19.06.25

OPENCAMPUS.sh

Introduction to data Science and Machine Learning

MISSING VALUES

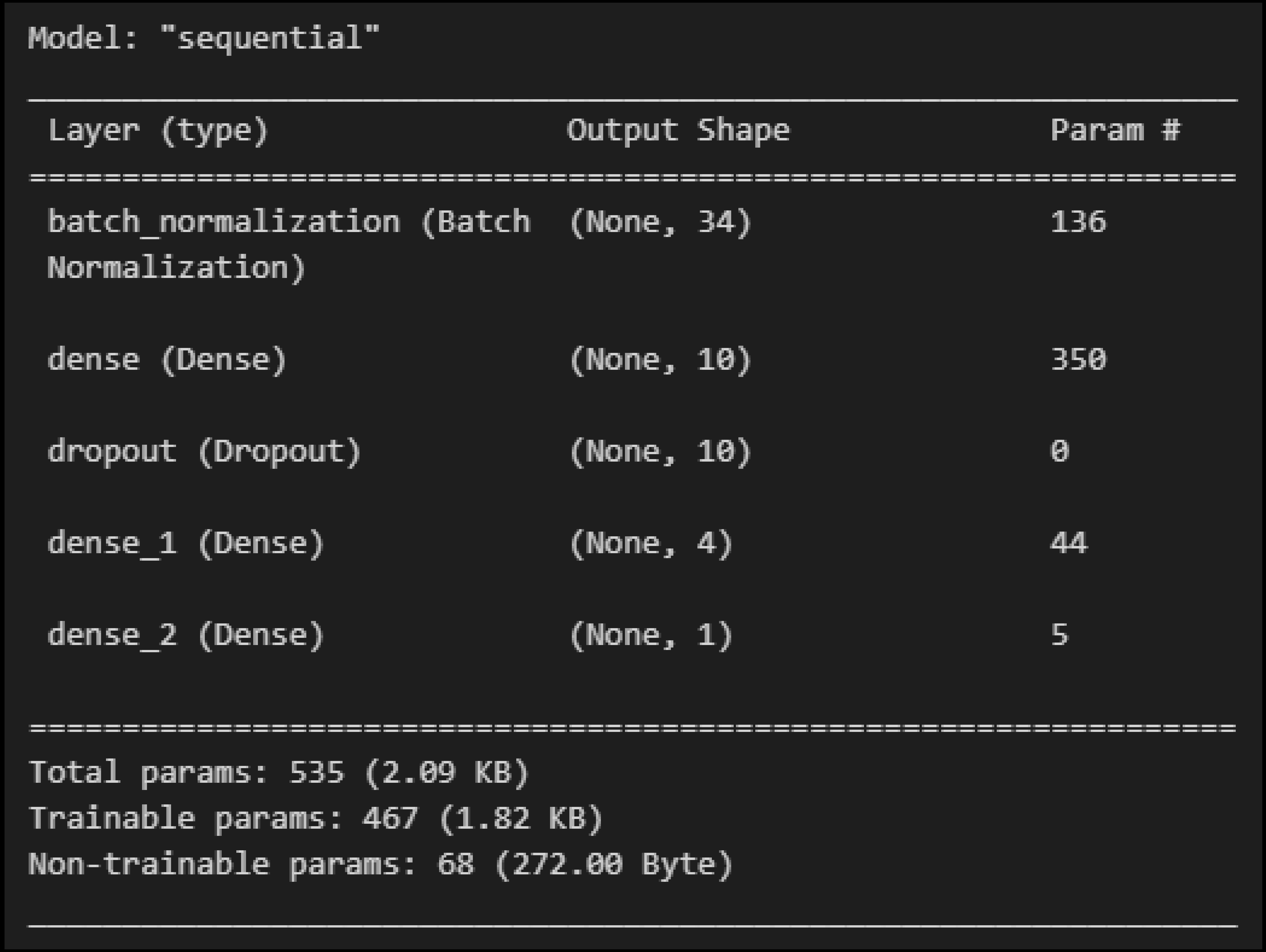
- Quiz
- Use of dropout layers
- Handling of missing values
- (Support Vector Machines)

QUIZ

NEURAL NET WITH DROPOUT LAYER

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import InputLayer, Dense, BatchNormalization, Dropout
from tensorflow.keras.optimizers import Adam

model = Sequential([
  InputLayer(input_shape=(training_features.shape[1], )),
  BatchNormalization(),
  Dense(10, activation='relu'),
  Dropout(.3),
  Dense(4, activation='relu'),
  Dense(1)
])
```



