Final Report:

Data Analysis of the US Health Insurance Dataset

ISYS 812 Programming and Applications for Data Analytics

Padmasree Sappa

Submission Date: May 24, 2024

Table of Contents

Introduction:

The project is a comprehensive analysis utilizing the US Health Insurance Dataset to understand the dynamics of health insurance charges. As healthcare costs continue to rise, it becomes increasingly important for stakeholders including policymakers, insurance companies, and individuals to grasp the underlying factors that drive these expenses. This study leverages data analytics to explore how demographic and health-related characteristics influence the cost of health insurance. Insights derived from this project are intended to aid in developing more equitable and efficient health insurance policies.

Project Objectives:

Our project aims to uncover underlying trends and determinants influencing health insurance charges in the United States. We seek to provide actionable insights that could influence policy-making and personal health decisions by analyzing demographic, health, and lifestyle data.

1. Data Understanding and Exploration:

- **Comprehensive Dataset Review**: The US Health Insurance Dataset, includes detailed entries for 1,338 individuals covering 7 key attributes like age, sex, BMI, number of children, smoker status, region, and charges.
- **Descriptive Analysis**: Conduct an initial review of the data to establish a baseline understanding of the distributions and ranges of the attributes, ensuring that all necessary data quality checks are performed to prepare for deeper analysis.

2. Exploratory Data Analysis (EDA):

- **Correlation Analysis**: Investigate how different attributes such as age, BMI, and smoking status correlate with insurance charges. This will involve statistical tests and visualizations to identify significant predictors of health insurance costs.
- **Impact of Lifestyle Choices**: Assess how lifestyle choices, specifically smoking and BMI, impact insurance costs. This analysis will help understand the financial implications of personal health choices on insurance premiums.

3. Predictive Modelling:

- **Model Development**: Build predictive models using linear regression, random forests, and gradient-boosting machines to estimate future health insurance charges based on demographic and health-related attributes.

- **Feature Engineering**: Introduce new features, such as risk scores calculated from critical factors including smoking status and BMI, to enhance model accuracy and relevance.

4. Model Evaluation and Optimization:

- **Performance Metrics**: Evaluate the models based on metrics such as Mean Squared Error (MSE) and R-squared to determine the best model for predicting insurance charges.
- **Model Refinement**: Refine models through iterative improvements and tuning of parameters to ensure optimal performance in predicting the insurance charges.

5. Reporting and Insights:

- **Insightful Narratives**: Develop narratives and insights that highlight the key drivers of health insurance costs. This will include detailed data analysis and visualization, illustrating how different demographic and health-related factors impact insurance charges.
- **Policy Recommendations**: Provide recommendations based on the analysis to help stakeholders make informed decisions about health insurance policies and personal health management.

Why This Project:

1. Addressing Healthcare Affordability: Understanding factors that drive insurance costs can help develop strategies to make healthcare more affordable.

- **Cost Reduction Insights:** This project aims to uncover the specific factors that most significantly influence health insurance charges by delving into the US Health Insurance Dataset. Understanding these factors can provide critical insights for insurance companies and policymakers to devise strategies to reduce premiums, especially for high-risk demographics. This could include tailoring health plans that mitigate high costs for individuals with certain risk factors such as high BMI or smoking habits.
- **Targeted Policy Implementation:** With detailed insights into how various factors like age, region, and lifestyle choices impact insurance costs, more effective and targeted healthcare policies can be developed. These policies could focus on preventive care measures in regions or demographic groups most affected by high insurance charges, thereby making healthcare more accessible and affordable.

2. Practical Application: This analysis offers us practical experience in healthcare analytics, which is pivotal in today's data-driven policy and business environments.

- **Real-World Data Analysis Skills:** The project provides hands-on experience with real-world data, allowing the team to apply theoretical knowledge in data analytics to practical problems in healthcare. This experience is invaluable in today's healthcare industry, which relies increasingly on data-driven decisions to optimize outcomes and resource allocation.
- **Career Advancement:** This project serves as a crucial stepping stone for aspiring data analysts, demonstrating their ability to handle complex datasets and extract meaningful insights that can influence real-world decisions in healthcare economics and insurance. Such practical experience is highly regarded in the job market and necessary for roles in data science and analytics within the healthcare and insurance industries.

Dataset Overview:

Name: US Health Insurance Dataset
(https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data)

Source: The dataset is publicly available on Kaggle, and it comprises detailed entries about individuals with health insurance in the United States.

Size and Scope:

- **Entries:** The dataset contains 1,338 individual records.
- **Attributes:** It includes 7 key attributes per entry.

Attributes Description:

- **Age:** The age of the insured individual, ranging from 18 to 64 years. This is a continuous variable that helps analyze how age affects insurance costs.
- **Sex:** The gender of the insured (male or female). This categorical attribute can be used to explore any gender disparities in insurance charges.
- **BMI:** Body Mass Index (BMI) of the individual, a key health metric that indicates the body fat based on height and weight. Values range from 15.96 to 53.13, which can be categorized into underweight, healthy weight, overweight, and obese.
- **Children:** The number of children/dependents covered under the insurance plan of the individual, ranging from 0 to 5. This variable helps us understand how family size affects insurance costs.

- **Smoker:** Indicates whether the individual is a smoker (yes or no). This is a crucial factor as smoking status significantly impacts health risks and, consequently, insurance premiums.
- **Region:** The residential region of the insured, is categorized into four areas: northeast, southeast, southwest, and northwest. This allows analysis of regional variations in insurance charges.
- **Charges:** The individual insurance charges billed to the person, which vary widely from approximately $1,122 to $63,770. This is the dependent variable in most analyses and models to determine how other factors influence these charges.

Data Quality:

- The dataset is clean with no missing values, ensuring robust analyses and modelling. Outliers exist, particularly in the BMI and insurance charges attributes, which are important for understanding extremes in insurance costs.

Usage in Analysis:

- The dataset is used extensively for exploratory data analysis to identify correlations and patterns.
- It serves as the basis for predictive modelling, where the goal is to predict health insurance charges based on demographic and health-related attributes.
- Advanced statistical techniques and data visualization tools are employed to extract actionable insights and to demonstrate the impact of various factors on insurance costs.

Significance:

- This dataset provides valuable insights into factors influencing health insurance charges. Understanding these factors is crucial in addressing healthcare affordability and accessibility issues.
- Analysing this dataset enables practical experience in healthcare analytics, a field with significant real-world applications and implications.

Analysis Goals:

1. Identify the main factors influencing health insurance charges:

- **Objective:** Determine which attributes like age, BMI, smoker status, number of children, gender, and region have significant impacts on the costs of health insurance.

- **Approach:** Analyse the dataset to explore correlations and trends between these factors and insurance charges. This involves using statistical tools to see which attributes are strongly associated with higher or lower insurance premiums.
- **Benefit:** Understanding these factors helps in recognizing what drives insurance costs, which is crucial for both consumers and providers in managing and setting insurance rates.

2. Predict charges based on demographic and health-related attributes:

- **Objective:** Use the data about individuals to develop a model that can accurately predict health insurance charges based on demographic and health-related attributes.
- **Approach:** Build and test several predictive models using techniques like linear regression or more complex algorithms depending on the data's characteristics. These models will be evaluated to find the best one for accurate predictions.
- **Benefit:** Creating a predictive model allows stakeholders to estimate insurance costs based on a person's specific data, making insurance more predictable and manageable.

What to Expect:

1. Exploratory Data Analysis (EDA):

Conduct a detailed exploration to uncover how different personal and health factors (like age, BMI, and smoking) are linked to insurance costs.

2. Insights Discussion:

Present and discuss the key findings that have practical implications, showing how these insights can be applied in real-world scenarios.

3. Predictive Modelling:

Develop models that can accurately forecast health insurance charges based on someone's demographic and health information.

Driving Question:

"What are the main factors influencing health insurance charges, and how can we predict charges based on demographic and health-related attributes?"

Data Preparation:

The preparation involved organizing the data to ensure it was in an optimal state for analysis. This included addressing any quality issues and preparing the dataset for detailed exploratory and predictive analyses.

