

Session overview

1	Fri, Sep 6, 2024	VSC, virtual environments
2	Mon, Sep 9, 2024	Version control
3	Tue, Sep 10, 2024	EDA & feature creation in python
4	Thu, Sep 12, 2024	EDA & feature creation in python (continued)
5	Fri, Sep 13, 2024	RStudio, importing and exploring data (EDA) in R
6	Tue, Sep 17, 2024	EDA live challenge

Truths and lies trackers

Alejandro	Delgado Tello
Aleksandr	Smolin
Anastasiia	Chernavskaia
Angad Singh	Sahota
Blanca	Jimenez
Deepak	Malik
Denis	Shadrin
Enzo	Infantes
Ferran	Boada Bergadà
Hannes	Schiemann
Julián	Romero
Lucia	Sauer
Maria	Simakova Mariukha

Maria Jose	Aleman Hernandez
Marta	Sala
Matias	Borrell
Moritz	Peist
Nicolas	Rauth
Noemi	Lucchi
Pablo	Fernández
Simon	Vellin
Soledad	Monge
Tarang	Kadyan
Viktoria	Gagua
Wei	Sun

Sessions 3:

EDA & feature creation in python

DSDM Brushup Course - Coding - September 2024
Margherita Philipp

Truths and lies interlude

Nicely done!

Average

8.91 / 10 points

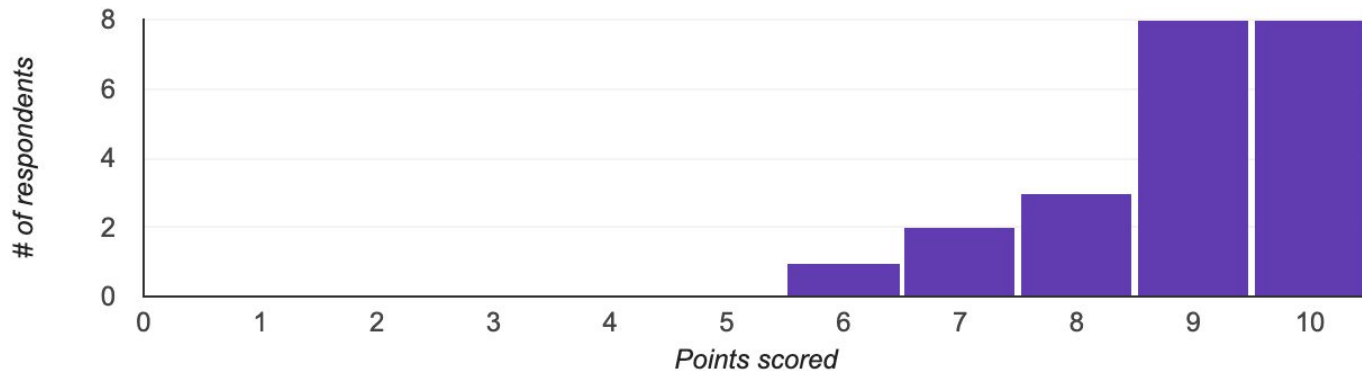
Median

9 / 10 points

Range

6 - 10 points

Total points distribution



Task: EDA 1

1. Create a folder called data under brushup_files
 - Extra: do it via the command line
2. Create a jupyter notebook called EDA_basics (not in data but in the root folder)
3. Import the WB data (can you do it via a relative path)
4. Display the head, check for missing values
5. Find the min and max values - overall and just for 2023
 - Which countries do they belong to?
6. Inspect the values in the “Country Code” column
7. Which county has seen the greatest population growth from the start to the end of the timeline?
 - In absolute terms?
 - In relative terms?
8. Show just the rows for Spain and your country/ countries of origin
 - Can you create a new row that shows the difference in population over time?
 - Can you plot this?
9. Done?
 - See if someone might benefit from your help.
 - Are there any outliers? What could you do with them?
 - Can you plot a histogram and/ or kernel density for the values in 2023?

