RICE | D2K
Data to Knowledge Lab

Welcome to DSCI 101

Introduction to Data Science

Week 4 Recap

- Pandas: tabular data in Python

- Pandas: data manipulation and basic operations
  - Import data and create dataframe
  - simple utility operations
  - indexing with df[], df.loc[] and df.iloc[]
  - indexing with Boolean: filtering
  - df.sort_values and df.sort_index

Week 5 Preview

- Data summaries:
  - df.groupby and df.pivot_table: summarize categorical variables
  - df.describe: summarize numerical variables
- Merge two dataframes
- Missing data

Review on variable types

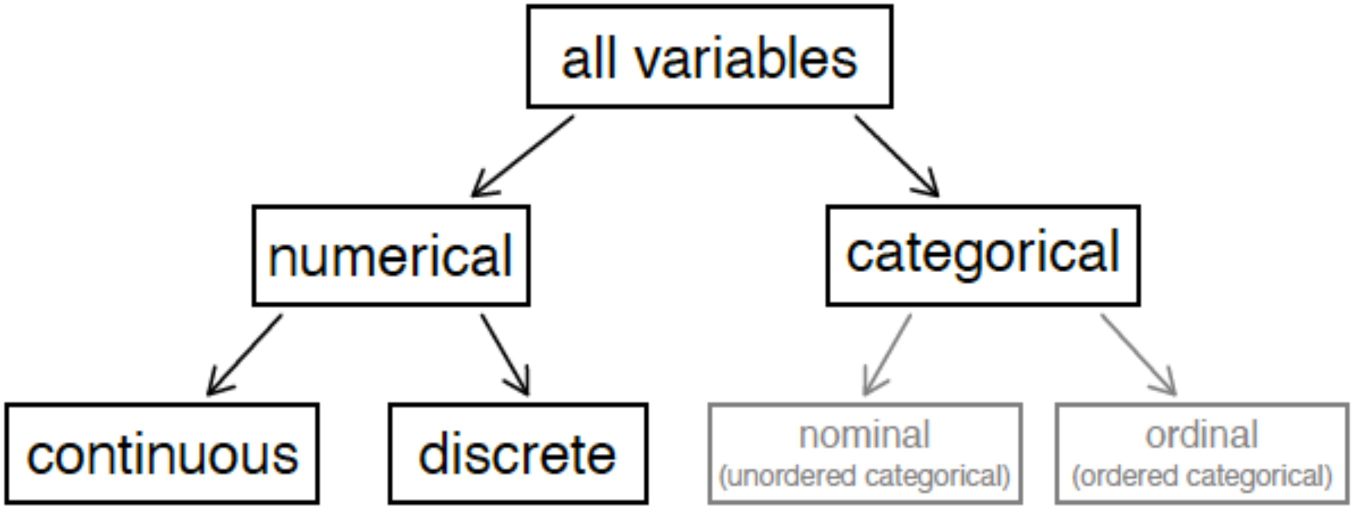


Figure 1.7: Breakdown of variables into their respective types.

Numerical variable

- Summary statistics by ***df.describe()***:
  - Count
  - Mean
  - Std
  - Min
  - 25%
  - 50%
  - 75%
  - Max
- Review definitions on descriptive statistics

Categorial variable

- Distribution of a categorical variable:
    - How many categories are there?
    - How many count for each category?

- ***Series.value_counts()***
    - ***df.column_name.value_counts()*** or
    - ***df[column_name].value_counts()***

Categorial and numerical variables

- A categorical variable has several categories
  - define subgroups of population

- Analysis on each subgroup
  - with respect to one or more numerical variables
  - compare summary statistic among subgroups

- Aggregate the numerical variable for each subgroups
  - sum, mean, min, max, std, …

Group by

- Information best summarized by ***df.groupby***
  - group the rows that belonging to one category together

- Counting how many records belonging to each category
  - ***df.groupby(cat_column_name).count()*** same as
  - ***df.column_name.value_counts()*** or
  - ***df[column_name].value_counts()***

- Look at other variable values for each category
  - need an aggregate function
  - ***df.groupby(cat_column_name)[[num_column_name]].aggfunc()***

