NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Lecture 8

Visualization I – matplotlib and seaborn

Reminders

- Survey 2
  - Monday October 07

Reminders

- Survey 2
- Homework 2
  - Friday October 04

Reminders

- Survey 2
- Homework 2
- Project 1
  - Sunday October 20

Reminders

- Survey 2
- Homework 2
- Project 1
- Forum
    - General
    - Lecture

Reminders

- Survey 2
- Homework 2
- Project 1
- Forum
- Final Exam
    - 6-8pm on Monday December 16

Agenda

- Review
  - Formats, Grouping, Joining

Agenda

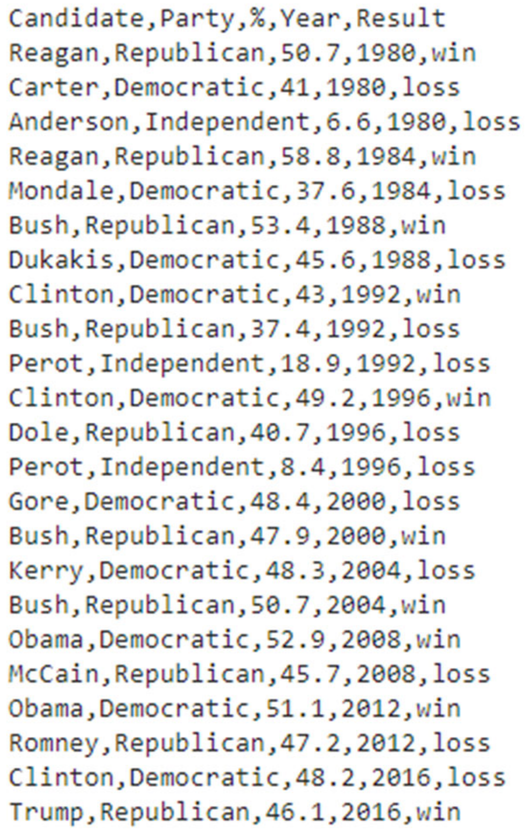
- Review
- Lesson
  - Plotting Categorical Data

Agenda

- Review
- Lesson
- Demo
    - Bar Charts

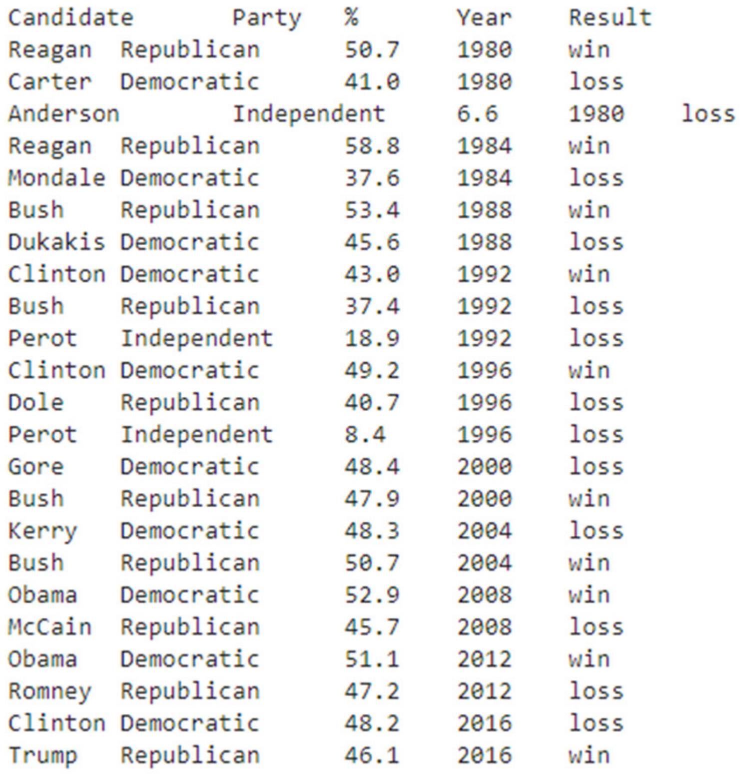
Flat Files

csv



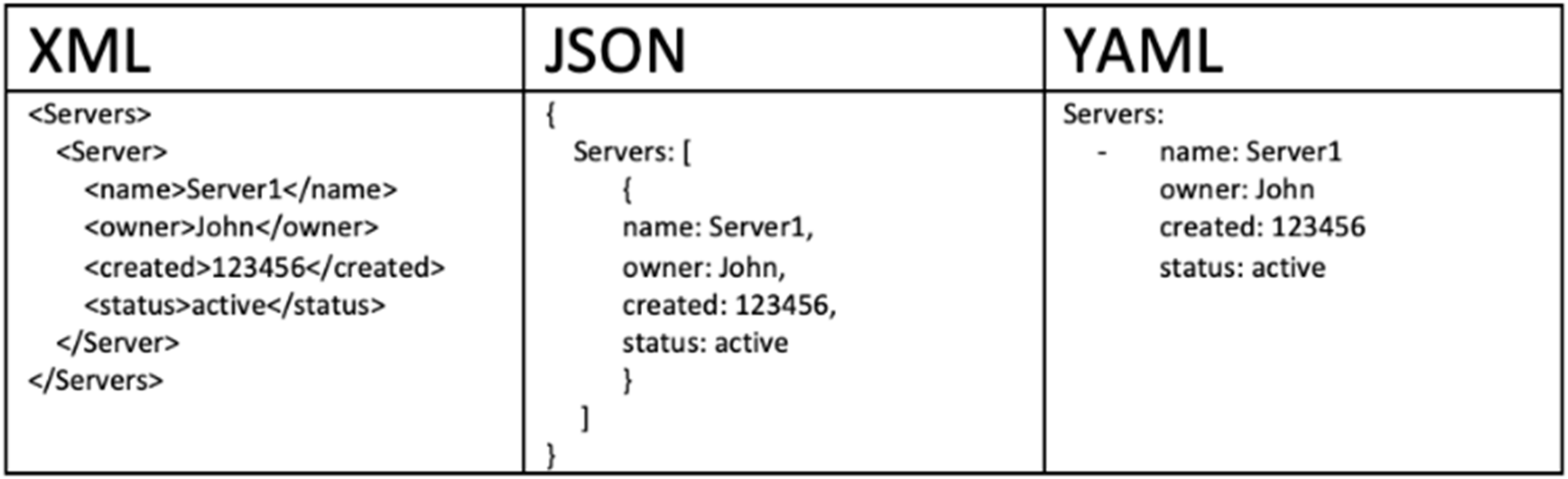
```
Candidate,Party,%,Year,Result
Reagan,Republican,50.7,1980,win
Carter,Democratic,41,1980,loss
Anderson,Independent,6.6,1980,loss
Reagan,Republican,58.8,1984,win
Mondale,Democratic,37.6,1984,loss
Bush,Republican,53.4,1988,win
Dukakis,Democratic,45.6,1988,loss
Clinton,Democratic,43,1992,win
Bush,Republican,37.4,1992,loss
Perot,Independent,18.9,1992,loss
Clinton,Democratic,49.2,1996,win
Dole,Republican,40.7,1996,loss
Perot,Independent,8.4,1996,loss
Gore,Democratic,48.4,2000,loss
Bush,Republican,47.9,2000,win
Kerry,Democratic,48.3,2004,loss
Bush,Republican,50.7,2004,win
Obama,Democratic,52.9,2008,win
McCain,Republican,45.7,2008,loss
Obama,Democratic,51.1,2012,win
Romney,Republican,47.2,2012,loss
Clinton,Democratic,48.2,2016,loss
Trump,Republican,46.1,2016,win
```

tsv



```
Candidate       Party   %       Year    Result
Reagan  Republican      50.7    1980    win
Carter  Democratic      41.0    1980    loss
Anderson        Independent     6.6     1980    loss
Reagan  Republican      58.8    1984    win
Mondale Democratic      37.6    1984    loss
Bush    Republican      53.4    1988    win
Dukakis Democratic      45.6    1988    loss
Clinton Democratic      43.0    1992    win
Bush    Republican      37.4    1992    loss
Perot   Independent     18.9    1992    loss
Clinton Democratic      49.2    1996    win
Dole    Republican      40.7    1996    loss
Perot   Independent     8.4     1996    loss
Gore    Democratic      48.4    2000    loss
Bush    Republican      47.9    2000    win
Kerry   Democratic      48.3    2004    loss
Bush    Republican      50.7    2004    win
Obama   Democratic      52.9    2008    win
McCain  Republican      45.7    2008    loss
Obama   Democratic      51.1    2012    win
Romney  Republican      47.2    2012    loss
Clinton Democratic      48.2    2016    loss
Trump   Republican      46.1    2016    win
```

