

Unirank: Data Preview (head, pipe and csvlook)

WE'LL COVER THE FOLLOWING



- Learning objectives
- Data download
- Dataset preview (use the 'head' command)
- Do you want to know more?

The ranking of universities has become a common task performed by many institutions, each of them proposes a different ranking based in several weighted categories. Examples of those rankings are: [Webometrics Ranking of World Universities](#), [THES - QS World Universities Rankings](#) and [Academic Ranking of World Universities](#). The first ranking measures the visibility of the universities and their global performance in the web. The last two attempt to measure the performance of the universities based in categories like prizes received by members, citations, and publications. Employers, especially from the multinational organisations use rankings to find universities to source graduates, so attending a high-ranking university can help in a competitive job market.

In this lesson we will use a simple (publicly available) dataset obtained from the [data.world](#) called: US News Universities Rankings 2017 edition. From this data, using Bash we will explore different features and finally find an interesting fact about the correlation of tuition fees and uni rank. This simple dataset contains the following fields.

- **Name** - institution name
- **Location** - City, State where located
- **Rank** - read methodology [here](#).
- **Description** - a snippet of text overview from U.S. News.

- **Tuition and fees** - combined tuition and fees.
- **Undergrad Enrollment** - number of enrolled undergraduate students

In each project described in this book, we will attempt to learn a few Bash commands and tricks.

Learning objectives

By completing this, you will learn to use the following Bash commands:

- `head` – output the first part of files
- `tail` – opposite to head
- `cat` – concatenate and print files
- `sort` – sort file contents
- `uniq` – remove duplicate entries

Data download

You should download the data from [here](#) , as we have slightly simplified the data and let's save the data as: `unirank.csv` . I believe my course would be incomplete without video demos of the commands I am showing here. Therefore, I will add a video demo with each lesson. Watch and enjoy it before you proceed to the next part:

Dataset preview (use the ‘head’ command)

This dataset is small (toy) and we could in principle open it in a text editor or in Excel. However, real-world datasets are often larger and cumbersome to open in their entirety. Let’s assume as if it were a Big Data (and unstructured) and we want to get a sneak peak of the data. This is often the first thing to do when you get your hands on new data- previewing; it is important to get a sense for what it contains, how it is organized, and whether the data makes sense in the first place.

To help us get a preview of the data, we can use the command `head`, which as the name suggests, shows the first few lines of a file (the `unirank.csv` dataset has been already been stored onto the course storagespace, thanks to educative.io team). Simply press “Run” and see the output!

```
#!/bin/bash  
head unirank.csv
```



However, you will find the outputs are not very interesting on the first place, therefore we install a tool called `csvkit`, which is a suite of command-line tools for converting to and working with CSV (install: `sudo pip install csvkit`).

This will greatly help our future analyses. After we have installed the `csvkit`, we re-run the `head` command, but outputs piped (`|`, which basically chains the output of the first command to the input of the next, soon we’ll learn about it) through the `csvlook` command from the `csvkit` suit:

```
#!/bin/bash  
head unirank.csv | csvlook
```

