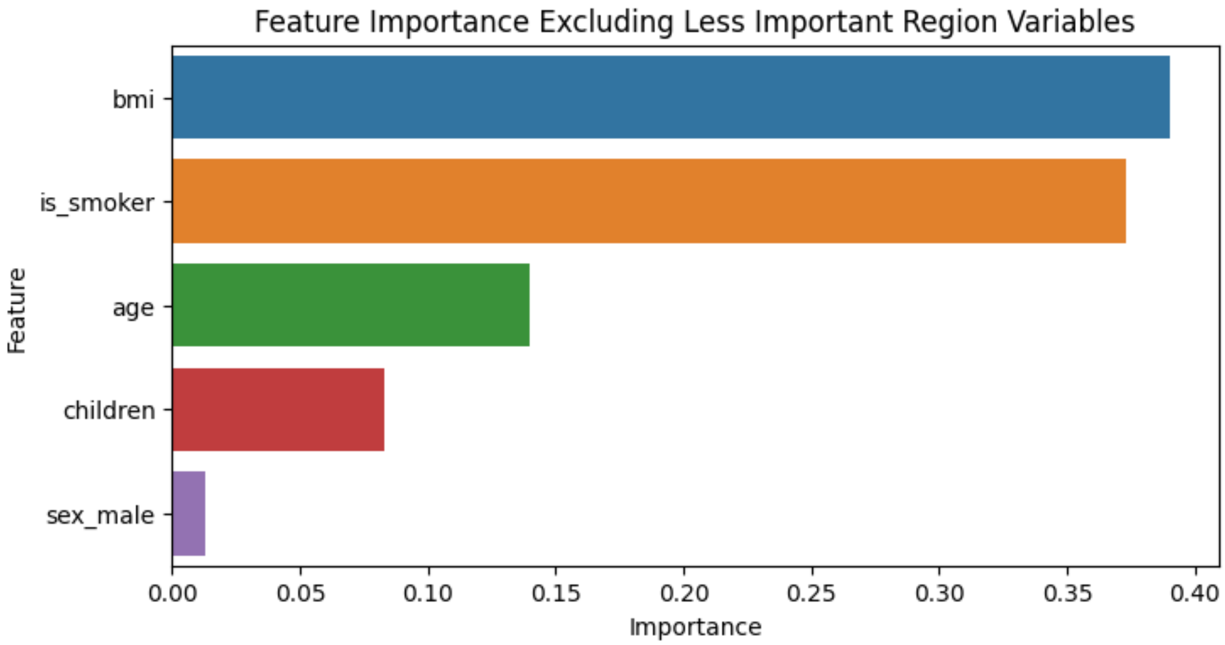
- **Findings:** This model demonstrated high accuracy, indicating that the original set of variables (age, BMI, children, smoking status, region, and sex) provided robust predictive power without the need for additional interaction terms. The key features identified were Smoking Status and BMI, which were consistently important across all analyses. The high accuracy and balanced performance across different charge groups suggest that the model captures the underlying patterns effectively without overfitting.

Model 2: Classification Excluding Less Important Region Variables

- **Accuracy:** 0.912
- **Classification Report:**

	precision	recall	f1-score	support
Low	0.86	0.91	0.89	196
Medium	0.89	0.86	0.88	196
High	0.98	0.96	0.97	196
accuracy			0.91	588
macro avg	0.91	0.91	0.91	588
weighted avg	0.91	0.91	0.91	588

- **Important Features:** bmi (0.3901), is_smoker (0.3734), age (0.1399), children (0.0833), sex_male (0.0133).



- **Findings:** By excluding the less important region variables, the model's accuracy slightly decreased. However, the performance for the medium charge group improved marginally. This suggests that while the region variables contribute to the model's overall predictive power, their exclusion simplifies the model and helps to reduce potential overfitting. The important features remained consistent, emphasizing the critical role of BMI and Smoking Status.