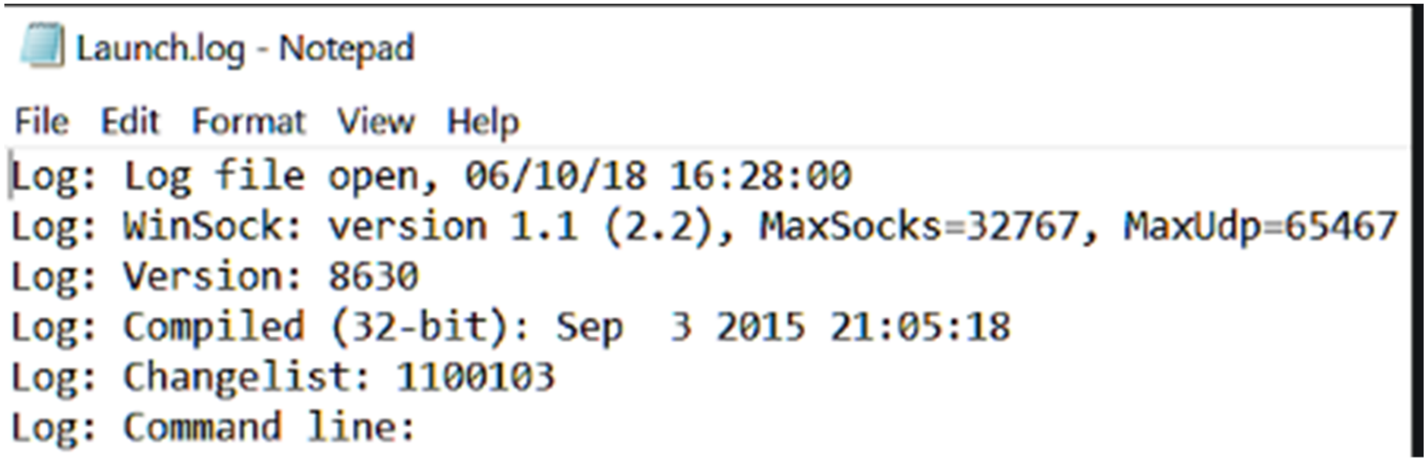
Nested Files



XML	JSON	YAML

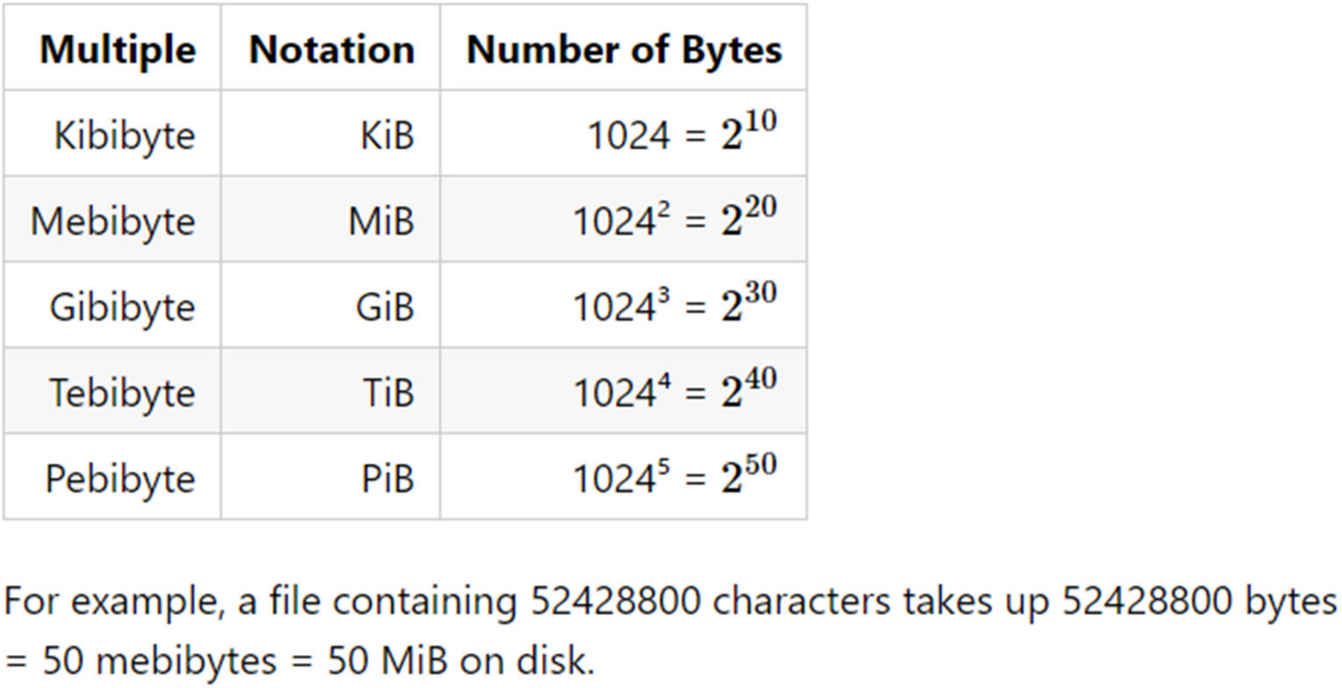
<Servers>
<Server>
<name>Server1</name>
<owner>John</owner>
<created>123456</created>
<status>active</status>
</Server>
</Servers>
	{
	Servers: [
	{
	name: Server1,
	owner: John,
	created: 123456,
	status: active
	}
	]
	}
		Servers:
		- name: Server1
		owner: John
		created: 123456
		status: active

Unstructured Files



```
Log: Log file open, 06/10/18 16:28:00
Log: WinSock: version 1.1 (2.2), MaxSocks=32767, MaxUdp=65467
Log: Version: 8630
Log: Compiled (32-bit): Sep  3 2015 21:05:18
Log: Changelist: 1100103
Log: Command line:
```

File Size



Multiple	Notation	Number of Bytes
Kibibyte	KiB	1024 = 2^{10}
Mebibyte	MiB	1024^2 = 2^{20}
Gibibyte	GiB	1024^3 = 2^{30}
Tebibyte	TiB	1024^4 = 2^{40}
Pebibyte	PiB	1024^5 = 2^{50}

For example, a file containing 52428800 characters takes up 52428800 bytes = 50 mebibytes = 50 MiB on disk.

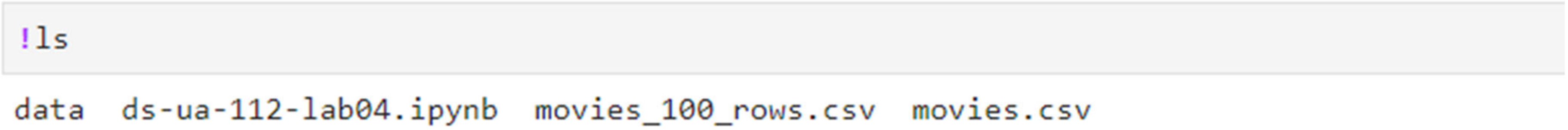
File Size

- When to read file?
  - pandas requires double the file size in available memory
  - **Example:** Reading in a 1 GiB file will typically require at least 2 GiB of available memory.
- How can we determine the file size before reading it?
  - Shell Interpreter
  - Command-line interface (CLI)

File Size

- When to read file?
  - pandas requires double the file size in available memory
  - **Example:** Reading in a 1 GiB file will typically require at least 2 GiB of available memory.