```
Model: "sequential"

_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 batch_normalization (Batch  (None, 34)                136       
 Normalization)                                                  
                                                                 
 dense (Dense)               (None, 10)                350       
                                                                 
 dropout (Dropout)           (None, 10)                0         
                                                                 
 dense_1 (Dense)             (None, 4)                 44        
                                                                 
 dense_2 (Dense)             (None, 1)                 5         
                                                                 
=================================================================
Total params: 535 (2.09 KB)
Trainable params: 467 (1.82 KB)
Non-trainable params: 68 (272.00 Byte)
_________________________________________________________________
```

DROPOUT LAYER CHARACTERISTICS

- **Sets individual activations in the previous layer to zero at each iteration step with the defined dropout probability.**
- **Introduces redundancy into the network.**
- **Helps to prevent overfitting.**
- **Is only applied during training; during inference, all neurons are always used.**

HANDLING OF MISSING VALUES

REASONS FOR MISSING VALUES

- **Missing responses in surveys**
- **Merging data from different sources with varying variable categories or time steps**
- **Technical issues in data collection or recording**

TYPES OF MISSING VALUES

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing not at Random (MNAR)

BREAKOUT

Discuss solutions for the following possible cases in the weather dataset:

- **Temperature data for a month with missing data for two days:**
  `[20, 19, 23, 19, 17, 17, NA, 24, 16, 20, 22, 21, 20, 19, 17, 22, 24, 21, 23, 15, 18, 18, 21, 19, 19, 21, 21, 19, 23, NA]`

- **Temperature data for a month with missing data for a week:**
  `[18, 15, 21, 15, 24, 16, 21, 16, 22, 18, 17, 25, 22, 21, 16, 19, 17, 23, NA, NA, NA, NA, NA, NA, NA, 21, 20, 20, 16, 15]`

- **Weather code data for 20 days with missing data for one day:**
  `[10, 60, NA, 95, 61, 1, 29, 81, 21, 25, 25, 80, 80, 63, 81, 80]`

HANDLING OF MISSING VALUES

- Listwise deletion of affected cases

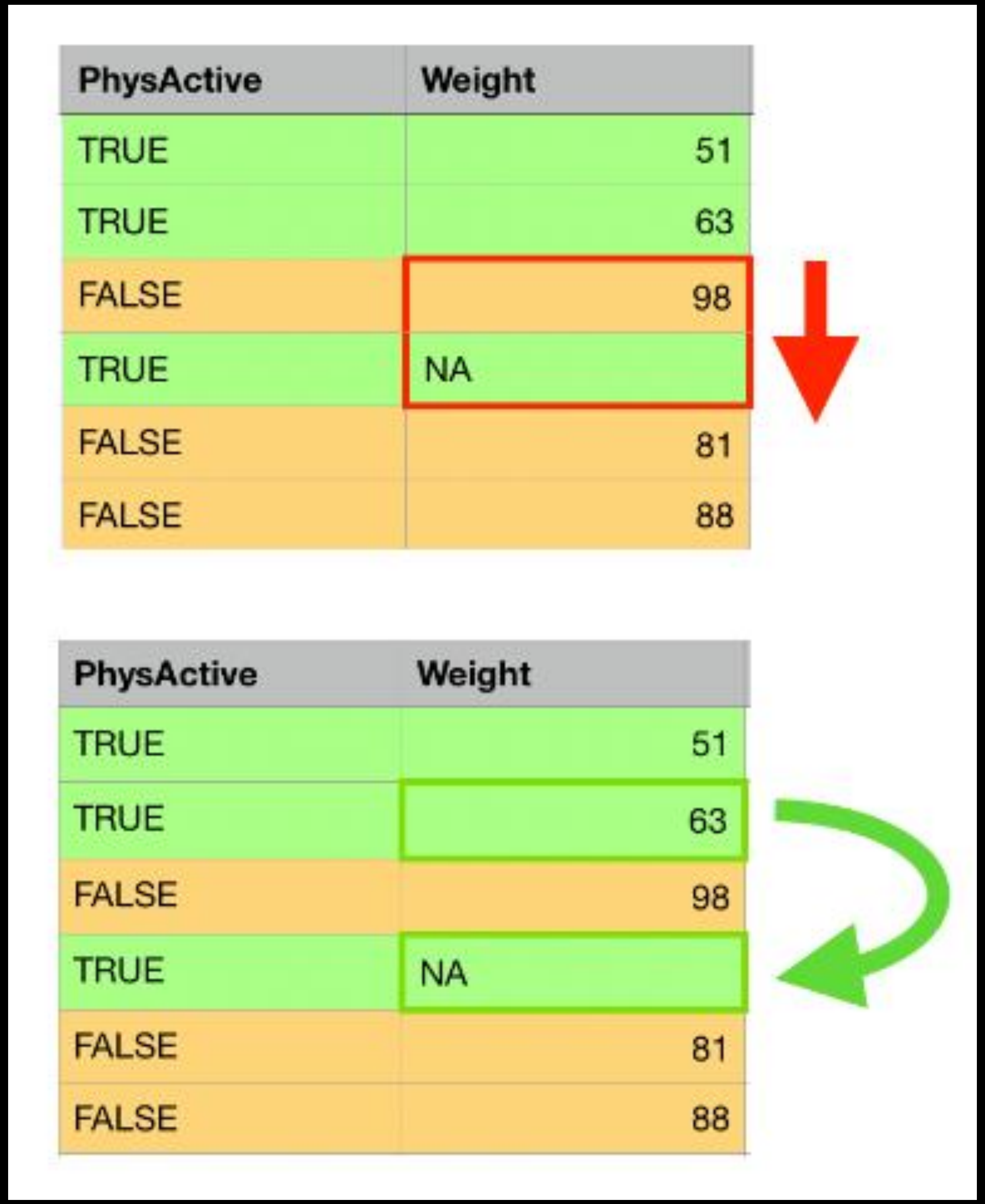
- Simple donor-based imputation:
  - Mean imputation (or median or mode)
  - Based on "similarity" (hot-deck imputation)
  - By minimal distance (k-nearest neighbors)

- Simple model-based imputation
  - Iterative regression

