LISÄTEHTÄVÄT viikko 35
Jenni Ylisirniö

Tehtävä 1

Tallenna oheinen diabetes.csv ja lataa se pandasin dataframeen. Tulosta muutamia tunnuslukuja mm. count, mean, min, max, std.

panda-exercise.py > ...

```python
import pandas as pd

#load data from csv
healthData = pd.read_csv('diabetes.csv')

print('Count values:')
print(healthData.count()) # print count of every column of the dataset
print('Count of pregnancies:', healthData.Pregnancies.count()) # print count of pregnancies

print('Max values:')
print(healthData.max()) # print max values of every column
print('Max BMI:',healthData.BMI.max()) # print max BMI

print('Min values:')
print(healthData.min()) # print minimum values of every column

print('Mean values:')
print(healthData.mean()) # average values of given dataset

print('Std:') # standard deviation values of given dataset
print(healthData.std())
```

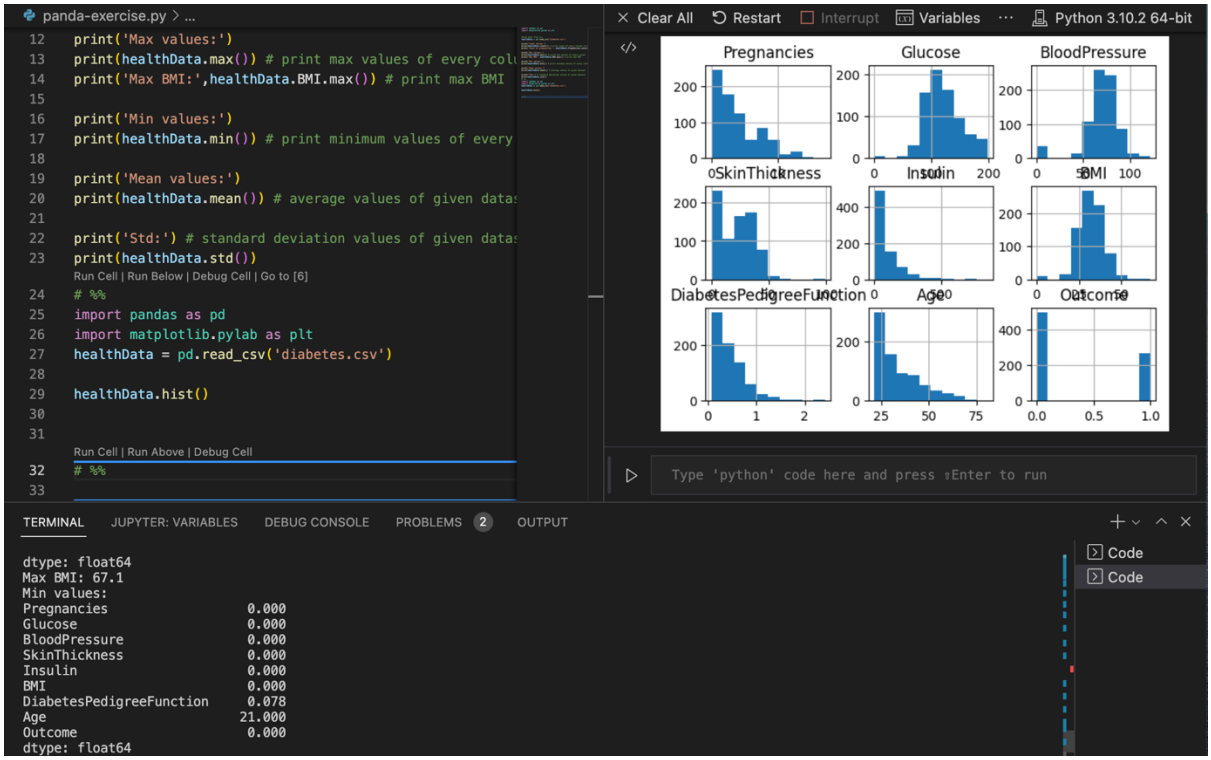
TERMINAL JUPYTER DEBUG CONSOLE PROBLEMS OUTPUT

```
Age                         33.240885
Outcome                      0.348958
dtype: float64
Std:
Pregnancies                  3.369578
Glucose                     31.972618
BloodPressure               19.368155
SkinThickness               15.952218
Insulin                    115.244002
BMI                          7.884160
DiabetesPedigreeFunction     0.331329
Age                         11.760232
Outcome                      0.476951
```