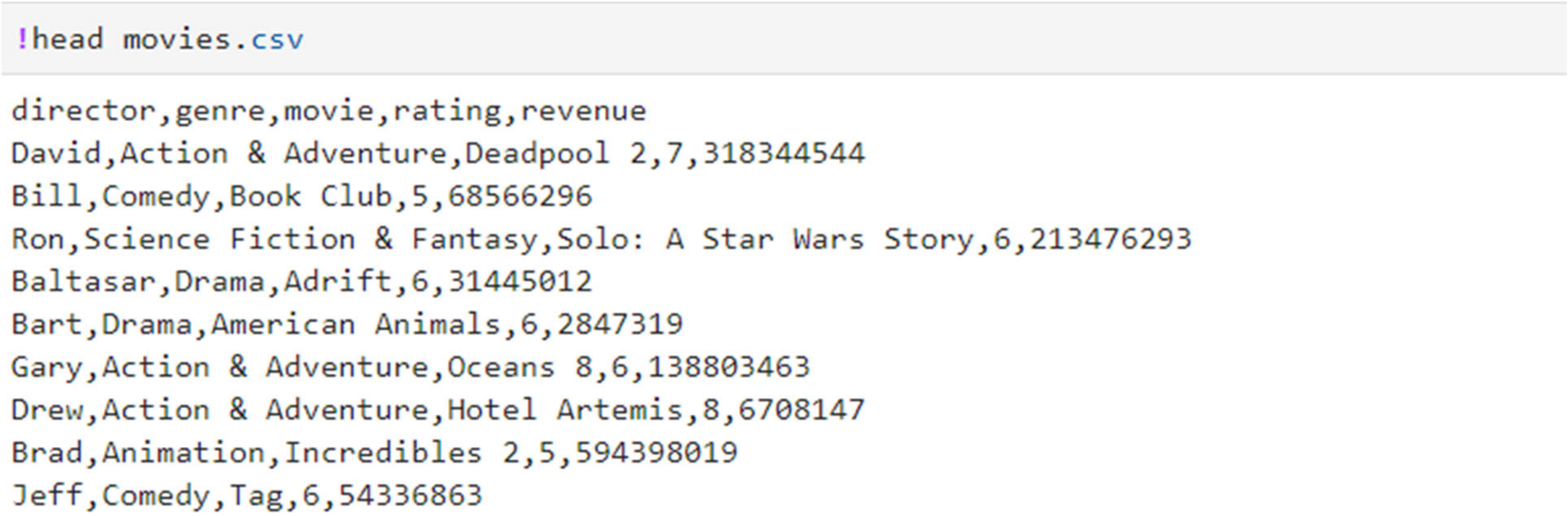
File Size: ls command



```
!ls
```

```
data  ds-ua-112-lab04.ipynb  movies_100_rows.csv  movies.csv
```

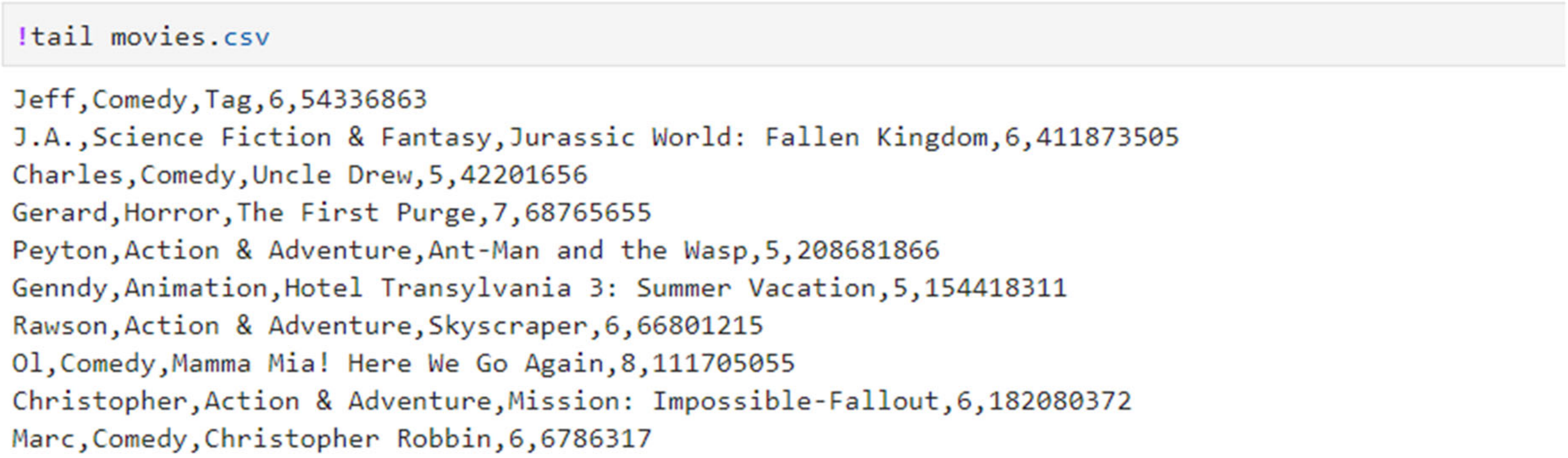
File Size: head, tail, cat commands



```
!head movies.csv
```

```
director,genre,movie,rating,revenue
David,Action & Adventure,Deadpool 2,7,318344544
Bill,Comedy,Book Club,5,68566296
Ron,Science Fiction & Fantasy,Solo: A Star Wars Story,6,213476293
Baltasar,Drama,Adrift,6,31445012
Bart,Drama,American Animals,6,2847319
Gary,Action & Adventure,Oceans 8,6,138803463
Drew,Action & Adventure,Hotel Artemis,8,6708147
Brad,Animation,Incredibles 2,5,594398019
Jeff,Comedy,Tag,6,54336863
```

