

**HEALTH INSURANCE**

**SAHIBA SHAHAB**

**12287**

Submitted to

Dr. Muhammad Arif Hussain

**College of Computing & Information Sciences**

**KIET Main/North Campus, Karachi Pakistan**

**ABSTRACT**

This study analyzes health insurance claims to identify trends and build predictive models. Key factors such as smoking, BMI, and age significantly influence claims. The dataset underwent preprocessing, including handling missing values, encoding categorical data, and removing duplicates. Using Linear and Polynomial Regression, the models demonstrated reliable predictions, supported by strong evaluation metrics. Visualizations highlighted important patterns, such as smokers incurring higher claims. The findings provide valuable insights for insurers and propose scalable methods for claim forecasting, with potential applications in optimizing insurance policies and risk assessment. Future work could explore advanced machine learning techniques for improved accuracy.

Table of Contents

[1. Introduction 4](#_Toc183513052)

[2. Methods 4](#_Toc183513053)

[3. Discussion 5](#_Toc183513054)

[4. Conclusion 5](#_Toc183513055)

[5. Recommendations 5](#_Toc183513056)

[6. References 6](#_Toc183513057)

[7. Appendices 6](#_Toc183513058)

[8. GRAPHS 7](#_Toc183513059)

[9. Train/Validation Data Boxplots 18](#_Toc183513060)

[10. Scatter Matrix of Features and Claim 20](#_Toc183513061)

# Introduction

Health insurance claims are a critical aspect of the insurance industry, directly impacting financial planning and risk management. This study investigates health insurance claim data to identify patterns and predict claim amounts using statistical and machine learning techniques. The objective is to aid insurers in understanding the key factors influencing claims and develop predictive models for better decision-making. The scope includes data preprocessing, feature selection, exploratory data analysis (EDA), and model evaluation. This report focuses on the effectiveness of Linear and Polynomial Regression models for claim prediction, using insights derived from real-world data.

# Methods

* **Data Collection:**
* Dataset: 1651277648862\_healthinsurance.csv
* Data includes variables such as age, BMI, smoking status, and claim amounts.
* **Preprocessing:**
* Handled missing values by imputing means for numerical variables.
* Encoded categorical variables using label encoding.
* Removed duplicate entries to ensure data quality.
* **Exploratory Data Analysis (EDA):**
* Visualized distributions and relationships using histograms, scatter plots, and box plots.
* Analyzed correlations using a heatmap.
* **Feature Selection:**
* Used SelectKBest with f\_regression to identify top predictors influencing claim amounts.
* **Modeling:**
* Developed Linear Regression and Polynomial Regression models.
* Evaluated models on training, validation, and test datasets using Mean Absolute Error (MAE) and R² scores.
* **Tools and Libraries:**
* Python libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.

# Discussion

The analysis revealed key trends in health insurance claims:

* **Factors Influencing Claims:**
  + Smoking status and BMI were found to significantly impact claim amounts.
  + Smokers had higher average claims compared to non-smokers.
  + BMI showed a positive correlation with claim amounts, suggesting higher claims for individuals with higher BMI.
* **Outlier Analysis:**
  + Identified extreme claim values using the IQR method. Outliers were analyzed separately to understand their characteristics.
* **Model Performance:**
  + Linear Regression achieved strong predictive performance with reliable evaluation metrics.
  + Polynomial Regression with degree optimization provided better accuracy, especially for non-linear relationships.
* **Visualization Insights:**
  + Scatter plots revealed patterns like increased claims with age and smoking status.
  + Histograms and count plots effectively highlighted distributions and categorical variable impacts.

# Conclusion

* + - * Key predictors of health insurance claims include smoking status, BMI, and age.
      * Linear and Polynomial Regression models provide reliable predictions, with Polynomial Regression showing better performance for complex relationships.
      * Proper data preprocessing and feature selection are crucial for enhancing model accuracy.

# Recommendations

* Explore advanced models like Random Forest, Gradient Boosting, or Neural Networks for improved accuracy.
* Incorporate additional features like lifestyle habits or medical history to enhance predictions.
* Extend the analysis to larger, more diverse datasets for better generalization.
* Develop user-friendly applications to assist insurers in real-time claim prediction and policy optimization.

# References

**Research paper**

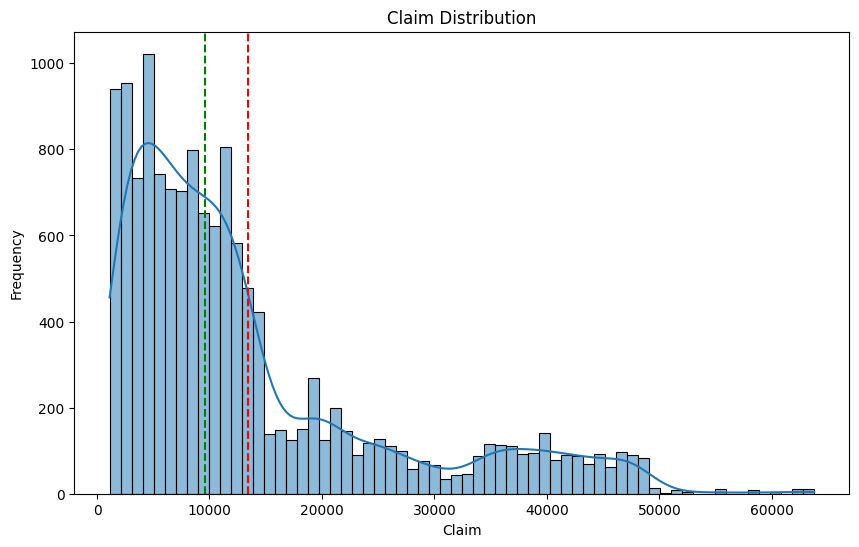
[**https://jespublication.com/upload/2019-V10-I9-34.pdf**](https://jespublication.com/upload/2019-V10-I9-34.pdf)