Before proceeding to actual analysis and modelling, the data needs to be prepared:

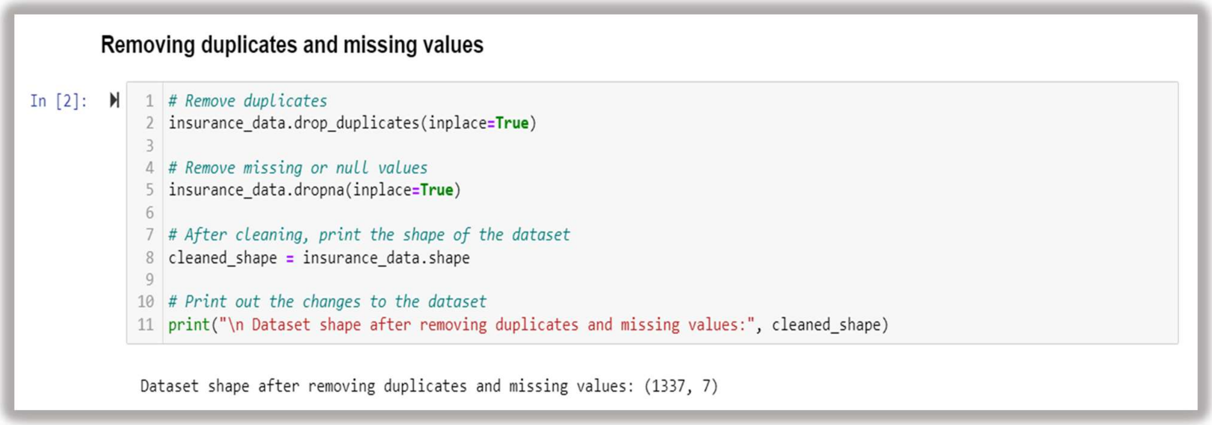
- Handling outliers, especially in BMI and charges.

- Encoding categorical variables like Sex, Smoker Status, and Region for model ingestion.

Data Cleaning:

Data Cleaning: Duplicates and Missing Values

- **Duplicates:** A check for duplicates revealed one duplicate entry. This entry was removed to prevent any skew in the analysis. Duplicate entries can misrepresent the data, leading to inaccurate conclusions.
- **Missing Values:** An examination of the dataset showed no missing values across all columns. This is advantageous as it means there is no need for imputation, which can sometimes introduce bias or distort the true characteristics of the data.



Removing duplicates and missing values

In [2]:

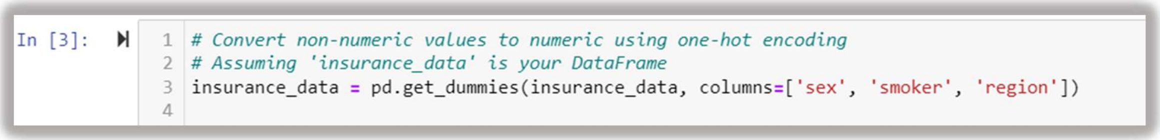
```python
# Remove duplicates
insurance_data.drop_duplicates(inplace=True)

# Remove missing or null values
insurance_data.dropna(inplace=True)

# After cleaning, print the shape of the dataset
cleaned_shape = insurance_data.shape

# Print out the changes to the dataset
print("\n Dataset shape after removing duplicates and missing values:", cleaned_shape)
```

```
 Dataset shape after removing duplicates and missing values: (1337, 7)
```

Data Manipulation:

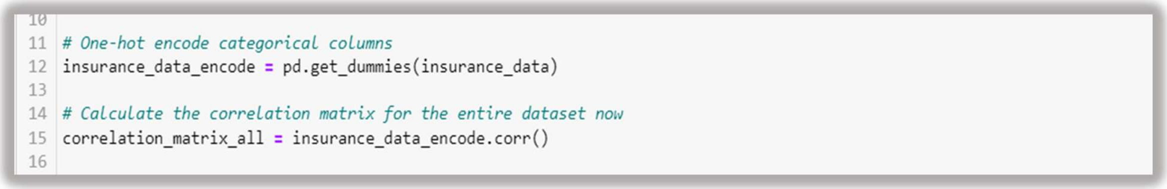
Data Encoding and Transformation

- **Action:** The categorical variables 'sex', 'smoker', and 'region' were transformed into numeric formats through one-hot encoding. This process is crucial for preparing the dataset for machine learning algorithms, which typically require numerical input. One-hot encoding prevents an ML model from learning a precedence order in categorical variables and thus treats them as nominal values instead of ordinal values.
- **Result:** This encoding effectively modifies the dataset to include binary columns for each category, enabling the incorporation of these variables into statistical and machine-learning models.

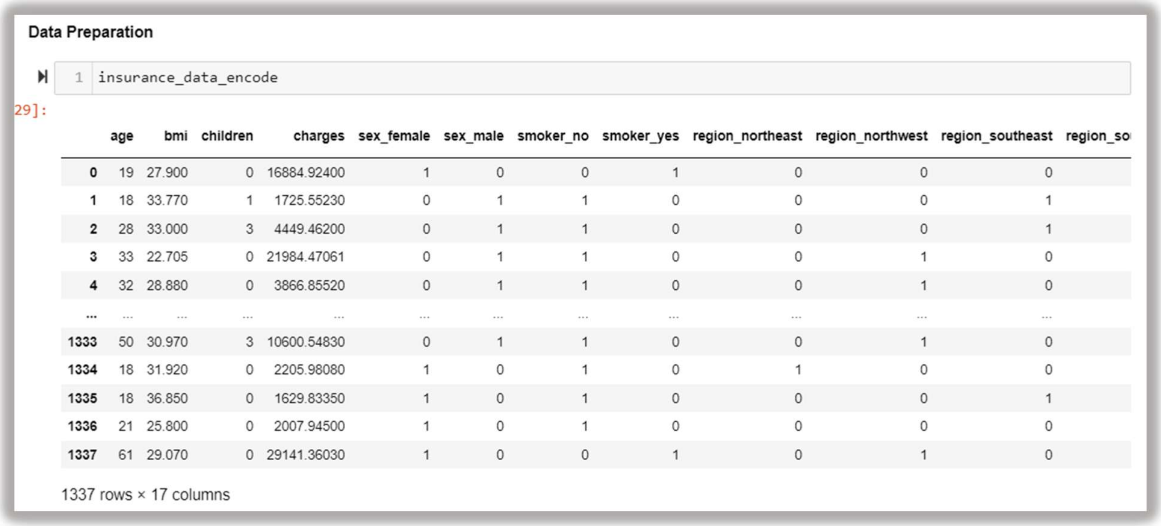


In [3]:

```python
# Convert non-numeric values to numeric using one-hot encoding
# Assuming 'insurance_data' is your DataFrame
insurance_data = pd.get_dummies(insurance_data, columns=['sex', 'smoker', 'region'])

```



```
# One-hot encode categorical columns
insurance_data_encode = pd.get_dummies(insurance_data)

# Calculate the correlation matrix for the entire dataset now
correlation_matrix_all = insurance_data_encode.corr()
```



Data Preparation

```
insurance_data_encode
```

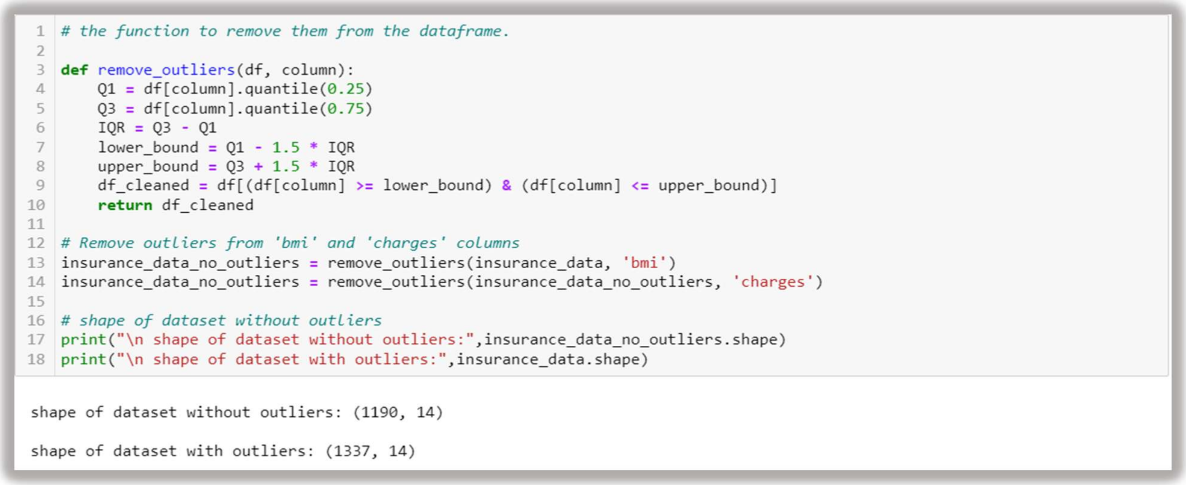
29]:

	age	bmi	children	charges	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	region_so
0	19	27.900	0	16884.92400	1	0	0	1	0	0	0	
1	18	33.770	1	1725.55230	0	1	1	0	0	0	1	
2	28	33.000	3	4449.46200	0	1	1	0	0	0	1	
3	33	22.705	0	21984.47061	0	1	1	0	0	1	0	
4	32	28.880	0	3866.85520	0	1	1	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	
1333	50	30.970	3	10600.54830	0	1	1	0	0	1	0	
1334	18	31.920	0	2205.98080	1	0	1	0	1	0	0	
1335	18	36.850	0	1629.83350	1	0	1	0	0	0	1	
1336	21	25.800	0	2007.94500	1	0	1	0	0	0	0	
1337	61	29.070	0	29141.36030	1	0	0	1	0	1	0	

1337 rows × 17 columns

Handling Outliers and Scaling

- **Outliers:** Outliers were identified in the 'BMI' and 'charges' columns using the Interquartile Range (IQR) method. Specifically, 9 outliers were noted for 'BMI' and 139 for 'charges'. The original dataset contained 1337 entries. After removing outliers from the 'BMI' and 'charges' columns, the dataset was reduced to 1190 entries. These outliers could represent extreme cases due to unique health conditions or data entry errors.
- **Scaling:** While specific scaling actions (such as standardization or normalization) aren't detailed, these steps are typically considered to ensure that all numerical data contribute equally to the analysis, avoiding bias towards variables with larger scales.
- **Decision on Outliers:** The decision to retain or adjust these outliers was postponed until after an initial visualization, to better understand their impact on the dataset's overall distribution.