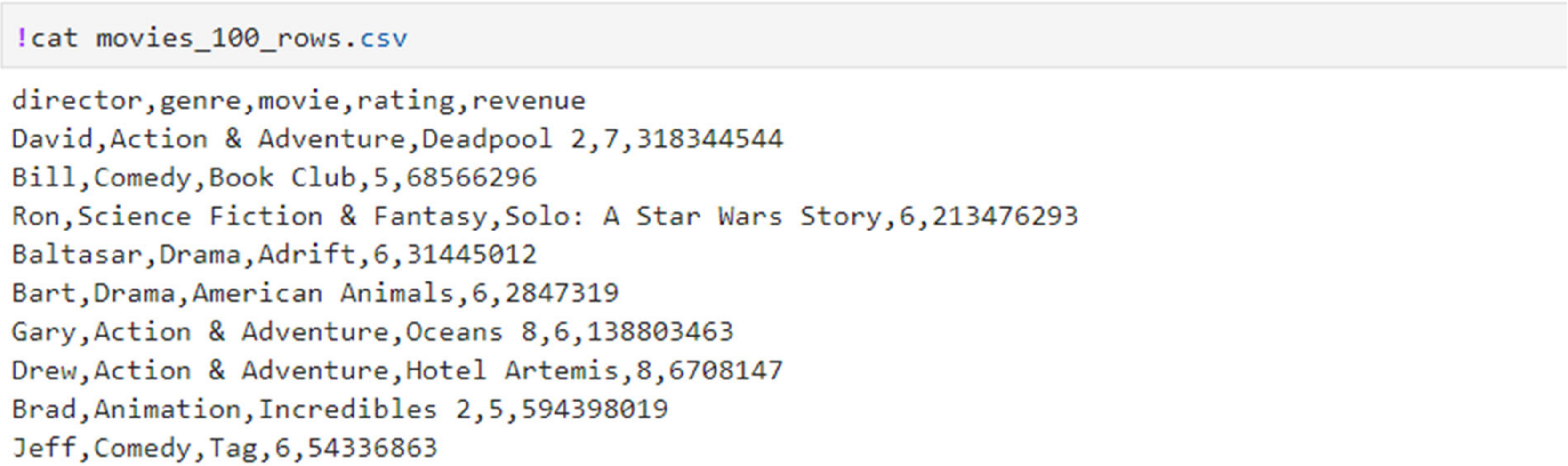
File Size: head, tail, cat commands



```
!tail movies.csv
```

```
Jeff,Comedy,Tag,6,54336863
J.A.,Science Fiction & Fantasy,Jurassic World: Fallen Kingdom,6,411873505
Charles,Comedy,Uncle Drew,5,42201656
Gerard,Horror,The First Purge,7,68765655
Peyton,Action & Adventure,Ant-Man and the Wasp,5,208681866
Genndy,Animation,Hotel Transylvania 3: Summer Vacation,5,154418311
Rawson,Action & Adventure,Skyscraper,6,66801215
Ol,Comedy,Mamma Mia! Here We Go Again,8,111705055
Christopher,Action & Adventure,Mission: Impossible-Fallout,6,182080372
Marc,Comedy,Christopher Robbin,6,6786317
```

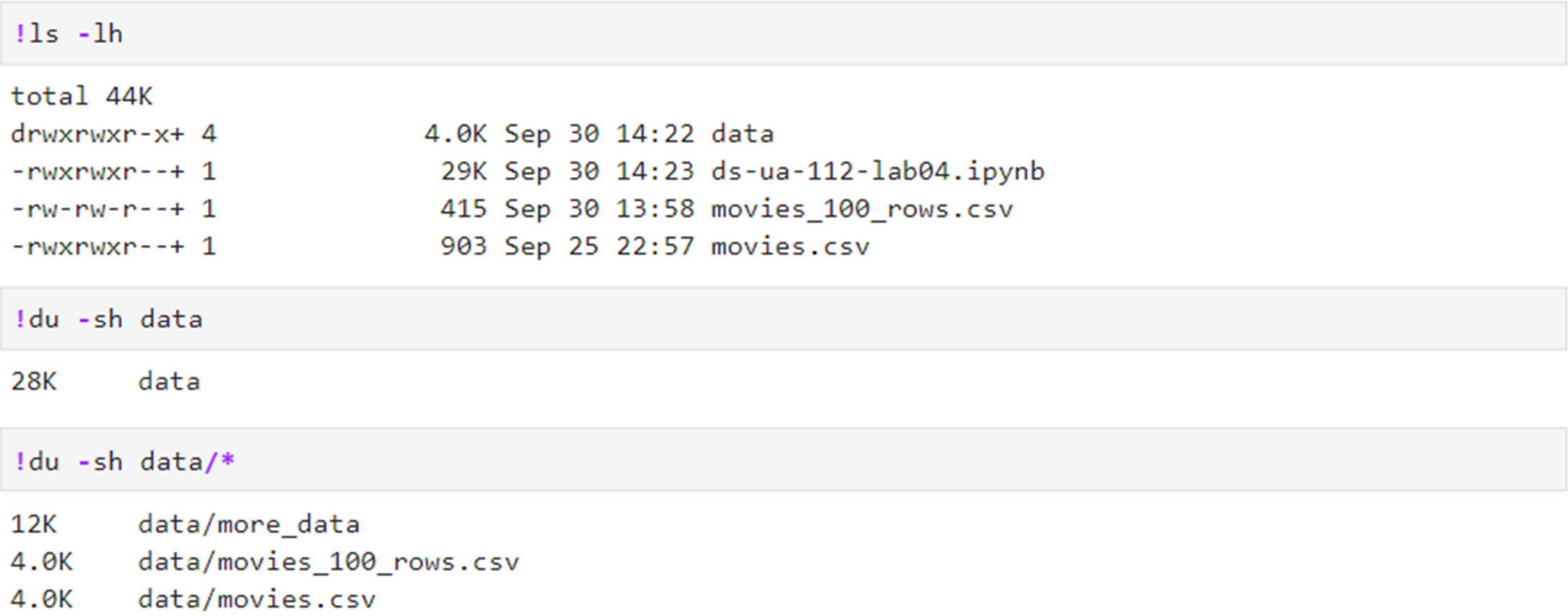
File Size: head, tail, cat commands



```
!cat movies_100_rows.csv
```

```
director,genre,movie,rating,revenue
David,Action & Adventure,Deadpool 2,7,318344544
Bill,Comedy,Book Club,5,68566296
Ron,Science Fiction & Fantasy,Solo: A Star Wars Story,6,213476293
Baltasar,Drama,Adrift,6,31445012
Bart,Drama,American Animals,6,2847319
Gary,Action & Adventure,Oceans 8,6,138803463
Drew,Action & Adventure,Hotel Artemis,8,6708147
Brad,Animation,Incredibles 2,5,594398019
Jeff,Comedy,Tag,6,54336863
```

File Size: du command



```
!ls -lh
```

```
total 44K
drwxrwxr-x+ 4                4.0K Sep 30 14:22 data
-rwxrwxr--+ 1                 29K Sep 30 14:23 ds-ua-112-lab04.ipynb
-rw-rw-r--+ 1                 415 Sep 30 13:58 movies_100_rows.csv
-rwxrwxr--+ 1                 903 Sep 25 22:57 movies.csv
```

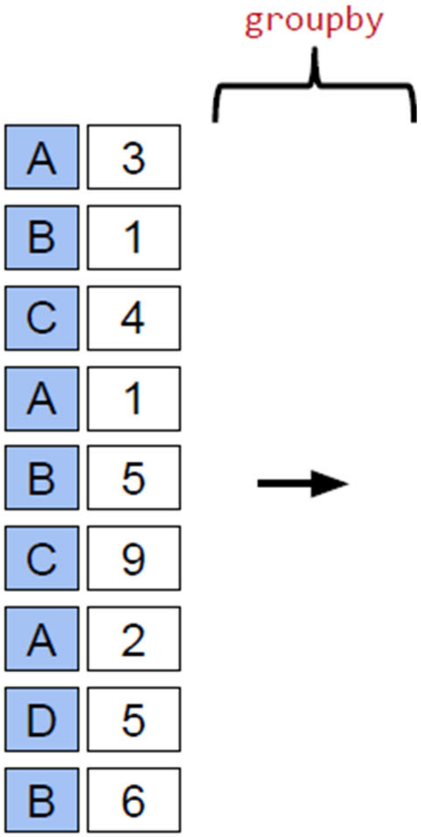
```
!du -sh data
```

```
28K     data
```

```
!du -sh data/*
```

