

## ✓ Seatwork 6.1: Exploratory Data Analysis on Your Own Dataset

CPE311 Computational Thinking with Python

**Name:** De los Reyes, Jann Moises Nyll B.

**Section:** CPE22S3

**Performed on :** 03/11/2024

**Submitted on :** 03/11/2024

**Submitted to :** 03/06/2024

Pandas for Data Analysis\*\*

Load the diabetes.csv file. Convert the diabetes.csv into dataframe Perform the following tasks in the diabetes dataframe:

1. Identify the column names
2. Identify the data types of the data
3. Display the total number of records
4. Display the first 20 records
5. Display the last 20 records
6. Change the Diagnosis column to Lab Result
7. Create a new column Classification by diagnosis that display "Diabetes" if the value of outcome is 1 , otherwise "No Diabetes"
8. Create a new dataframe "withDiabetes" that gathers data with diabetes
9. . Create a new dataframe "noDiabetes" thats gathers data with no diabetes
10. Create a new dataframe "Pedia" that gathers data with age 0 to 19
11. Create a new dataframe "Adult" that gathers data with age greater than 19
12. Use numpy to get the average age and glucose value.
13. Use numpy to get the median age and glucose value.
14. Use numpy to get the middle values of glucose and age
15. Use numpy to get the standard deviation of the skinthickness.

Link of the dataset:

<https://www.kaggle.com/datasets/avarice02/urinalysis-test-results>

```
1 #Uploading urinalysis_test.csv file and converting dataframe
2
3 filepath = '/content/urinalysis_tests.csv'
4 import numpy as np
5 import pandas as pd
6
7 data = pd.read_csv(filepath)
8 data
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity	WBC	RBC	Epithelial Cells	Mucous Threads	Amorphous Urates
0	0	76.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.010	1-3	0-2	OCCASIONAL	RARE	
1	1	9.00	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.030	1-3	0-2	RARE	FEW	
2	2	12.00	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.030	0-3	0-2	RARE	FEW	MODERATE
3	3	77.00	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.020	5-8	LOADED	RARE	RARE	
4	4	29.00	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.025	1-4	0-2	RARE	RARE	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1431	1431	0.06	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.015	1-2	0-1	RARE	FEW	
1432	1432	42.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.010	0-2	0-2	RARE	NONE	

Next steps: [View recommended plots](#)

```
1 #converting the data as dataframe
2
3 df = pd.DataFrame(data)
4 df
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity	WBC	RBC	Epithelial Cells	Mucous Threads	Amorphous Urates
0	0	76.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.010	1-3	0-2	OCCASIONAL	RARE	
1	1	9.00	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.030	1-3	0-2	RARE	FEW	
2	2	12.00	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.030	0-3	0-2	RARE	FEW	MODERATE
3	3	77.00	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.020	5-8	LOADED	RARE	RARE	
4	4	29.00	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.025	1-4	0-2	RARE	RARE	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1431	1431	0.06	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.015	1-2	0-1	RARE	FEW	
1432	1432	42.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.010	0-2	0-2	RARE	NONE	

Next steps: [View recommended plots](#)

```
1 #1. Identify the column names
2 df =pd.DataFrame(data)
3 df.columns

Index(['Unnamed: 0', 'Age', 'Gender', 'Color', 'Transparency', 'Glucose',
      'Protein', 'pH', 'Specific Gravity', 'WBC', 'RBC', 'Epithelial Cells',
      'Mucous Threads', 'Amorphous Urates', 'Bacteria', 'Diagnosis'],
      dtype='object')
```

```
1 #2. Identify the data types of the data
2 df =pd.DataFrame(data)
3 df.dtypes
```

```
Unnamed: 0      int64
Age            float64
Gender         object
Color          object
Transparency   object
Glucose        object
Protein        object
pH            float64
```

```
Specific Gravity    float64
WBC                 object
RBC                 object
Epithelial Cells    object
Mucous Threads      object
Amorphous Urates    object
Bacteria            object
Diagnosis            object
dtype: object
```

```
1 #3.Display the total number of records
```

```
2 df =pd.DataFrame(data)
```

```
3 len(df)
```

```
1436
```

```
1 #4.Display the first 20 records
```

```
2 df.iloc[:20]
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity
0	0	76.0	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.010
1	1	9.0	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.030
2	2	12.0	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.030
3	3	77.0	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.020
4	4	29.0	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.025
5	5	50.0	FEMALE	DARK YELLOW	HAZY	NEGATIVE	2+	6.0	1.020
6	6	3.0	MALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	7.0	1.005
7	7	23.0	FEMALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	NEGATIVE	7.0	1.005
8	8	8.0	MALE	YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	6.0	1.030
9	9	2.0	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	7.5	1.010
10	10	65.0	MALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.010
11	11	62.0	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.020
12	12	10.0	FEMALE	LIGHT	CLEAR	NEGATIVE	NEGATIVE	6.0	1.020