- Multiple imputation

HOT-DECK IMPUTATION

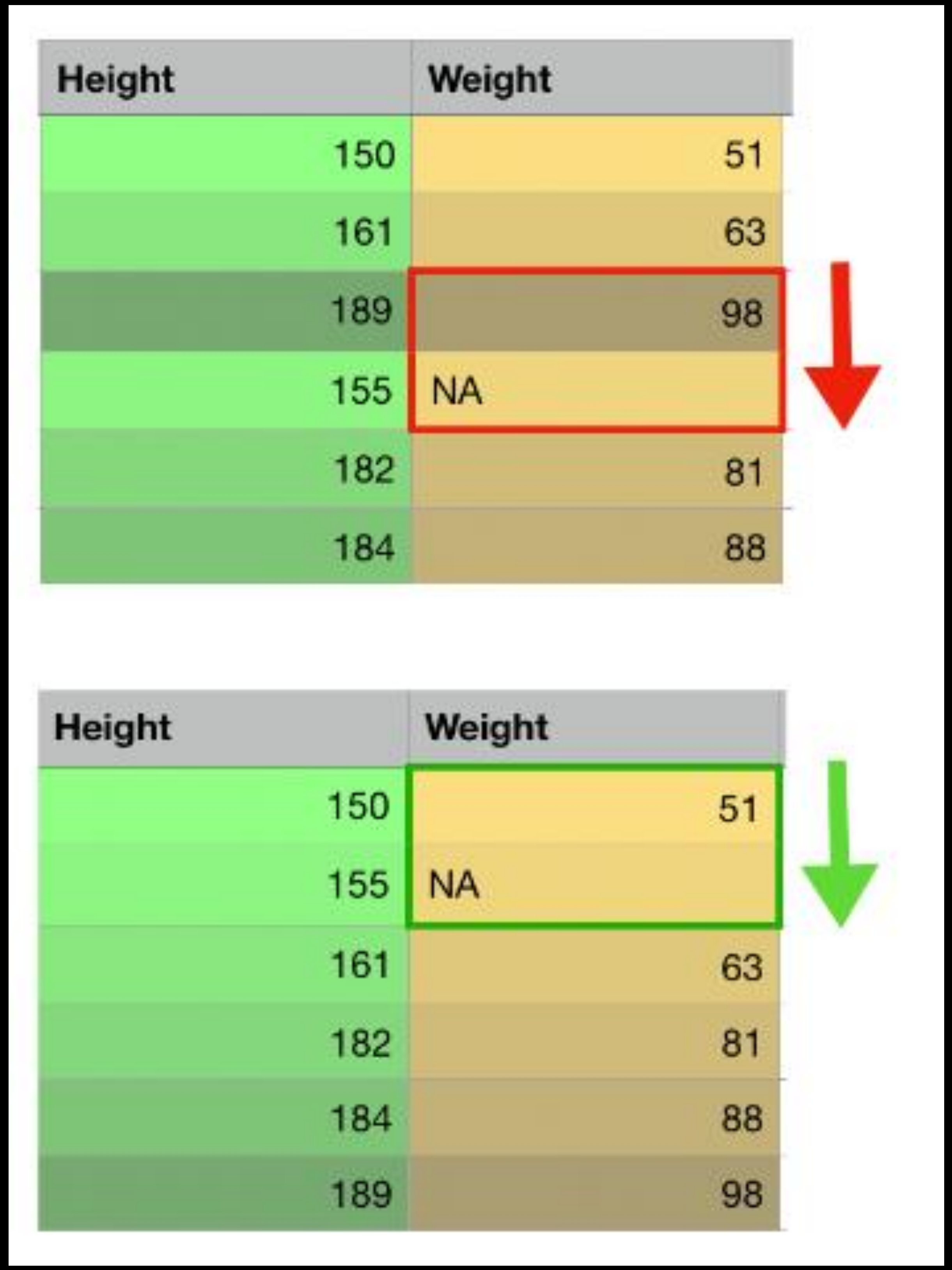
By Domains



PhysActive	Weight
TRUE	51
TRUE	63
FALSE	98
TRUE	NA
FALSE	81
FALSE	88

PhysActive	Weight
TRUE	51
TRUE	63
FALSE	98
TRUE	NA
FALSE	81
FALSE	88

By Correlation



Height	Weight
150	51
161	63
189	98
155	NA
182	81
184	88

Height	Weight
150	51
155	NA
161	63
182	81
184	88
189	98

Oleszak, M. (2021). Hot-deck imputation | R. Retrieved December 6, 2021, from Datacamp website: https://campus.datacamp.com/courses/handling-missing-data-with-imputations-in-r/donor-based-imputation?ex=5

K-NEAREST NEIGHBORS (KNN)

Search for the k cases with the minimal distance

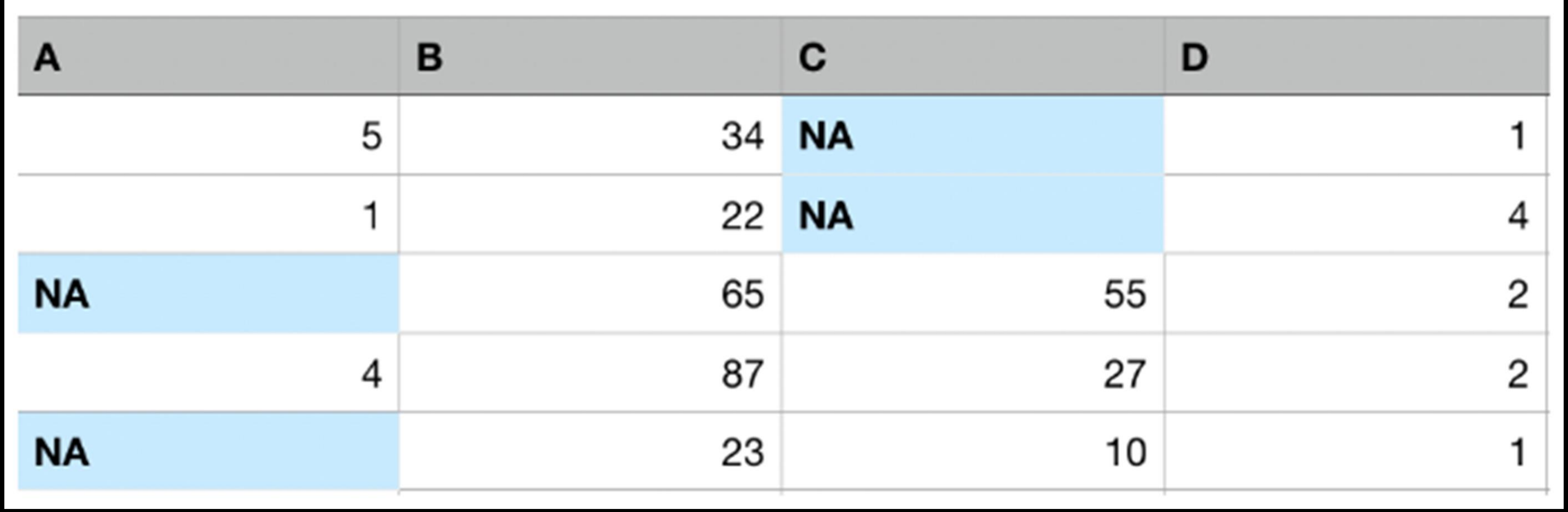
- Different distance measurements depending on the variable type
- Aggregation of distances using a sum function

Various approaches to calculate the imputation value:

- The value with the minimal distance is taken (1NN)
- Random selection from the k cases
- Calculation from the k cases using the (weighted mean)

ITERATIVE REGRESSION

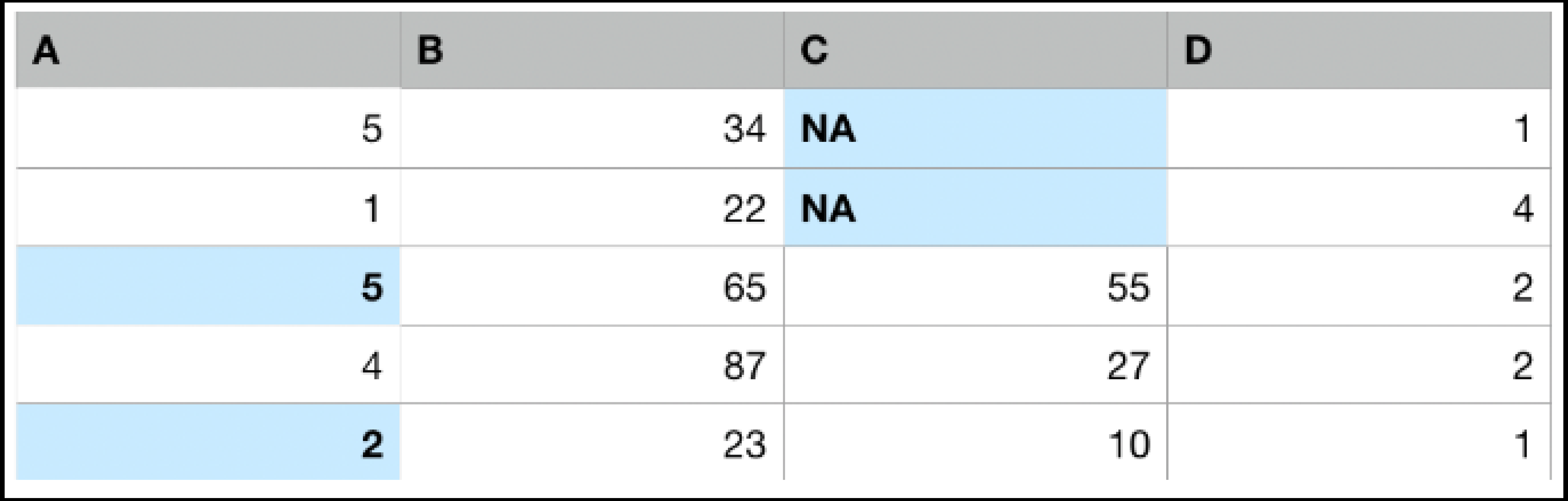
1) Prediction of missing values in A



A	B	C	D
5	34	NA	1
1	22	NA	4
NA	65	55	2
4	87	27	2
NA	23	10	1

ITERATIVE REGRESSION

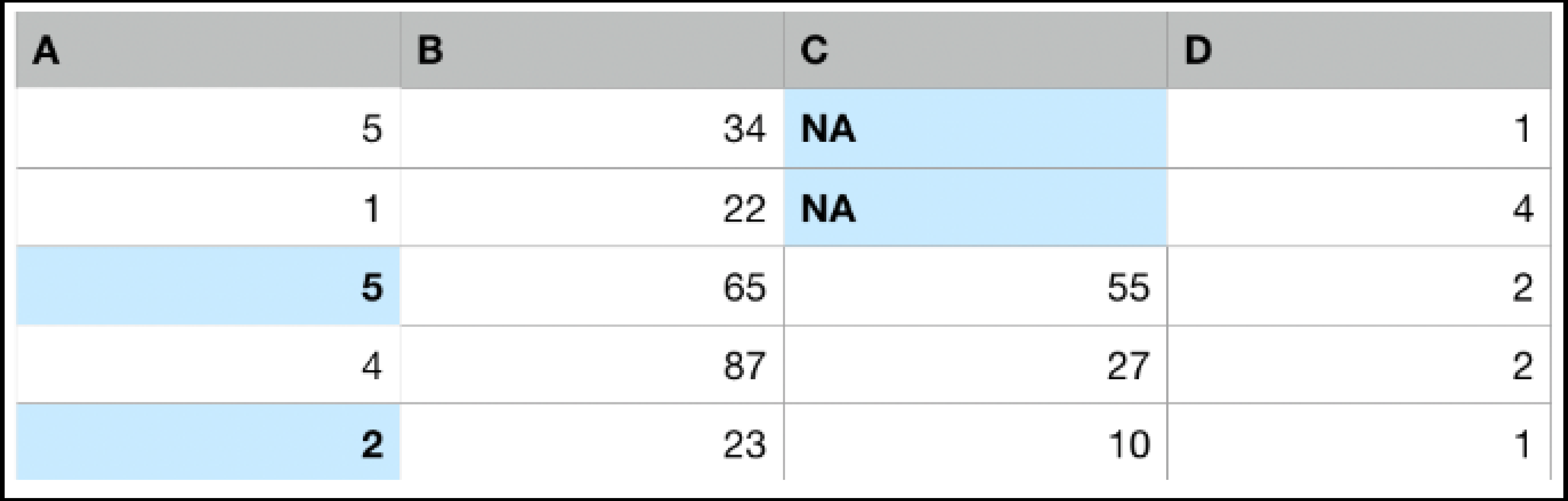
1) Prediction of missing values in A



A	B	C	D
5	34	**NA**	1
1	22	**NA**	4
5	65	55	2
4	87	27	2
2	23	10	1

ITERATIVE REGRESSION

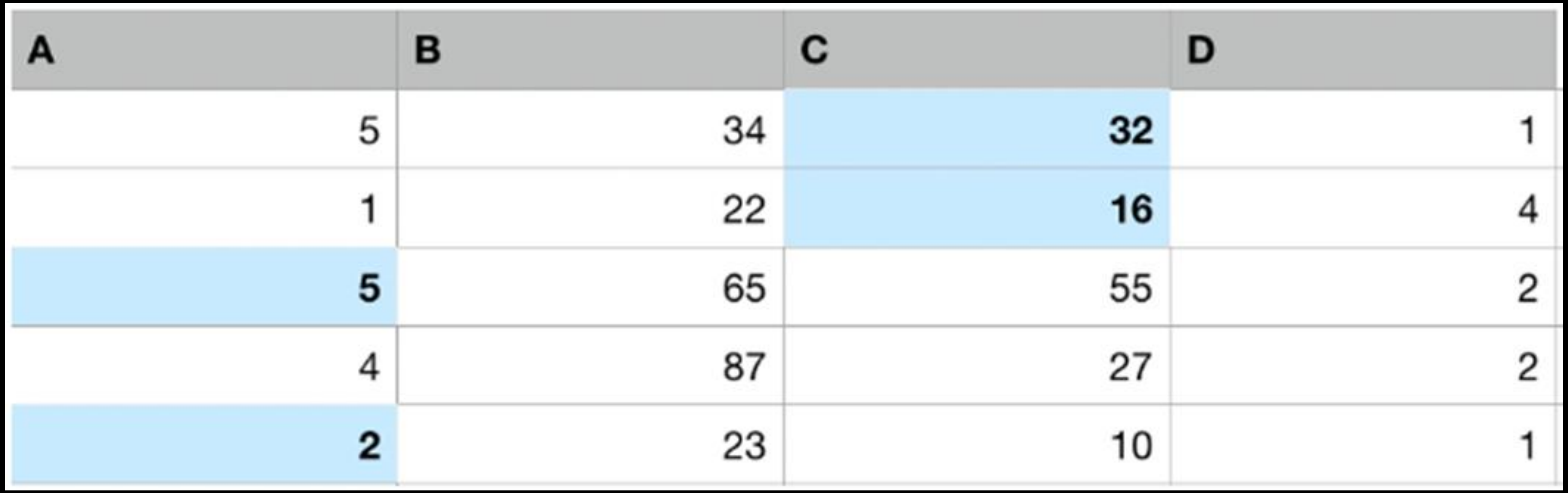
2) Prediction of missing values in C using the imputed values from A



A	B	C	D
5	34	**NA**	1
1	22	**NA**	4
5	65	55	2
4	87	27	2
2	23	10	1

ITERATIVE REGRESSION

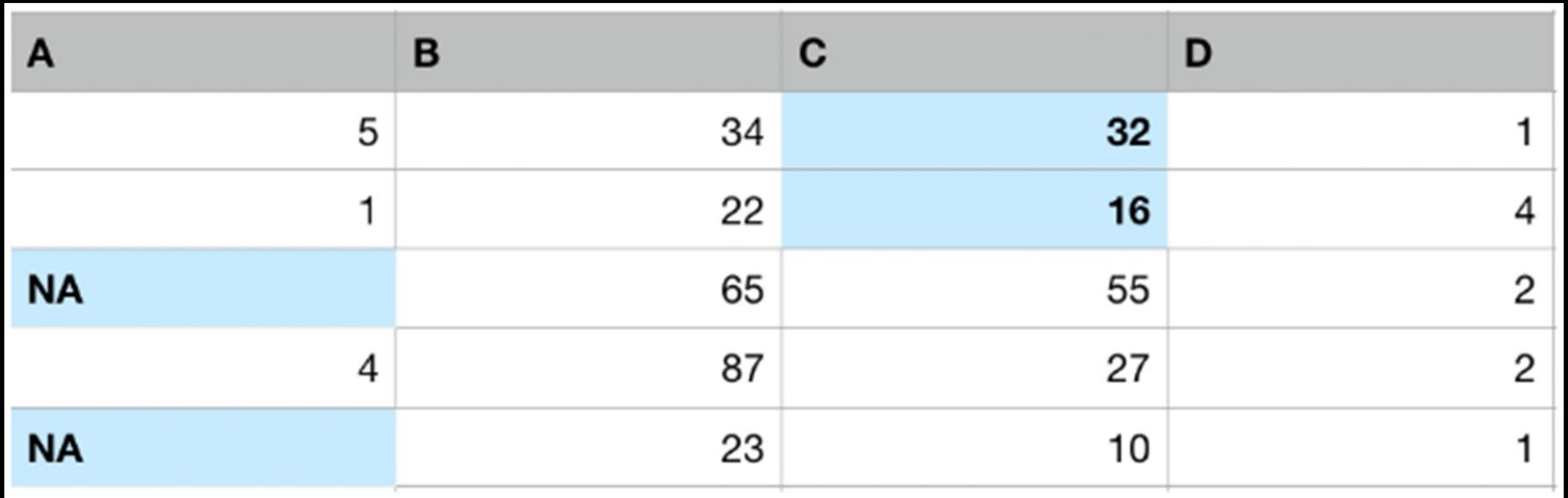
2) Prediction of missing values in C using the imputed values from A



A	B	C	D
5	34	**32**	1
1	22	**16**	4
5	65	55	2
4	87	27	2
2	23	10	1

ITERATIVE REGRESSION