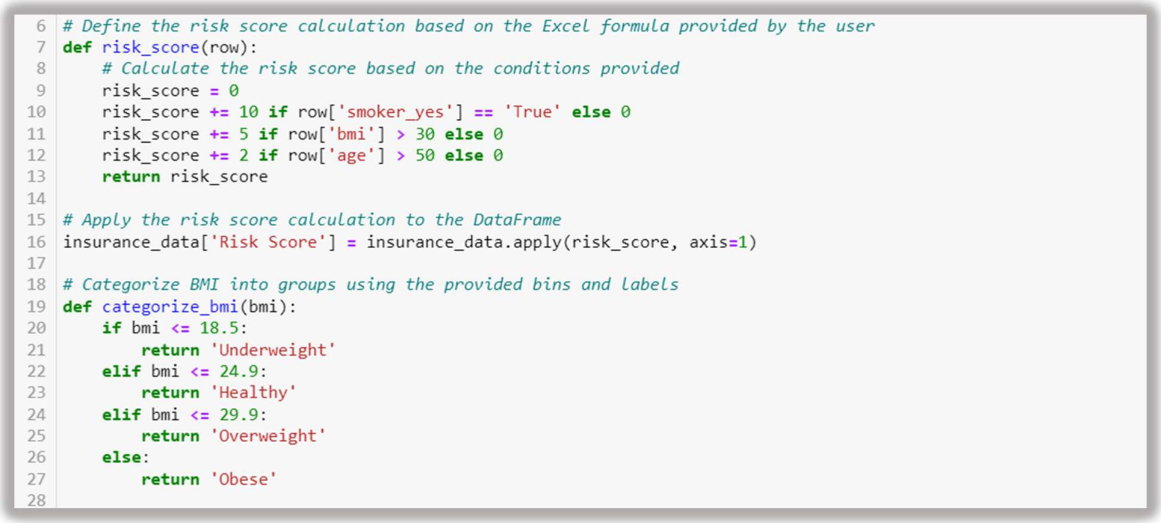


```
# the function to remove them from the dataframe.

def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df_cleaned = df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]
    return df_cleaned

# Remove outliers from 'bmi' and 'charges' columns
insurance_data_no_outliers = remove_outliers(insurance_data, 'bmi')
insurance_data_no_outliers = remove_outliers(insurance_data_no_outliers, 'charges')

# shape of dataset without outliers
print("\n shape of dataset without outliers:",insurance_data_no_outliers.shape)
print("\n shape of dataset with outliers:",insurance_data.shape)
```

```
shape of dataset without outliers: (1190, 14)

shape of dataset with outliers: (1337, 14)
```

Risk Score and Obesity Column

- **Risk Score Calculation:** A 'Risk Score' was calculated for each individual, taking into account smoking status, BMI, and age.
- **BMI Categorization:** Individuals' BMI values were categorized into 'Underweight', 'Healthy', 'Overweight', and 'Obese'.



```
# Define the risk score calculation based on the Excel formula provided by the user
def risk_score(row):
    # Calculate the risk score based on the conditions provided
    risk_score = 0
    risk_score += 10 if row['smoker_yes'] == 'True' else 0
    risk_score += 5 if row['bmi'] > 30 else 0
    risk_score += 2 if row['age'] > 50 else 0
    return risk_score

# Apply the risk score calculation to the DataFrame
insurance_data['Risk Score'] = insurance_data.apply(risk_score, axis=1)

# Categorize BMI into groups using the provided bins and labels
def categorize_bmi(bmi):
    if bmi <= 18.5:
        return 'Underweight'
    elif bmi <= 24.9:
        return 'Healthy'
    elif bmi <= 29.9:
        return 'Overweight'
    else:
        return 'Obese'

```

Utility of New Columns: The 'Risk Score' helps in identifying individuals who may require more medical services and thus higher insurance coverage. The 'BMI Category' facilitates quick assessment of health risks associated with body weight.

Exploratory Data Analysis (EDA):

Statistical Techniques:

- **Correlation Analysis:** The dataset was encoded using one-hot encoding to convert non-numeric categorical variables into a numerical format, facilitating correlation analysis with insurance charges.
- **Results:** The correlation analysis provided clear insights:
  1. **Age** has a moderate positive correlation (0.2983) with charges, suggesting insurance costs increase with age.
  2. **BMI** shows a moderate positive correlation (0.1984), indicating higher BMI is linked to higher charges.
  3. **Children** present a weak positive correlation (0.0674), slightly influencing charges.
  4. **Smoker status** exhibits a strong positive correlation (0.7872), highlighting a significant impact on charges.
  5. **Sex (Female vs. Male)** displays a weak negative correlation (-0.0580), suggesting females might incur slightly lower charges than males.
  6. **Risk Score** correlates moderately (0.2602) with charges, validating its use in assessing insurance costs.

Analytical Techniques:

- **Encoded Dataset Analysis:** Post one-hot encoding, the relationships between all variables were examined. The strong correlation of the risk score with charges underscores its effectiveness in evaluating potential insurance costs.

Visualization Techniques:

- **Heatmap Visualization:** A heatmap was used to visually represent the correlation matrix, aiding in the quick assessment of how various factors are interrelated.
- **Scatter Plots and Box Plots:** These visualizations focused on relationships like age versus charges and the impact of smoking on charges.
- **Histograms:** Used to analyze the distribution of insurance charges, providing insights into the skewness and presence of outliers.

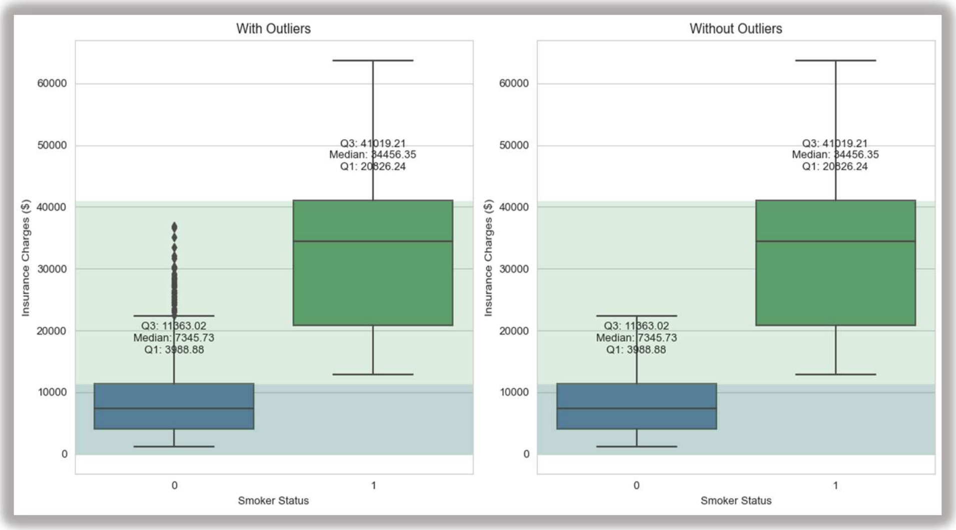
Dataset Observations:

- **With and Without Outliers:** Separate analysis was conducted for datasets with and without outliers to assess the impact of extreme values on insurance charges.

