

# **ANL252**

# **PYTHON FOR DATA ANALYTICS**

**End-of-Course Assessment**

# **July Semester 2023**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **YEO JIA JIE JACKSON** | **N2770191** |

**Tutorial Group: T03**

**Instructor’s Name: Prof. KUMAR MUNISH**

**Submission Date: 24/10/2023**

**Question 1**

# Import “pandas”, under alias “pd”.

import pandas as pd

# Execute 'pd.read\_csv' to load and read csv file.

medical\_costs\_df = pd.read\_csv("ECA.csv")

# Look at first 5rows of data to look at the various columns

medical\_costs\_df.head()

# Input .shape function to identify dimensions

medical\_costs\_df.shape

# the data has 1340 rows and 8 columns

# Execute .isnull().sum() function to identify variables (columns) with missing values

medical\_costs\_df.isnull().sum(axis = 0)

# From output, variable "age has 123 missing values.

# **Task1.** Deleting of rows containing missing values via .dropna()

medical\_costs\_df.dropna(axis = 0, how = "any", inplace=True)

# Input .shape function to identify new dimensions

medical\_costs\_df.shape

# the data now has 1217 rows and 8 columns

# **Task2.** Correcting the strings in sex column to standardise format

# Define the condition for selecting rows

condition1 = (medical\_costs\_df['sex'] == 'F')

condition2 = (medical\_costs\_df['sex'] == 'M')

# Correct the strings for the selected rows

medical\_costs\_df.loc[condition1, 'sex'] = 'female'

medical\_costs\_df.loc[condition2, 'sex'] = 'male'

# **Task3.** Finding and removing outliers in medical\_costs\_df.

q1 = medical\_costs\_df["charges"].quantile(q=.25)

q3 = medical\_costs\_df["charges"].quantile(q=.75)

iqr = q3-q1

treated\_medical\_costs = medical\_costs\_df[

~((medical\_costs\_df["charges"]<q1-1.5\*iqr) | (medical\_costs\_df["charges"]>q3+1.5\*iqr))

]

# Input .shape function to identify new dimensions for treated df

treated\_medical\_costs.shape

# After 3 pre-processing tasks done, the clean dataset has 1091 rows.

# export treated df to csv file

treated\_medical\_costs.to\_csv("treated\_medicalcosts.csv", index=False)

(213 words)

**Question 2**

**Plot 1: Histogram**

# import relevant libraries for histogram plot

import matplotlib.pyplot as plt

# Plot a histogram of bmi column

treated\_medical\_costs.hist(column = 'bmi')

# Set the title and axis labels

plt.title('Histogram of BMI from Medical costs data')

plt.xlabel('BMI (ideal 18.5 to 24.9)')

plt.ylabel('Frequency')

#show the plot

plt.show()

A graph of a blue line

Description automatically generated with medium confidence

The histogram shows that highest number of people incurring medical costs are in the bmi range of 30, which is above the ideal bmi. It is interesting to note that most of the people incurring medical costs are in the bmi range of 25 to 35, as compared those who are severely above the ideal bmi range, at between 35 to 50. Number of people who are within the ideal bmi range of 18.5 to 25 and incurring medical costs, is comparable to those who are severely above the ideal bmi range, at between 35 to 50.

**Plot 2: Scatter Plot**

# Plot a scatter plot to show relationship between age and charges

x = treated\_medical\_costs['age']

y = treated\_medical\_costs['charges']

plt.scatter(x, y)

# Labelling the plot

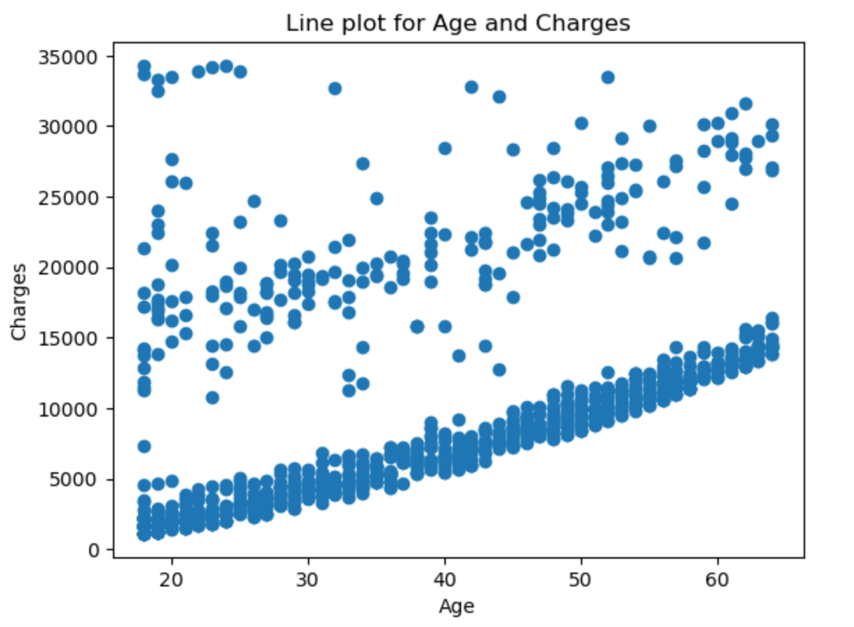
plt.title('Line plot for Age and Charges')

plt.xlabel('Age')

plt.ylabel('Charges')