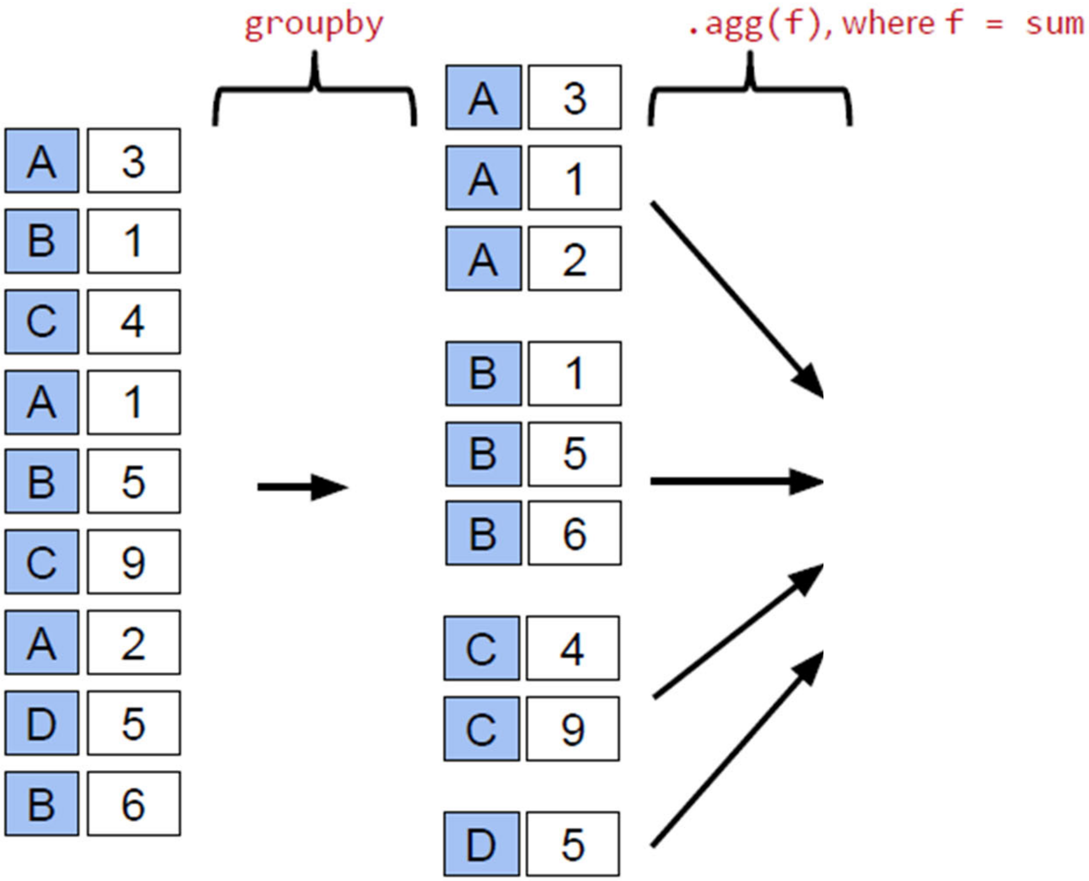
```
12K     data/more_data
4.0K    data/movies_100_rows.csv
4.0K    data/movies.csv
```