# Appendices

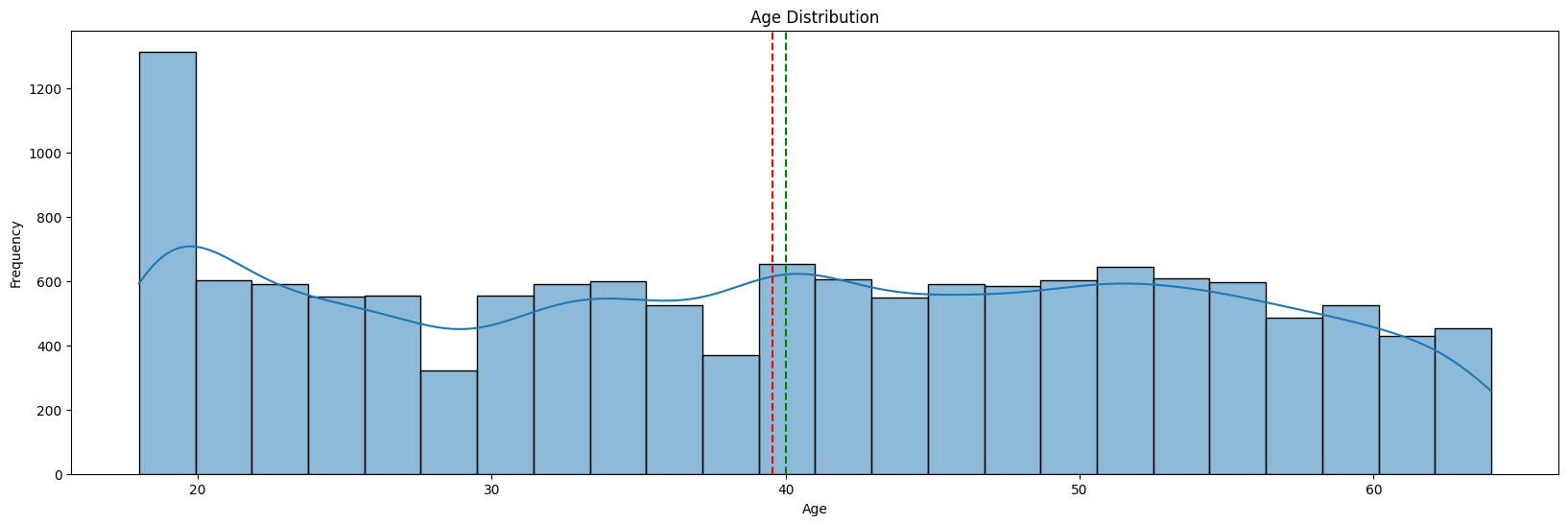
* **Python Scripts:**
* Preprocessing and EDA scripts.
* Model training and evaluation code.
* **Visualizations:**
* Histograms, scatter plots, and heatmaps.
* **Evaluation Metrics:**
* Training and validation MAE and R² scores.
* **Dataset Information:**
* Summary statistics of the dataset after preprocessing.

# GRAPHS

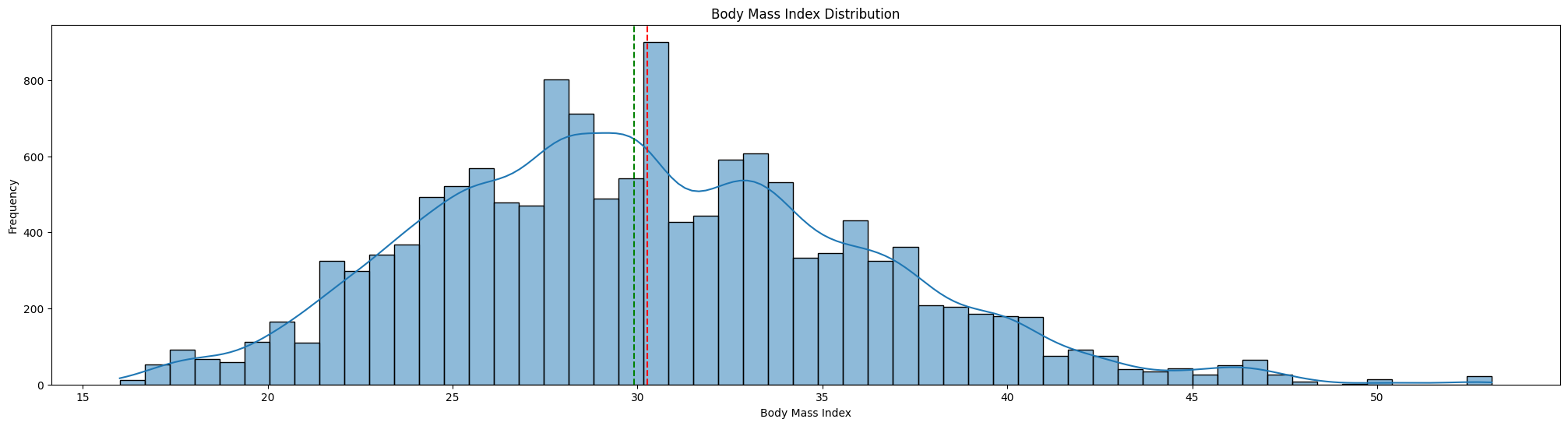
* 1. **Distribution and Relationship Analysis**
* Claim Distribution



