

**ANL252 Python for Data Analytics**

**End-of-Course Assessment July 2023**

**Submission Date: 3 November 2023**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| Chen Shi Yun | Q2311793 |

Q1

Explanation of the Data Pre-processing Code for the ECA Medical Costs Dataset

The code provided aims to address three primary data pre-processing concerns: handling missing values, removing duplicates, and ensuring data type consistency.

Loading the Dataset:

The first step is to load the ECA dataset using the pandas library. This is done with the read\_csv function, which reads the data from a CSV file into a DataFrame.

Handling Missing Values:

Missing data can introduce bias, reduce the statistical power of a dataset, and lead to a loss of information. In the ECA dataset, the 'age' column had missing values. To address this, the code calculates the median of the 'age' column, which is then used to replace any missing values. The median is often chosen over the mean in such scenarios to prevent the impact of outliers from skewing the imputed values.

Removing Duplicates:

Duplicate rows can occur due to various reasons, such as data entry errors or system glitches. They can distort analyses and lead to incorrect conclusions. The code identifies and removes any duplicate rows in the dataset using the drop\_duplicates method.

Checking and Adjusting Data Types:

Consistent data types are essential for ensuring accurate data analysis. The 'age' column's values were initially of the float data type, which is not typical for age values. The code converts this column to an integer data type using the astype method, making it more intuitive and appropriate for age-related analyses.

Lastly, the cleaned dataset is optionally saved to a new CSV file using the to\_csv method. This step provides a cleaned version of the dataset for future use and analyses.

In essence, this code provides a systematic approach to cleaning and preparing a dataset, ensuring that it's primed for accurate and meaningful analyses. (293 words)