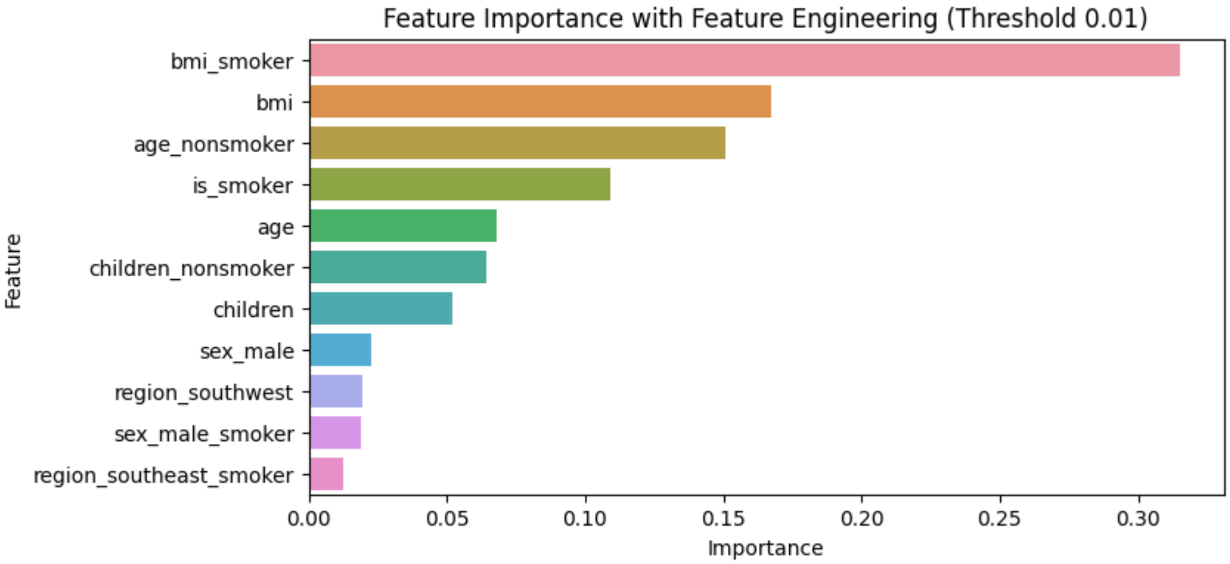
Model 3: Classification with Feature Engineering (Threshold 0.01)

- **Accuracy:** 0.910
- **Classification Report:**

	precision	recall	f1-score	support
Low	0.85	0.92	0.89	196
Medium	0.90	0.84	0.87	196
High	0.99	0.96	0.98	196

accuracy			0.91	588
macro avg	0.91	0.91	0.91	588
weighted avg	0.91	0.91	0.91	588

- **Important Features:** bmi_smoker (0.3152), bmi (0.1671), age_nonsmoker (0.1509), is_smoker (0.1092), age (0.0680), children_nonsmoker (0.0640), children (0.0521), sex_male (0.0228), region_southwest (0.0194), sex_male_smoker (0.0186), region_southeast_smoker (0.0124).



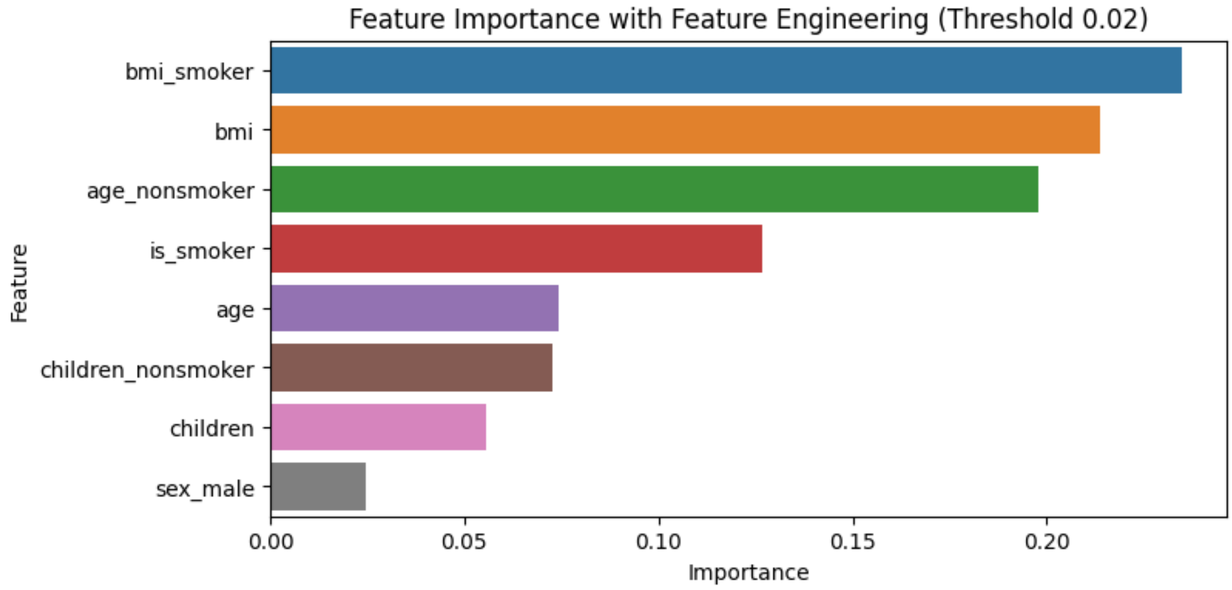
- **Findings:** Introducing interaction terms and additional features through feature engineering resulted in a slightly lower accuracy compared to Model 1. The performance for the medium charge group dropped, indicating potential overfitting due to the inclusion of interaction terms. Despite this, the model highlighted important interaction terms like bmi_smoker and age_nonsmoker, which align with the observed patterns in the data. This model shows the critical role of interaction effects, particularly for smokers and non-smokers, but also underscores the complexity added by such features.