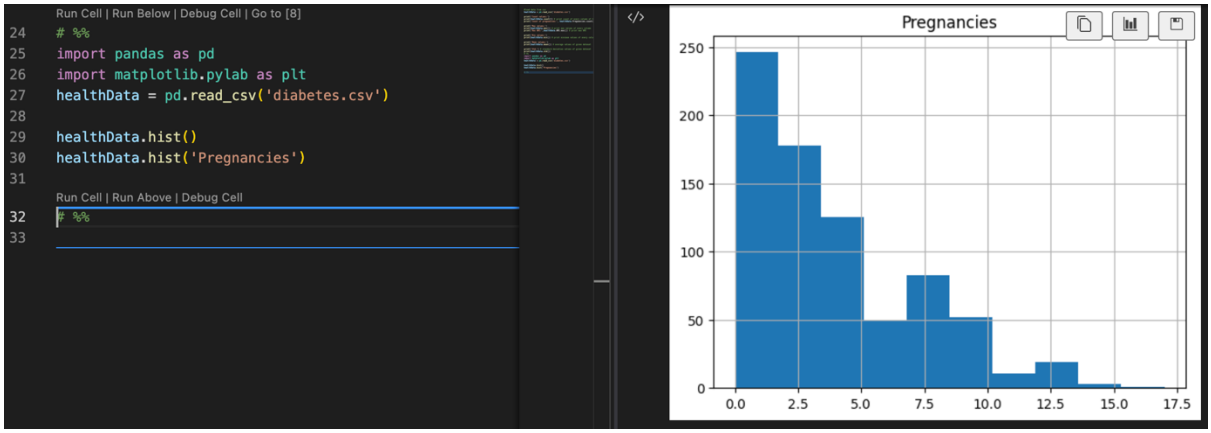
Screenshotissa näkyy vain pieni osa ohjelman ajosta, mutta ajo toimii täydellisesti.

Tehtävä 2

Piirrä histogrammi kuvaaja datasta:



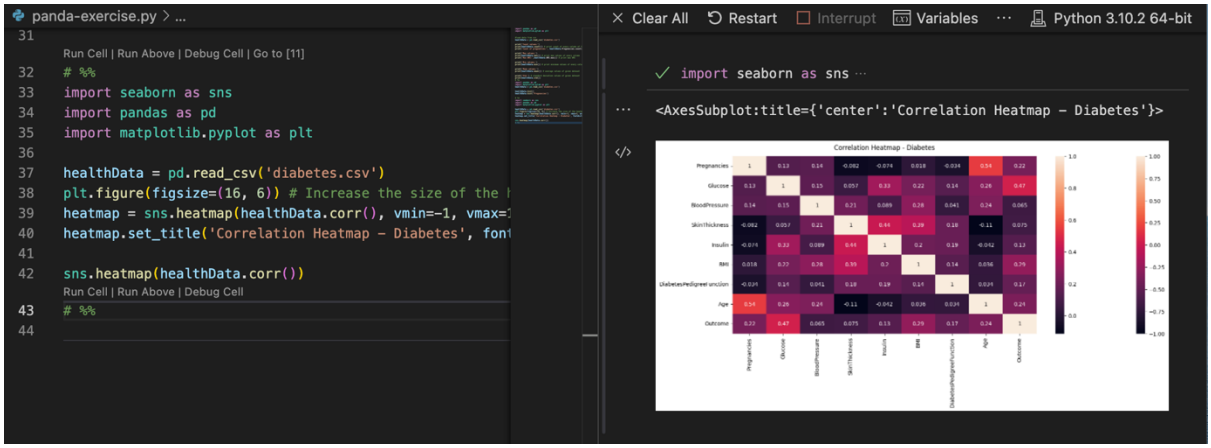
Yksi histogrammi näyttää vähän siistimmältä:

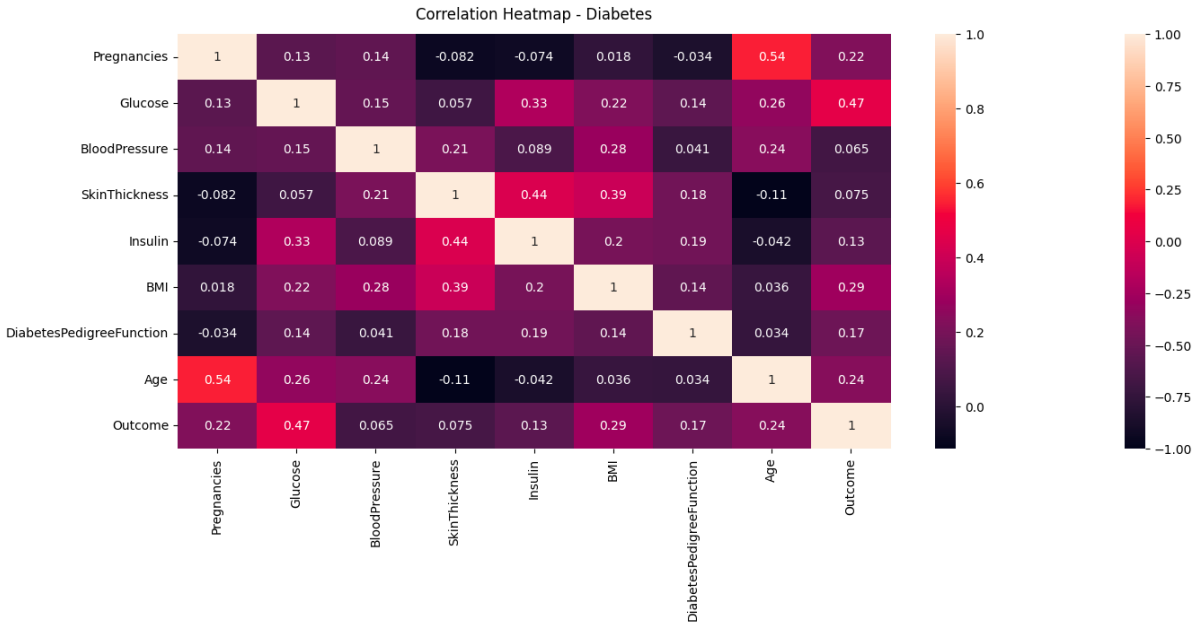


Tehtävä 3

Piirrä korrelaatioheatmap datasta:

```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

healthData = pd.read_csv('diabetes.csv')
plt.figure(figsize=(16, 6)) # Increase the size of the heatmap.
heatmap = sns.heatmap(healthData.corr(), vmin=-1, vmax=1, annot=True)
heatmap.set_title('Correlation Heatmap - Diabetes', fontdict={'fontsize':12},
pad=12)

sns.heatmap(healthData.corr())
```