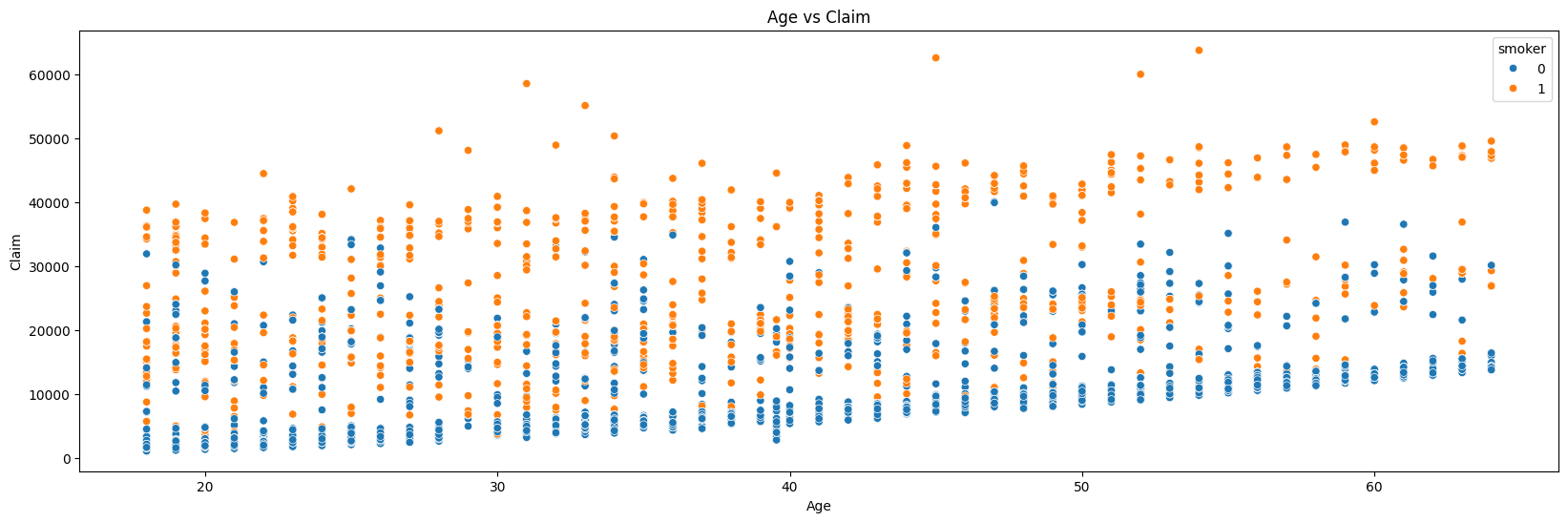
* Age Distribution

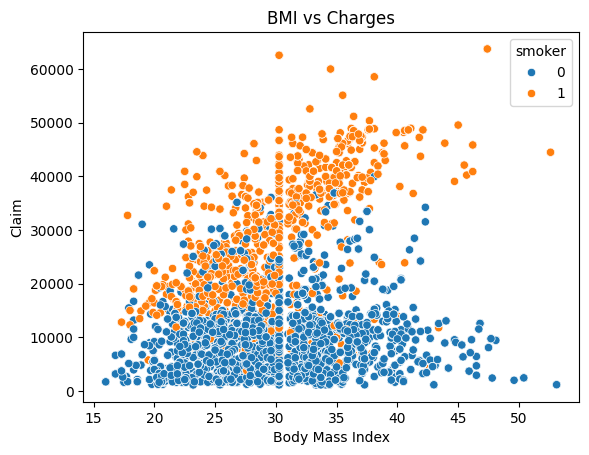


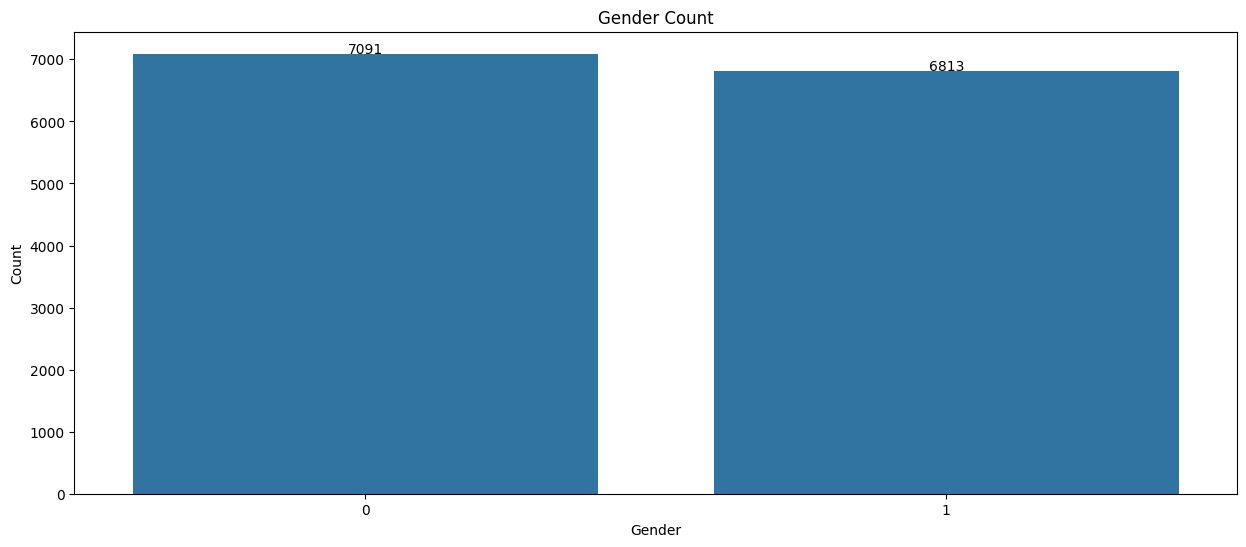
* BMI Distribution



