w

**ANL252**

**Python For Data Analytics**

# **End-of-Course Assessment**

**July 2023 Semester**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Lau Wei Lian** | M2273019 |

**Tutorial Group: ­­­­­­­­­­­­ T03**

**Submission Date: 28/10/2023**

Table of Contents

[Question 1 3](#_Toc147993958)

[Question 2 5](#_Toc147993959)

[Question 3 10](#_Toc147993960)

[Question 4 11](#_Toc147993961)

[Question 5 13](#_Toc147993962)

[References 14](#_Toc147993963)

# Question 1

The first pre-processing task is to look for missing data.

From the raw dataset, we need to identify if there are any missing values or if there are any outliers to ensure data consistency.

Once we have identified any missing data, we must drop them off from the data frame or change them into the data required to ensure accuracy and relevance in analysis.

In this dataset, there are no outliers after checking on the raw data set in the excel. Therefore, there isn't any need to use Python code to match.

We identify missing values under column "AGE" with the following codes and drop those rows from the data frame.

import pandas as pd

import numpy as np

eca = pd.read\_csv(‘eca.csv’)

*#check for null value in dataframe*

eca.isnull().sum(axis=0)

eca\_1 = eca.dropna()

*#check if there are still any more null value after changes*

eca1.isnull().sum(axis=0)

The second pre-processing task is to ensure the consistency of data.

The raw dataset consists of a mixture of female, f, male, and m under the sex column, and this would create a discrepancy. Should we not make any changes, this would create a diverge of findings. Therefore, it is necessary to ensure consistency by reducing the data information from the current four types into two main categories.

Therefore, we need to divide data into female and male categories.

eca1.sex.unique()

replace\_dict = {‘F’:’female’,’M’:’male’}

eca1[‘sex’] = eca1[‘sex’].replace(replace\_dict)

*#check for updated dataframe*

eca1

The last pre-processing task is to make the values in both BMI and Charges into two decimal points instead of a mix of two or three decimal points. Similar to the above sex category, we would need to standardise the format of decimal points to ensure consistency in the dataset.

eca1[‘bmi’] = np.round(eca1[‘bmi’],decimals=2)

eca1[‘charges’] = np.round(eca1[‘charges’],decimals=2)

eca\_clean = eca1

*#check for updated dataframe*

eca\_clean

Exporting the DataFrame into Excel is always a good practice once we have completed the necessary preparation of the dataset for analysis.

*#export dataframe into Excel for analysis*

eca\_clean.to\_excel(‘eca\_clean.xlsx’, index = False)

Once we have completed the above preparation, this dataset is ready, and we can proceed to the analysis stage.

**Word count – 350 words**

# Question 2

Based on the dataset on medical costs, we are going to use it to analyse the following:

1. The average medical cost spent per region per person.
2. The average medical cost spent, smoker or non-smoker.
3. The average medical cost spent by BMI category.

A graph of blue bars

Description automatically generated

Graph 1 above shows the total medical cost billed by health insurance per person by region. It implies that people staying within the southeast regions incurred the highest per-person medical cost while those staying in the southwest have the lowest per-person medical cost.

Such could mean the people in the southwest are more health conscious than the other regions.

The codes that created the bar graph are as follows:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

eca = pd.read\_csv('eca\_clean.csv')

*#count the number of unique person in the region*

eca['person'] = eca['PersonID'].map(eca['PersonID'].value\_counts())

*#re-sort the table by summing both the charges and number of person in the region*

eca\_region = eca.groupby(by=["region"])['charges','person'].sum().reset\_index()

*#Average spend per person per region*

eca\_region['Average.charges'] = eca\_region['charges']/eca\_region['person']

eca\_region

plt.bar(eca\_region['region'], eca\_region['Average.charges'])

plt.xlabel('Region')

plt.ylabel('People')

plt.title("Graph 1 \n Charges by Regions Per Person")

plt.show()

A graph of a number of blue rectangular objects

Description automatically generated

Graph 2 above shows the total medical cost billed by health insurance by smokers or non-smokers. The graph shows very clearly that smokers incurred more medical costs by almost four times, reiterating the adverse effects of smoking on the body. The insurance companies could impose more insurance premiums for smokers based on the graph.

The codes that created the bar graph are as follows:

*#re-sort the table by summing both the charges and number of person in the region*

eca\_smoke = eca.groupby(by=["smoker"])['charges','person'].sum().reset\_index()

*# Average spend per person , Smoker or non smoker*

eca\_smoke['Average.charges'] = eca\_smoke['charges']/eca\_smoke['person']

eca\_smoke

plt.bar(eca\_smoke['smoker'], eca\_smoke['Average.charges'])

plt.xlabel('smoker')

plt.ylabel('People')

plt.title("Graph 2 \n Charges by Smoker or non Smoker")

plt.show()

A graph with blue rectangular bars

Description automatically generated

Graph 3 above shows the total medical cost billed by health insurance based on the BMI classification. The graph shows that people with higher BMI very clearly incurred the highest medical cost while the average medical cost for acceptable BMI hovers at 10,000. Based on the graph, insurance companies could impose more insurance premiums for people with higher BMI.