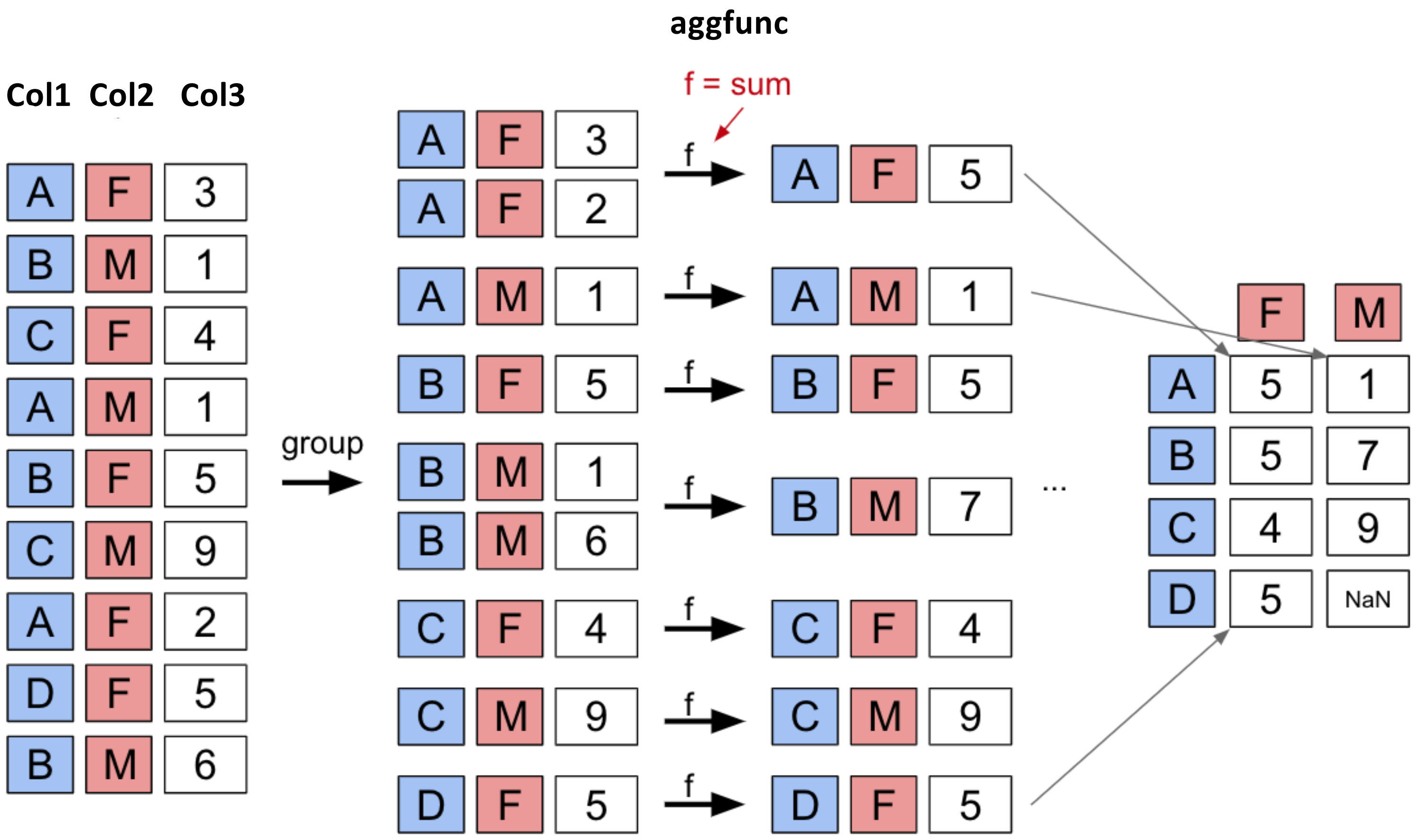
Pivot tables

- A visual way to summarize information for 3 variables
  - group by 2 categorical variables: col1, col2
  - summarize 1 numerical variable: col3

- Group rows into categories based on col1 and col2, summarize col3 by aggfunc.

- ***df.pivot_table(columns=[“col1”], index=[“col2"], values=“col3", aggfunc="mean")***

Pivot table



aggfunc

f = sum

Col1	Col2	Col3
A	F	3
B	M	1
C	F	4
A	M	1
B	F	5
C	M	9
A	F	2
D	F	5
B	M	6

group

A	F	3
A	F	2
A	M	1
B	F	5
B	M	1
B	M	6
C	F	4
C	M	9
D	F	5

f

A	F	5
A	M	1
B	F	5
B	M	7
C	F	4
C	M	9
D	F	5

...