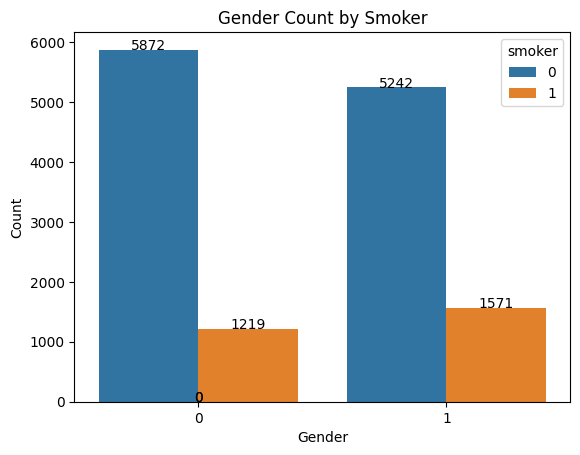
* Age vs. Claim Scatter Plot



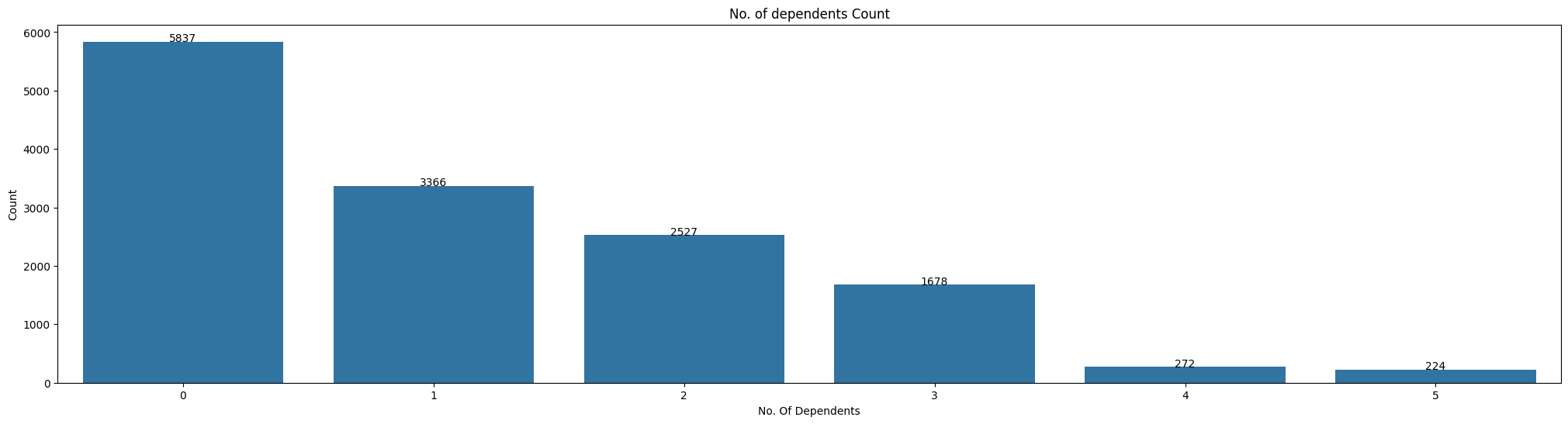
* BMI vs. Claim Scatter Plot 
  1. **Categorical Data Analysis**
* Gender Count