**import pandas as pd**

**# Load the dataset**

**eca\_data = pd.read\_csv("/mnt/data/ECA.csv")**

**# Task 1: Handling Missing Values**

**# Impute missing values in the 'age' column with its median**

**median\_age = eca\_data['age'].median()**

**eca\_data['age'].fillna(median\_age, inplace=True)**

**# Task 2: Removing Duplicates**

**# Remove duplicate rows**

**eca\_data.drop\_duplicates(inplace=True)**

**# Task 3: Checking Data Types**

**# Convert the 'age' column to int64**

**eca\_data['age'] = eca\_data['age'].astype(int)**

**# Save the cleaned dataset (optional)**

**eca\_data.to\_csv("/mnt/data/cleaned\_ECA.csv", index=False)**

Q2

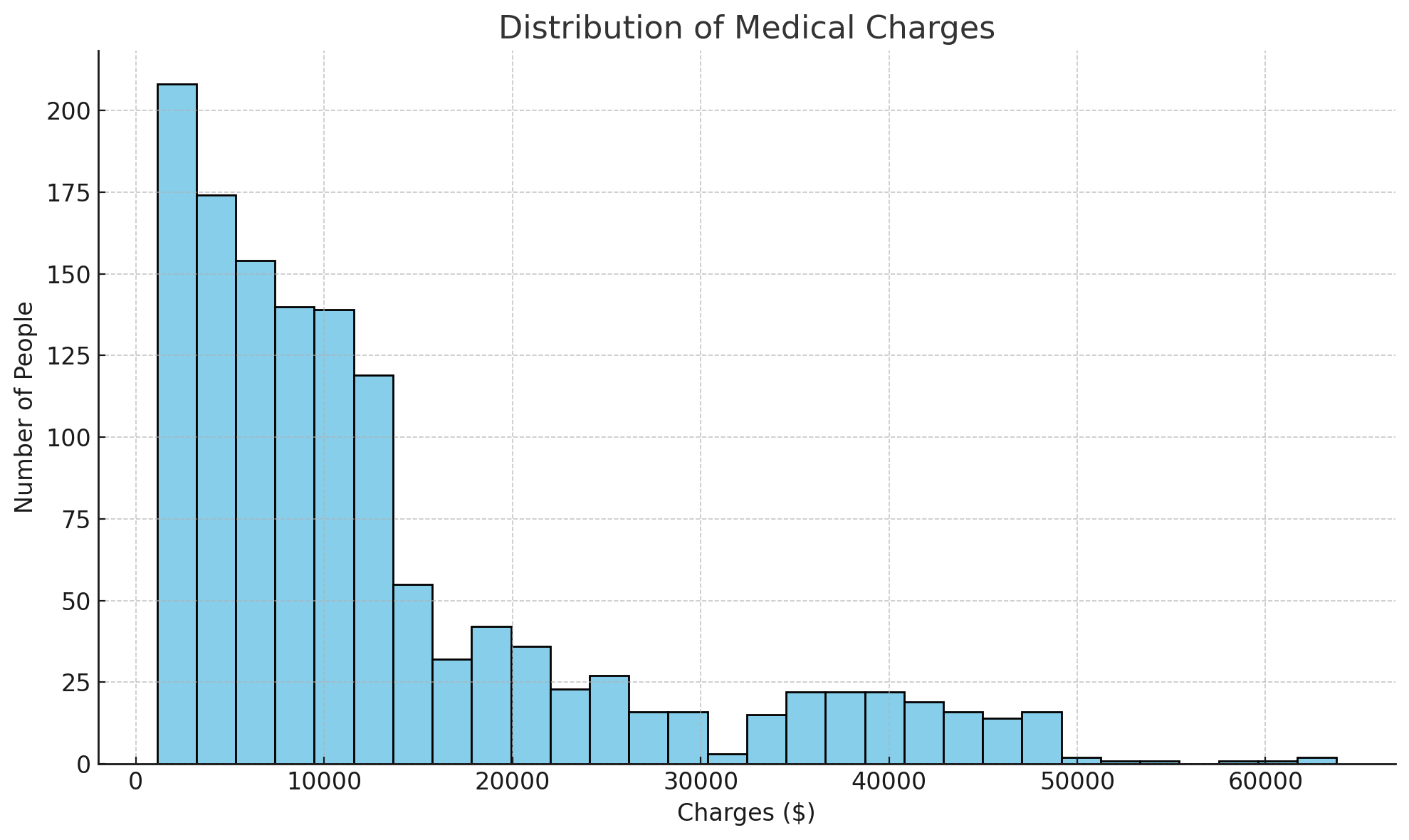


Figure 1: Distribution of Medical Charges

Insights:

The majority of individuals have medical charges that fall in the lower range (below $10,000).

There's a noticeable peak around the lower charge range, suggesting that many people have only minor or routine medical expenses.

There are fewer individuals with higher medical charges, but there's a noticeable tail extending towards the higher charge range, indicating that while rarer, there are people with significantly higher medical expenses.