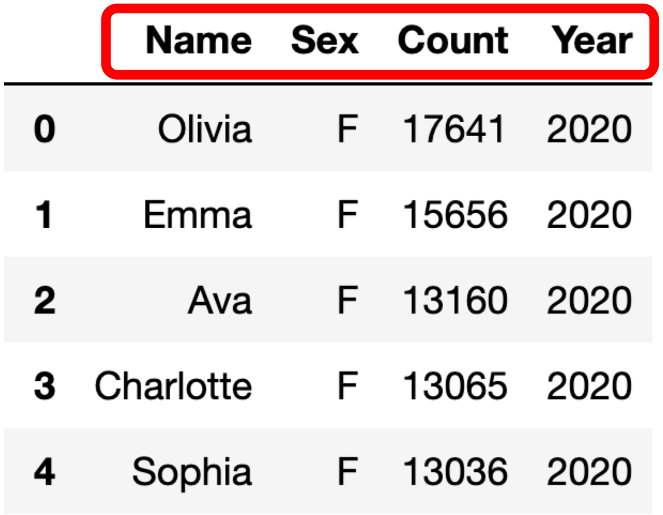
	F	M
A	5	1
B	5	7
C	4	9
D	5	NaN

More Pandas: merge two dataframes

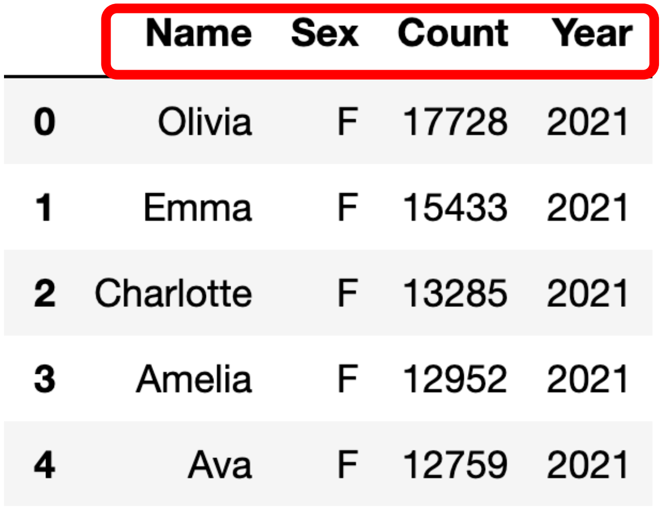
- It would be nice if we have a giant data table with everything we need, but….

- ***df1.merge(df2, how=“ ”, left_on=“ ”, right_on=“ ”)***
  - how: inner, outer, left, right
  - on: column_with_same_name from df1 and df2
  - left_on: column_x_from_df1
  - right_on: column_y_from_df2

df.merge v.s. df.concat

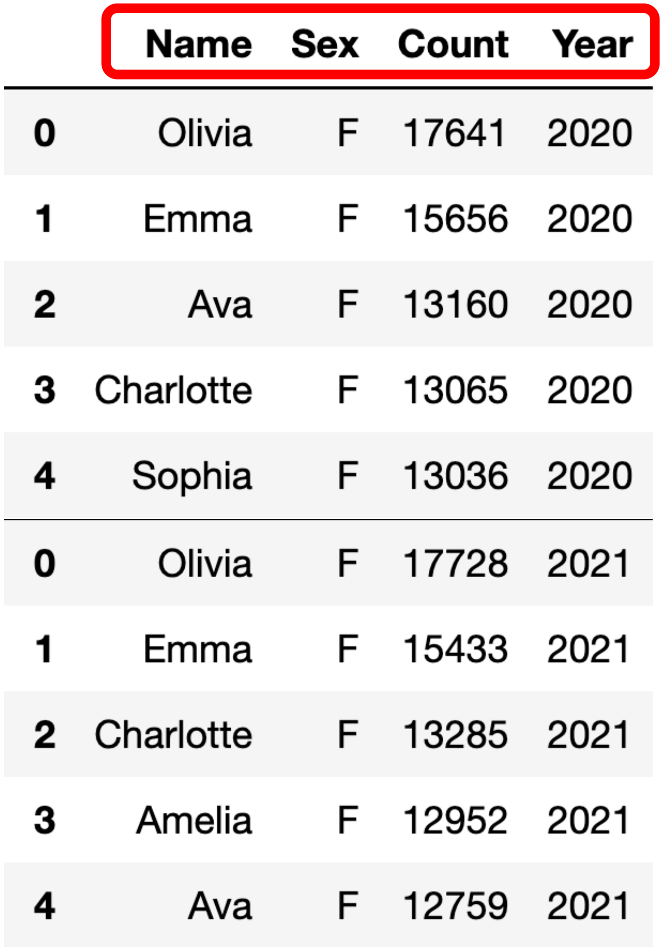


	Name	Sex	Count	Year
0	Olivia	F	17641	2020
1	Emma	F	15656	2020
2	Ava	F	13160	2020
3	Charlotte	F	13065	2020
4	Sophia	F	13036	2020