Model 4: Classification with Feature Engineering (Threshold 0.02)

- **Accuracy:** 0.903
- **Classification Report:**

	precision	recall	f1-score	support
Low	0.84	0.93	0.88	196
Medium	0.90	0.82	0.86	196
High	0.98	0.95	0.97	196
accuracy			0.90	588
macro avg	0.91	0.90	0.90	588
weighted avg	0.91	0.90	0.90	588

- **Important Features:** bmi_smoker (0.2346), bmi (0.2137), age_nonsmoker (0.1980), is_smoker (0.1266), age (0.0742), children_nonsmoker (0.0725), children (0.0557), sex_male (0.0246).



- **Findings:** With a stricter threshold for feature importance, this model had a slightly lower accuracy. However, it demonstrated a more streamlined set of features, reducing complexity. The most important features remained consistent, with bmi_smoker and bmi showing high importance. This model suggests that a more conservative approach to feature selection can help generalize better to unseen data, although it might lose some nuanced information.

Evaluation and Insights

- Performance and Accuracy:

- **Model 1:** Achieved the highest accuracy (0.925), suggesting that the original set of variables (Age, BMI, Number of Children, Smoking Status, Region, and Sex) is sufficient for robust predictions. The high accuracy and balanced performance across different charge groups indicate that this model effectively captures the underlying patterns without overfitting.
- **Model 2:** Excluded less important region variables and had a slightly lower accuracy (0.912) compared to Model 1. This model simplifies the feature set while maintaining good predictive capability, emphasizing the critical role of BMI and Smoking Status.
- **Model 3 (Threshold 0.01):** Achieved an accuracy of 0.910. The inclusion of interaction terms (e.g., bmi_smoker and age_nonsmoker) captured more complex relationships between variables. Although this model showed a slightly lower accuracy compared to Model 1, it provided deeper insights into the interaction effects, which may introduce some complexity.
- **Model 4 (Threshold 0.02):** Had an accuracy of 0.903. This model used a stricter threshold for feature importance, resulting in a simpler and more conservative set of features. While it had slightly lower accuracy compared to Model 3, it highlights the key features with less complexity.

- Key Findings:

- **Feature Importance and Consistency:**.Across all models, Smoking Status and BMI consistently emerged as the most important features, underscoring their critical role in predicting medical charges. Age also remained a significant predictor, especially for non-smokers. The inclusion of interaction terms in Models 3 and 4 revealed specific interactions (e.g., bmi_smoker and age_nonsmoker) that offer deeper insights but also introduce complexity. These findings further confirmed the significant interaction effects of BMI for Smokers and Age for Non-Smokers on Medical Charges.

- **Model Performance:** Model 1's highest accuracy suggests that the original variables are sufficient for robust predictions. Model 2's slightly lower accuracy but more streamlined performance indicates that excluding less important variables can simplify the model while maintaining predictive power. Models 3 and 4, which include feature engineering, offer additional insights into variable interactions. However, this comes at the expense of increased complexity and does not improve accuracy.

Summary of Classification Analysis

The classification analysis highlights the critical factors influencing Medical Charges. Smoking Status, BMI, and Age consistently emerged as significant predictors across all models, aligning with previous statistical and correlation findings.

Model 1 achieved the highest accuracy (0.925), demonstrating that the original set of variables is sufficient for robust predictions. Model 2 simplifies the model by excluding less important region variables, resulting in a slightly lower accuracy (0.912).

Feature engineering in Models 3 and 4 provided deeper insights into specific interactions, such as the combined effects of BMI and Smoking Status (bmi_smoker) and Age and Non-Smoking Status (age_nonsmoker). However, this added complexity did not improve accuracy and could potentially introduce the risk of overfitting.

In this analysis, the simple Model 1 achieved the highest accuracy. The inclusion of all original variables provided the highest predictive power, as the Random Forest model effectively considered inherent interactions. Introducing additional interaction effects in Models 3 and 4 did not enhance performance and posed the risk of overfitting.

Overall, these findings underscore the effectiveness of Model 1's simplicity and accuracy. They highlight the importance of focusing on the most impactful features, demonstrating that a straightforward model can offer robust predictions without the added complexity of interaction terms.

