US MEDICAL INSURANCE COST:

Goal/Objective: In this project, I will attempt to gain insight on different attributes surrounding Medical Insurance Costs in the U.S.

First, I will import modules:

Import pandas as pd – rather than use import csv(I will create a dataframe to help assist organize data and perform analysis)

Import numpy as py

Import seaborn as sns

Import matplotlib.pyplot as plt

**First, get familiar with the dataset – What are the data types? Are they what you expected? Is there anything that needs to be changed? Is there any missing data? Etc.**

In this dataset we have a collection of patient information such as, age, sex, BMI (Body Mass Index), number of children, smoker status, region, and charges.

Based off observation, age Is an integer – sex is an object/string – BMI is a float, children is an integer, smoker is an object( – which could be a bool – ) region so far has 3 options and are object/strings and charges are floats.

---- I want to retrieve information about the current environment so I will use the .info() function – this information will display the data type, operation system etc. I preferred to use the .info() function rather than the .dtypes function because I want to include information regarding missing values. – the .info() function isn’t always guaranteed. I have 1338 entries from index 0 – 1337 and there are no missing values.

According to this information I want to further confirm that this is true. I will created a variable – and use the is.null() function and add those missing variable up using .sum().

* The datatypes were what I expected and there were no missing values. I can move forward to gather insight on the different attributes from the dataset.

I will start by getting a summary statistics of the data by passing .describe(include=’all’) – which will help assist me with generating questions I have:

I would like to primarily focus on the smoking attributes in the data but before I delve further into the smoker analysis, I’d like to perform some basic functions and task. Below are the areas I plan to attack:

1. What is the average age?
   1. According to the summary stats – 39
      1. Def analyze\_age():
         1. Boxplot
2. Is there a correlation between age and the yearly average cost?
   1. According to the data No – there seems to be a number of factor that affect a persons cost of insurance but age is not one them.
      1. Def group\_by\_age
3. Is the dataset evenly representative amongst male and female?
   1. Yes – there is 664 females and 676 males
      1. Def analyze\_sex():
         1. Bar chart and Pie chart
4. What is the average charge?
   1. The average charge is $13270.42
      1. Def anaylze\_average\_charges():
         1. No Graph to display this.

Main Focus ---

1. What is the average cost for patients that smoke? / **What is the average cost of patients who do not smoke?** / is there a correlation between high insurance cost and individuals that smoke?  **---** Compare the average cost of insurance of patients who smoke and don’t? Is there a difference?
   1. The average charge for patients who smoke is $16, 884.92 / **The average cost of patients who do not smoke is $8,434.27 ---** Patients that smoke pay approximately, $8, 450.65 more in medical insurance
      1. Def analyze\_smoker\_charges():, def analyze\_non\_smoker\_charges():, Def analyze\_smoker():
         1. Box plot (scatter plot to further display correlation)
2. Men Vs women smokers? Is there a significant difference?
   1. Perform further analysis
      1. Pass [Histoplot]
3. How many underage smokers are there?
   1. Perform further analysis
      1. Pass
4. How many underage smokers are in each region?
   1. Perform further analysis
      1. Pass [Histoplot]

WORKFLOW:

1. Project Obectives:

• Clearly outline the goals and objectives. What specific insights are we looking to gain from understaning the cost of medical insurance?

• What is the average cost of insurance?

• How much do male and females pay? Is there a significant difference?

• Does smoking impact insurance cost?

• Is there a relationship between the cost of insurance and age?

• Does BMI have an impact on cost?

• Does the number of children determine whether a patient pays more in insurance?

• What is the average cost of insurance per region?

2. Data Acquistion:

•insurance.csv file (provided by codecademy)

3. Data Explortation

•Load dataset into Jupyter notebook

•Explore the dataset to understand its structure, size, and features

•Check for missing values, outliers and data types

•Visualize key statsistics and distrubutions

• Age Distribution

• Sex Distribution

• BMI Distribution

• Smoker Proportion

• Distribution of Patients in each Region

• Distribution of Insurance Cost

4. Descriptive Statistics

•Calculate basic desriptive statistics such as mean, median, mode and standard deviation for relevant variables

•Analyze the distribution of insurance costs

5. Exploratory Data Analysis(EDA):

•Perform EDA to uncover patterns, trends, and relationships in data.

•Use visualization to aid in understanding

A. Feature relationships

•Explore relationships between different features and the insurance costs.

• Ages/Cost of Insurance

• Sex/Cost of Insurance

• BMI/Cost of Insurance

• Children/Cost of Insurance

• Smoker/Cost of Insurance

• Region/Cost of Insurance

•Conduct correlation analysis to identify potential associatons

• Age/Cost of Insurance

• BMI/Cost of Insurance

B. Demographic Analysis:

•Analyze the impact of demographic factors on insurance costs.

• Smoking Status/Charges

• BMI/Sex

• Count of patient in Region

• BMI/Region (average BMI per region)

•Create visualizations to illustrate these relationships.

C. Cost Variation Analysis:

•Identify factors contributng to variations in insurance costs.

•Break down cost by different catergories (e.g smoker status, # of children, bmi)

• Ages/Cost of Insurance

• Sex/Cost of Insurance

• BMI/Cost of Insurance

• Children/Cost of Insurance

• Smoker/Cost of Insurance

• Region/Cost of Insurance

D. Regional Analysis:

•Analyze regional variations in insurance cost.

• Examine how factors such as charges, age and BMi vary across regions

•Use visualizations to illustrate regional differences

6. Conclusions and insights:

•Summarize your key findings and insights from the analysis.

•Discuss notable trends/patterns discovered.