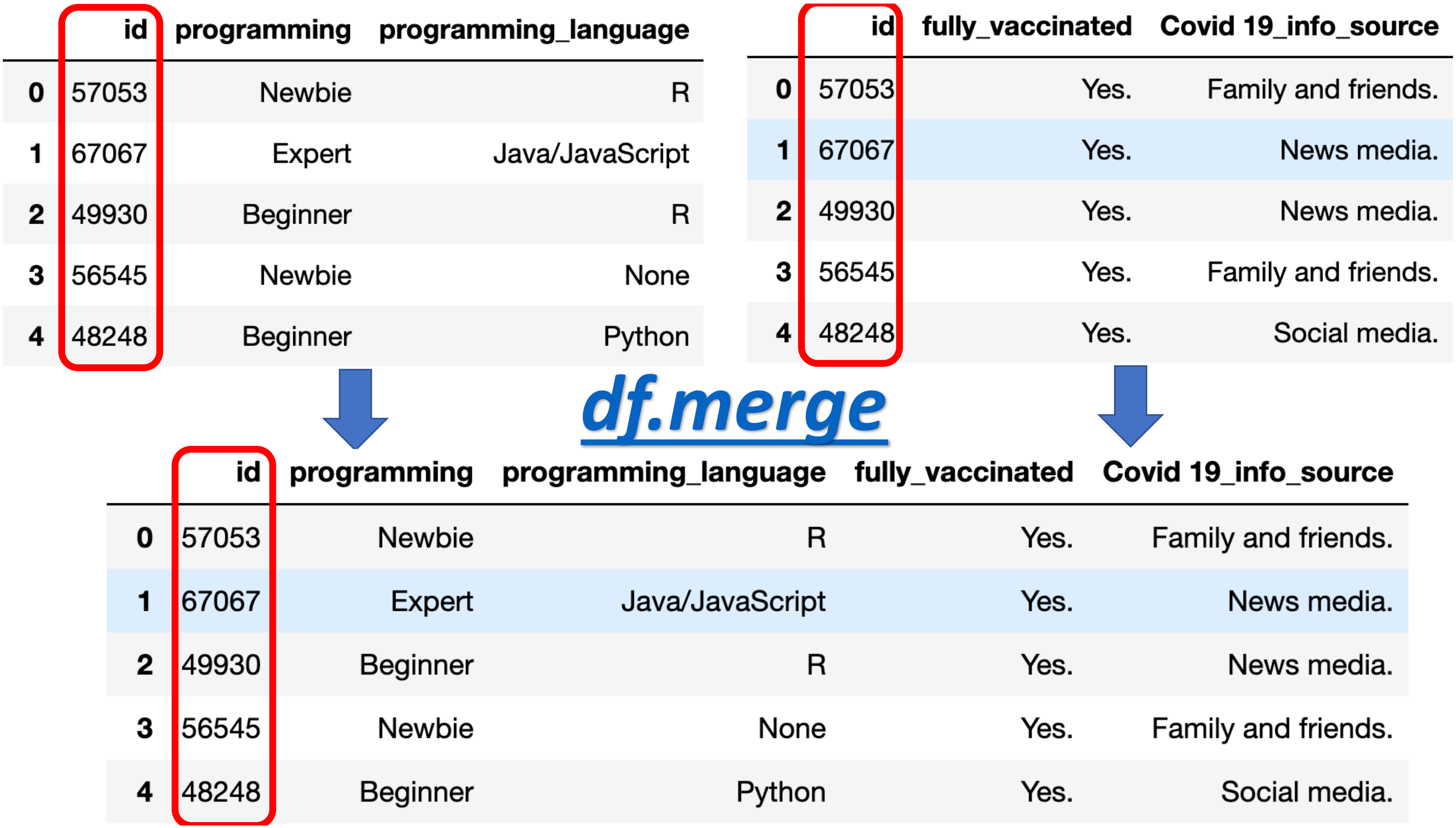


	Name	Sex	Count	Year
0	Olivia	F	17728	2021
1	Emma	F	15433	2021
2	Charlotte	F	13285	2021
3	Amelia	F	12952	2021
4	Ava	F	12759	2021

concatenate



	Name	Sex	Count	Year
0	Olivia	F	17641	2020
1	Emma	F	15656	2020
2	Ava	F	13160	2020
3	Charlotte	F	13065	2020
4	Sophia	F	13036	2020
0	Olivia	F	17728	2021
1	Emma	F	15433	2021
2	Charlotte	F	13285	2021
3	Amelia	F	12952	2021
4	Ava	F	12759	2021