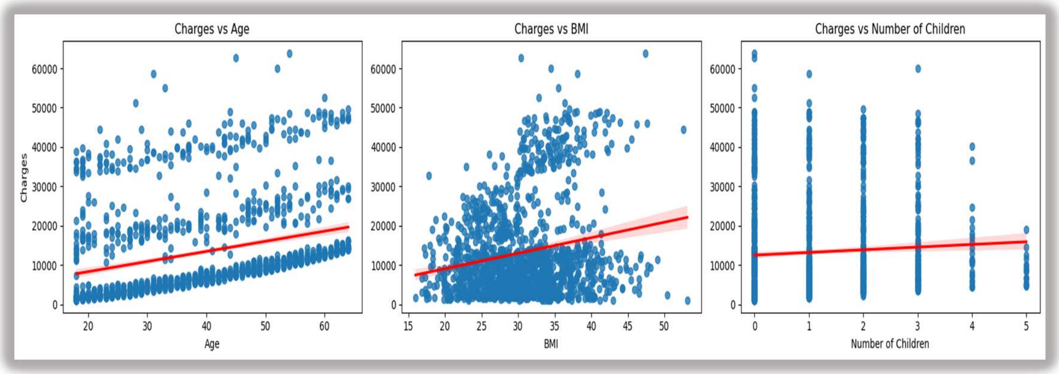
1. **Smoker status:** Both scenarios showed a surging high insurance charge for smokers, justifying the strong correlation with charges.
2. **Age:** Both scenarios showed that older individuals tend to face higher insurance charges.
3. **BMI:** A positive correlation with charges was observed, though it was less pronounced compared to age.
4. **Children:** The number of children did not show a strong correlation with charges, suggesting it's a less significant predictor.

Dataset with Outliers:

- **Charges vs Smoker:** The box plot shows the outliers for non-smokers having a high insurance charge, while the smoker's insurance trend is similar to the non-outliers box plot.

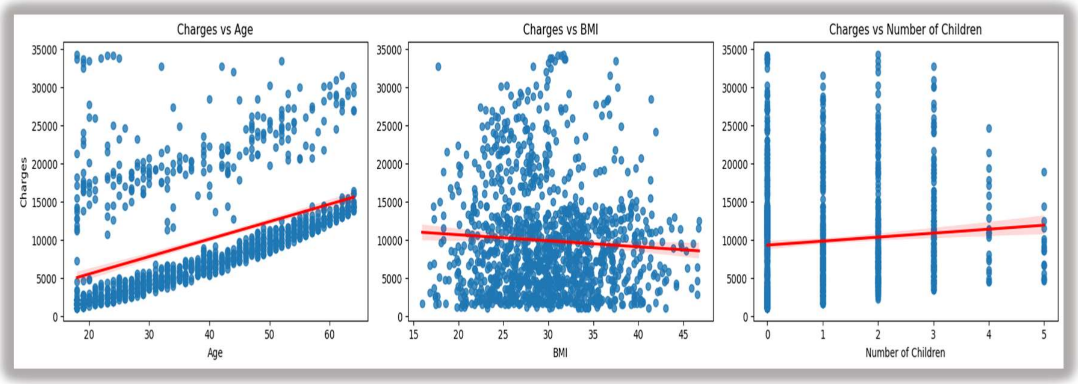


- **Charges vs Age:** The inclusion of outliers shows a similar trend to the dataset without outliers, but with more variation in charges, especially in the higher age bracket, which could be due to complex medical needs or outlier cases with high medical expenses.
- **Charges vs BMI:** With outliers, the scatter plot reveals that high BMI values can be associated with very high charges, likely reflecting extreme cases or individuals with significant health issues.
- **Charges vs Number of Children:** Again, more variation is observed with outliers present. The data shows that individuals with three or more children can have high insurance charges, possibly reflecting the higher costs associated with larger families.



Dataset without Outliers:

- **Charges vs Smoker Status:** The box plot shows that the non-outliers drop high insurance charges for non-smokers having a high insurance charge, while the smoker's insurance charges trend is similar for both cases. The high charges for non-smokers could be caused by some other reasons such as age, high BMI, or maybe something else.
- **Charges vs Age:** A clear upward trend suggests that as age increases, so do the insurance charges. This aligns with the expectation that older individuals may require more medical care and thus incur higher insurance costs.
- **Charges vs BMI:** The trend is less pronounced, but there appears to be a positive correlation between BMI and charges, albeit weaker than age. This might indicate that individuals with higher BMI, potentially at greater risk for health issues, could face higher charges.
- **Charges vs Number of Children:** The relationship here seems to be more dispersed, with no clear trend observed. This suggests that the number of children might not be a strong predictor of insurance charges, or other factors may have a more significant impact.



Thus, based on the above patterns, we decided to move with the original dataset that contains the outliers. Based on the small dataset that we have this would give us and

Implications and Conclusions:

- **Risk Score Utility:** The correlation findings validate the risk score as a crucial metric for insurance companies to assess individual risks and adjust premiums.
- **Health Risks and Insurance Costs:** The analysis confirms that lifestyle factors like smoking and health metrics such as BMI significantly influence insurance charges.
- **Policy Design and Lifestyle Changes:** These insights can guide insurance companies in risk stratification and help individuals make informed decisions about lifestyle changes to potentially lower their insurance costs.

Visualizations and Graphical Insights:

- **Scatter Plot (Age vs. Charges by Smoking Status):** Illustrated how age and smoking status influence charges, with smokers generally facing higher costs.
- **Box Plot (Charges for Smokers vs. Non-Smokers):** Highlighted the stark contrast in charges between smokers and non-smokers, reinforcing the financial impact of smoking.
- **Histogram (Distribution of Insurance Charges):** Offered a detailed view of how charges are distributed among the insured, emphasizing the range and skewness of the data.

Data Analysis:

In this phase, we seek to find how the above-mentioned parameters influence the insurance charges and what valuable insights can be drawn from those trends. Each insight is broken down into sections, such as:

- **Data Analysis Questions** – Questions we seek to answer through the insight.
- **Finding** – Our observations from the insight.
- **Narrative** – Our understanding from the analysis and findings.
- **Interpretation** – Explanation provided by the insight.
- **Policy Implication** – Steps that can be taken by the insurance company and society to improve the conditions.

Insight – 1: Significant Impact of Smoking on Insurance Charges

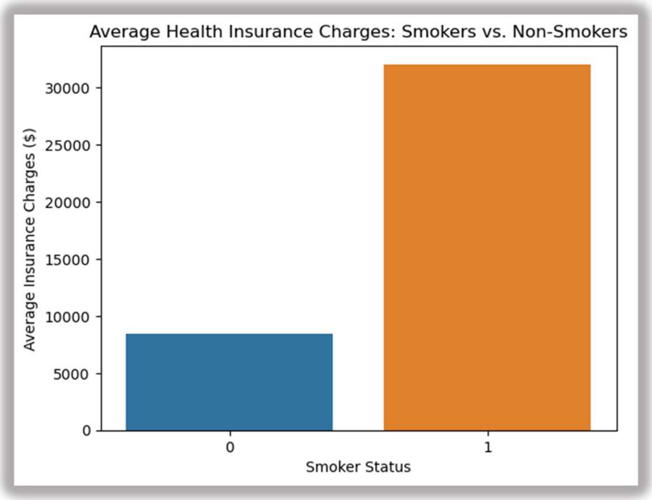
i. **Data Analysis Questions:**

   a. Comparison of insurance charges for smokers and non-smokers?

b. How does age influence health insurance charges across different smoking statuses?

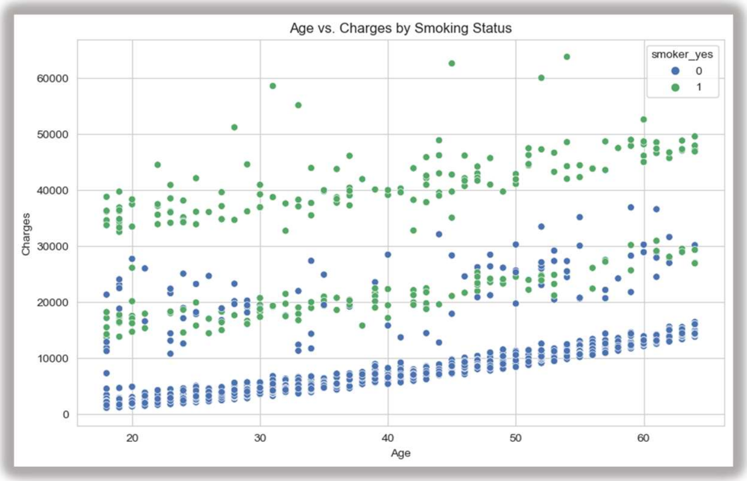
c. How do smoking habits influence insurance charges across different age groups and BMI categories? (Please refer to the Python code file for the graphs)

ii. **Finding:** Smokers incur dramatically higher insurance charges, averaging $32,050 compared to $8,440 for non-smokers



iii. **Narrative:** Smoking significantly elevates health risks, leading to substantial increases in medical and insurance costs. This disparity not only reflects health risks but also underscores potential savings from smoking cessation.

iv. **Interpretation:** The analysis demonstrates a clear, quantifiable impact of smoking on insurance costs, with smokers facing nearly four times the charges of non-smokers, consistently across various age groups and irrespective of BMI levels.



v. **Policy Implication:** These findings support policies for integrated smoking cessation programs within health insurance plans, potentially adjusting premiums to encourage and reward non-smoking behaviors. Public health initiatives could also focus more on smoking cessation to reduce healthcare costs universally.