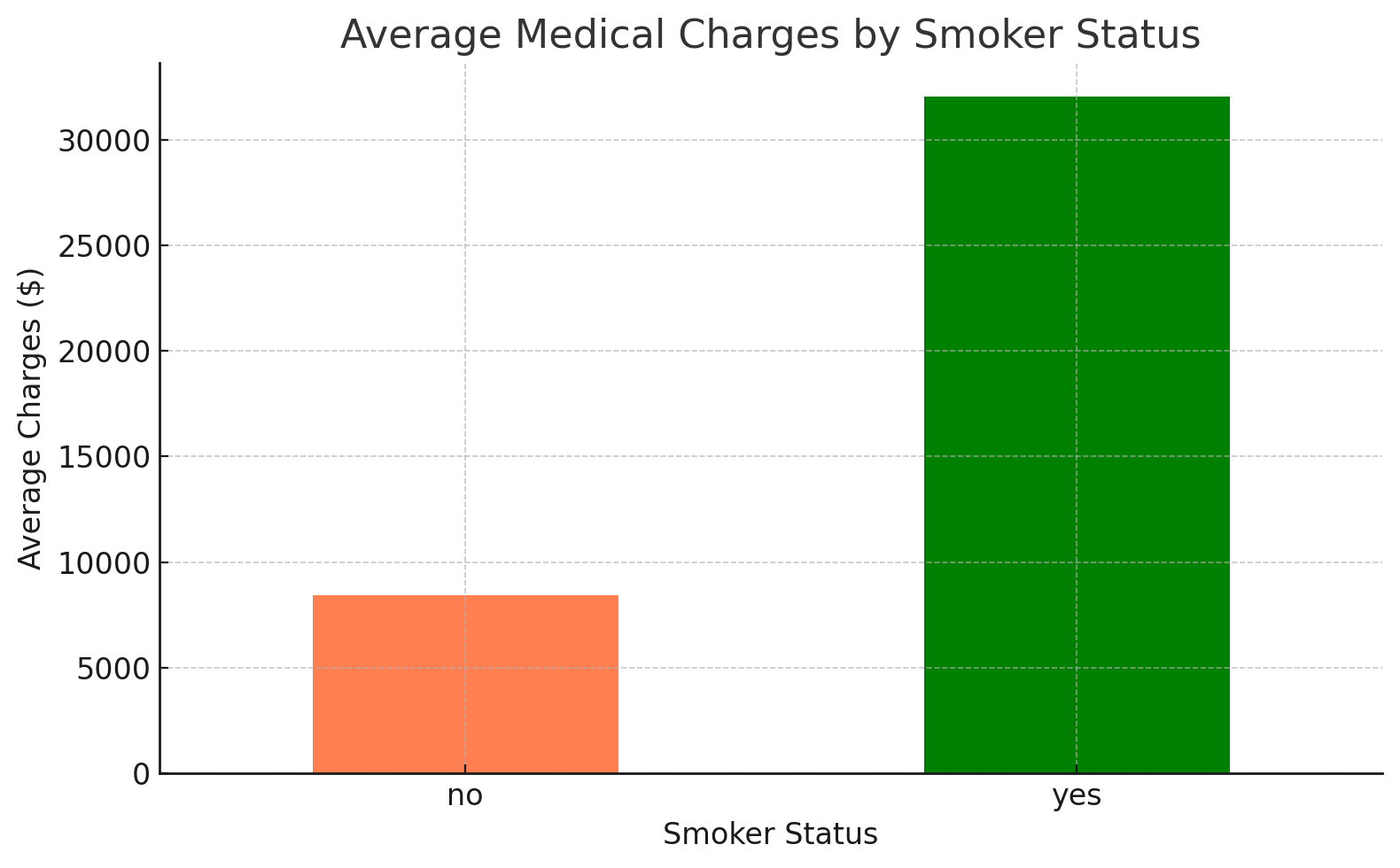


Figure 2: Average Medical Charges by Smoker Status

Insights:

Individuals who are smokers have, on average, significantly higher medical charges compared to non-smokers.

The difference in average charges between smokers and non-smokers is substantial, suggesting that smoking has a pronounced impact on medical expenses.

This observation aligns with the general understanding that smoking is associated with various health issues, leading to higher medical costs.

