ST1 Capstone Programming Project

This project will be analysing the attributes of a medical insurance dataset to create a model for prediction of incurred charges.

My approach to this task will be incorporating the Python libraries **tkinter**, **pandas**, **seaborn**, and **matplotlib**.

Therefore, before we move on we shall import these libraries:

import tkinter as tk

from tkinter import filedialog, messagebox

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

**Step 1.**

*Reading the Dataset*

# Load dataset

data = pd.read\_csv('medical\_insurance.csv')

# Display first few rows of the dataset in consoel

print(data.head())

# Display general information and basic statistics in console

print(data.info())

print(data.describe())

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer code

Description automatically generatedA close up of words

Description automatically generated

**Key observations from Step 1**

* The dataset is a sample of 2772 people with medical insurance
* The attributes include age, bmi, number of children, charges, gender, smoker, and region.
* Gender, smoker and region attributes are functionally Booleans, being True or False. For example, sex\_male = False would indicate a female.

**Step 2.**

*Problem Statement Definition*

* Creating a prediction model in order to predict the price of medical insurance based on given attributes.

**Step 3.**

*Target Variable Identification*

* Target Variable: Charges (Affected by: Age, Sex, Smoker, Region, etc)

**Step 4.**

*Choosing appropriate Algorithm for Data Analysis*

* The target variable is Continuous, and we shall be proceeding with a Ridge regression model as we expect a linear relationship.

*Visualising the distribution of Target Variable*

# Pair plot to examine relationships among numerical features

sns.pairplot(data)

A graph of a number of people

Description automatically generated with medium confidenceplt.show()

* We can see that the data has a relatively linear relationship, with most charges being weighted towards the lower end.
* We shall proceed and investigate how the other variables affect this.

**Step 5.**

*Data Exploration at the Basic Level*

* As per Step 1.

# Display first few rows of the dataset in consoel

print(data.head())

# Display general information and basic statistics in console

print(data.info())

A close up of words

Description automatically generatedA screenshot of a computer code

Description automatically generatedA screenshot of a computer screen

Description automatically generatedprint(data.describe())

* We expect that the target variable will be affected by number of *children*, *sex*, whether they are a *smoker*, and their *region*. BMI could be a potential factor.
* Age – Continuous
* BMI – Continuous
* Children – Continuous
* Charges – Continuous
* Sex – Categorical
* Smoker – Categorical
* Region – Categorical

**Step 6.**

*Identifying and Rejecting unwanted columns*

* Given our variables, there should be no need to remove any columns
* I suspect they shall all have some effect

**Step 7.**

*Visual Exploratory Data Analysis*

# Pair plot to examine relationships among numerical features

sns.pairplot(data)

A collage of graphs

Description automatically generatedplt.show()

**Observations from Step 7.**

* This pairplot shows the non-categorical data from the dataset.
* The variables are represented on both X and Y axis
* We can see how some variables such as BMI form a nice bell curve indicating a sufficient and diverse sample of people
* Some charts are less relevant although still do have features that indicate a relationship.

**Step 8.**

*Feature Selection based on data distribution*

* Despite some graphs being rather dense due to the dataset density, all show signs of being relevant to the target variable.
* Therefore we shall move on and analyse further

**Step 9.**

*Outliers and Missing Values*

# Calculates the Interquartile Range for charges

Q1 = data['charges'].quantile(0.25)

Q3 = data['charges'].quantile(0.75)

IQR = Q3 - Q1

# Define bounds

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

# Filter the dataset to remove outliers

filtered\_data = data[(data['charges'] >= lower\_bound) & (data['charges'] <= upper\_bound)]

# Check the size of the filtered dataset compared to the original dataset

print("Original dataset size:", len(data))

print("Filtered dataset size:", len(filtered\_data))

* After running this code, we have trimmed down the dataset from 2772 to 2476, which is a moderate amount of “outliers” given the total and suggests that the data is relatively consistent.