Group: Split



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

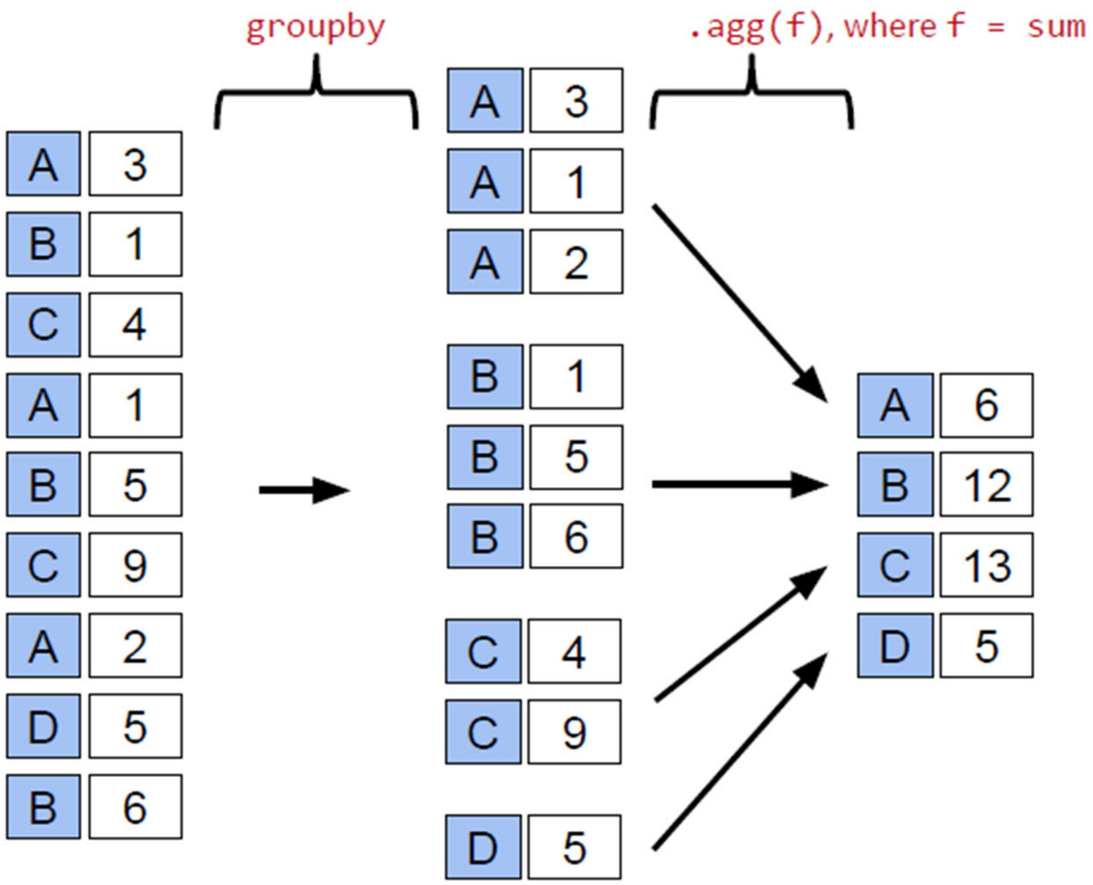
Group: Apply



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

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

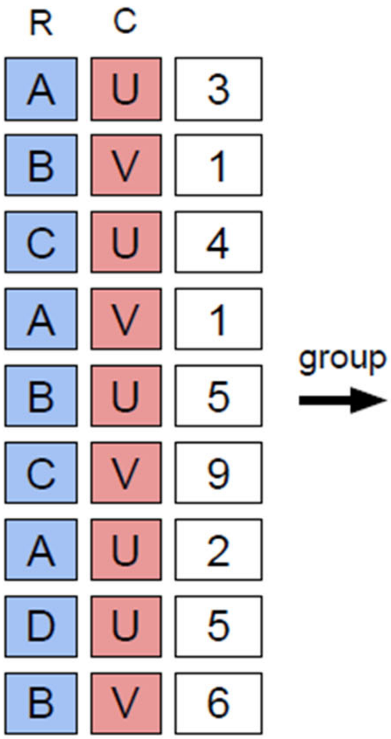
Group: Combine



groupby

.agg(f), where f = sum

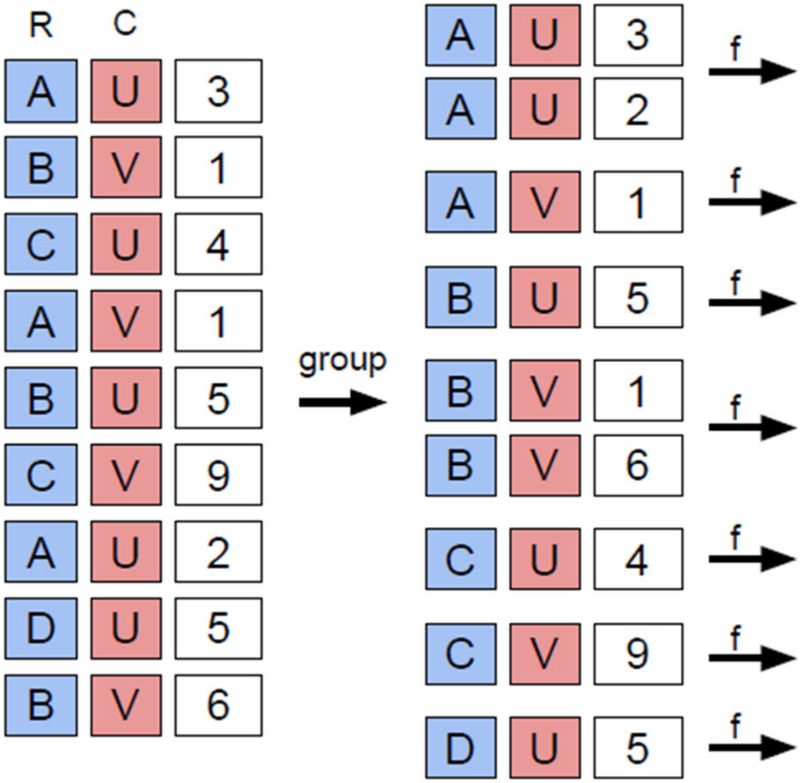
Pivot



R	C	
A	U	3
B	V	1
C	U	4
A	V	1
B	U	5
C	V	9
A	U	2
D	U	5
B	V	6

group

Pivot

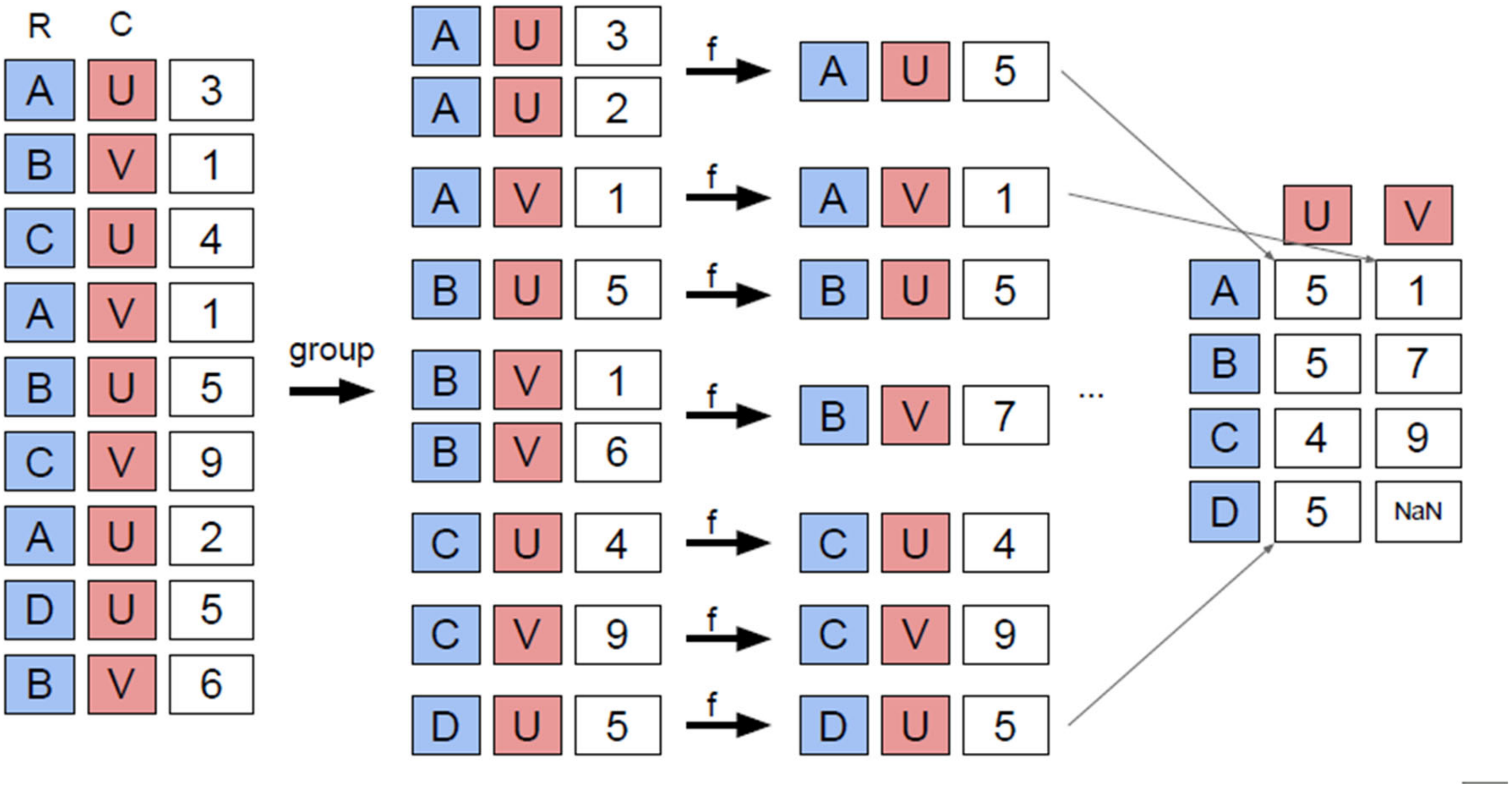


R	C	
A	U	3
B	V	1
C	U	4
A	V	1
B	U	5
C	V	9
A	U	2
D	U	5
B	V	6

group

A	U	3
A	U	2
A	V	1
B	U	5
B	V	1
B	V	6
C	U	4
C	V	9
D	U	5

Pivot



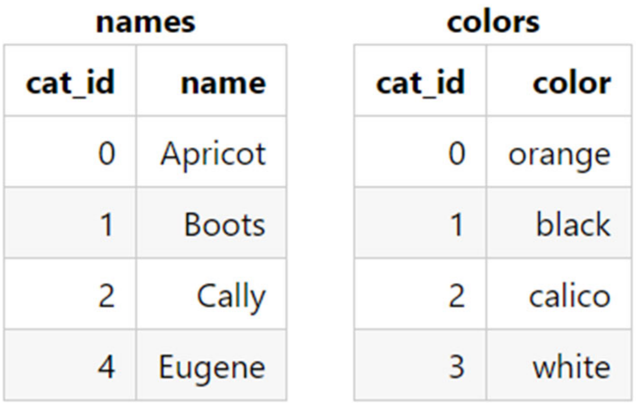
R	C	
A	U	3
B	V	1
C	U	4
A	V	1
B	U	5
C	V	9
A	U	2
D	U	5
B	V	6

A	U	3
A	U	2
A	V	1
B	U	5
B	V	1
B	V	6
C	U	4
C	V	9
D	U	5

A	U	5
A	V	1
B	U	5
B	V	7
C	U	4
C	V	9
D	U	5

	U	V
A	5	1
B	5	7
C	4	9
D	5	NaN

Join



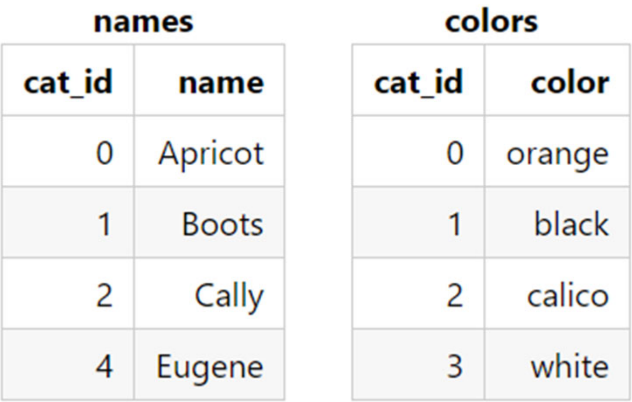
names

cat_id	name
0	Apricot
1	Boots
2	Cally
4	Eugene

colors

cat_id	color
0	orange
1	black
2	calico
3	white

Join: Inner

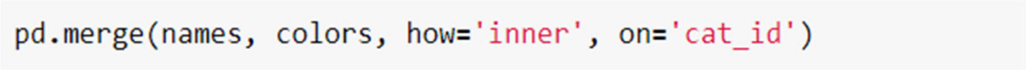


names

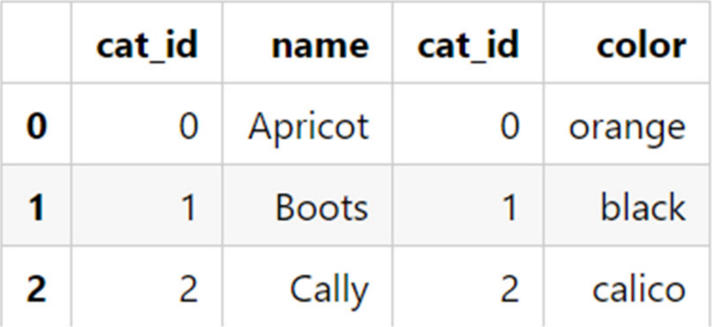
cat_id	name
0	Apricot
1	Boots
2	Cally
4	Eugene