The codes that created the bar graph are as follows:

def categorize\_bmi(bmi):

if bmi > 24.9:

return "Above Ideal Range"

elif bmi > 18.5:

return "Ideal Range"

else:

return "Under Ideal Range"

eca['bmi\_category'] = eca['bmi'].apply(categorize\_bmi)

eca\_bmi = eca.groupby(by=["bmi\_category"])['charges','person'].sum().reset\_index()

eca\_bmi['Average.charges'] = eca\_bmi['charges']/eca\_bmi['person']

plt.bar(eca\_bmi['bmi\_category'], eca\_bmi['Average.charges'])

plt.xlabel('Category')

plt.ylabel('Charges')

plt.title("Graph 3 \n Charges by BMI Category")

plt.show()

**Word count – 401 words**

# Question 3

In order to use a decision tree, which is a sample classification technique to explore the dataset further, there is a need to do the following.

1. All data in the dataset must be clean and formatted adequately into numerical values. Therefore, the header "Sex" and "Region" must be recoded into numerical or dummy values.
2. To divide the dataset into two, namely, training and testing datasets. Such a separation is crucial to the decision tree model's performance, whereby one evaluates its accuracy while the other trains the model.
3. Setting up the model of the decision tree is the core step. The scikit-learn library would serve as the base to set up the tree classifier for the most suitable machine-learning model task. Additionally, Entropy and Gini measure the homogeneity in the classification tree.

* Since we are using "smoker" as the dependent variable, all other columns within the data set would be the basis for the model to make predictions.
* We need the DecisionTreeClassifier to be initialised under a random state to ensure the ability to repeat the predictions.
* Also, in machine learning, the model further enhances its capability to segregate the data based on the patterns of what could be why individuals are "smokers."

1. The testing dataset can indicate how many predictions the model got right and is one of the key elements for classification issues.
2. Once the model is mature, we can take the dataset further to analyse and visualise it to gain insight into the decision-making process.

In summary, the proper approach to using decision trees requires data reprocessing, splitting the dataset for test and training, building a tree model, and lastly evaluating the accuracy of the model.

**Word count – 281 words**

# Question 4

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

Based on the plotted decision tree, we can see that the accuracy level is 95% when we change the 'test\_size' in 3 different numbers, 0.2, 0.5 & 0.8.

Additionally, by computing the mean accuracy as well as the standard deviation for all three sizes, it yields the same result of 0.96 ( mean accuracy) and 0.0258 ( standard deviation ).

Such result means

* The model is stable and is not sensitive to changes in training and testing size.
* The model is suitable for the dataset, and it is well suited to the specific variables of smokers.
* The model has captured the data relationship correctly and can process its decision-making consistently regardless of size.

In summary, the consistent standard deviation and mean accuracy reflect good stability and suitability of the decision tree model. The models are also well-balanced and able to provide consistent and reliable predictions.

**Word count – 145 words**

# Question 5

Exploratory data analysis, or EDA, is one method scientists use to analyze and investigate datasets. Data scientists use this method to summarize the dataset characteristics and explore ways to manipulate it best to discover patterns or check assumptions. (What is exploratory data analysis?, Unknown)

A decision tree, on the other hand, is a supervised learning algorithm for classification and regression tasks, and it's suitable for making predictions from a dataset. (What is a Decision Tree?, Unknown)

Decision trees are just one of the many tools that EDA can use, with the pros of having data visualization for practical analysis complementing EDA. Visualizing the tree can provide a direction of how the data is being pushed down the nodes, giving the scientist a better chance of understanding the patterns. Pattern recognition is one of the pros of a decision tree because, through the tree's structure, it unveils how different features can influence the outcome.

Decision trees are also sensitive to outliers as they will influence spitting decisions. When branches in the tree show distinctive limited data points, this indicates an outlier.

The downside of the decision tree is the possibility of overfitting data when the dataset is large, causing a disconnect in the relationship in the data resulting in errors in reading the output.

Additionally, trees are susceptible to data imbalance and can affect the interpretation of a tree when one category is significantly more.

In summary, decision trees can be valuable tools for EDA, particularly regarding interpretability and non-linearity detection. However, they should be used to avoid overfitting and to ensure that the tree structure reflects meaningful insights into the data rather than just predictive performance. Depending on the dataset and the specific goals of EDA, scientists can use decision trees to complement other techniques to provide a comprehensive understanding of the data..

**Word count – 291 words**

# References

What is a Decision Tree? (Unknown). Retrieved from IBM: https://www.ibm.com/topics/decision-trees

What is exploratory data analysis? (Unknown). Retrieved from IBM: https://www.ibm.com/topics/exploratory-data-analysis