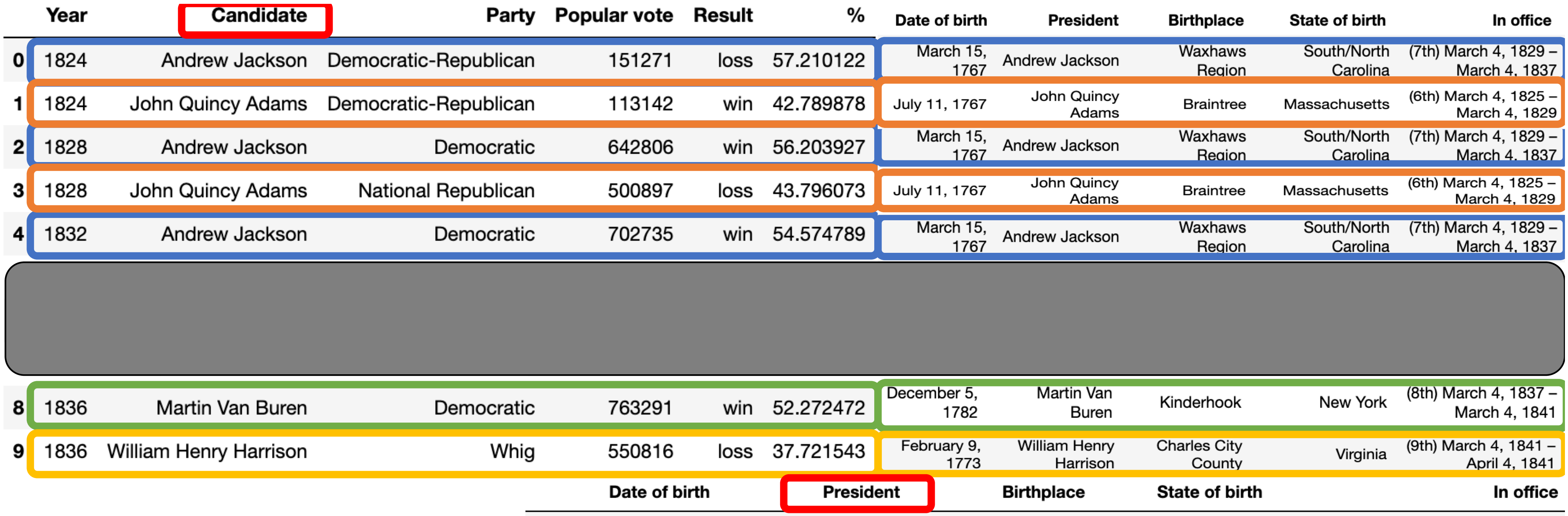


	id	programming	programming_language
0	57053	Newbie	R
1	67067	Expert	Java/JavaScript
2	49930	Beginner	R
3	56545	Newbie	None
4	48248	Beginner	Python

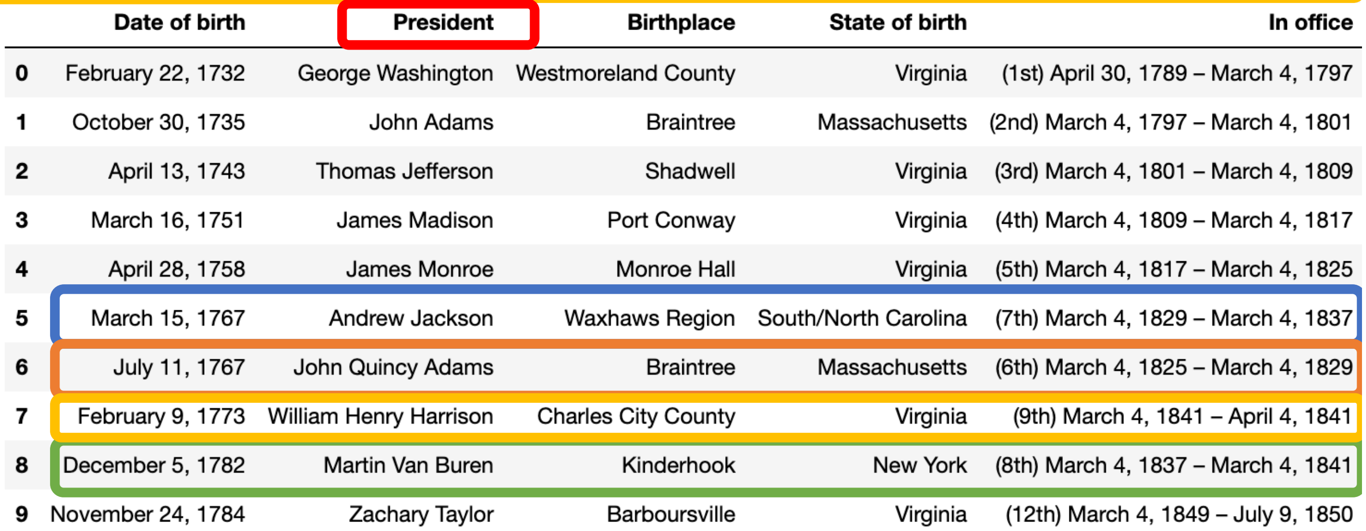
	id	fully_vaccinated	Covid 19_info_source
0	57053	Yes.	Family and friends.
1	67067	Yes.	News media.
2	49930	Yes.	News media.
3	56545	Yes.	Family and friends.
4	48248	Yes.	Social media.

df.merge

	id	programming	programming_language	fully_vaccinated	Covid 19_info_source
0	57053	Newbie	R	Yes.	Family and friends.
1	67067	Expert	Java/JavaScript	Yes.	News media.
2	49930	Beginner	R	Yes.	News media.
3	56545	Newbie	None	Yes.	Family and friends.
4	48248	Beginner	Python	Yes.	Social media.



