Introduction to Data Science

Data Wrangling Continued - Multiple Data Frames

Joanna Bieri DATA101

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 Click Here for Joanna's Schedule

Announcements

Exam 1 will be handed out Next Class

Please come to office hours to get help!

Day 7 Assignment - same drill.

Join or Merge

Here is some Fake Data:

	id	data1
0	1	x1
1	2	x2
2	3	x3

	id	data1
0	1	y1
1	2	y2
2	4	y4

Grammar of the Merge/Join

```
pd.merge(data frame 1,
         data frame 2,
         on='column name',
         how=type of merge )
```

Left Join - how="left"

GIF - LEFT JOIN

```
pd.merge(DF_fake1, DF_fake2, on='id',how='left')
```

	id	data1_x	data1_y
0	1	x1	y1
1	2	x2	y2
2	3	x3	NaN

Right Join - how="right"

GIF - RIGHT JOIN

	id	data1_x	data1_y
0	1	x1	y1
1	2	x2	y2
2	4	NaN	y4

Outer Join - how="outer"

This type of join, also called a **full join**.

GIF - OUTER JOIN

	id	data1_x	data1_y
0	1	x1	y1
1	2	x2	y2
2	3	x3	NaN
3	4	NaN	y4

Inner Join - how="inner"

GIF - INNER JOIN

	id	data1_x	data1_y
0	1	x1	y1
1	2	x2	y2

Join and keep an indicator

```
pd.merge(DF_fake1,
         DF_fake2,
         on='id',
         how='outer',
         indicator=True)
```

	id	data1_x	data1_y	_merge
0	1	x1	y1	both
1	2	x2	y2	both
2	3	x3	NaN	left_only
3	4	NaN	y4	right_only

Three Data Sets - Dates:

	name	birth_year	death_year
0	Janaki Ammal	1897	1984.0
1	Chien-Shiung Wu	1912	1997.0
2	Katherine Johnson	1918	2020.0
3	Rosalind Franklin	1920	1958.0
4	Vera Rubin	1928	2016.0
5	Gladys West	1930	NaN
6	Flossie Wong-Staal	1947	NaN
7	Jennifer Doudna	1964	NaN

Three Data Sets - Profession:

	name	profession
0	Ada Lovelace	Mathematician
1	Marie Curie	Physicist and Chemist
2	Janaki Ammal	Botanist
3	Chien-Shiung Wu	Physicist
4	Katherine Johnson	Mathematician
5	Rosalind Franklin	Chemist
6	Vera Rubin	Astronomer
7	Gladys West	Mathematician
8	Flossie Wong-Staal	Virologist and Molecular Biologist
9	Jennifer Doudna	Biochemist

Three Data Sets - Work:

	name	known_for
0	Ada Lovelace	first computer algorithm
1	Marie Curie	theory of radioactivity, discovery of element...
2	Janaki Ammal	hybrid species, biodiversity protection
3	Chien-Shiung Wu	confim and refine theory of radioactive beta d...
4	Katherine Johnson	calculations of orbital mechanics critical to ...
5	Vera Rubin	existence of dark matter
6	Gladys West	mathematical modeling of the shape of the Eart...
7	Flossie Wong-Staal	first scientist to clone HIV and create a map ...
8	Jennifer Doudna	one of the primary developers of CRISPR, a gro...

Combining Three Data Sets

```
DF_scientists = pd.merge(DF_professions,
                         DF_works,
                         on='name',
                         how='left')
DF_scientists = pd.merge(DF_scientists,
                         DF_dates,
                         on='name',
                         how='left')
DF_scientists
```

Combining Three Data Sets

	name	profession	known_for
0	Ada Lovelace	Mathematician	first computer a
1	Marie Curie	Physicist and Chemist	theory of radioa
2	Janaki Ammal	Botanist	hybrid species, l
3	Chien-Shiung Wu	Physicist	confim and refir
4	Katherine Johnson	Mathematician	calculations of c
5	Rosalind Franklin	Chemist	NaN
6	Vera Rubin	Astronomer	existence of dar
7	Gladys West	Mathematician	mathematical m
8	Flossie Wong-Staal	Virologist and Molecular Biologist	first scientist to
9	Jennifer Doudna	Biochemist	one of the prima

Practice Exam

The homework for today includes a practice exam. Some things you will notice:

1. I do not give code in the practice exam, but you can copy and paste the code from your other assignments or the lecture notes.
2. The exam will be open notes, open book, open basic internet search - but you should not use AI to generate your answer. It is usually very easy to see when code has been generated by AI - it will have commands we have not learned or a style of programming that is clearly AI. YOU MUST UNDERSTAND ALL THE CODE YOU SUBMIT!
3. The questions get progressively more involved. Just do one at a time.
4. I expect you to explain your results. I should not have to interpret what your numbers or code outputs mean. Add Markdown cells to describe what YOU see in the results.