Regression Analysis

Introduction

The regression analysis aims to complement the findings from the exploratory data analysis, correlation analysis, statistical analysis, and classification analysis by predicting the actual medical charges based on identified significant features. While classification analysis helped us understand the grouping of medical charges and identify key predictive factors, regression analysis provides a quantitative estimation of the charges themselves. This approach allows us to not only determine which factors are important but also to quantify their impact on the actual medical charges.

Methodology

For our regression analysis, we developed two models to compare their effectiveness in predicting medical charges: one using the original feature set (Regression with Original Feature Set) and another excluding less important Region variables (Regression Excluding Less Important Region Variables).

- Data Preparation

The data preparation process was consistent across both models:

1. **Encoding Categorical Variables**: Categorical variables (Sex, Region, Smoker) were encoded into binary variables using one-hot encoding.
2. **Normalization**: All variables were normalized to ensure they contributed equally to the model.

- Models Developed

- **Model 1: Regression with Original Feature Set**
  - **Feature Set:** Original variables including Smoking Status, BMI, Age, Number of Children, Sex, and Region.
- **Model 2: Regression Excluding Less Important Region Variables**
  - **Feature Set:** Original variables excluding the region variables identified as less important in the classification analysis and initial regression analysis.

- Model Training and Evaluation

1. **Regressor**: A Random Forest Regressor was used for both models due to its robustness and ability to capture complex interactions between variables.

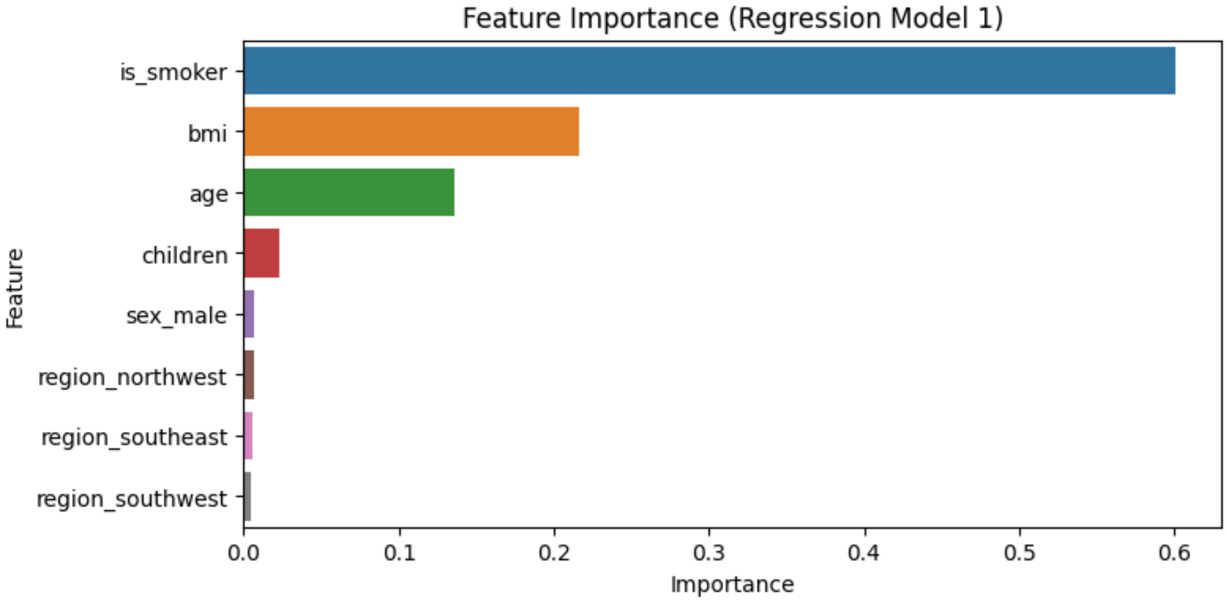
2. **Cross-Validation**: Cross-validation was performed on the same split train/test dataset for both regression models to ensure a fair comparison. We used a 3-fold cross-validation strategy.
3. **Evaluation Metrics**: Model performance was assessed using Mean Squared Error (MSE) and R² Score. These metrics provided a quantitative estimation of the model's predictive capabilities. Feature importance scores were also calculated to identify the most significant predictors in each model.

This structured approach allowed us to systematically compare the effectiveness of different feature sets in predicting medical charges, ensuring robust and fair evaluation of each model's performance.