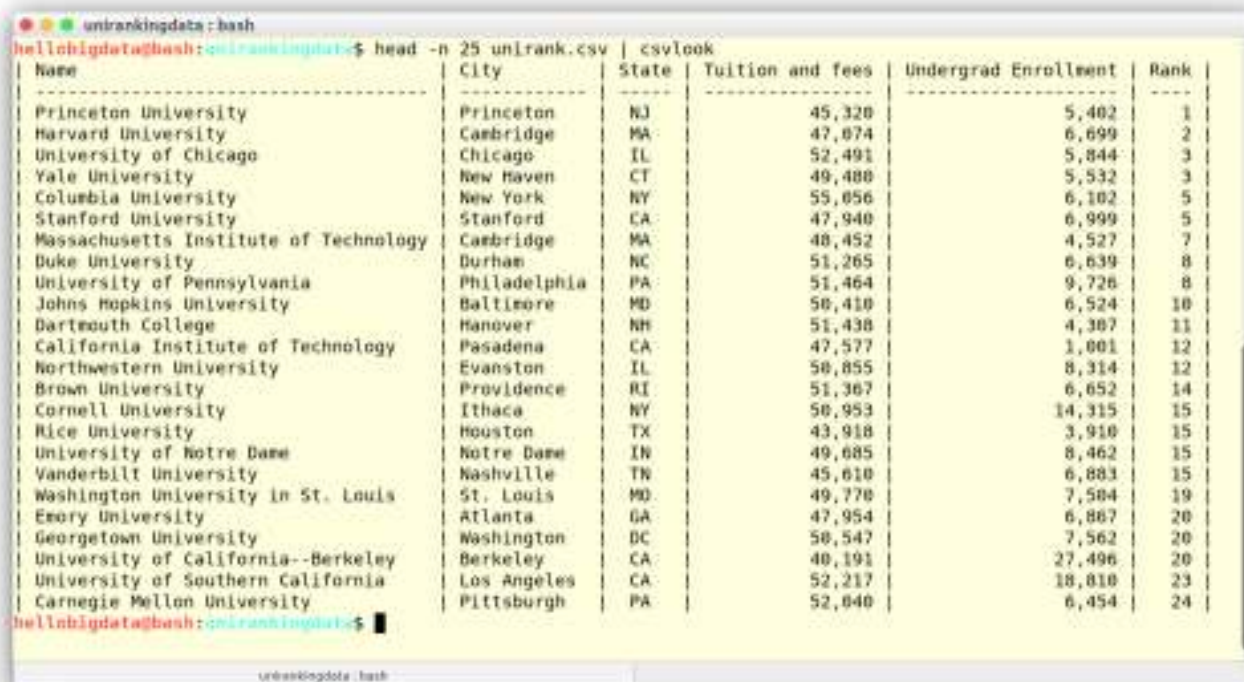


You should see the first 10 lines of the file output onto the screen, to see more than the first 10 lines, e.g. the first 25, use the `-n` option:

```
#!/bin/bash
head -n 25 unirank.csv | csvlook
```



Here, the dataset name `unirankingdata.csv` is a **command-line** argument that is given to the command `head` and the `-n` is an option which allows us to overwrite the 10-line default. Such command-line options are typically specified with a dash followed by a string, a space, and the value of the option (e.g. `-n 25`). However, often the options don't require a value but instead are made for toggling a feature on or off, for example `top -h` shows the help page for the command `top` that shows off all the running process and apps.



```
unrankingdata: bash
hellinbigdata@bash:~$ head -n 25 unirank.csv | csvlook
```

Name	City	State	Tuition and fees	Undergrad Enrollment	Rank
Princeton University	Princeton	NJ	45,320	5,402	1
Harvard University	Cambridge	MA	47,074	6,699	2
University of Chicago	Chicago	IL	52,491	5,844	3
Yale University	New Haven	CT	49,480	5,532	3
Columbia University	New York	NY	55,056	6,102	5
Stanford University	Stanford	CA	47,940	6,999	5
Massachusetts Institute of Technology	Cambridge	MA	48,452	4,527	7
Duke University	Durham	NC	51,265	6,639	8
University of Pennsylvania	Philadelphia	PA	51,464	9,726	8
Johns Hopkins University	Baltimore	MD	50,410	6,524	10
Dartmouth College	Hanover	NH	51,438	4,307	11
California Institute of Technology	Pasadena	CA	47,577	1,001	12
Northwestern University	Evanston	IL	50,855	8,314	12
Brown University	Providence	RI	51,367	6,652	14
Cornell University	Ithaca	NY	50,953	14,315	15
Rice University	Houston	TX	43,918	3,910	15
University of Notre Dame	Notre Dame	IN	49,685	8,462	15
Vanderbilt University	Nashville	TN	45,610	6,883	15
Washington University in St. Louis	St. Louis	MO	49,770	7,504	19
Emory University	Atlanta	GA	47,954	6,867	20
Georgetown University	Washington	DC	50,547	7,562	20
University of California--Berkeley	Berkeley	CA	40,191	27,496	20
University of Southern California	Los Angeles	CA	52,217	18,010	23
Carnegie Mellon University	Pittsburgh	PA	52,040	6,454	24

From the first 25 lines of the file, we can infer that the data is formatted as a file with separated values. From the first line (often called a header line) and the first few lines of data, we can infer the column contents: `Name`, `City`, `State`, `Tuition and fees`, `Undergrad Enrollment` and `Rank`.

Do you want to know more? #

So, you want to know more?

Read the attached man pages:



'head' man page



'csvlook' man page



'pipe' tldp man page

