List 01. Into to Python

Nikita V. Artamonov

February 14, 2025

Contents

1 Pandas

#**1**. Load a dataset from `sleep75.csv` file.

1. evaluate the dataset size (a number of observation & a number of variables)

2. Fill the table of values of variables

Obs	sleep	totwrk	age	male
0				
5				
100				
700				

3. Calculate the correlation matrix for the following variables: sleep, totwrk, age

4. the table of values of variables

Desc.Stat	sleep	totwrk	age	hrwage
max				
min				
mean				
median				
st.dev				
var (unbiased)				
var (biased)				
1st quartile				
3rd quartile				

Remark: 1st/3rd quartiles are 25%/75% quantiles respectively.

5. How many observations in the dataset
   (a) with sleep>3000
   (b) with totwrk<2000
   (c) with age>40
   (d) with age<30

6. How many observations in the dataset with totwrk=0? Who is this people?

7. Do we have evidence for missing values in the dataset? How many do we have?

#**2**. Load a dataset from `Electricity.csv` file.

1. evaluate the dataset size (a number of observation & a number of variables)

2. Fill the table of values of variables

Obs	cost	q	pl	pk	pf
1					
15					
48					
87					

3. Calculate the correlation matrix for the following variables: cost, q, pl, pk, pf

4. the table of values of variables

Desc.Stat	cost	q	pl	pk	pf
max					
min					
mean					
median					
st.dev					
var (unbiased)					
var (biased)					
1st quartile					
3rd quartile					

   Remark: 1st/3rd quartiles are 25%/75% quantiles respectively.

5. How many observations in the dataset

   (a) with cost>40

   (b) with q<5000

   (c) with q>4000

   (d) with 20<cost<50

6. Do we have evidence for missing values in the dataset? How many do we have?

#**3**. Load a dataset from `wage2.csv` file.

1. evaluate the dataset size (a number of observation & a number of variables)

2. Fill the table of values of variables

Obs	wage	hour	IQ	educ	exper	age
1						
25						
179						
800						

3. Calculate the correlation matrix for the following variables: wage, hour, IQ, educ, exper
4. the table of values of variables

Desc.Stat	wage	hour	IQ	educ	exper	wage
max						
min						
mean						
median						
st.dev						
var (unbiased)						
var (biased)						
1st quartile						
3rd quartile						

   Remark: 1st/3rd quartiles are 25%/75% quantiles respectively.

5. How many observations in the dataset
   (a) with wage>1000
   (b) with age<40
   (c) with exper>10
   (d) with 100<IQ<130
6. Do we have evidence for missing values in the dataset? How many do we have?

#**4**. Load a dataset from `Labour.csv` file. Create a new DataFrame containing the log-variables of the initial dataset.

#**5**. Load a dataset from `Electricity.csv` file. Create a new DataFrame containing the log-variables of the initial dataset.

2 NumPy

#**1**. Consider matrices

$$A = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -1 \\ -2 & 0 & 2 \end{pmatrix} \quad B = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} \quad C = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 3 & -1 \end{pmatrix} \quad D = \begin{pmatrix} 5 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}$$

Calculate

$$\det(A) \qquad A^{-1} \qquad AB \qquad CA \qquad B^\top A \qquad 2C \pm 3D \qquad C^\top D \qquad D^\top C$$

#**2**. Consider matrices

$$A = \begin{pmatrix} 0 & 1 & -2 & 6 \\ 1 & 0 & 1 & -1 \\ -1 & 1 & -2 & 0 \\ 1 & -1 & 2 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 2 \\ -1 \\ 0 \\ 4 \end{pmatrix} \qquad C = \begin{pmatrix} 2 & 0 & 1 & 1 \\ 1 & 3 & -1 & 3 \end{pmatrix}$$

Calculate

$$\det(A) \qquad A^{-1} \qquad A^2 \qquad A^3 \qquad AB \qquad CA \qquad CB \qquad B^\top A$$

#**3**. Consider 1-D arrays:

$$\begin{aligned} x^\top &= \begin{pmatrix} 1 & 2.8 & 1.8 & 3 & 0.5 & 1.5 \end{pmatrix} \\ y^\top &= \begin{pmatrix} 3.8 & 2.1 & 5.3 & 3.4 & 0.2 & 0.1 \end{pmatrix} \\ z^\top &= \begin{pmatrix} -0.4 & 0 & -4.3 & 6.8 & -3.3 & 2.7 \end{pmatrix} \end{aligned}$$

Perform the following manipulations: $x+2$, $x-3$, $x+3y$, z^2, z^3, $\log(x)$, $\sqrt{x}$, $|z|$, $\log(|z|+1)$

#**4**. Solve a system of linear equation in matrix form $Ax = b$ with matrices

$$A, b = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad A, b = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad A, b = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$$

3 Visualization

#**1**. Load a dataset from `sleep75.csv` file.

1. Draw histogram for variables sleep, totwrk, age, hrwage, educ
2. Draw stacked histogram for sleep across male dummy
3. Draw stacked histogram for totwrk across south dummy
4. Draw stacked histogram for totwrk across smsa dummy

5. Draw scatter plot sleep vs totwrk
6. Draw scatter plot sleep vs totwrk with grouping by male dummy
7. Draw scatter plot sleep vs age
8. Draw scatter plot sleep vs age with grouping by south dummy
9. Draw scatter plot sleep vs edu
10. Draw scatter plot sleep vs edu with grouping by smsa dummy

#**2**. Load a dataset from `Labour.csv` file.

1. Draw histogram for variables output, capital, labour, wage
2. Draw histogram for log-variables output, capital, labour, wage
3. Draw scatter plots output vs another variables
4. Draw scatter plots log(output) vs log of another variables

#**3**. Load a dataset from `Electricity.csv` file.

1. Draw histogram for variables cost, q, pf, pk, pl
2. Draw histogram for log-variables cost, q, pf, pk, pl
3. Draw scatter plots cost vs another variables
4. Draw scatter plots log(cost) vs log of another variables

#**4**. Load a dataset from `diamonds.csv` file.

1. Draw histogram for variables price, carat
2. Draw histogram for log-variables price, carat
3. Draw stacked histogram for price across cut
4. Draw stacked histogram for carat across clarity
5. Draw stacked histogram for log(price) across color
6. Draw stacked histogram for log(carat) across color

7. Draw scatter plot price vs carat
8. Draw scatter plot log-price vs log-carat
9. Draw scatter plot log-price vs log-carat with grouping by cut
10. Draw scatter plot log-price vs log-carat with grouping by color
11. Draw scatter plot log-price vs log-carat with grouping by clarity

#**5**. Load a dataset from `Diamond.csv` file.

1. Draw histogram for variables price, carat
2. Draw histogram for log-variables price, carat
3. Draw stacked histogram for price across certification
4. Draw stacked histogram for carat across clarity
5. Draw stacked histogram for log(price) across colour
6. Draw stacked histogram for log(carat) across colour
7. Draw scatter plot price vs carat
8. Draw scatter plot log-price vs log-carat
9. Draw scatter plot log-price vs log-carat with grouping by certification
10. Draw scatter plot log-price vs log-carat with grouping by colour
11. Draw scatter plot log-price vs log-carat with grouping by clarity