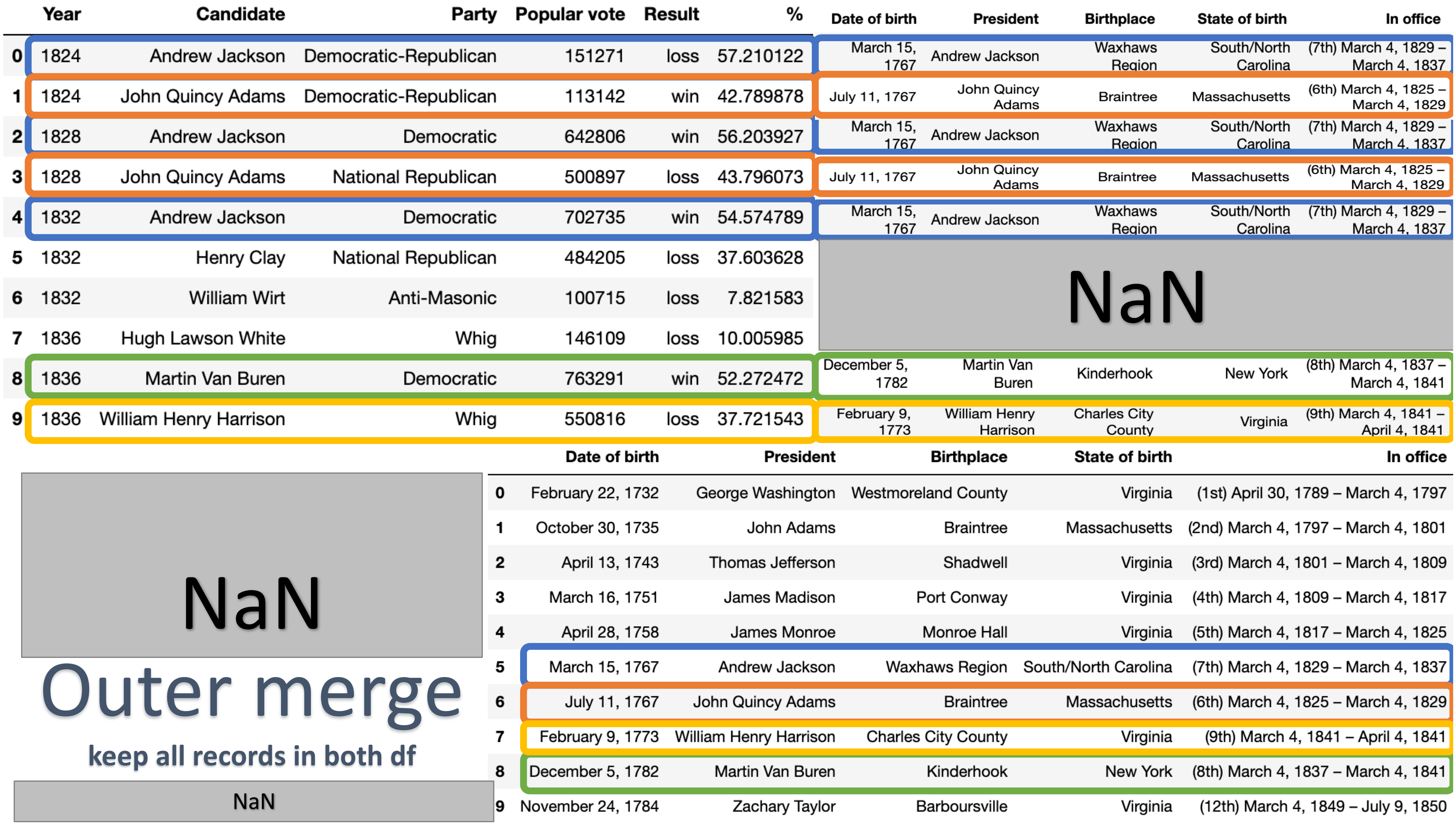
	Year	Candidate	Party	Popular vote	Result	%	Date of birth	President	Birthplace	State of birth	In office
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
2	1828	Andrew Jackson	Democratic	642806	win	56.203927	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
4	1832	Andrew Jackson	Democratic	702735	win	54.574789	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
8	1836	Martin Van Buren	Democratic	763291	win	52.272472	December 5, 1782	Martin Van Buren	Kinderhook	New York	(8th) March 4, 1837 – March 4, 1841
9	1836	William Henry Harrison	Whig	550816	loss	37.721543	February 9, 1773	William Henry Harrison	Charles City County	Virginia	(9th) March 4, 1841 – April 4, 1841



	Date of birth	President	Birthplace	State of birth	In office
0	February 22, 1732	George Washington	Westmoreland County	Virginia	(1st) April 30, 1789 – March 4, 1797
1	October 30, 1735	John Adams	Braintree	Massachusetts	(2nd) March 4, 1797 – March 4, 1801
2	April 13, 1743	Thomas Jefferson	Shadwell	Virginia	(3rd) March 4, 1801 – March 4, 1809
3	March 16, 1751	James Madison	Port Conway	Virginia	(4th) March 4, 1809 – March 4, 1817
4	April 28, 1758	James Monroe	Monroe Hall	Virginia	(5th) March 4, 1817 – March 4, 1825
5	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
6	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
7	February 9, 1773	William Henry Harrison	Charles City County	Virginia	(9th) March 4, 1841 – April 4, 1841
8	December 5, 1782	Martin Van Buren	Kinderhook	New York	(8th) March 4, 1837 – March 4, 1841
9	November 24, 1784	Zachary Taylor	Barboursville	Virginia	(12th) March 4, 1849 – July 9, 1850