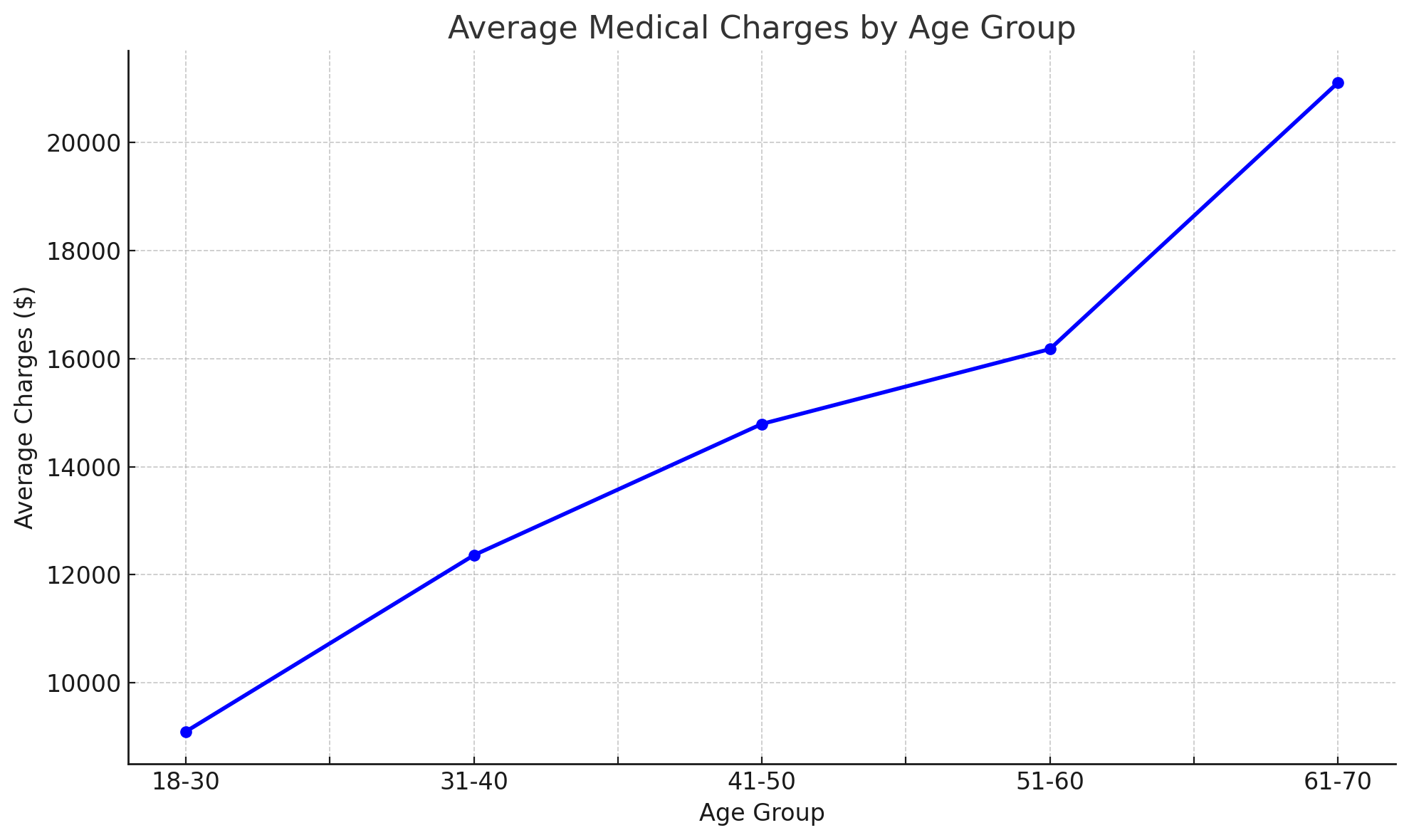


Figure 3: Average Medical Charges by Age Group

Insights:

There's a clear upward trend in average medical charges as age increases. This suggests that older individuals, on average, tend to have higher medical expenses compared to younger ones.

The steepest increase in average charges is observed between the age groups 18-30 and 31-40, indicating that entering the 30s might be associated with a significant rise in medical costs.

The gradual increase in costs across age groups reflects the general understanding that as individuals age, they are more likely to encounter health issues, leading to increased medical expenses.

In summary, the three figures provide valuable insights into the medical charges distribution and its relationship with factors like smoking status and age. The findings emphasize the economic implications of smoking and the natural increase in healthcare costs as one ages. These insights can guide health insurance companies in premium determination and individuals in making informed health-related decisions. (292 words)