Comparison of Regression Model Results

Model 1: Regression with Original Feature Set

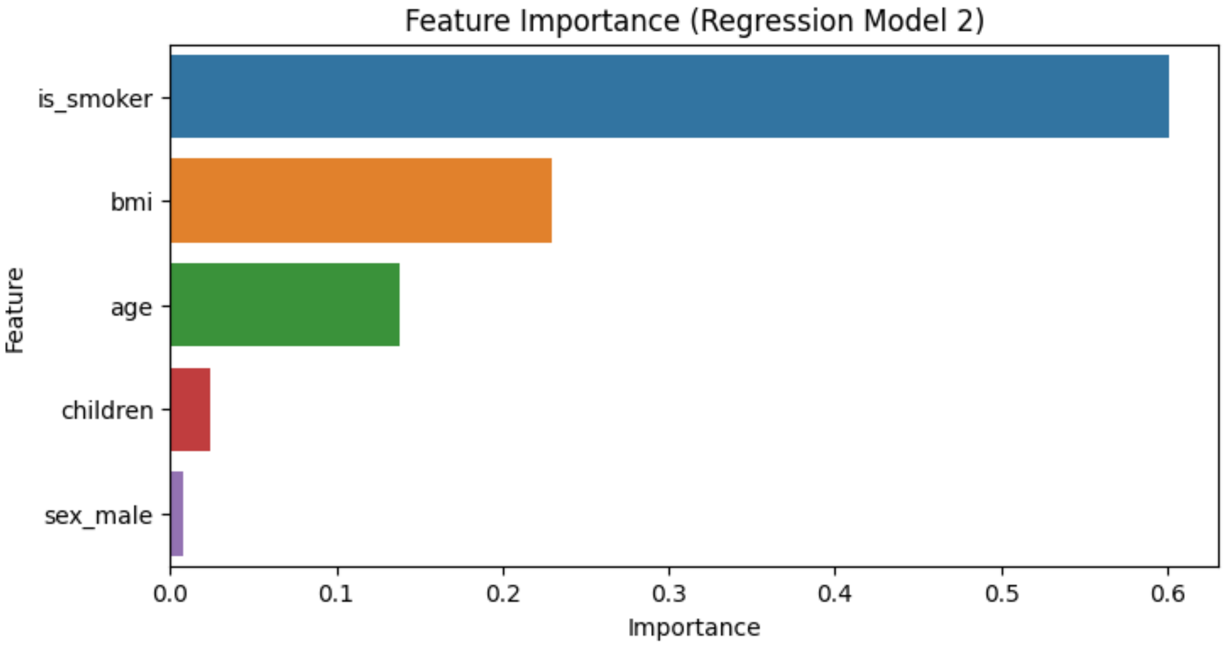
- **Cross-validation (MSE):** [25.68M, 22.23M, 25.24M] (Average: 24.39M)
- **Train MSE:** 3.59M, **Test MSE:** 22.21M
- **Train R²:** 0.974, **Test R²:** 0.879
- **Important Features:** is_smoker (0.6008), bmi (0.2161), age (0.1357), children (0.0229), sex_male (0.0071), region_northwest (0.0068), region_southeast (0.0060), region_southwest (0.0046)



- **Findings:** This model achieved a high train R² score (0.974), indicating strong predictive power with the original feature set. The significant importance of is_smoker (0.6008) and bmi (0.2161) highlights their impact on medical charges. While the test R² score (0.879) shows a slight overfitting, the overall performance remains strong.

Model 2: Regression Excluding Less Important Region Variables

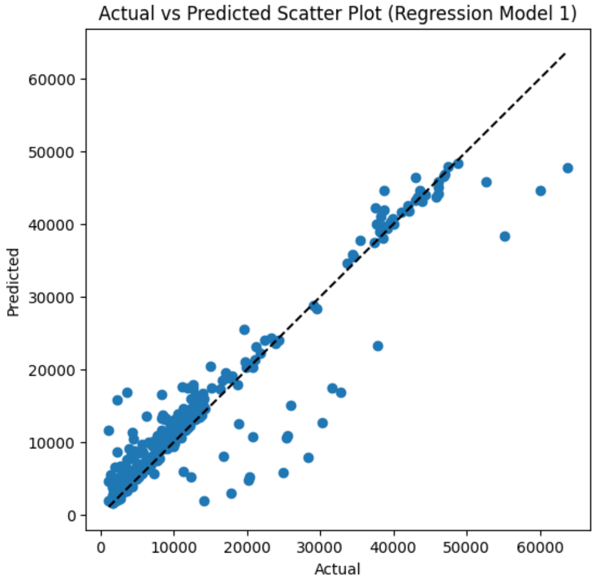
- **Cross-validation (MSE):** [25.82M, 22.39M, 26.59M] (Average: 24.93M)
- **Train MSE:** 3.75M, **Test MSE:** 23.24M
- **Train R²:** 0.973, **Test R²:** 0.874
- **Important Features**: is_smoker (0.6008), bmi (0.2295), age (0.1379), children (0.0242), sex_male (0.0076)