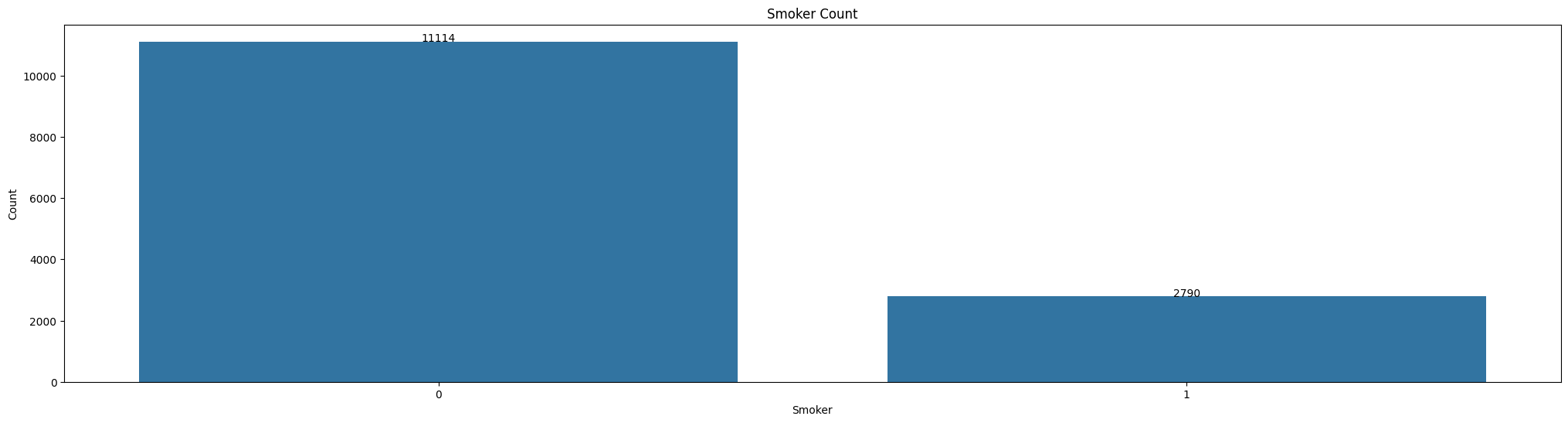
* Gender Count by Smoker



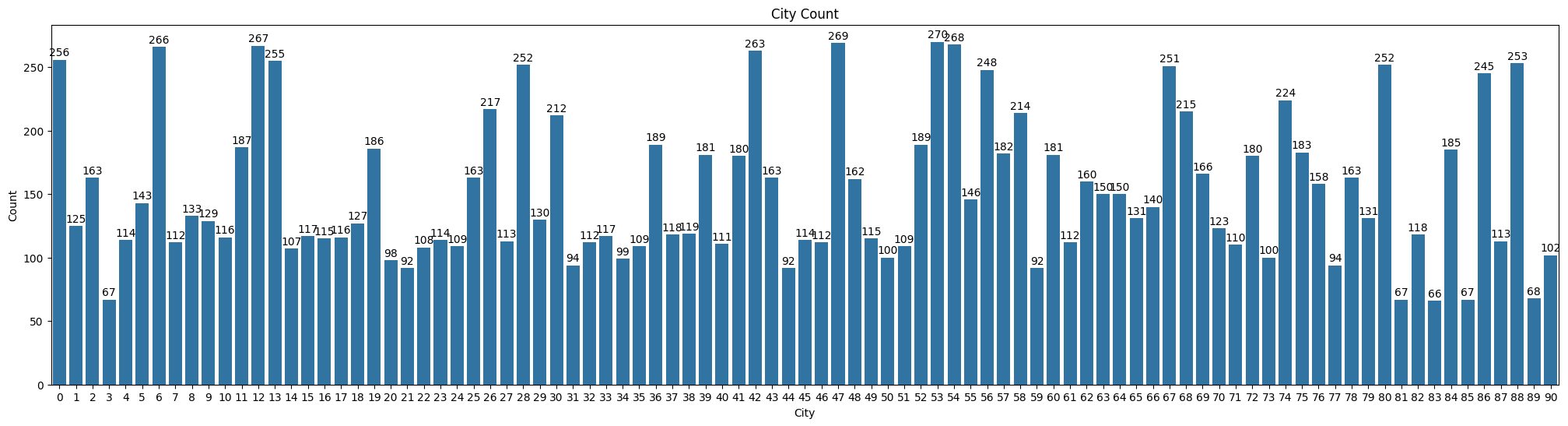
* No. of Dependents Count



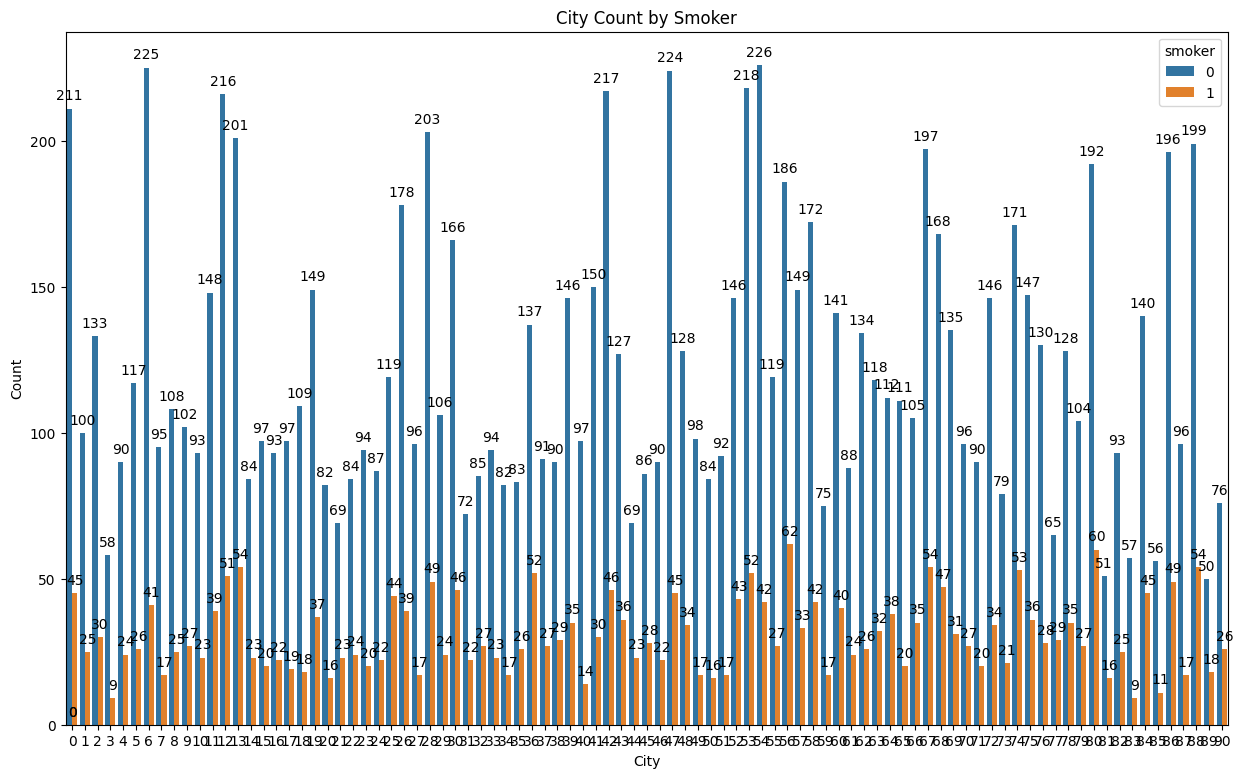
* Smoker Count



