

**ANL252 (Online)**

**Python for Data Analytics**

# **End-of-Course Assessment**

**July 2023 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Poh Sze Ying** | **M2170235** |

**Question 1**

|  |
| --- |
| *#to import pandas*  **import** pandas **as** pd  *#to import data from excel file*  smoker\_data = pd.read\_csv("ECA.csv")  *#check for missing data* smoker\_data.isnull().any()  *#remove missing data* smoker\_data.dropna(subset = ['age'], inplace = **True**)  *#round up the charges to two decimal places* smoker\_data['charges'] = smoker\_data.charges.round(2)  *#Replace data with 'M' and 'F' to 'male' or 'female'*  smoker\_data['sex'].replace(['M','F'],['male', 'female'], inplace = **True**)  *#Create check point file to check if changes are made*  smoker\_data.to\_csv("smoker\_data.csv", index = **False**) |

The first clean up was to remove the missing data in the dataset. By using the isnull().any(), we are able to check that there are missing data in the age variable. The second clean up was to round up the medical cost to two decimal places instead of three decimal places as monetary amount are usually in two decimal places. The last clean up was to replace the gender of the insurance gender to ‘male’ or ‘female’ instead of ‘M’ or ‘F’ for easier sorting of data at the later stage.

**Question 2**

|  |
| --- |
| *#to import graphs*  **import** pandas **as** pd  **import** matplotlib.pyplot **as** plt  **import** numpy **as** np  *#import data from excel file*  mc\_data = pd.read\_csv("smoker\_data.csv")  *#set width of bar and size of the bar*  barWidth = 0.25  fig = plt.subplots(figsize = (8,4))  *#set selected data from excel file*  age\_loc = mc\_data['age']  gen\_loc = mc\_data['sex']  Male = [0, 0, 0, 0]  Female = [0, 0, 0, 0]  amt\_age\_loc = len(age\_loc)  *#Position 0 = teens*  **for** x **in** range(amt\_age\_loc):  **if** age\_loc[x] < 20:  **if** gen\_loc[x] == 'female':  Female[0]+=1  **else**:  Male[0]+=1  *#Position 1 = young adult*  **elif** age\_loc[x] > 19 **and** age\_loc[x] < 26:  **if** gen\_loc[x] == 'female':  Female[1]+=1  **else**:  Male[1]+=1    *#Position 2 = Adult*  **elif** age\_loc[x] > 25 **and** age\_loc[x] < 41:  **if** gen\_loc[x] == 'female':  Female[2]+=1  **else**:  Male[2]+=1    *#Position 3 = Pioneer*  **elif** age\_loc[x] > 40:  **if** gen\_loc[x] == 'female':  Female[3]+=1  **else**:  Male[3]+=1    *#Set position of bar on X axis*  br1 = np.arange(len(Male))  br2 = [x + barWidth **for** x **in** br1]  *#Make the plot*  plt.bar(br1, Male, color ='m', width = barWidth, label ='Male')  plt.bar(br2, Female, color ='y', width = barWidth, label ='Female')  *#Adding Xticks*  plt.xlabel('Gender', fontweight ='bold', fontsize = 10)  plt.ylabel('Number of Beneficiary', fontweight ='bold', fontsize = 10)  plt.title("Insurance by different age group")  plt.xticks([r + (barWidth/2) **for** r **in** range(len(Male))],  ['< 19', '20-25', '26-40', '> 41'])  plt.legend()  plt.show() |

A graph of different colored rectangles

Description automatically generated

The number of recipients from each age group for both males and females is depicted in the graph above. According to the aforementioned data, beneficiaries who are 41 years of age or older purchase more insurance than those who are 19 years of age or younger.

|  |
| --- |
| #set width of bar  charges\_loc = mc\_data['charges']  gen\_loc = mc\_data['sex']  Male = [0, 0, 0]  Female = [0, 0, 0]  amt\_chrg\_loc = len(charges\_loc)  *#For charges less than 5,000*  **for** x **in** range(amt\_chrg\_loc):  **if** charges\_loc[x] < 5000:  **if** gen\_loc[x] == 'female':  Female[0]+=1  **else**:  Male[0]+=1  *#For charges between 5,000 to 10,000*  **elif** charges\_loc[x] > 4999 **and** charges\_loc[x] < 10001:  **if** gen\_loc[x] == 'female':  Female[1]+=1  **else**:  Male[1]+=1    *#Position 2 = Adult*  **elif** charges\_loc[x] > 10000:  **if** gen\_loc[x] == 'female':  Female[2]+=1  **else**:  Male[2]+=1 |

A graph of a number of people

Description automatically generated with medium confidence

In the second graph, we are comparing the amount that each gender spent on medical expenses that were reimbursed by their health insurance.

|  |
| --- |
| #set width of bar  smoker\_loc = mc\_data['smoker']  gen\_loc = mc\_data['sex']  Male = [0, 0]  Female = [0, 0]  amt\_smk\_loc = len(smoker\_loc)  *#For charges less than 5,000*  **for** x **in** range(amt\_smk\_loc):  **if** smoker\_loc[x] == ‘yes’:  **if** gen\_loc[x] == 'female':  Female[0]+=1  **else**:  Male[0]+=1  *#For charges between 5,000 to 10,000*  **elif** smoker\_loc[x] == 'no':  **if** gen\_loc[x] == 'female':  Female[1]+=1  **else**:  Male[1]+=1 |

A graph of different colored squares

Description automatically generated

In the third graph, we are comparing the number of smokers between the different genders.

**Question 3**

The first method involved changing all of the numeric columns—including age, sex, bmi, children, region, and charges—into strings because all of the data must be in numeric form. Next, we’ll assign the different data in each column with a numeric value and make use of the map() function to return the data in the dataframe for each non-numeric column (W3Schools, n.d.). As the dependent variable is ‘smoker’, we will separate the other categories such as age, sex, bmi, and charges. The rationale behind leaving region and children out of the features is that they have no bearing on the costs that the insurance must pay. The decision tree would then be constructed using the X and Y operators.

**Question 4**

The decision tree's insights were obtained and divided into 11 rows while testing against the other four categories, with the "smoker" serving as the dependant and the other categories as the "test subject." The first row starts by comparing the charges of the insurance before comparing against the age and bmi of the insurer.

**Question 5**

Exploratory data analysis requires the initial data to be able to use graphical representations and summary statistics to find trends, identify anomalies, test hypotheses, and verify assumptions (Patil, 2022). However, if a large dataset is given, a decision tree would not be effective as it would create a lot of branches which can make the dataset look complicated to understand (Master’s in Data Science with edX, 2023). Exploratory data analysis involves user decision-making, an example could be a customer wants to buy mixed rice for lunch and has a budget in mind which limits the number of ingredients he can choose from. With so many options available to him, more data had to be gathered to produce the results he was hoping for. The purpose of a decision tree was to make a complex dataset easier to understand and able to provide the necessary useful data to the user (Master’s in Data Science with edX, 2023).

**Reference**

W3Schools. (n.d.). *Python Machine Learning Decision Tree*. Retrieved October 23, 2023, from https://www.w3schools.com/python/python\_ml\_decision\_tree.asp

Patil, P. (2022, May 30). What is Exploratory Data Analysis? - Towards Data Science. *Medium*. <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

Master’s in Data Science with edX. (2023, June 13). *What is a decision tree?* CORP-MIDS1 (MDS). <https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/>