**import matplotlib.pyplot as plt**

**# Load the dataset (assuming it's already cleaned as per Q1)**

**eca\_data = pd.read\_csv("/mnt/data/cleaned\_ECA.csv")**

**# Figure 1: Distribution of Charges**

**plt.figure(figsize=(10, 6))**

**plt.hist(eca\_data['charges'], bins=30, color='skyblue', edgecolor='black')**

**plt.title('Distribution of Medical Charges')**

**plt.xlabel('Charges ($)')**

**plt.ylabel('Number of People')**

**plt.grid(axis='y', linestyle='--', alpha=0.7)**

**plt.tight\_layout()**

**plt.show()**

**# Figure 2: Average Charges by Smoker Status**

**avg\_charges\_by\_smoker = eca\_data.groupby('smoker')['charges'].mean()**

**plt.figure(figsize=(8, 5))**

**avg\_charges\_by\_smoker.plot(kind='bar', color=['coral', 'green'])**

**plt.title('Average Medical Charges by Smoker Status')**

**plt.xlabel('Smoker Status')**

**plt.ylabel('Average Charges ($)')**

**plt.xticks(rotation=0)**

**plt.grid(axis='y', linestyle='--', alpha=0.7)**

**plt.tight\_layout()**

**plt.show()**

**# Segmenting age into groups for Figure 3**

**bins = [18, 30, 40, 50, 60, 70]**

**labels = ['18-30', '31-40', '41-50', '51-60', '61-70']**

**eca\_data['age\_group'] = pd.cut(eca\_data['age'], bins=bins, labels=labels, right=False)**

**# Figure 3: Average Charges by Age Group**

**avg\_charges\_by\_age = eca\_data.groupby('age\_group')['charges'].mean()**

**plt.figure(figsize=(10, 6))**

**avg\_charges\_by\_age.plot(kind='line', marker='o', color='blue', linewidth=2)**

**plt.title('Average Medical Charges by Age Group')**

**plt.xlabel('Age Group')**

**plt.ylabel('Average Charges ($)')**

**plt.grid(True, linestyle='--', alpha=0.7)**

**plt.tight\_layout()**

**plt.show()**

Q3

Approach Taken:

Data Preparation: Converted categorical variables to a numerical format using one-hot encoding and split the data into training and testing sets.

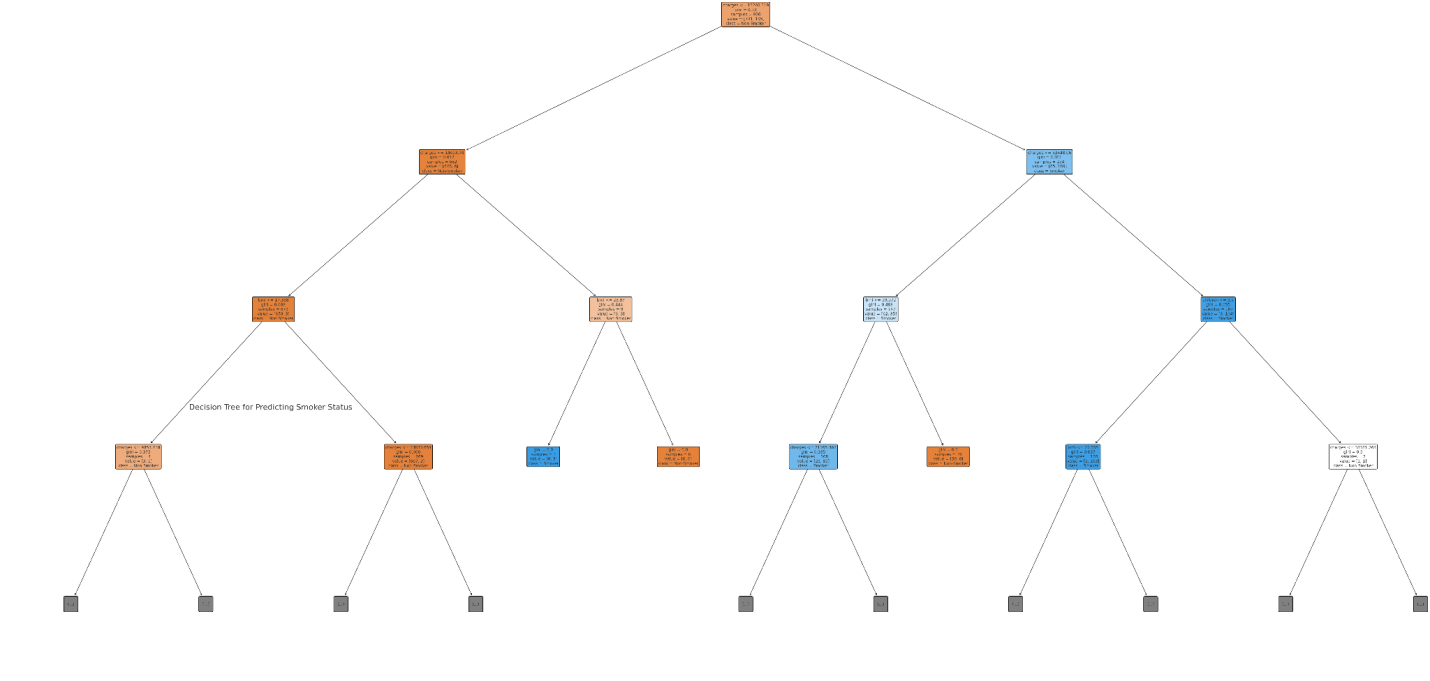
Model Training: Trained a decision tree classifier on the training data.