Sessions 4:

EDA & feature creation in python

DSDM Brushup Course - Coding - September 2024
Margherita Philipp

Truths and lies interlude

The last question had several right answers...

Average

3.64 / 5 points

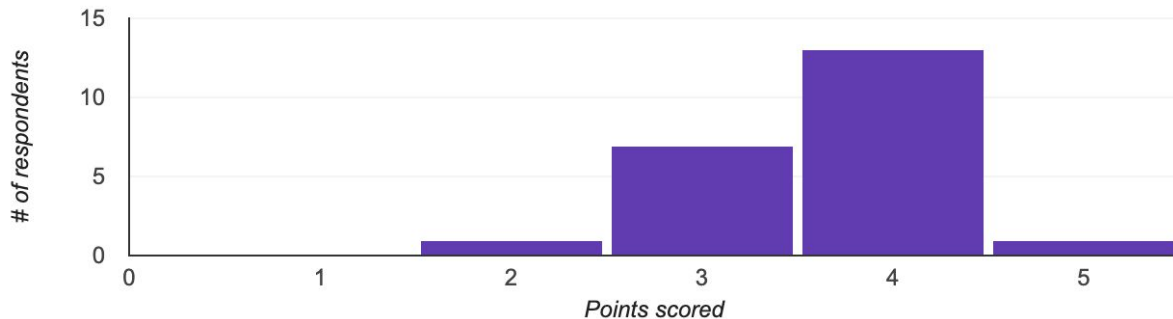
Median

4 / 5 points

Range

2 - 5 points

Total points distribution



How to best learn/ “teach” coding?

- No single best way of doing things
- Get exposure to a variety of approaches
- Read other peoples’ code
- Know what to look for
- Know where to look
- Practice, practice, practice

20m Task: Reading other people's code

1. What are different ways to remove a row from a data frame?
 - Which one is most useful under which circumstances?
2. What does “transposing” a dataframe do?
 - And why is it useful?
3. What are the three most useful new things you learned?
 - Write them down! I'll ask you about them in today's quiz
4. Done?
 - Can you explain how I cleaned the columns in the dataframe?
 - Does someone else need some help reading the notebook?

Text as data

- You'll be doing more on this with Hannes

Regex cheat sheet

Quantifiers

<code>a b</code>	Match either "a" or "b"
<code>?</code>	Match either "a" or "b"
<code>+</code>	One or more
<code>*</code>	Zero or more
<code>*?</code>	Zero or more, but stop after first match
<code>{N}</code>	Exactly N number of times (Where N is number)
<code>{N, M}</code>	From N to M number of times (Where N and M are numbers)

Pattern Collections

<code>[A-Z]</code>	Match any uppercase character from "A" to "Z"
<code>[a-z]</code>	Match any lowercase character from "a" to "z"
<code>[0-9]</code>	Match any number
<code>[asdf]</code>	Match any character that's either "a", "s", "d", or "f"
<code>[^asdf]</code>	Match any character that's not any of the following: "a", "s", "d", or "f"

General Tokens

<code>.</code>	Any character
<code>\n</code>	Newline character
<code>\t</code>	Tab character
<code>\s</code>	Any whitespace character (Including \t, \n, etc)
<code>\S</code>	Any non-whitespace character
<code>\w</code>	Any word character (Upper/lowercase letters, 0-9, _)
<code>\W</code>	Any non-word character
<code>\b</code>	Word boundary (Matches between characters)
<code>\B</code>	Non-word boundary
<code>^</code>	The start of a line
<code>\$</code>	The end of a line
<code>\\</code>	The literal character "\"

Regex cheat sheet

Flags

<code>g</code>	Global, match more than once
<code>m</code>	Force <code>\$</code> and <code>^</code> to match each newline individually
<code>i</code>	Make the regex case-insensitive

Groups

<code>(...)</code>	Capture group (Matches any 3 characters)
<code>(?: ...)</code>	Non-capture group (Matches any 3 characters)
<code>(?<name> ...)</code>	Named capture group Group is called "name"

Named Back Reference

<code>\k<name></code>	Reference named capture group "name" in query
-----------------------------	---