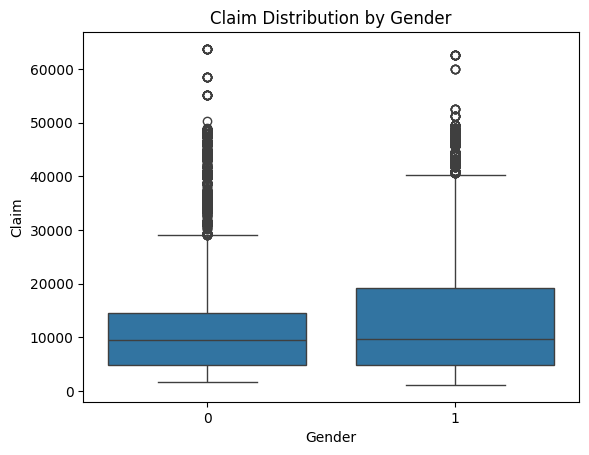
* City Count



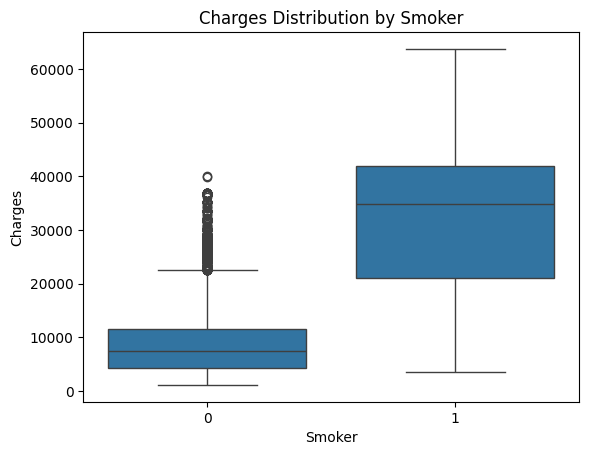
* City Count by smoker



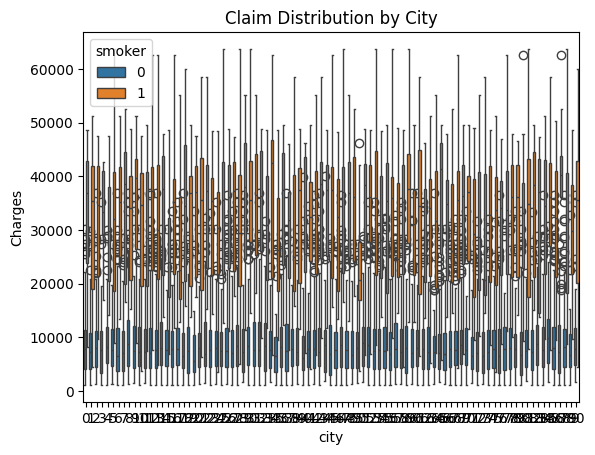
* 1. **Boxplots**
* Claim Distribution by Gender