Inner merge

Only keep matching records

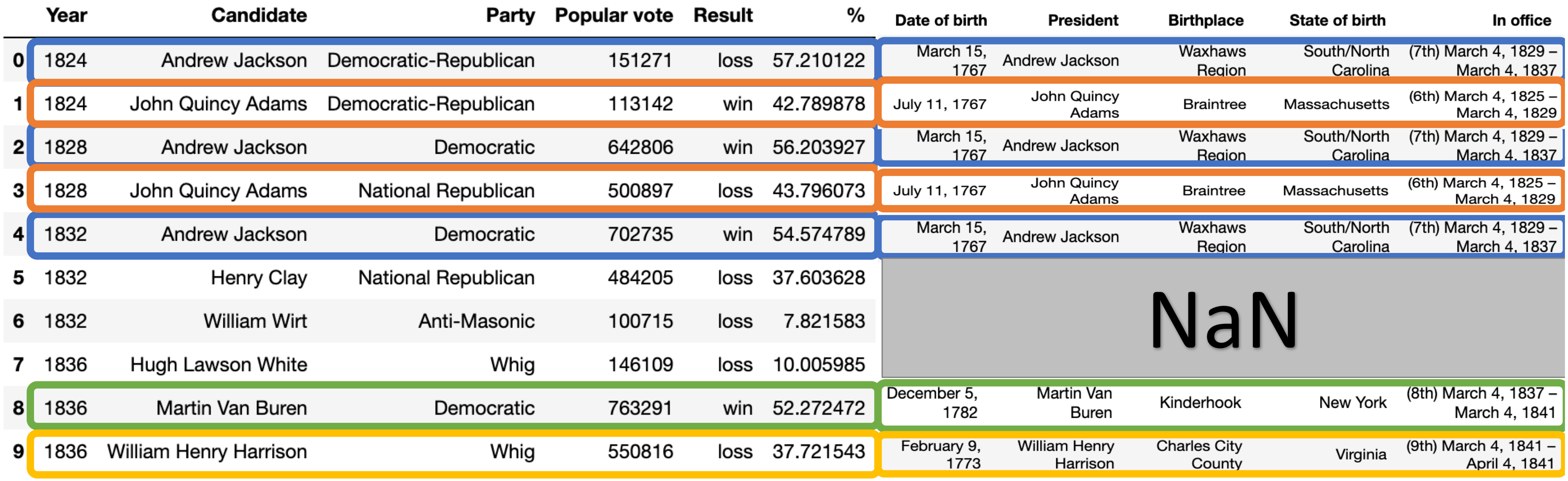


	Year	Candidate	Party	Popular vote	Result	%	Date of birth	President	Birthplace	State of birth	In office
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
2	1828	Andrew Jackson	Democratic	642806	win	56.203927	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
4	1832	Andrew Jackson	Democratic	702735	win	54.574789	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
5	1832	Henry Clay	National Republican	484205	loss	37.603628	NaN				
6	1832	William Wirt	Anti-Masonic	100715	loss	7.821583					
7	1836	Hugh Lawson White	Whig	146109	loss	10.005985					
8	1836	Martin Van Buren	Democratic	763291	win	52.272472	December 5, 1782	Martin Van Buren	Kinderhook	New York	(8th) March 4, 1837 – March 4, 1841
9	1836	William Henry Harrison	Whig	550816	loss	37.721543	February 9, 1773	William Henry Harrison	Charles City County	Virginia	(9th) March 4, 1841 – April 4, 1841

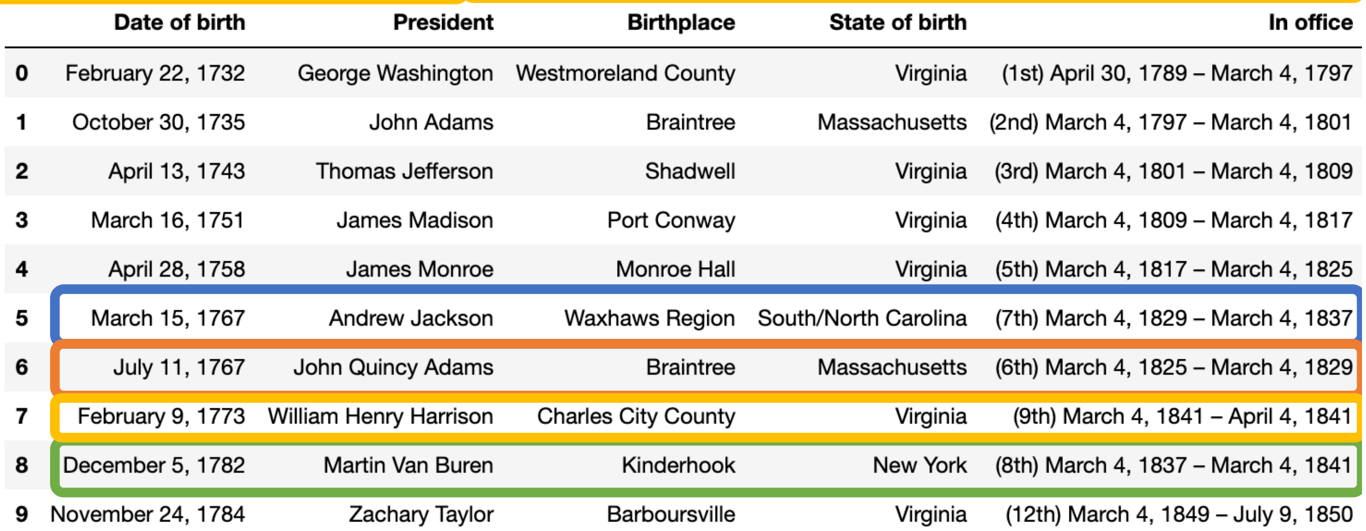
		Date of birth	President	Birthplace	State of birth	In office
NaN	0	February 22, 1732	George Washington	Westmoreland County	Virginia	(1st) April 30, 1789 – March 4, 1797
	1	October 30, 1735	John Adams	Braintree	Massachusetts	(2nd) March 4, 1797 – March 4, 1801
	2	April 13, 1743	Thomas Jefferson	Shadwell	Virginia	(3rd) March 4, 1801 – March 4, 1809
	3	March 16, 1751	James Madison	Port Conway	Virginia	(4th) March 4, 1809 – March 4, 1817
	4	April 28, 1758	James Monroe	Monroe Hall	Virginia	(5th) March 4, 1817 – March 4, 1825
	5	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
	6	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
	7	February 9, 1773	William Henry Harrison	Charles City County	Virginia	(9th) March 4, 1841 – April 4, 1841
	8	December 5, 1782	Martin Van Buren	Kinderhook	New York	(8th) March 4, 1837 – March 4, 1841
NaN	9	November 24, 1784	Zachary Taylor	Barboursville	Virginia	(12th) March 4, 1849 – July 9, 1850