3) **Prediction of missing values in A using the imputed values from C**



A	B	C	D
5	34	**32**	1
1	22	**16**	4
NA	65	55	2
4	87	27	2
NA	23	10	1

➔ **Repeat until no further changes occur**

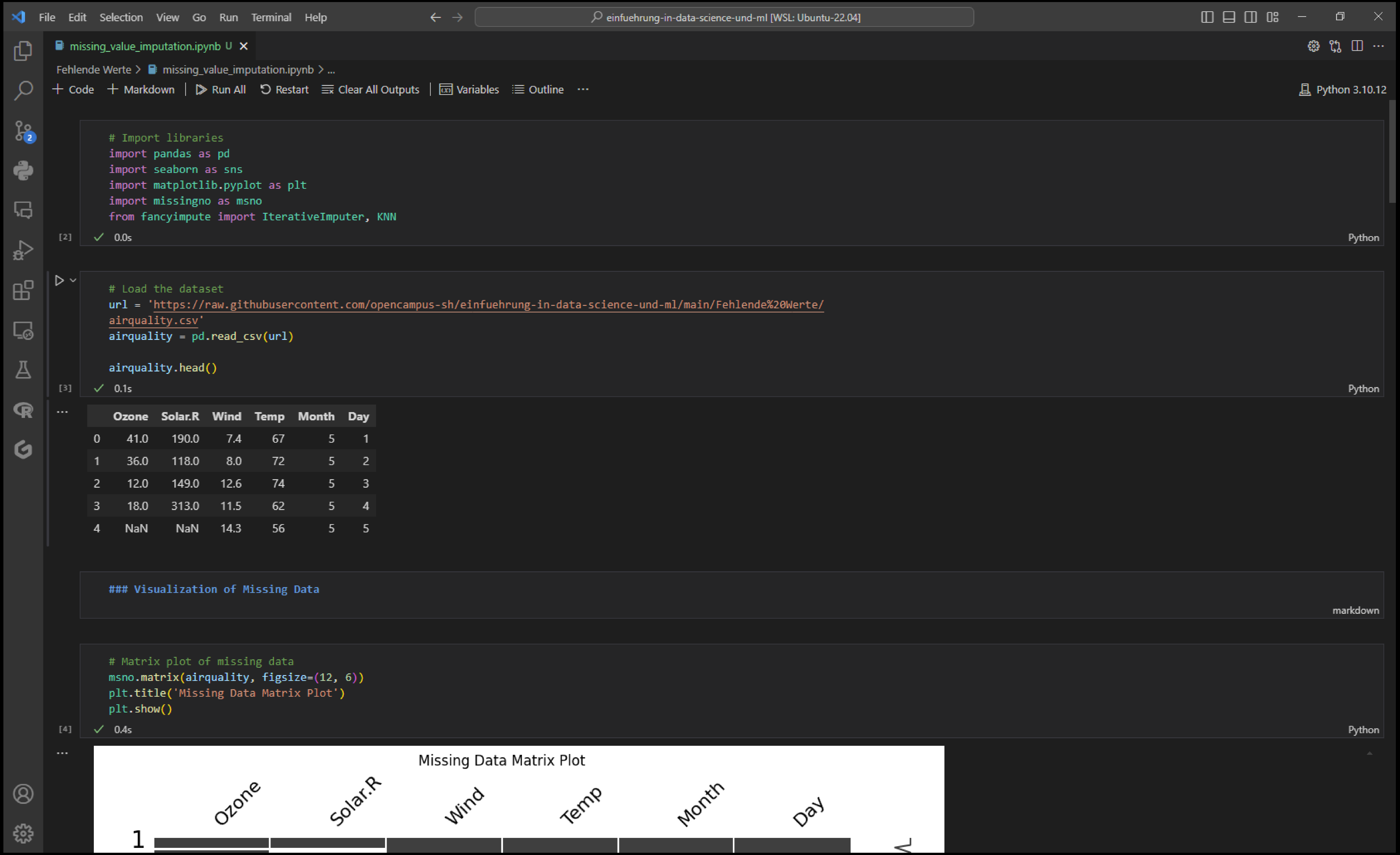
ITERATIVE REGRESSION

1) Go through all variables of the dataset step by step.
2) For each variable, build a regression model based on all other variables.
3) Predict all missing values.

- Now repeat steps 1) to 3) again and re-estimate the missing values—this time using the already imputed missing values.

- Repeat this process until the imputed values no longer change.

IMPUTATION EXAMPLES


File
Edit
Selection
View
Go
Run
Terminal
Help
einfuehrung-in-data-science-und-ml [WSL: Ubuntu-22.04]
missing_value_imputation.ipynb U
Fehlende Werte > missing_value_imputation.ipynb > ...
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
Python 3.10.12