Insight – 2: Progressive Increase in Insurance Charges with Age

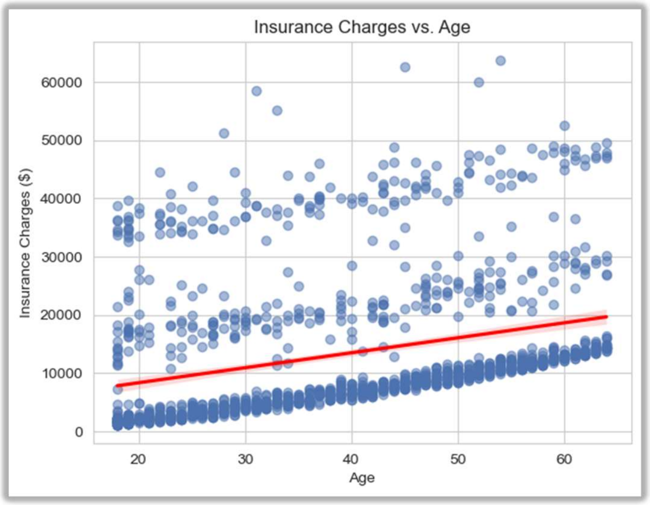
i. **Data Analysis Question:**

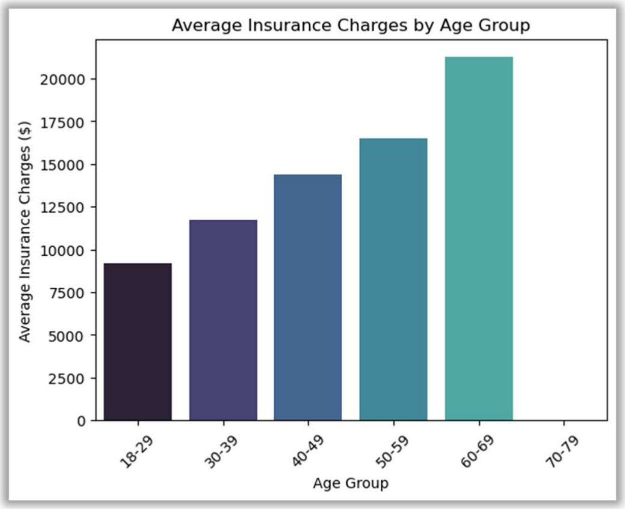
   a. Correlation coefficient between age and charges

   b. Scatter plot with a regression line for Insurance Charges vs. Age

   c. Average Insurance Charges by Age Group

ii. **Finding:** Insurance charges show a moderate positive correlation (r=0.30) with age, indicating progressively higher costs as beneficiaries age.





iii. **Narrative:** The increasing insurance charges with age reflect rising health risks and medical needs as people get old. This trend highlights the critical need for early preventive measures and health maintenance.

iv. **Interpretation:** The correlation underscores the financial implications of aging on healthcare costs, emphasizing the necessity for healthcare policies that focus on enhancing the quality of life through preventative care and support for healthy aging.

v. **Policy Implication:** Insights suggest that insurance companies and policymakers should consider age-adjusted premium structures and promote early interventions that could mitigate health issues later in life. Programs encouraging active and healthy lifestyles could be beneficial in reducing future medical expenses.

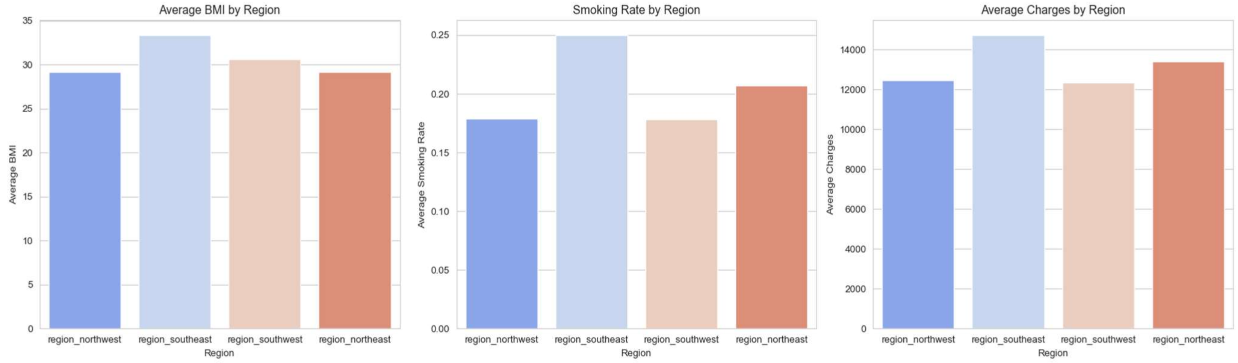
Insight – 3: Regional Differences in Insurance Charges

i. **Data Analysis Question:**

   a. Which US region is susceptible to a higher insurance charge?

   b. What can be the crucial reasons behind the surging charges?

ii. **Finding:** The Southeast region has the highest average insurance charges at $14,735, with noticeable differences from other regions.



iii. **Narrative:** Variations in insurance charges across regions could be influenced by regional lifestyle differences, access to healthcare services, and local regulatory environments.

iv. **Interpretation:** This insight highlights the complex interplay between geographic factors and health insurance costs, suggesting that regional policies and healthcare accessibility play significant roles in determining insurance charges.

v. **Policy Implication:** Addressing regional disparities in insurance charges could involve examining and potentially reforming healthcare access and insurance regulations in high-cost regions. Enhancing healthcare infrastructure and accessibility could also be targeted to balance out regional insurance costs.

Insight – 4: Correlation Between BMI and Insurance Charges

i. **Data Analysis Question:**

   a. Are there any observable trends in insurance charges over different BMI categories when adjusted for smoking status?

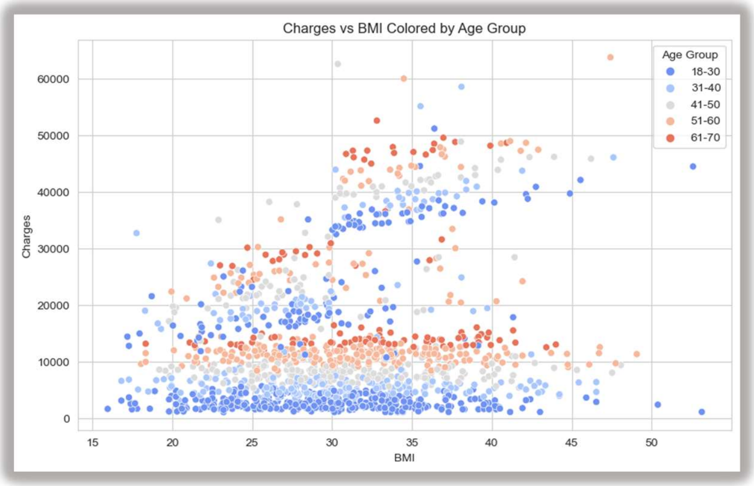
   b. Correlation coefficient between BMI & insurance charges.

   c. Relationship Between BMI and Insurance Charges

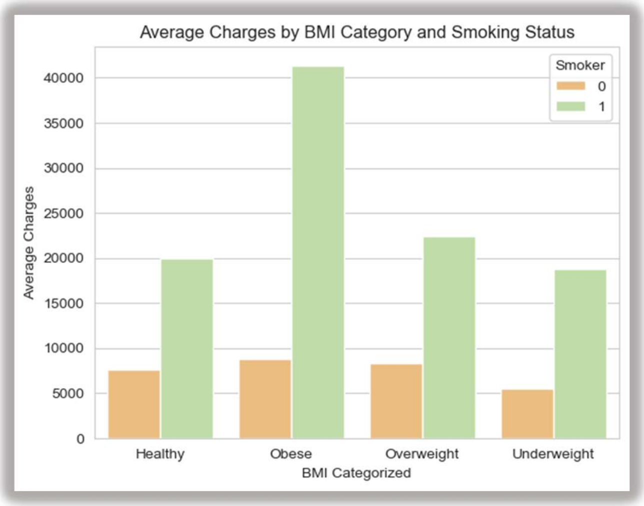
   d. How do age and BMI together influence the insurance charges?

   e. Impact of extreme BMI values on health insurance charges across different demographics

ii. **Finding:** There is a positive correlation (r=0.20) between BMI and insurance charges, indicating that higher BMI is associated with higher insurance costs.



iii. **Narrative:** This relationship emphasizes the financial impact of obesity on insurance premiums, highlighting the broader economic consequences of higher BMI levels.



iv. **Interpretation:** The correlation between BMI and insurance charges reaffirms the health risks associated with obesity, including increased chronic disease risk which directly translates into higher healthcare costs.

v. **Policy Implication:** Health policy initiatives could focus on preventive health measures to reduce obesity, such as promoting physical activity and healthier eating habits. Insurance companies might also consider offering incentives for weight loss and maintenance programs to reduce the burden of obesity-related health costs.