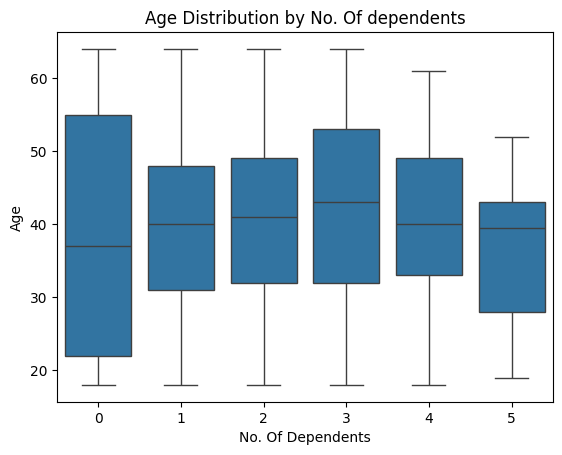
* Claim Distribution by Smoker



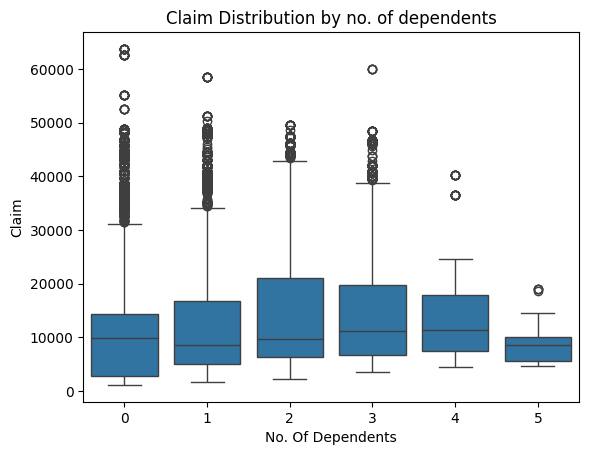
* Claim Distribution by City



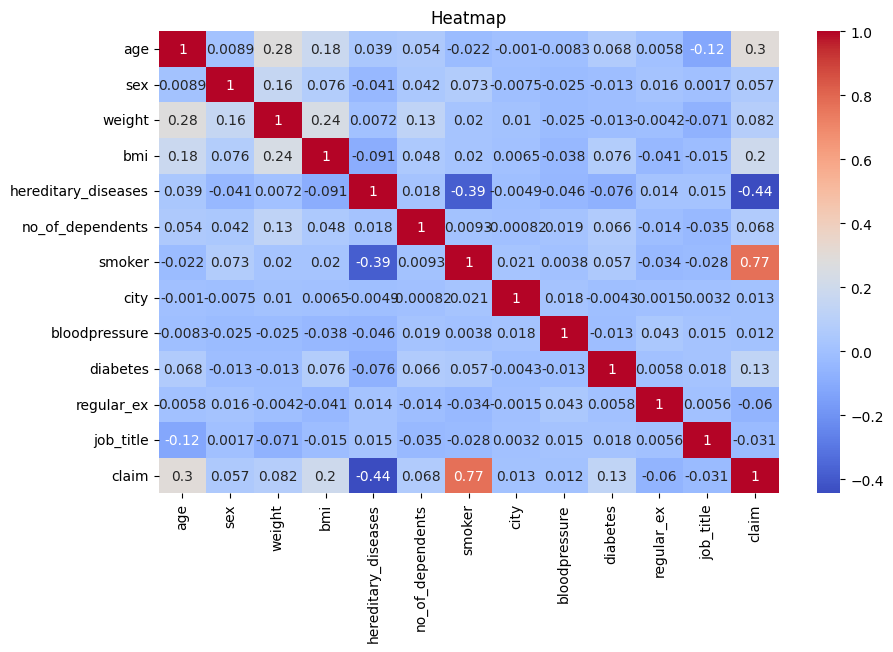
* Age Distribution by Number of Dependents



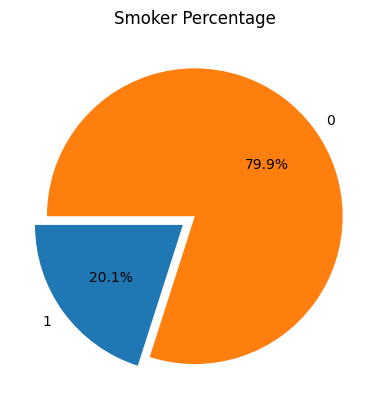
* Claim Distribution by Number of Dependents



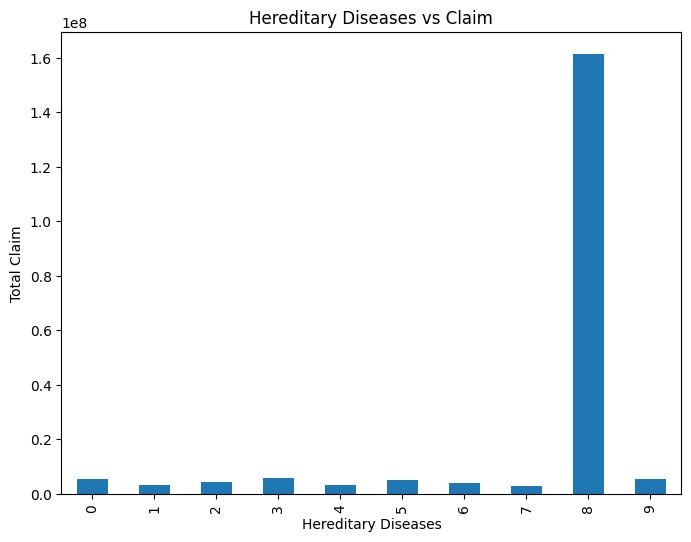
* 1. **Correlation and Heatmap**
* Correlation Heatmap



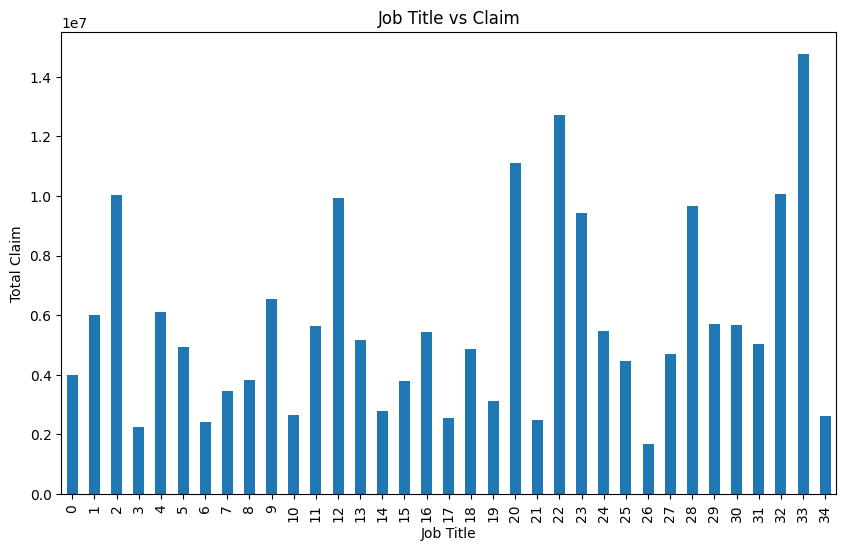
* 1. **Other Visualizations**
* Smoker Percentage Pie Chart