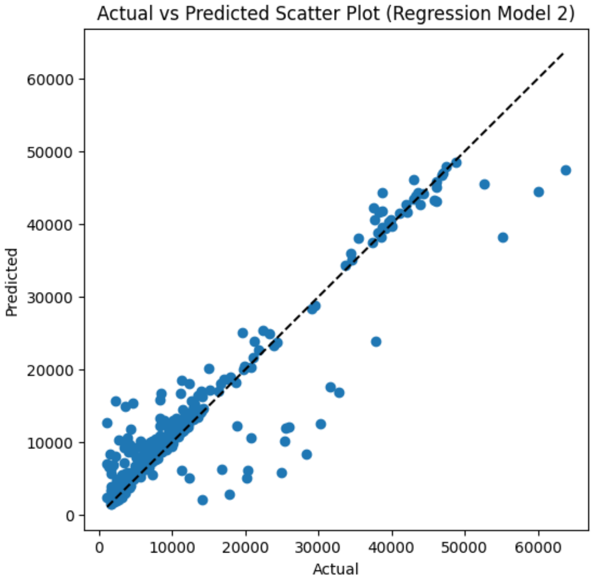


- **Findings:** Removing the less important region variables resulted in a slight decrease in the train R² score (0.973) and test R² score (0.874), indicating a negligible drop in predictive power. The importance scores for is_smoker (0.6008) and bmi (0.2295) remained high, reinforcing their significance. This model simplified the feature set and reduced complexity while maintaining similar performance.

Residual Analysis

- Residuals vs. Predicted Scatter Plot





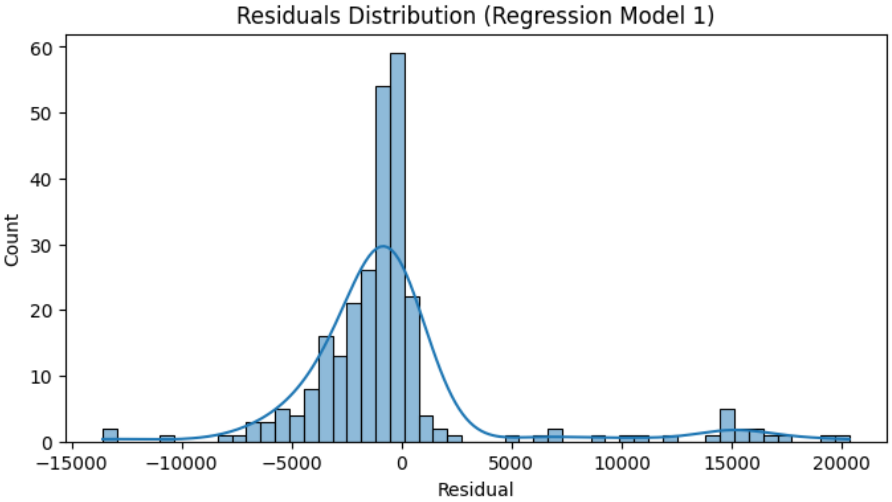
The Residuals vs. Predicted Scatter Plots help to identify non-linearity, unequal error variances, and outliers by showing how residuals are spread across the range of predicted values.

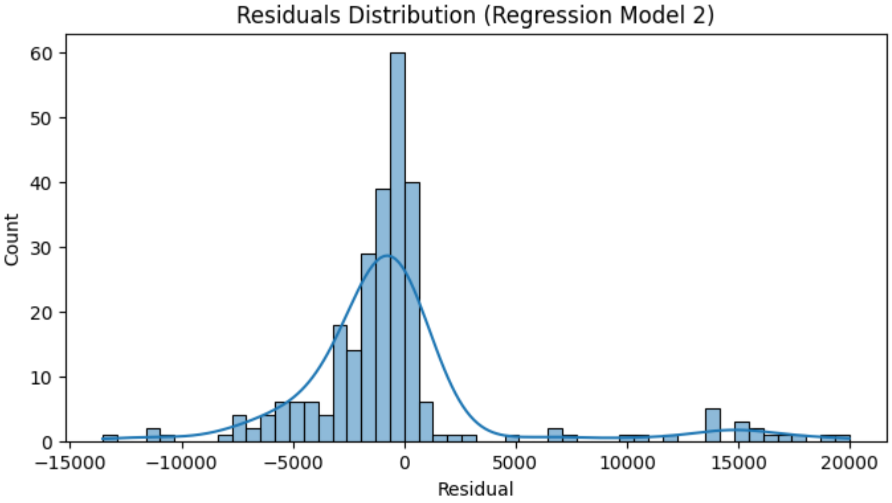
Both models show a good fit, with most data points close to the ideal dashed line. However, there are some specific observations worth noting:

- **Deviation from the Line**: For both models, data points at lower charge values (below 20,000) are closely clustered around the dashed line, indicating accurate predictions. However, as charges increase beyond 20,000, we start to see more noticeable deviations from the line. This suggests that while the models predict low to moderate charges well, they struggle more with higher charges.
- **Prediction Errors**: The scatter plots show that both models have prediction errors, especially for higher charge values. These errors are more dispersed and tend to deviate from the dashed line, highlighting areas where the model's predictions are less reliable.
- **General Trend Capture**: Despite these deviations, the overall pattern shows that both models capture the general trend of the data well. The majority of data points align closely with the dashed line, indicating that the models effectively

capture the relationship between the features and the target variable for most of the data range.

- ## Residual Distribution Plot





The Residual Distribution Plots check if the residuals are approximately normally distributed. These plots help identify skewness, kurtosis, and the presence of outliers, which are crucial for validating the assumptions of the regression model. By analyzing these plots, we can assess how well the model's errors adhere to the assumptions of normality and unbiased predictions, as detailed in the following observations:

- **Centered Around 0**: For both models, the residuals are centered around 0, indicating no systematic bias in the predictions. This means that the models are not consistently overestimating or underestimating the charges.
- **Positive Skew**: Both models show a slight positive skew in the residuals distribution. This suggests that the models tend to slightly underestimate higher charges, as indicated by the longer tail on the right side of the distribution.
- **Residual Spread**: The spread of residuals for higher charges is wider, indicating greater prediction errors at higher charge values. This is consistent with the observations from the scatter plots, where higher charges showed more deviation from the ideal line.
- **Normality and Outliers**: The residuals distribution also helps in identifying any potential outliers. Both models show a relatively normal distribution of residuals, but with some outliers that could be affecting the model's performance.

Evaluation and Insights

- Performance and Accuracy:

Model 1 (Original Feature Set):

- High R^2 score (Train: 0.974, Test: 0.879), indicating strong predictive power.
- Shows some overfitting, as indicated by the difference between train and test R^2 scores.

Model 2 (Excluding Less Important Region Variables):

- Slightly lower R^2 score (Train: 0.973, Test: 0.874) but with a negligible difference compared to Model 1.
- Simplified feature set with reduced complexity while maintaining strong predictive power.

- Key Findings:

- **Feature Importance and Consistency:**.
  - Smoking Status and BMI consistently emerged as the most important features in both models.
  - Age remained an important predictor.
  - The exclusion of Region variables in Model 2 did not significantly affect the importance of Smoking Status, BMI, and Age, confirming their critical role in predicting Medical Charges.

- **Model Performance:**
  - **Model 1:** Achieved the highest accuracy with a Train R^2 of 0.974 and a Test R^2 of 0.879. This indicates that the original set of variables provided strong predictive power but exhibited slight overfitting.
  - **Model 2:** Had a Train R^2 of 0.973 and a Test R^2 of 0.874. This model simplified the feature set by excluding less important Region variables, resulting in almost no difference in performance compared to Model 1. This suggests that Region might not be a significant variable for predicting Medical Charges.

Summary of Regression Analysis

The regression analysis complemented previous findings by quantifying medical charges. Two models were compared: **Model 1 (Original Feature Set)**, which included Smoking Status, BMI, Age, Number of Children, Sex, and Region, and **Model 2 (Excluding Less Important Region Variables)**, which excluded less important Region variables.

Model 1 showed strong predictive power but some overfitting (Train R^2: 0.974, Test R^2: 0.879). **Model 2** had a slightly lower R^2 (Train R^2: 0.973, Test R^2: 0.874) but simplified the model with a negligible difference compared to Model 1.

Smoking Status and BMI were consistently the most important features, with Age also significant. Both models performed well for lower to medium charges but struggled with higher charges. Residual analysis confirmed these findings, showing increased variability for higher charges and a slight positive skew in residuals.

The results suggest that Region might not be a significant variable for predicting medical charges, as excluding these variables had almost no impact on model performance.

Summary of Feature Importance

The results from the EDA, correlation analysis, segmentation and statistical analysis, classification, and regression analyses are consistent. Here's a detailed look at how each significant feature aligns with these analyses:

- **Smoking Status**

  - **Correlation Analysis:** Smoking Status had the highest correlation with Medical Charges (0.79), indicating a strong relationship.
  - **Classification Analysis:** Consistently showed high importance across all models, with an importance score of 0.3722 in the model without feature engineering.
  - **Regression Analysis:** Highest importance score (0.6008), reinforcing its dominant role in predicting Medical Charges.