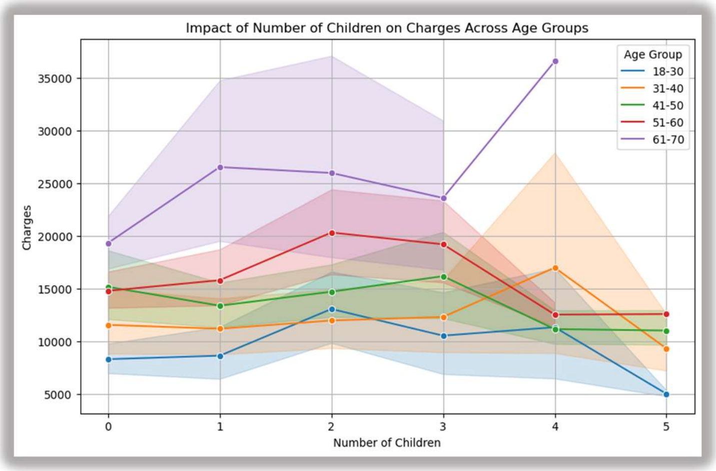
Insight – 5: The Influence of Children on Insurance Charges

i. **Data Analysis Question:**

a. What is the impact of the number of children on insurance charges across different age groups?

b. Correlation coefficient between number of children and insurance charges

c. Visualizing the relationship between the number of children and insurance charges



ii. **Finding:** There's a weak positive correlation (r=0.07) between the number of children a beneficiary has and their insurance charges.

iii. **Narrative:** Adding dependents to a policy slightly increases charges due to the higher potential for medical care needs, but the impact is relatively minor.

iv. **Interpretation:** The modest increase in charges with additional children suggests that while family size does influence insurance costs, other factors such as the policyholder’s age and personal health status are more significant determinants.

v. **Policy Implication:** The findings could guide the design of family-based insurance plans that provide cost-effective coverage for larger families, possibly including family discounts or caps on premiums based on the number of children covered.

Predictive Modelling:

Detailed Process and Results

Model Selection:

For this section, we performed a comparative study of the performance of 3 different ML models. We studied how each model performed on our dataset. Here are the different models, we studied:

1. **Linear Regression:** It is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-

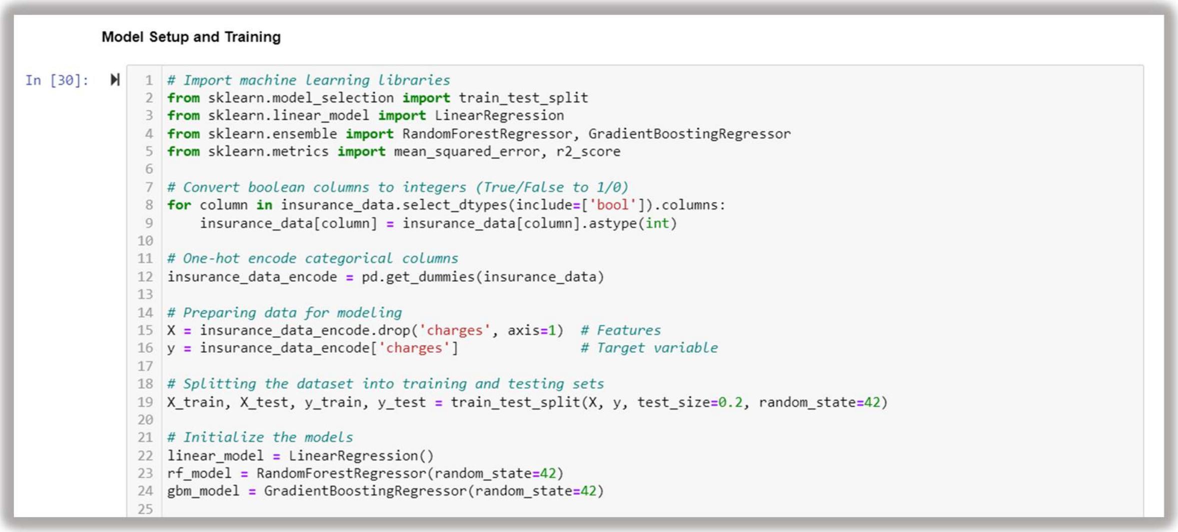
fitting straight line (the regression line) that predicts the dependent variable based on the independent variables. The goal of linear regression is to minimize the sum of the squared differences between the observed values and the values predicted by the line.

2. **Random Forest:** It is an ensemble learning method used for classification and regression tasks. It constructs multiple decision trees during training and merges their results to improve accuracy and control overfitting. Each tree is built from a random subset of the data and features, making the forest robust and less sensitive to noise in the dataset. The final prediction is typically made by averaging the predictions of the individual trees (regression) or by majority voting (classification).
3. **Gradient Boosting Machines (GBM):** Gradient boosting is an ensemble learning technique used for regression and classification tasks. It successively builds a model by combining the predictions of several weak learners, typically decision trees. Each new tree is trained to correct the errors made by the previous trees by focusing on the residuals (differences between actual and predicted values). The method iteratively adds trees, with each one improving the model's performance. The final model is a weighted sum of all the individual trees, resulting in a strong overall predictive model.

Model Training:

Preparatory steps for modelling included:

- **Data Preparation:** Encoding categorical variables, splitting the dataset into an 80/20 training-testing ratio, and applying necessary data normalization.



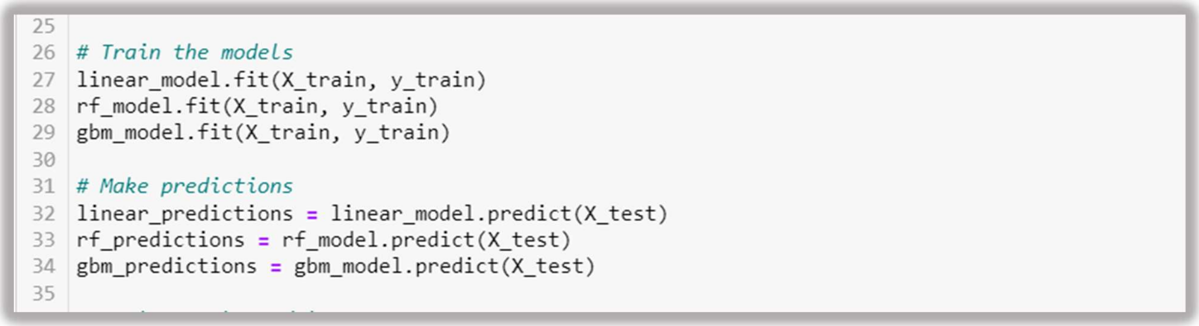
Model Setup and Training

In [30]:

```
# Import machine learning libraries
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.metrics import mean_squared_error, r2_score

# Convert boolean columns to integers (True/False to 1/0)
for column in insurance_data.select_dtypes(include=['bool']).columns:
    insurance_data[column] = insurance_data[column].astype(int)

# One-hot encode categorical columns
insurance_data_encode = pd.get_dummies(insurance_data)

# Preparing data for modeling
X = insurance_data_encode.drop('charges', axis=1)  # Features
y = insurance_data_encode['charges']               # Target variable

# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the models
linear_model = LinearRegression()
rf_model = RandomForestRegressor(random_state=42)
gbm_model = GradientBoostingRegressor(random_state=42)

```

- **Training Process:**
    - Linear Regression was set up to explore straightforward linear correlations.

- Random Forest was initialized with multiple decision trees to leverage its ensemble method benefits.
- GBM was sequentially trained to refine accuracy progressively.

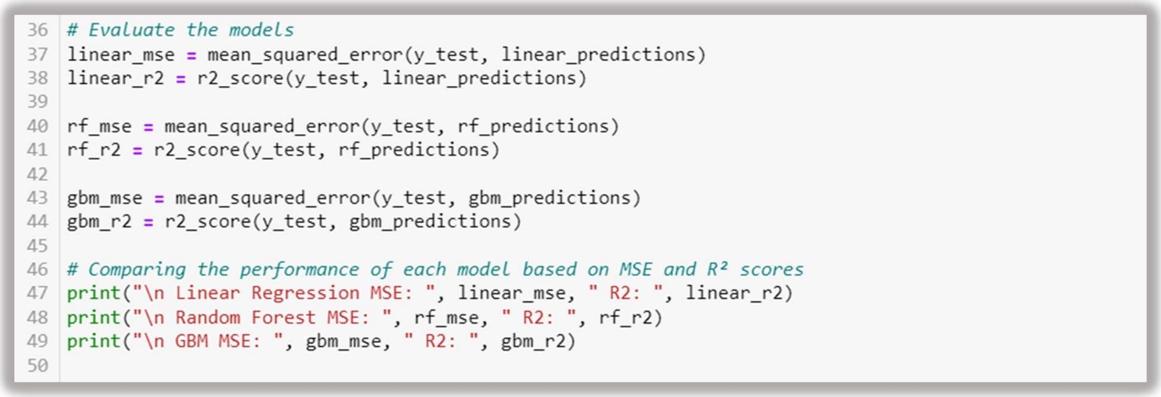


```
# Train the models
linear_model.fit(X_train, y_train)
rf_model.fit(X_train, y_train)
gbm_model.fit(X_train, y_train)

# Make predictions
linear_predictions = linear_model.predict(X_test)
rf_predictions = rf_model.predict(X_test)
gbm_predictions = gbm_model.predict(X_test)

```

Model Evaluation:

Performance metrics used were:

- **Mean Squared Error (MSE):** Key in quantifying the average of the squares of the errors—lower values indicate better accuracy.
- **R-squared (R²):** Reflects the proportion of variance in the dependent variable predictable from the independent variables—higher values suggest a better model fit.