* Hereditary Diseases vs. Claim

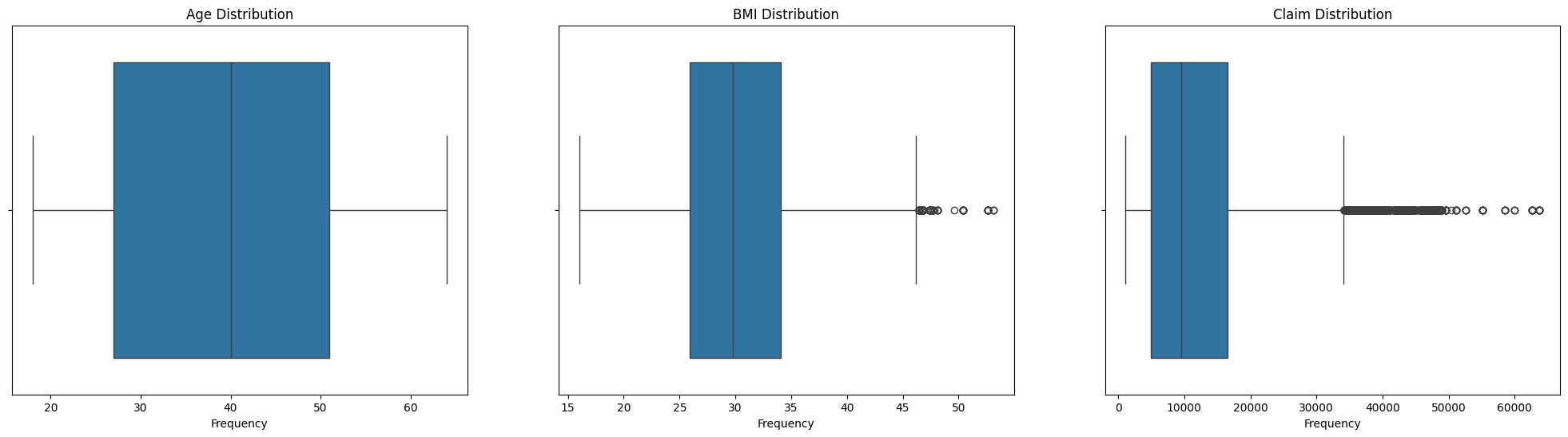


* Job Title vs. Claim

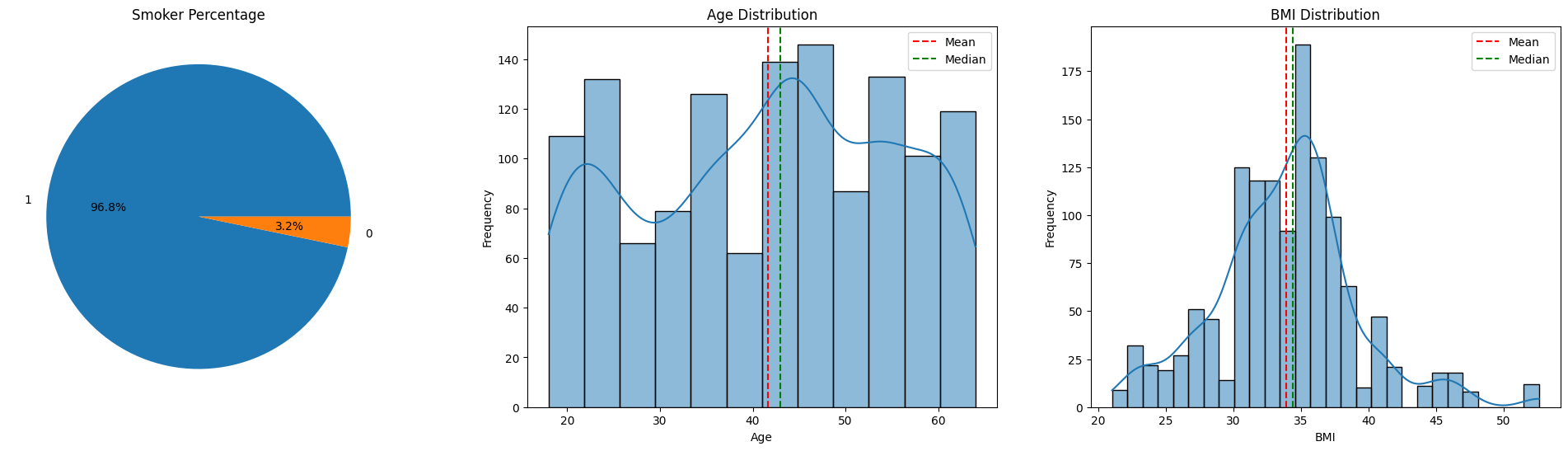


# Train/Validation Data Boxplots

* 1. **Age, claim and BMI distribution**



* 1. **Outlier analysis**
* Smoker percentage, Claim, Age distribution



# Scatter Matrix of Features and Claim