```
1 #5.Display the last 20 records
```

```
2 df.iloc[-20:]
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity
1416	1416	67.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.0
1417	1417	32.00	FEMALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	7.0	1.0
1418	1418	8.00	FEMALE	YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.0
1419	1419	4.00	FEMALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.0
1420	1420	4.00	FEMALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.0
1421	1421	22.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.0
1422	1422	22.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.0
1423	1423	20.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.5	1.0
1424	1424	26.00	FEMALE	YELLOW	CLEAR	TRACE	NEGATIVE	6.0	1.0
1425	1425	7.00	MALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	5.0	1.0
1426	1426	4.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.0
1427	1427	60.00	MALE	YELLOW	CLEAR	3+	NEGATIVE	5.0	1.0

```
1 #change the column name 'Diagnosis' to 'Lab Result'
2 df.rename(columns={"Diagnosis":"Lab Result"})
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity	WBC	RBC	Epithelial Cells	Mucous Threads	Amount
0	0	76.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.010	1-3	0-2	OCCASIONAL	RARE	
1	1	9.00	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.030	1-3	0-2	RARE	FEW	
2	2	12.00	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.030	0-3	0-2	RARE	FEW	MODERATE
3	3	77.00	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.020	5-8	LOADED	RARE	RARE	
4	4	29.00	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.025	1-4	0-2	RARE	RARE	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1431	1431	0.06	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.015	1-2	0-1	RARE	FEW	
1432	1432	42.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.010	0-2	0-2	RARE	NONE SEEN	
1433	1433	47.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.030	2-4	0-2	MODERATE	MODERATE	

```
1 #7. Create a new column Classification by diagnosis that display "UTI" if the value of outcome is 1 , otherwise "NO UTI"
2 df['Classification by diagnosis'] = df['Diagnosis'].apply(lambda x: 'UTI' if x == "POSITIVE" else 'No UTI')
3 df
4
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specif Gravi
0	0	76.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.0
1	1	9.00	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.0
2	2	12.00	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.0
3	3	77.00	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.0
4	4	29.00	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.0
...	...	...	...	...	...	...	...	...	...
1431	1431	0.06	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.0
1432	1432	42.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.0
1433	1433	47.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.0
1434	1434	57.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	5.0	1.0

Next steps: [View recommended plots](#)

```
1 #8. Create a new dataframe "withUTI" that gathers data with UTI patients.
2 withUTI = df[df['Classification by diagnosis']=='UTI']
3 withUTI
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specifi Gravit
9	9	2.0	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	7.5	1.01
39	39	23.0	FEMALE	STRAW	TURBID	NEGATIVE	NEGATIVE	5.0	1.01
57	57	70.0	FEMALE	YELLOW	HAZY	NEGATIVE	1+	6.0	1.01
59	59	23.0	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.5	1.01
69	69	26.0	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	7.0	1.00
...	...	...	...	...	...	...	...	...	...
1170	1170	16.0	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.03
1213	1213	4.0	FEMALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.02
1221	1221	78.0	MALE	YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.01
1398	1398	3.0	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.01

Next steps: [View recommended plots](#)

```
1 #9. Create a new dataframe "noUTI" thats gathers data with no UTI
2 noUTI = df[df['Classification']=='No UTI']
3 noUTI
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity	WBC	RBC	Epithelial Cells	Mucous Threads	Amorphous Urates
0	0	76.00	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.010	1-3	0-2	OCCASIONAL	RARE	
1	1	9.00	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.030	1-3	0-2	RARE	FEW	
2	2	12.00	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.030	0-3	0-2	RARE	FEW	MODERATE
3	3	77.00	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.020	5-8	LOADED	RARE	RARE	
4	4	29.00	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.025	1-4	0-2	RARE	RARE	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1431	1431	0.06	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.015	1-2	0-1	RARE	FEW	
1432	1432	42.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.010	0-2	0-2	RARE	NONE SEEN	
1433	1433	47.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.030	2-4	0-2	MODERATE	MODERATE	
1434	1434	57.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	5.0	1.030	0-2	0-2	PLENTY	PLENTY	
1435	1435	3.00	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.020	0-1	0-2	RARE	RARE	

1355 rows x 18 columns

Next steps: [View recommended plots](#)

```
1 #10. Create a new dataframe "Minor" that gathers data with age 0 to 19
2 Minor = df[df['Age'] <= 19]
3 Minor
```