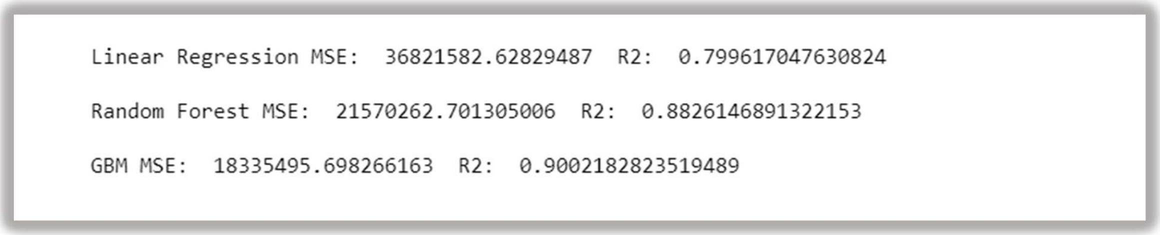
```
# Evaluate the models
linear_mse = mean_squared_error(y_test, linear_predictions)
linear_r2 = r2_score(y_test, linear_predictions)

rf_mse = mean_squared_error(y_test, rf_predictions)
rf_r2 = r2_score(y_test, rf_predictions)

gbm_mse = mean_squared_error(y_test, gbm_predictions)
gbm_r2 = r2_score(y_test, gbm_predictions)

# Comparing the performance of each model based on MSE and R² scores
print("\n Linear Regression MSE: ", linear_mse, " R2: ", linear_r2)
print("\n Random Forest MSE: ", rf_mse, " R2: ", rf_r2)
print("\n GBM MSE: ", gbm_mse, " R2: ", gbm_r2)

```

Model Comparison:

The comparative results are as follows:

- **Linear Regression:** Reported an MSE of 36,821,582.63 and an R² of 0.7996, indicating a decent model but with limitations due to the simplicity of assuming linear relationships.
- **Random Forest:** Achieved a more favorable MSE of 21,313,038.13 and an R² of 0.8840, showing improved accuracy over Linear Regression by addressing non-linear factors effectively.

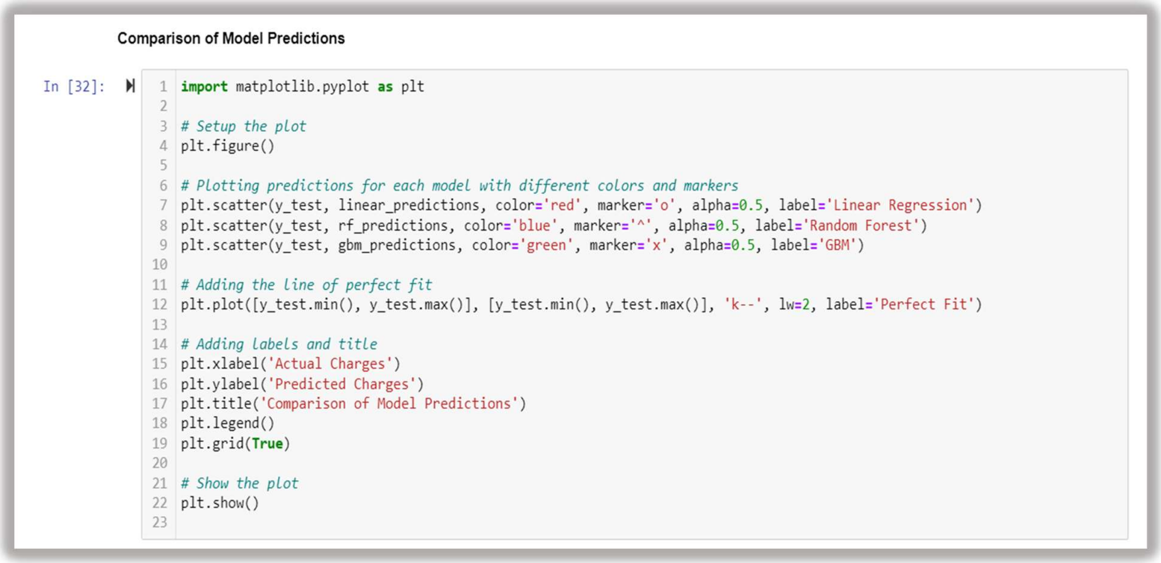
- **Gradient Boosting Machines (GBM):** Demonstrated the best performance with the lowest MSE of 18,253,459.65 and the highest R² of 0.9007, validating its efficacy in dealing with complex data interactions.



```
Linear Regression MSE:  36821582.62829487  R2:  0.799617047630824

Random Forest MSE:  21570262.701305006  R2:  0.8826146891322153

GBM MSE:  18335495.698266163  R2:  0.9002182823519489
```

Results Interpretation:

- **Insights from GBM:** The strong performance of GBM highlights its capability to handle various data structures and complex relationships, making it highly suitable for predictive analytics in insurance charge estimations.

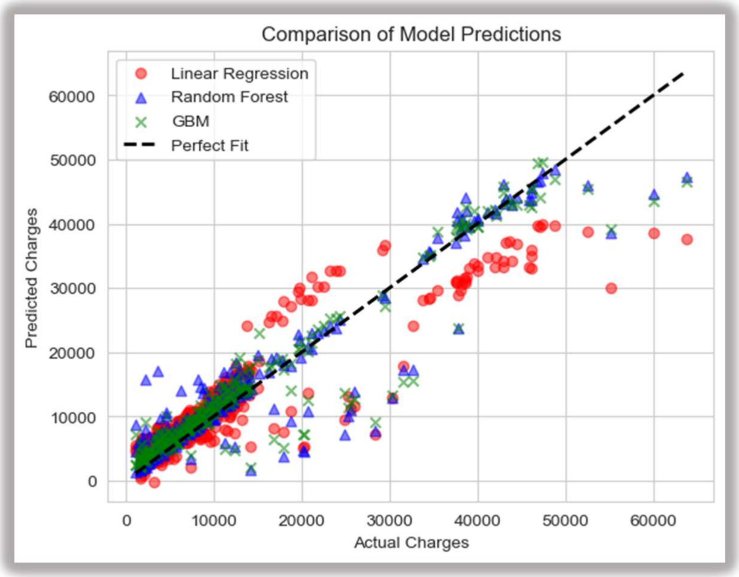


Comparison of Model Predictions

In [32]:

```
import matplotlib.pyplot as plt

# Setup the plot
plt.figure()

# Plotting predictions for each model with different colors and markers
plt.scatter(y_test, linear_predictions, color='red', marker='o', alpha=0.5, label='Linear Regression')
plt.scatter(y_test, rf_predictions, color='blue', marker='^', alpha=0.5, label='Random Forest')
plt.scatter(y_test, gbm_predictions, color='green', marker='x', alpha=0.5, label='GBM')

# Adding the line of perfect fit
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2, label='Perfect Fit')

# Adding labels and title
plt.xlabel('Actual Charges')
plt.ylabel('Predicted Charges')
plt.title('Comparison of Model Predictions')
plt.legend()
plt.grid(True)

# Show the plot
plt.show()

```

- **Random Forest Considerations:** The robust nature of Random Forest makes it a reliable choice, particularly useful for datasets like the insurance dataset, which may include outliers and varied feature scales.
- **Assessment of Linear Regression:** While it provides a valuable baseline, Linear Regression's performance confirms that it is less equipped to handle the complexities of this particular dataset, which includes non-linear interactions.



Conclusions and Implications:

- **Policy and Strategy:** These models, especially GBM, can aid insurance companies in refining how premiums are structured and risks are assessed, leading to more precise risk management.
- **Support for Preventive Health Measures:** Identifying significant predictors of insurance charges like smoking and high BMI can inform targeted health interventions, potentially reducing future insurance charges.

Conclusion: Summative Insights and Future Directions

Project Summary:

This project leveraged extensive data analysis to uncover significant factors influencing health insurance charges. Through exploratory data analysis (EDA) and predictive modelling, it became clear that smoking, BMI, age, family size, and regional disparities are key variables affecting insurance costs. The use of statistical and machine learning techniques allowed for an in-depth understanding of how these factors interplay to shape insurance premiums.