Lookahead and Lookbehind

<code>(?!)</code>	Negative lookahead
<code>(?=)</code>	Positive lookahead
<code>(?<!)</code>	Negative lookbehind
<code>(?<=)</code>	Positive lookbehind





20m Task: Regex & combining data

1. What do these do: `findall(r'...')`, `contains(r'...')`, `count(r'...')`
 - Add a brief description as a comment above each one.
 - Can you try anything additional?
2. Create a new sentence example
 - Try some of the regex things shown.
3. Compare the different ways of combining data.
 - Play around with the “how” parameter of the merge function. What changes? Which option might be appropriate when?
 - Why does the first concatenate function not specify `axis=0`?
 - What needs to be true of the data frames for the two concat options to work?
 - Create two data frames of your own to show what happens when there are columns with the same name in each of the frames?

Moving from exploration to manipulation

- What is a feature?

Sooooooooo many resources

	Real World Data Cleaning in Python Pandas (Step By Step) Ryan & Matt Data Science 72K views · 1 year ago
	How to use the Pandas GroupBy function Pandas tutorial Misra Turp 29K views · 2 years ago
	Complete Python Pandas Data Science Tutorial! (Reading... Keith Galli ✓ 3.1M views · 5 years ago
	Exploratory Data Analysis with Pandas Python Rob Mulla ✓ 467K views · 2 years ago


10m Task: Melting vs pivoting

1. What does melting the data frame do?
 - How have the dimensions (shape) of the df changed?
 - Run `count_values` on the Series Name. What do you notice?
 - What are `id_vars`?
 - What happens to the `value_vars`?
 - Replace `any_name` with a name of your choice
2. What does pivoting the data frame do?
 - Why might you be getting an error message? Can you fix it?
 - What are the index values?
 - How does the function know what to use as columns?
 - What happens if you don't reset the index?

Melt vs pivot

From long to wide:

Property ID	Attribute	Score
House 1	Price	0.8
House 1	Space	0.6
House 1	Location	0.99
House 1	Aesthetics	0.5
House 2	Price	0.4
House 2	Space	0.3
House 2	Location	0.2
House 2	Aesthetics	0.6
House 3	Price	0.8
House 3	Space	0.7
House 3	Location	0.4
House 3	Aesthetics	0.8



Property ID	Price	Space	Location	Aesthetics
House 1	0.8	0.6	0.99	0.5
House 2	0.4	0.3	0.2	0.6
House 3	0.8	0.7	0.4	0.8

Pandas Pivot and Melt

From wide to long:

item_code	1-4-20	2-4-20	3-4-20	4-4-20
A	20	15	15	10
B	10	12	8	15
C	5	6	5	3



item_code	Date	Quantity
A	1-4-20	20
B	1-4-20	10
C	1-4-20	5
A	2-4-20	15
B	2-4-20	12
C	2-4-20	6
A	3-4-20	15
B	3-4-20	8
C	3-4-20	5
A	4-4-20	10
B	4-4-20	15
C	4-4-20	3

Melt vs pivot

From long to wide:

PIVOT



Property ID	Attribute	Score
House 1	Price	0.8
House 1	Space	0.6
House 1	Location	0.99
House 1	Aesthetics	0.5
House 2	Price	0.4
House 2	Space	0.3
House 2	Location	0.2
House 2	Aesthetics	0.6
House 3	Price	0.8
House 3	Space	0.7
House 3	Location	0.4
House 3	Aesthetics	0.8

Property ID	Price	Space	Location	Aesthetics
House 1	0.8	0.6	0.99	0.5
House 2	0.4	0.3	0.2	0.6
House 3	0.8	0.7	0.4	0.8



Pandas Pivot and Melt

From wide to long:

MELT



item_code	1-4-20	2-4-20	3-4-20	4-4-20
A	20	15	15	10
B	10	12	8	15
C	5	6	5	3

item_code	Date	Quantity
A	1-4-20	20
B	1-4-20	10
C	1-4-20	5
A	2-4-20	15
B	2-4-20	12
C	2-4-20	6
A	3-4-20	15
B	3-4-20	8
C	3-4-20	5
A	4-4-20	10
B	4-4-20	15
C	4-4-20	3



30m Task: Missingness, groupby and features

1. Renaming

- Use a dictionary to give the columns better names.
- Set the country code as the index.