```
# Import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import missingno as msno
from fancyimpute import IterativeImputer, KNN
```

[2] ✓ 0.0s Python

```
# Load the dataset
url = 'https://raw.githubusercontent.com/opencampus-sh/einfuehrung-in-data-science-und-ml/main/Fehlende%20Werte/
airquality.csv'
airquality = pd.read_csv(url)

airquality.head()
```

[3] ✓ 0.1s Python

	Ozone	Solar.R	Wind	Temp	Month	Day
0	41.0	190.0	7.4	67	5	1
1	36.0	118.0	8.0	72	5	2
2	12.0	149.0	12.6	74	5	3
3	18.0	313.0	11.5	62	5	4
4	NaN	NaN	14.3	56	5	5

```
### Visualization of Missing Data
```

markdown

```
# Matrix plot of missing data
msno.matrix(airquality, figsize=(12, 6))
plt.title('Missing Data Matrix Plot')
plt.show()
```

[4] ✓ 0.4s Python

BREAKOUT

- **Perform an initial imputation on your dataset.**
- **Use the code provided in the example notebook to assist you.**

DISCUSSION

- What might be a problem in evaluating a model when parts of the data is imputed?

CALCULATION OF THE IMPUTATION ERROR

1. Creating a complete dataset (“reference dataset”)
2. Randomly removing data
3. Imputing the missing data using the chosen method (possibly several methods for comparison)
4. Comparing the imputed data with the original data, e.g., by calculating
   - the mean squared error (MSE) or
   - the absolute error
5. Evaluating the error (and adjusting the imputation method if needed)

ERROR ANALYSIS

- Analysis of error by product group, day of the week, or similar factors.

- Selection of cases with particularly large prediction errors, such as those
  - exceeding 50% error or
  - the 1% of predictions with the largest errors.

  → Identification of common patterns.

TASKS

- Choose one (or several) methods to replace the missing values in your dataset.

- Divide the tasks well within your team:
  Who will work on data optimization, and who on model optimization?