colors

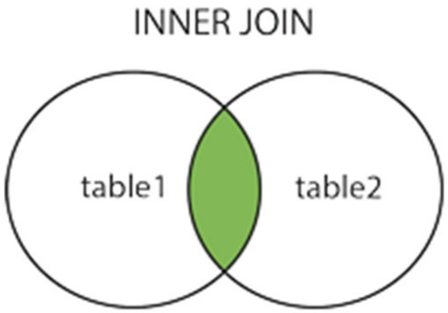
cat_id	color
0	orange
1	black
2	calico
3	white



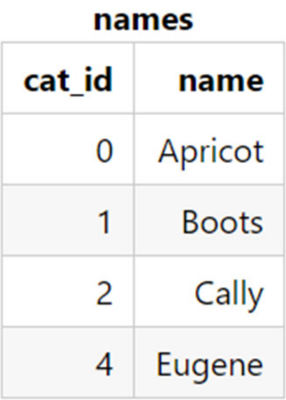
```
pd.merge(names, colors, how='inner', on='cat_id')
```



	cat_id	name	cat_id	color
0	0	Apricot	0	orange
1	1	Boots	1	black
2	2	Cally	2	calico

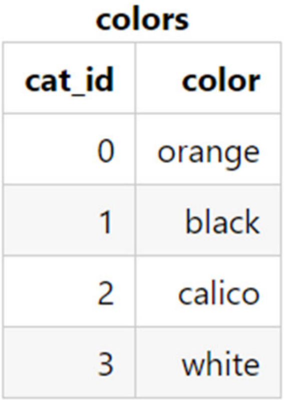


Join: Outer



names

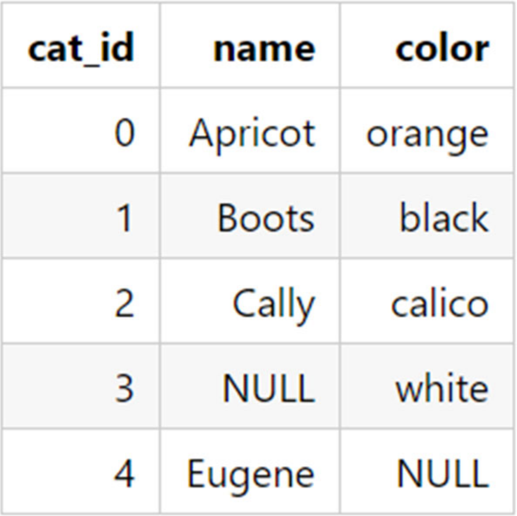
cat_id	name
0	Apricot
1	Boots
2	Cally
4	Eugene



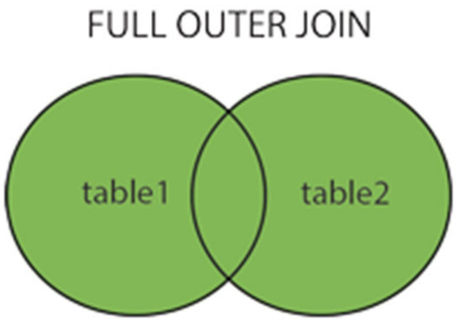
colors

cat_id	color
0	orange
1	black
2	calico
3	white

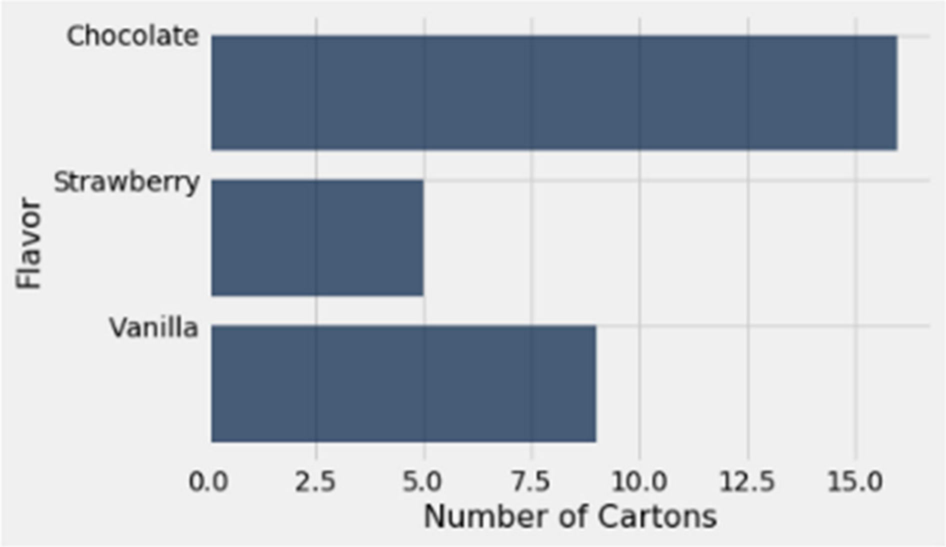
```
pd.merge(names, colors, how='outer', on='cat_id')
```

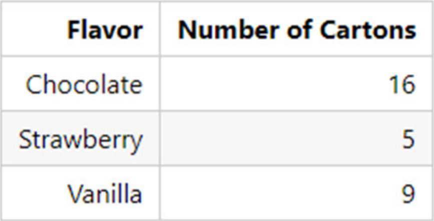


cat_id	name	color
0	Apricot	orange
1	Boots	black
2	Cally	calico
3	NULL	white
4	Eugene	NULL