#show the plot

plt.show()



It can be seen that there is an upward trend between the age of person, and the medical charges incurred. The minimum charges of someone in the 20s starts from less than $5,000, while the minimum charges of someone is the 60s starts from above $10,000. However, it can also be seen the highest medical charges incurred, in the region of $30,000 to $35,000, are actually incurred by people between 20 to 30 years old.

**Plot 3: Bar Graph**

# Plot a bar grap to show average charges in different regions

#import seaborn for bar graph plot

import seaborn as sns

ax = sns.barplot(treated\_medical\_costs, x="region", y="charges", estimator="mean", errorbar=None)

ax.bar\_label(ax.containers[0], fontsize=10);

# Labelling the plot

plt.xlabel("Region in the US")

plt.ylabel("Medical costs")

plt.title("Average medical costs based on regions in US")

#show the plot

plt.show()

A graph of different colored rectangular bars

Description automatically generated

From the bar graphs, it can be seen that the people in the northeast region of the US incurs the highest average medical costs. The northwest region is not too far off with just a difference of around $500 compared to the northeast. As a whole, the north region of the US incurs higher average medical costs, compared to the south region.

(379 words)

**Question 3**

# change string values into numerical values in sex column

sex = {'female': 0, 'male': 1}

treated\_medical\_costs['sex'] = treated\_medical\_costs['sex'].map(sex)

# double check the change

treated\_medical\_costs.head()

# change string values into numerical values in smoker column

smoker = {'yes': 0, 'no': 1}

treated\_medical\_costs['smoker'] = treated\_medical\_costs['smoker'].map(smoker)

# double check the change

treated\_medical\_costs.head()

# change string values into numerical values in region column

region = {'southwest': 0, 'southeast': 1, 'northwest': 2, 'northeast': 3}

treated\_medical\_costs['region'] = treated\_medical\_costs['region'].map(region)

# double check the change

treated\_medical\_costs.head()

# import relevant libraries for decision tree

from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

# extract independent variables and dependent variable (smoker) for decision tree

independent = ['age', 'sex', 'region', 'charges']

X = treated\_medical\_costs[independent]

y = treated\_medical\_costs['smoker']

medical\_costs\_dtree = DecisionTreeClassifier()

# fit decision tree

medical\_costs\_dtree = medical\_costs\_dtree.fit(X, y)

For decision tree, we need to first change non-numerical variables to numeric values so that machine learning can handle the dataset. As smoker is the dependent variable, smoker is the parameter y. The rest of the variables will be the independent variables, which is parameter X. However, I have chosen not to include the column Person ID, and children as I do not think these are important variables to studying this dataset. For this decision tree, the purpose will be to help us predict from the dataset, if someone is a smoker or not, given the medical charges incurred, age of person, gender of the person and also the region where the person is living in.

(244 words)

**Question 4**

# plot decision tree

tree.plot\_tree(medical\_costs\_dtree)

A diagram of a structure

Description automatically generated

From the decision tree plotted, it can be seen that there are a total of 61 leaf nodes.

(23 words)

**Question 5**

In my opinion, I think decision trees can be effectively used for exploratory data analysis, moving beyond their traditional role in making predictions.

Firstly, decision trees can provide us with a measure of feature importance, where the higher the value, the more important is the feature. (Stacey Ronaghan. May 2018) This allows us to know which independent variables will have a more significant impact on the dependent variable. These will be the key factors in our dataset.

Secondly, decision trees can help us to detect outliers. As the decision tree is being plotted, outliers may be detected by leaf nodes with little samples.

Lastly, the structure of the decision tree can tell us possible patterns of the data. As decision tree plot is very visual, from the structure and paths of the tree, relationships between variables may be determined.

In summary, other than make predictions, decision trees can be effective for exploratory data analysis.

(154 words)

Reference:

1. Stacey Ronaghan. 12 May 2018. The mathematics of decision trees, random forests and feature importance in Scikit-learn and Spark.

<https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3#:~:text=Feature%20importance%20is%20calculated%20as,the%20total%20number%20of%20samples>.