2. Missingness

- What do you notice about missing values?
- Can you insert np.nan values?
- If you remove all missing values: do you still have a complete country-year of data for each country?
- Keep only the latest year of data for each country without missing values

3. Can you use groupby to find the min, max, median and mean GNI for each country over time?

- You might want to run .info to inspect data types...

4. Create some features for each country-year

- Total number of school-aged children
- Share of school-aged children relative to the population

5. Done?


- Keep only the latest year of data: Which countries have the least up to date data?
- Use map and/ or apply to create additional features, e.g. a categorical variable of income group
- Plot the relationships between GNI and population, GNI per capita and the share of children out of primary school

- `'Population, total'`
- `'Children out of school, primary'`
- `'Children out of school (% of primary school age)'`
- `'GNI, Atlas method (current US$)'`
- `'GNI per capita, Atlas method (current US$)'`

For tomorrow

1. Be able to open R-file (preferably RStudio)
2. Try mounting Google collab (see demo notebook shared)
3. Push your new files to your github (submit screenshot by end of Sunday)
4. Explore the two notebooks and the .py file

Google Colab

 Collab demo.ipynb ☆

File Edit View Insert Runtime Tools Help

+ Code + Text

1s

[1] import pandas as pd

23s

from google.colab import drive

drive.mount('/content/drive', force_remount=True)

Mounted at /content/drive

+ Code + Text

1s

[3] df = pd.read_csv('/content/drive/MyDrive/EconAI_private/Teaching/Brushup_DSDM/data/WB_pop_clean.csv')

#df = pd.read_csv('/content/drive/MyDrive/WB_pop_clean.csv')

0s

df.head()

	Series Name	Series Code	Country Name	Country Code	2001	2002	2003	2011	2012	2013	2021	2022	2023
0	Population, total	SP.POP.TOTL	Afghanistan	AFG	19688632	21000256	22645130	29249157	30466479	31541209	40099462	41128771	42239854
1	Population, total	SP.POP.TOTL	Albania	ALB	3060173	3051010	3039616	2905195	2900401	2895092	2811666	2777689	2745972
2	Population, total	SP.POP.TOTL	Algeria	DZA	31200985	31624696	32055883	36543541	37260563	38000626	44177969	44903225	45606480
3	Population, total	SP.POP.TOTL	American Samoa	ASM	58324	58177	57941	54310	53691	52995	45035	44273	43914
4	Population, total	SP.POP.TOTL	Andorra	AND	67820	70849	73907	70567	71013	71367	79034	79824	80088

Next steps: [Generate code with df](#) [View recommended plots](#) [New interactive sheet](#)

Using GitHub desktop

The screenshot shows the GitHub web interface for the repository 'MargheritaPhilipp / brushup'. The repository is public and has 1 branch and 0 tags. The 'Code' dropdown menu is open, showing options for cloning the repository. The 'Local' option is selected, and the 'Clone' button is highlighted. The 'Clone' button is located in the 'Local' section of the dropdown menu. The 'Clone' button is located in the 'Local' section of the dropdown menu. The 'Clone' button is located in the 'Local' section of the dropdown menu.

Code

Issues Pull requests Actions Projects Wiki Security Insights Settings

brushup Public

main 1 Branch 0 Tags

Go to file

Add file Code

Local Codespaces

Clone

HTTPS SSH GitHub CLI

https://github.com/MargheritaPhilipp/brushup

Clone using the web URL.

Open with GitHub Desktop

Download ZIP

MargheritaPhilipp Initial commit

.gitignore Initial commit

LICENSE Initial commit

README MIT license

Add a README

Help people interested in this repository understand your project by adding a README.

About

No description, website, or topics provided.

Activity

0 stars

1 watching

0 forks

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package