Future Research:

- **In-depth Analysis of Underexplored Variables:** Future studies could explore additional variables such as employment status, income levels, or pre-existing health conditions that may also impact insurance charges.
- **Longitudinal Data Analysis:** Analysing changes over time could help understand trends and the long-term impact of policy changes on insurance costs.

- **Comparative Studies:** Examining data from other regions or countries could provide comparative insights that highlight effective insurance and health policy strategies used globally.

Final Thoughts:

The analysis provided robust evidence that lifestyle choices and demographic factors significantly impact health insurance costs. Tailoring health policies to address these specific factors could lead to more effective management of healthcare costs and improved public health outcomes.

Overall Conclusion:

The findings from this comprehensive analysis illuminate the multifaceted nature of health insurance charges. Smoking and high BMI are confirmed as major contributors to higher charges, emphasizing the need for targeted public health interventions. Aging populations and larger family sizes also play crucial roles, necessitating policies that support these demographic groups.

Policy Implications:

- **Tailored Health Policies:** The data supports the creation of targeted health initiatives such as smoking cessation programs and obesity prevention strategies, which could substantially reduce healthcare costs.
- **Family-Friendly Insurance Plans:** Insights into the impact of family size on insurance charges can guide the development of plans that offer better coverage options for families, potentially including discounts or capped rates for larger families.
- **Regional Healthcare Equity:** Addressing regional disparities in insurance charges could involve reassessing how healthcare resources are distributed and funded, ensuring that all regions have equitable access to affordable healthcare.

Insurance Strategy:

For insurance companies, these insights underline the importance of integrating individual risk assessments into premium calculations. Offering personalized insurance products that account for personal health behaviors and regional factors could make premiums fairer and encourage healthier lifestyles among policyholders.

Public Health Strategy:

Public health officials are encouraged to use these findings to prioritize interventions targeting significant cost drivers. By focusing on areas like smoking and obesity, and tailoring strategies to fit regional and demographic specifics, public health initiatives can more effectively reduce overall healthcare expenses and improve population health outcomes.

Challenges Encountered:

The project presented numerous challenges, starting with the need to structure the code clean, neat, and logically ordered. This required researching new concepts like one-hot encoding, correlation, covariance, and various types of visualizations and graphs. Writing the code for these elements demanded significant effort and learning. Handling outliers was another major challenge. We identified outliers in the 'BMI' and 'charges columns using statistical methods and faced the critical decision of whether to remove them. To ensure robust conclusions, we conducted analyses both with and without outliers to understand their impact on our dataset. Calculating the risk score was complex, requiring an understanding of factors contributing to health risks such as smoking status, BMI, and age. Developing an accurate and functional risk score model involved both theoretical research and practical coding skills.

The predictive modelling phase was particularly time-consuming and challenging. We explored various models, including Linear Regression, Random Forest, and Gradient Boosting Machines, to determine the best fit for our dataset. This required extensive research, effective training, and careful interpretation of results. Interpreting the results to provide actionable insights was also demanding. We needed to translate complex statistical outputs and model predictions into clear, understandable conclusions to inform policy recommendations and public health strategies. Each insight, such as the impact of smoking on insurance charges, the increase in charges with age, regional differences in charges, the influence of BMI, and the effect of having children, required careful analysis and presentation.

The project involved a challenging yet rewarding mix of coding, data analysis, and result interpretation. These challenges not only enhanced our learning experience but also significantly contributed to the robustness and reliability of our findings.

Benefits Gained:

This project significantly improved our data analysis skills. Working with real-world healthcare data provided practical experience in data cleaning, manipulation, and exploration. We learned how to preprocess data effectively for thorough analysis, sharpening our ability to manage and visualize large datasets using statistical tools and Python libraries, essential in data analytics. Our understanding of modelling techniques also advanced. By building and comparing predictive models such as Linear Regression, Random Forest, and Gradient Boosting Machines, we learned to train, evaluate, and tune models for optimal performance. This experience was invaluable in applying theoretical concepts to practical problems. We derived actionable insights from the data, such as the impact of smoking and BMI on insurance charges. These insights can inform policy recommendations and public health initiatives, showcasing the practical application of data analytics.

The project fostered collaboration and teamwork. We divided tasks, shared findings, and integrated various parts of the project, enhancing our project management and communication skills. This collaborative effort was crucial in overcoming challenges and ensuring the project's success. Our communication and reporting skills improved as we developed comprehensive reports and presentations that effectively communicated complex analyses and results. We learned to convey data-driven insights clearly and actionable, making our work accessible to a broader audience.

Finally, we formulated policy and strategy recommendations based on our data insights, such as targeted health campaigns and premium adjustments for smokers and individuals with high BMI. This demonstrated the potential of data analytics to influence health policy and insurance strategies, contributing to more informed decision-making. The project highlighted the role of data in driving strategic decisions.

Personal Statement:

During this project, I made several significant contributions that were crucial to its success. I began by conducting a thorough data preparation phase, which involved a detailed review and descriptive analysis of the US Health Insurance Dataset. I meticulously handled outliers, transformed categorical variables through one-hot encoding, and ensured data quality by addressing duplicates and verifying the absence of missing values. This foundational work was essential for accurate analysis and highlighted the importance of data integrity. Additionally, I developed and implemented the risk score calculation. This involved researching how to quantify health risks based on smoking status, BMI, and age, and integrating this risk score into our predictive models. This task enhanced both my theoretical understanding and practical skills, providing a valuable metric for assessing individual health risks.

In the exploratory data analysis (EDA) phase, I examined relationships between age, BMI, smoking status, and insurance charges using advanced visualization tools like scatter plots, box plots, and heatmaps. These visualizations uncovered key insights into the factors driving insurance costs and emphasized the importance of clear data presentation. In the predictive modelling phase, I built and compared models such as Linear Regression, Random Forest, and Gradient Boosting Machines. I tuned parameters and evaluated model performance, which significantly boosted my confidence in using advanced analytical tools. Furthermore, I was responsible for interpreting the results to provide actionable insights. I translated complex statistical outputs and model predictions into clear, understandable conclusions that informed policy recommendations and public health strategies. Each insight, including the impact of smoking, age, BMI, and having children on insurance charges, was carefully analyzed and presented, ensuring that the findings were robust and informative.

ISYS 812 Programming and Application for Data Analytics

Project: Part 1

Name: Durga Naga Padmasree Sappa

Dataset Description:

a. Dataset Name: US Health Insurance Dataset
(https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data)

b. Number of Columns and Rows: This dataset comprises 1,338 entries (rows) and 7 key attributes (columns), including Age, Sex, BMI, Children, Smoker Status, Region, and Charges.

c. Description of the Data: The US Health Insurance Dataset provides information about individuals with health insurance coverage in the United States. Each row represents a different insured individual, and the columns contain demographic and health-related information about them.

The columns include:

Age: Age of the beneficiary.
Sex: Gender of the beneficiary (male or female).
BMI: Body Mass Index (BMI) of the beneficiary.
Children: Number of children/dependents covered by the insurance.
Smoker: Whether the beneficiary is a smoker or not (yes or no).
Region: The beneficiary's residential area (northeast, southeast, southwest, northwest).
Charges: Medical insurance costs are billed to individuals.

Summary: The dataset on health insurance in the USA contains the most individual records of persons covered by different health insurance plan coverages in the United States. The data included details like factors such as pandemic information, body mass index (BMI) of the person insured, the number of dependents on him, and his smoking status. It also includes the geographical region and charges pertinent to the health insurance plan. These databases facilitate the interaction between insurance prices and health-related factors, therefore revealing underlying trends that impact healthcare accessibility and affordability.

d. Reason for Choosing this Dataset: The US Health Insurance Dataset provides valuable insights into factors influencing health insurance charges. Understanding these factors is crucial in addressing healthcare affordability and accessibility issues. As aspiring data analysts, our interest lies in exploring the interplay between demographic attributes (such as age, gender, and region), lifestyle choices (like smoking habits), and health metrics (BMI) with insurance costs. Moreover, analyzing this dataset enables us to acquire practical experience in healthcare analytics, a field with significant real-world applications and implications.

Questions for Data Analysis:

1. What are the main factors influencing health insurance charges?
2. Is there a significant difference in charges between different regions?
3. How does BMI correlate with health insurance charges?
4. Can we predict health insurance charges based on demographic and health-related attributes?
5. Is there any relationship between smoking habits and health insurance charges?
6. How do health insurance charges vary across different age groups?
7. Is there a gender-based disparity in health insurance charges?
8. What is the impact of having children on health insurance charges?
9. Can we identify underlying clusters or segments within the insured population based on their demographic and health attributes?
10. How do lifestyle factors, such as smoking habits, interact with BMI to affect health insurance charges?
11. Can we identify any significant trends or patterns in insurance charges over time?
12. Is there a relationship between the number of children and smoking habits?
13. Are there any demographic differences in smoking rates?
14. Does the region affect the distribution of BMI among individuals?