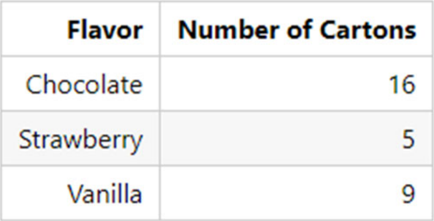
Categorical Data



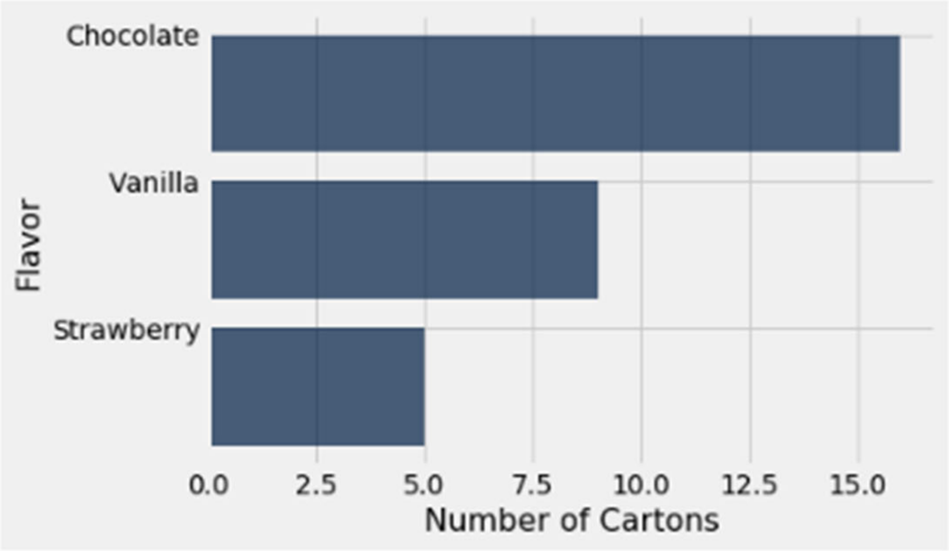


Flavor	Number of Cartons
Chocolate	16
Strawberry	5
Vanilla	9

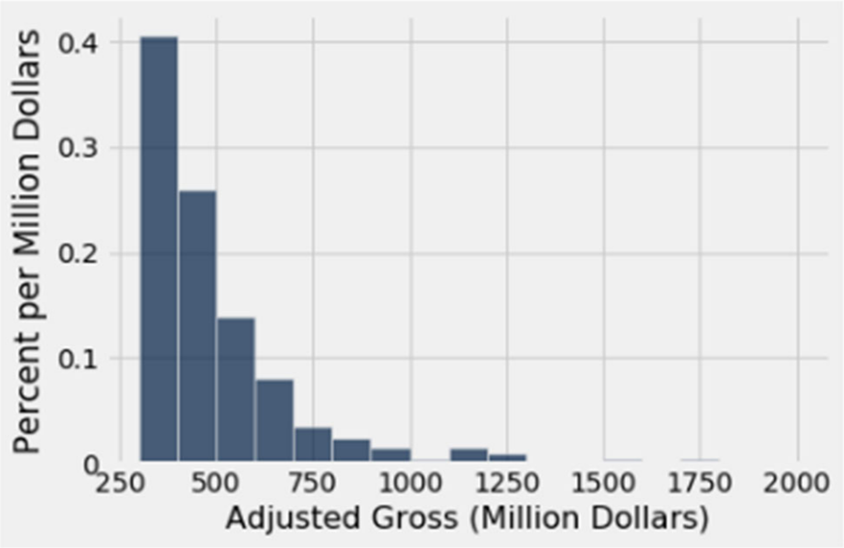
Categorical Data

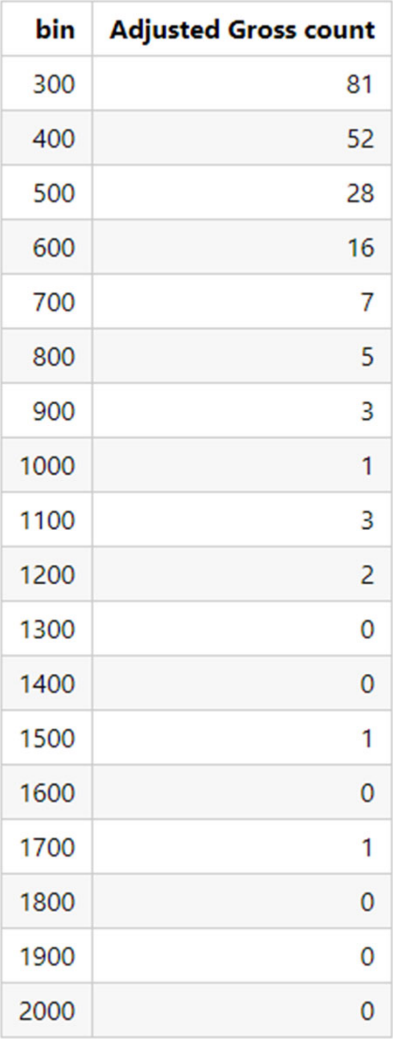


Flavor	Number of Cartons
Chocolate	16
Strawberry	5
Vanilla	9



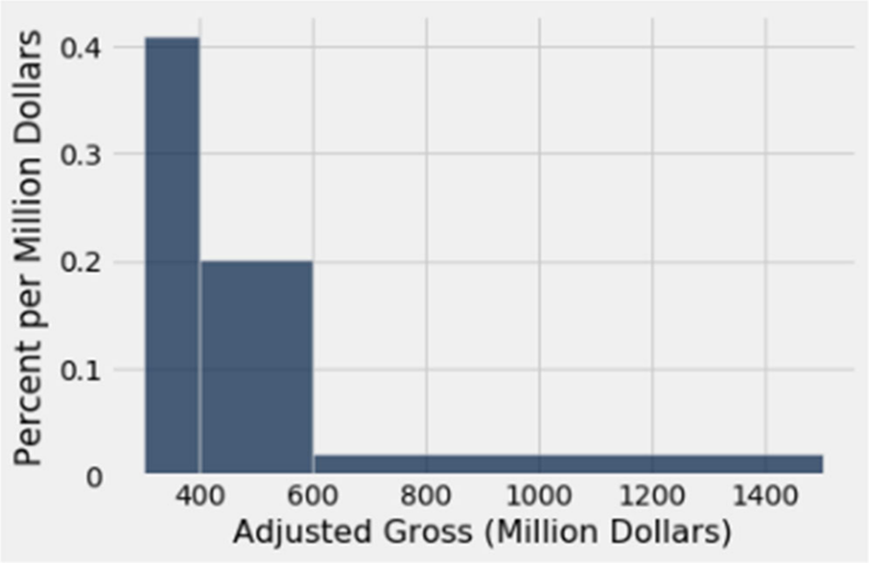
Categorical Data

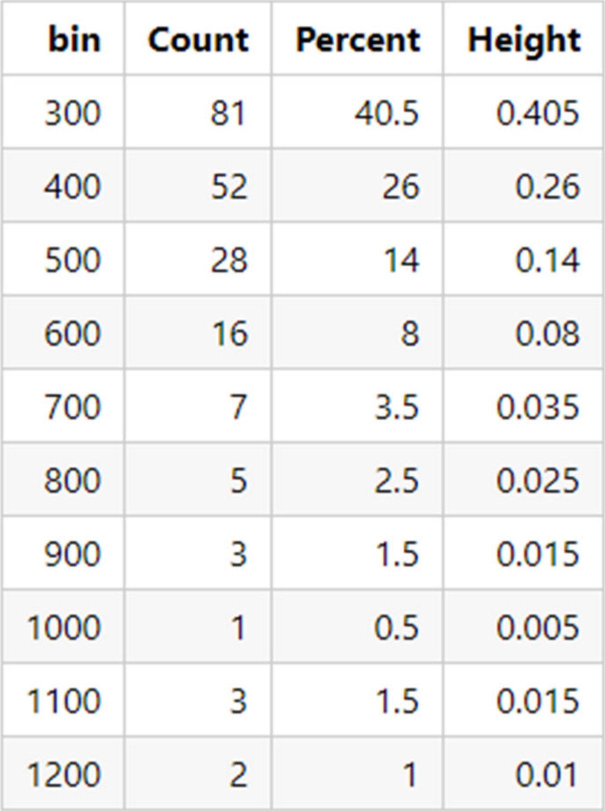




bin	Adjusted Gross count
300	81
400	52
500	28
600	16
700	7
800	5
900	3
1000	1
1100	3
1200	2
1300	0
1400	0
1500	1
1600	0
1700	1
1800	0
1900	0
2000	0

Categorical Data





bin	Count	Percent	Height
300	81	40.5	0.405
400	52	26	0.26
500	28	14	0.14
600	16	8	0.08
700	7	3.5	0.035
800	5	2.5	0.025
900	3	1.5	0.015
1000	1	0.5	0.005
1100	3	1.5	0.015
1200	2	1	0.01

Take-Aways

- File Size
  - kibi, mebi, gibi, tebi
  - ls, du, head, tail
- Split-Apply-Combine
  - Aggregate
  - Filter
  - Transform
- Join
  - Inner
  - Outer
- Plotting Categorical Data
  - Bar chart vs Histogram