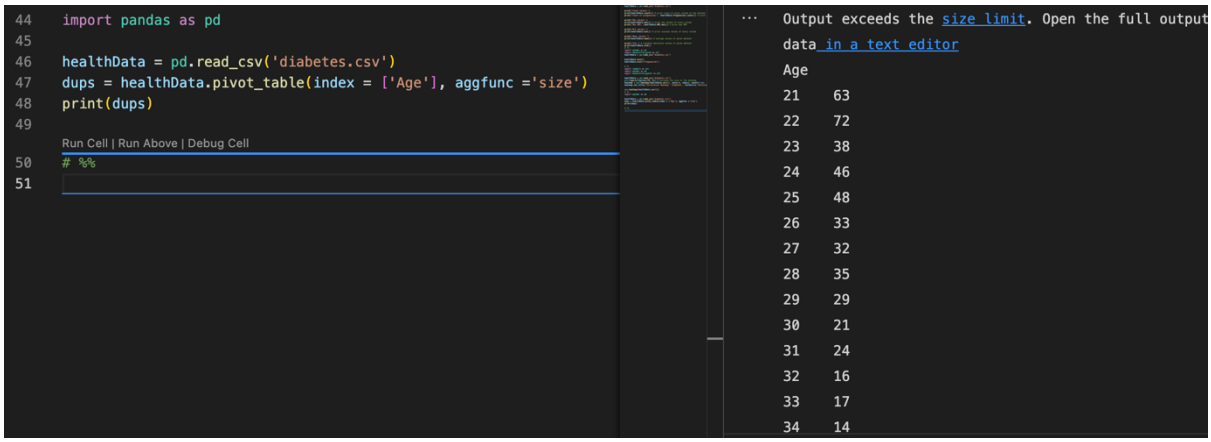




https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

Tehtävä 4

Laske potilaiden lukumäärä iän mukaan siten, että suurin lukumäärä on ensin.



```
import pandas as pd

healthData = pd.read_csv('diabetes.csv')
dups = healthData.pivot_table(index = ['Age'], aggfunc ='size')
print(dups)

Run Cell | Run Above | Debug Cell
# %%
```

```
Output exceeds the size limit. Open the full output data in a text editor
Age
21    63
22    72
23    38
24    46
25    48
26    33
27    32
28    35
29    29
30    21
31    24
32    16
33    17
34    14
```

Osasin laskea potilaiden lukumäärän iän mukaan, mutta järjestystä en osannut vaihtaa niin, että suurin lukumäärä olisi ensin. Yritin sort_values , mutta ei jostain syystä onnistunut.

Tulosta myös montako diabetestapausta (1) ja ei diabtesta (0) on aineistossa.

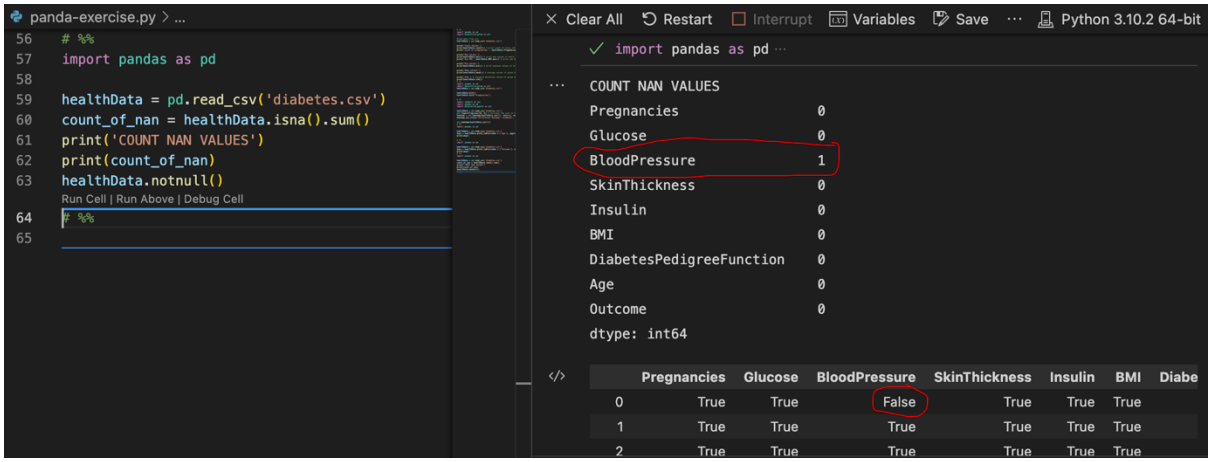
```
Run Cell | Run Above | Debug Cell | Go to [34]
# %%
import pandas as pd

healthData = pd.read_csv('diabetes.csv')
dups = healthData.pivot_table(index = ['Outcome'], aggfunc ='size')
print(dups)
Run Cell | Run Above | Debug Cell
# %%
```

```
dtype: int64

✓ import pandas as pd ···

Outcome
0    500
1    268
dtype: int64
```

Eli 500 ei ole (0) diabetestä ja 268 on (1) diabetes.

Tehtävä 5

Onko aineistossa nan arvoja? Jos on, missä sarakkeissa ja montako?



Aineistossa on yksi NaN (not a number) arvo. Se löytyy Blood Pressure -sarakkeesta, riviltä yksi.