**Step 10.**

*Visualising data after outlier removal*

# Pair plot to examine relationships among numerical features

sns.pairplot(filtered\_data)

plt.show()

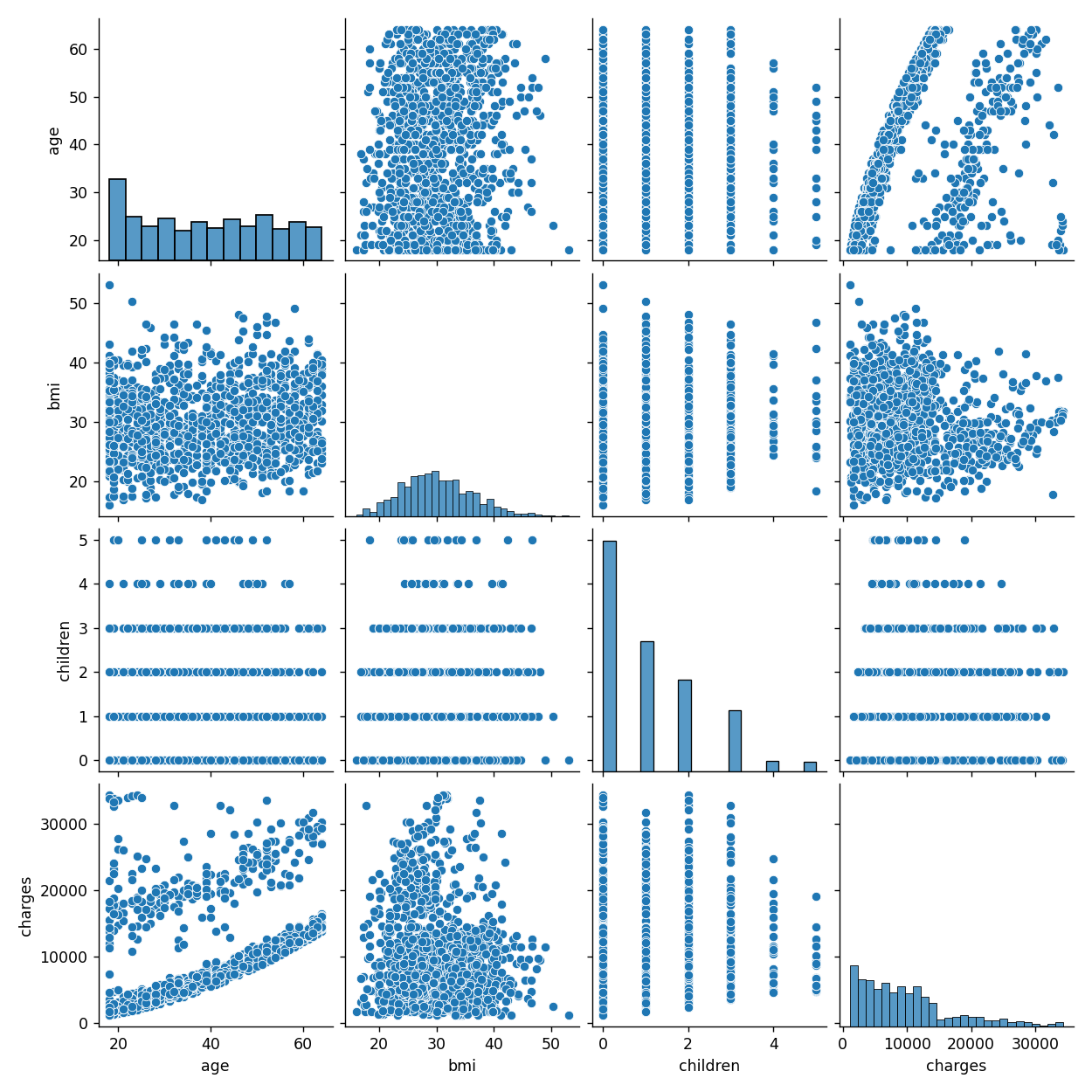
# Analyse categorical variables using boxplots

for column in ['sex', 'smoker', 'region']:

    sns.boxplot(x=column, y='charges', data=filtered\_data)

    plt.title(f"Charges vs {column}")

    plt.show()

* Data has been changed to filtered data for analysis

**Observations from Step 10.**

* There is little visible difference in the pairplot although certainly present
* Greatest change can be seen in the target variable ‘charges’
* The Interquartile Range filtering may have adverse effects on the overall data as now it potentially may be too consistent
* We no longer have charges approaching 20K
* Tail is still fairly linear