- **Body Mass Index (BMI)**

  - **Correlation Analysis:** Moderate correlation with Medical Charges (0.20). Stronger correlation for smokers (0.81), indicating its significance for Smokers.
  - **Classification Analysis:** Top predictor across all models. Highest importance score of 0.3901 in the model excluding less important Region variables. In models with feature engineering, bmi_smoker had the highest importance scores (0.3152 and 0.2346).
  - **Regression Analysis:** Second-highest importance score (0.2161), highlighting its significant impact.

- **Age**

  - **Correlation Analysis:** Moderate correlation with Medical Charges (0.30). Stronger correlation for non-smokers (0.63), indicating that older Non-Smokers tend to have higher charges.
  - **Classification Analysis:** Consistently showed high importance. Scores of 0.1275 and 0.1399 in models without feature engineering, and 0.1509 and 0.1980 in models with feature engineering, especially significant for Non-Smokers.
  - **Regression Analysis:** Importance score of 0.1357, confirming its relevance.

- **Number of Children**

  - **Correlation Analysis:** Almost no correlation with Medical Charges (0.07). Slightly stronger for Non-Smokers (0.14) than for Smokers (0.04).
  - **Segmentation and Statistical Analysis:** Significant for Non-Smokers, with a p-value of 0.0113.
  - **Classification Analysis:** Lower importance compared to other features. Scores of 0.1020 without feature engineering, and 0.0584 for children_nonsmoker with feature engineering.
  - **Regression Analysis:** Low importance score (0.0229), consistent with its moderate impact.

- **Sex**

  - **Correlation Analysis:** Almost no correlation with Medical Charges (0.06). Minor correlation for Smokers (0.10) and almost no correlation for Non-Smokers (-0.06).
  - **Segmentation and Statistical Analysis:** Borderline significant for smokers, with a p-value of 0.0376.
  - **Classification Analysis:** Relatively lower importance. Scores of 0.0261 without feature engineering, and 0.0186 for sex_male_smoker with feature engineering.
  - **Regression Analysis:** Very low importance score (0.0071), confirming its minimal impact.

- **Region**

  - **Correlation Analysis:** No correlation with Medical Charges (-0.01). Very minor correlation for smokers (0.09) and almost no correlation for non-smokers (-0.07).
  - **Segmentation and Statistical Analysis:** Borderline significant for smokers, with a p-value of 0.0486.
  - **Classification Analysis:** Often had the lowest importance scores (~0.01) or were removed with minimal impact on model performance.
  - **Regression Analysis:** Very low importance scores for region_northwest (0.0068), region_southeast (0.0060), and region_southwest (0.0046). Dropped in the second regression model, resulting in a simpler model with minimal impact on accuracy.

Conclusion and Limitations

Conclusion

The comprehensive analysis of medical charges has highlighted several key factors that significantly influence costs. Smoking Status, BMI, and Age consistently emerged as the most critical predictors, aligning with previous findings from statistical and correlation analyses. Smoking Status, in particular, showed the strongest relationship with medical charges, emphasizing the need for targeted interventions for smokers. BMI and Age also play significant roles, with BMI being particularly important for Smokers and Age for Non-Smokers.

The classification and regression models provided robust predictions. Results from both analyses indicated that the original variable set provides strong predictive power, achieving the best performance compared to models that removed variables with low importance scores or added complexity through additional interaction effects.

Simplifying the regression models by excluding Region variables did not significantly affect performance, suggesting that Region may not be crucial in predicting Medical Charges.

These findings highlight the potential influence of lifestyle factors, such as Smoking and BMI, on Medical Costs. Although our analysis does not establish causality, the insights suggest that healthcare policies and interventions targeting smoking cessation and weight management could potentially reduce medical expenses and improve overall health outcomes.

Limitations

While the analysis provides comprehensive insights, several limitations should be noted:

1. **Data Scope**: The analysis is based on a specific dataset, which may limit the generalizability of the findings to other populations or regions.
2. **Residual Variability**: Both models struggled to predict higher charges accurately, as indicated by the residual analysis. Further refinement of models is needed to address this issue.
3. **Unaccounted Variables**: There may be other significant factors influencing medical charges that were not included in the dataset. Future studies could explore additional variables to enhance the predictive models.

Overall, this analysis provides a solid foundation for understanding the driving factors behind medical charges and offers a basis for future research and policy-making to control healthcare costs effectively.