Evaluation: Assessed the model's accuracy on the test data.

Visualization: Visualized the decision tree to interpret its decision-making process.

In conclusion, the decision tree model offers an interpretable way to understand the factors influencing whether an individual is likely to be a smoker. The model's high accuracy suggests that it captures the underlying patterns in the data effectively. The visual representation provides clear insights into the decision-making process and the significance of various features in predicting smoker status. (112 words)

Q4



Key Insights:

The root node (the topmost node) splits the data based on the 'charges' feature. This suggests that medical charges are a significant determinant in predicting whether someone is a smoker. Higher medical charges are often associated with smokers, likely due to the health implications of smoking.

As we move down the tree, we can observe splits based on other features like 'age', 'bmi', and 'children'. Each split is a decision point that further classifies individuals based on the conditions set by the feature values.

The leaves (final nodes) represent the predicted class, either 'Non-Smoker' or 'Smoker', based on the path taken through the decision point. (107 words)

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.preprocessing import OneHotEncoder**

**from sklearn.tree import DecisionTreeClassifier, plot\_tree**

**from sklearn.metrics import accuracy\_score**

**import matplotlib.pyplot as plt**

**# Load the dataset (assuming it's already cleaned as per Q1)**

**eca\_data = pd.read\_csv("/mnt/data/cleaned\_ECA.csv")**

**# Data Preparation**

**# Convert categorical variables into numerical format using one-hot encoding**

**eca\_data\_encoded = pd.get\_dummies(eca\_data, drop\_first=True)**

**# Splitting data into training and testing sets**

**X = eca\_data\_encoded.drop('smoker\_yes', axis=1)**

**y = eca\_data\_encoded['smoker\_yes']**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)**

**# Model Training**

**clf = DecisionTreeClassifier(random\_state=42)**

**clf.fit(X\_train, y\_train)**

**# Predict on the test data and Evaluate the accuracy**

**y\_pred = clf.predict(X\_test)**

**accuracy = accuracy\_score(y\_test, y\_pred)**

**# Visualize the decision tree (limiting the depth for clarity)**

**plt.figure(figsize=(20, 10))**

**plot\_tree(clf, max\_depth=3, feature\_names=X\_train.columns, class\_names=['Non-Smoker', 'Smoker'], filled=True, rounded=True)**

**plt.title("Decision Tree for Predicting Smoker Status")**

**plt.show()**

Q5

Decision trees hold significant potential as tools for exploratory data analysis. One of their standout features is their ability to perform feature selection. The top nodes in a decision tree, which are determined early during the tree-building process, typically represent the most crucial variables in the dataset. By examining the sequence and hierarchy of these nodes, analysts can discern the relative importance of different features.

Furthermore, the visual representation provided by them offers an intuitive means to understand intricate relationships and interactions between various variables. This clarity in decision-making pathways allows for a comprehensive view of how different features influence the target variable. Another advantage is their capability to capture non-linear relationships within the data. Such relationships, which might be elusive through traditional exploratory methods, become evident as the tree segments data based on specific conditions.

Decision trees also serve as indicators of data quality. Features that remain absent from the tree might either be irrelevant or suffer from issues like missing values or noise. On the other hand, if a variable achieves a split very early in the tree but such a decision doesn't align with domain knowledge, it might be a sign of data anomalies or outliers. Additionally, trees are adept at highlighting interactions between variables. For instance, a tree might initially segment data based on one feature and subsequently split a resulting subgroup using another feature. Such patterns indicate interaction effects, offering insights into segment-specific behaviors or trends.

However, it's essential to note that while decision trees are instrumental, they shouldn't be the sole tool for exploratory data analysis. A comprehensive understanding of a dataset demands their usage in tandem with other exploratory techniques. In essence, while primarily viewed as prediction tools, decision trees' structured and visual nature still makes them indispensable for exploratory data analysis.

(299 words)