	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity	WBC	RBC	Epithelial Cells	Mucous Threads	Amorphous Urates
1	1	9.00	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	1+	5.0	1.030	1-3	0-2	RARE	FEW	FEW MODERATE
2	2	12.00	MALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	5.0	1.030	0-3	0-2	RARE	FEW	MODERATE
6	6	3.00	MALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	7.0	1.005	0-3	0-2	RARE	RARE	NONE SEEN
8	8	8.00	MALE	YELLOW	SLIGHTLY HAZY	NEGATIVE	TRACE	6.0	1.030	1-3	0-2	RARE	FEW	NONE SEEN
9	9	2.00	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	7.5	1.010	70-75	2-4	RARE	FEW	NONE SEEN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1426	1426	4.00	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.020	13-15	0-2	FEW	RARE	NONE SEEN
1428	1428	2.00	FEMALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	7.5	1.010	0-1	1-3	RARE	RARE	NONE SEEN
1430	1430	8.00	MALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.010	0-1	0-1	RARE	NONE SEEN	NONE SEEN
1431	1431	0.06	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.0	1.015	1-2	0-1	RARE	FEW	NONE SEEN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Next steps: [View recommended plots](#)

```
1 #11. Create a new dataframe "Adult" that gathers data with age greater than 19
2 Adult = df[df['Age'] > 19]
3 Adult
```



	Unnamed: 0	Age	Gender	Color	Transparency	Glucose	Protein	pH	Specific Gravity	WBC	RBC	Epithelial Cells	Mucous Threads	Amorph Ur:
0	0	76.0	FEMALE	LIGHT YELLOW	CLEAR	NEGATIVE	NEGATIVE	5.0	1.010	1-3	0-2	OCCASIONAL	RARE	NO SIGNS
3	3	77.0	MALE	BROWN	CLOUDY	NEGATIVE	1+	6.0	1.020	5-8	LOADED	RARE	RARE	NO SIGNS
4	4	29.0	FEMALE	YELLOW	HAZY	NEGATIVE	TRACE	6.0	1.025	1-4	0-2	RARE	RARE	NO SIGNS
5	5	50.0	FEMALE	DARK YELLOW	HAZY	NEGATIVE	2+	6.0	1.020	0-3	0-2	PLENTY	FEW	NO SIGNS
7	7	23.0	FEMALE	LIGHT YELLOW	SLIGHTLY HAZY	NEGATIVE	NEGATIVE	7.0	1.005	3-5	1-2	MODERATE	FEW	NO SIGNS
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1427	1427	60.0	MALE	YELLOW	CLEAR	3+	NEGATIVE	5.0	1.010	0-2	0-1	RARE	NONE SEEN	NO SIGNS
1429	1429	49.0	MALE	DARK YELLOW	SLIGHTLY HAZY	NEGATIVE	NEGATIVE	6.5	1.015	3-5	2-4	MODERATE	RARE	NO SIGNS
1432	1432	42.0	MALE	YELLOW	CLEAR	NEGATIVE	NEGATIVE	6.5	1.010	0-2	0-2	RARE	NONE SEEN	NO SIGNS
1433	1433	47.0	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	6.0	1.030	2-4	0-2	MODERATE	MODERATE	NO SIGNS
1434	1434	57.0	FEMALE	DARK YELLOW	CLEAR	NEGATIVE	TRACE	5.0	1.030	0-2	0-2	PLENTY	PLENTY	NO SIGNS

782 rows × 18 columns

Next steps:

☒ View recommended plots

```

1 #12. Use numpy to get the average age , pH Level
2 #and specific gravity
3 # Extraction of data such as age, ph level and specific gravity
4 # in the dataset
5
6 ages= np.array(df['Age'])
7 pH_Level = np.array(df['pH'])
8 specific_gravity = np.array(df['Specific Gravity'])

1 #Compute the following mean using numpy
2 ave_age = np.mean(ages)
3 ave_pH = np.mean(pH_Level)
4 ave_sg = np.mean(specific_gravity)

1 #Display the following mean where age as a whole number and
2 # glucose, pH level and specific gravity into 4 decimal value.
3
4 print(f"Average Age: {ave_age:.0f} years old")
5 print(f"Average pH Level: {ave_pH:.4f} ")
6 print(f"Average Specific Gravity: {ave_sg:.4f} ")

Average Age: 27 years old
Average pH Level: 6.0529
Average Specific Gravity: 1.0158

1 #13. Use numpy to get the median age and pH Level, and specific gravity
2 # Using the extraction above we can get the median
3
4 #Compute the median of age and value using numpy
5 median_age = np.median(ages)
6 median_ph = np.median(pH_Level)
7 median_sg= np.median(specific_gravity)
8
9 #Display the median
10 print(f"Median age:{median_age} years old")
11 print(f"Median pH Level: {median_ph:.2f}")
12 print(f"Median specific gravity:{median_sg:.2f}")

Median age:23.0 years old
Median pH Level: 6.00

```

Median specific gravity:1.01

```
1 #15.Use numpy to get the standard deviation of the ph Level and Specific Gravity.  
2  
3  
4 #compute for the std  
5 std_ph = np.std(ph_level, dtype=np.float32)
```