Outer merge

keep all records in both df



	Year	Candidate	Party	Popular vote	Result	%	Date of birth	President	Birthplace	State of birth	In office
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
2	1828	Andrew Jackson	Democratic	642806	win	56.203927	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
4	1832	Andrew Jackson	Democratic	702735	win	54.574789	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
5	1832	Henry Clay	National Republican	484205	loss	37.603628	NaN				
6	1832	William Wirt	Anti-Masonic	100715	loss	7.821583					
7	1836	Hugh Lawson White	Whig	146109	loss	10.005985					
8	1836	Martin Van Buren	Democratic	763291	win	52.272472	December 5, 1782	Martin Van Buren	Kinderhook	New York	(8th) March 4, 1837 – March 4, 1841
9	1836	William Henry Harrison	Whig	550816	loss	37.721543	February 9, 1773	William Henry Harrison	Charles City County	Virginia	(9th) March 4, 1841 – April 4, 1841



	Date of birth	President	Birthplace	State of birth	In office
0	February 22, 1732	George Washington	Westmoreland County	Virginia	(1st) April 30, 1789 – March 4, 1797
1	October 30, 1735	John Adams	Braintree	Massachusetts	(2nd) March 4, 1797 – March 4, 1801
2	April 13, 1743	Thomas Jefferson	Shadwell	Virginia	(3rd) March 4, 1801 – March 4, 1809
3	March 16, 1751	James Madison	Port Conway	Virginia	(4th) March 4, 1809 – March 4, 1817
4	April 28, 1758	James Monroe	Monroe Hall	Virginia	(5th) March 4, 1817 – March 4, 1825
5	March 15, 1767	Andrew Jackson	Waxhaws Region	South/North Carolina	(7th) March 4, 1829 – March 4, 1837
6	July 11, 1767	John Quincy Adams	Braintree	Massachusetts	(6th) March 4, 1825 – March 4, 1829
7	February 9, 1773	William Henry Harrison	Charles City County	Virginia	(9th) March 4, 1841 – April 4, 1841
8	December 5, 1782	Martin Van Buren	Kinderhook	New York	(8th) March 4, 1837 – March 4, 1841
9	November 24, 1784	Zachary Taylor	Barboursville	Virginia	(12th) March 4, 1849 – July 9, 1850

Left merge

keep all records in left df

Missing data

- Missing data is a very difficult problem!
  - missing completely at random: independent to all variables
  - missing at random: depends on other variables
  - missing NOT at random: depends on other values of the missing data variable

- How to check missing assumptions?
  - explore missing data pattern
  - can you predict “missing” v.s “not missing” based on other variables?
  - usually cannot prove missing completely at random

Identify missing values

- Nice missing values: np.nan
  - ***df.isna()***: return a df of Booleans
  - ***df.isna().any()***: return a Boolean for each column
  - ***df.isna().sum()***: number of missing values in each column

- Annoying missing values: “ “, “?”, “—”, *&^%$#@~......
  - ***missing_values = [“?”, “—”]***
    df = pd.read_csv(“data.csv”, na_values=missing_values)
  - ***df.replace({“?”: np.nan, “—”: np.nan}, inplace=True)***

What to do with missing values?

- Drop the records with missing values
  - most common
  - use with caution

- Imputation: inferring missing values
  - mean imputation: replace with mean / median / mode
  - hot deck imputation: replace with a random value

- more advanced imputation:
  - “predict” missing values based on other variables
  - model based

Handle missing values

- ***df.dropna(axis= , how= , inplace=)***
  - axis: 0 means drop by row, 1 means drop by column
  - how: “any” means drop if one missing, “all” means drop only if all missing
  - inplace: True means make changes in the original df

- ***df.fillna(value_to_replace_NA)***
  - ***df.fillna(0)***
  - ***col_mean = df[“column1”].mean()***
  - ***df[“column1”] = df[“column1”].fillna(col_mean)***