Lesson A1
Introduction to statistics, Visualisation

GEO1001.2020

Clara García-Sánchez, Stelios Vitalis

Resources adapted from:
- David M. Lane et al. (http://onlinestatbook.com)
- Allen B. Downey et al. (https://greenteapress.com/wp/think-stats-2e/)

Lesson A1

Introduction to Statistics

Overview

- Introduction
- Descriptive statistics
- Inferential statistics
- Percentiles
- Distributions
- Significance
- Mean and variance

Overview

- **Introduction**
- Descriptive statistics
- Inferential statistics
- Percentiles
- Distributions
- Significance
- Mean and variance

Introduction

The study of statistics involves:

Math + Calculation of numbers + How those numbers are chosen + How the results are interpreted

What is wrong with this affirmations:

1) A new ad for Ben and Jerry's ice cream introduced in late May resulted in 30% increase in ice cream sales for the following 3 months.

2) The more churches in a city, the more crime there is —> + churches == + crimes

3) 75% more international marriages are occurring this year than 25 years ago, so our society accepts international marriages.

Introduction

Statistics are not only facts and figures, but they refer to a range of techniques and procedures for analysing, interpreting, displaying, and making decisions based on data.

A statistical approach

To address the limitations of anecdotes, a statistical approach uses tools like:

1. Data collection

2. Descriptive Statistics

3. Exploratory data analysis

4. Hypothesis testing

5. Estimation

Overview

- Introduction
- **Descriptive statistics**
- Inferential statistics
  - Percentiles
  - Distributions
  - Significance
  - Mean and variance

Descriptive Statistics

Descriptive statistics are numbers that are used to summarise and describe data.

They just describe the data in numbers, and they **do not generalise beyond the numbers**

$112,760	pediatricians
$106,130	dentists
$100,090	podiatrists
$76,140	physicists
$53,410	architects,
$49,720	school, clinical, and counseling psychologists
$47,910	flight attendants
$39,560	elementary school teachers
$38,710	police officers
$18,980	floral designers

Average salaries in the US 1999.

Overview

- Introduction
- Descriptive statistics
- **Inferential statistics**
- Percentiles
- Distributions
- Significance
- Mean and variance

Inferential Statistics

They rely on a sample (small subset) of a larger set of data.

Inferential statistics are the **mathematical procedures** whereby we **convert information** about the **sample** into intelligent **guesses** about the **population**.

Statistics over Europe population

Finite individuals from Europe

Choosing the sample is crucial, but why?

Inferential Statistics

Inferential statistics are based on:

1. Assumption of random sampling

Every member of the population needs to have an equal chance of being selected into the sample.

2. Sample is large enough to represent the population

More complex sampling:

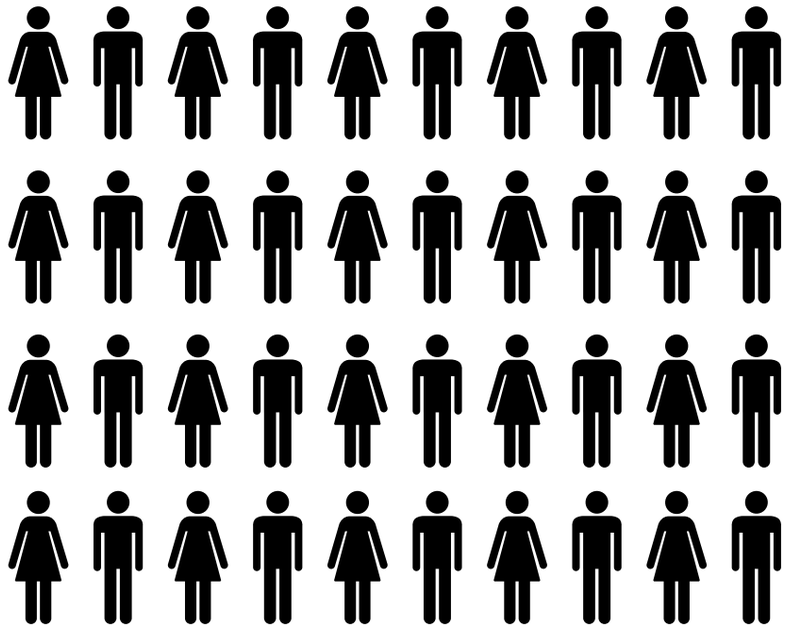
Random assignment (medical treatments where the sample is divided in two groups), **stratified sampling** (samples from groups with sizes that represent the population)

Practice

To figure out who wants to remain in EU and who wants to quit the EU from the country member states

Statistics over Europe population

Finite individuals from Europe



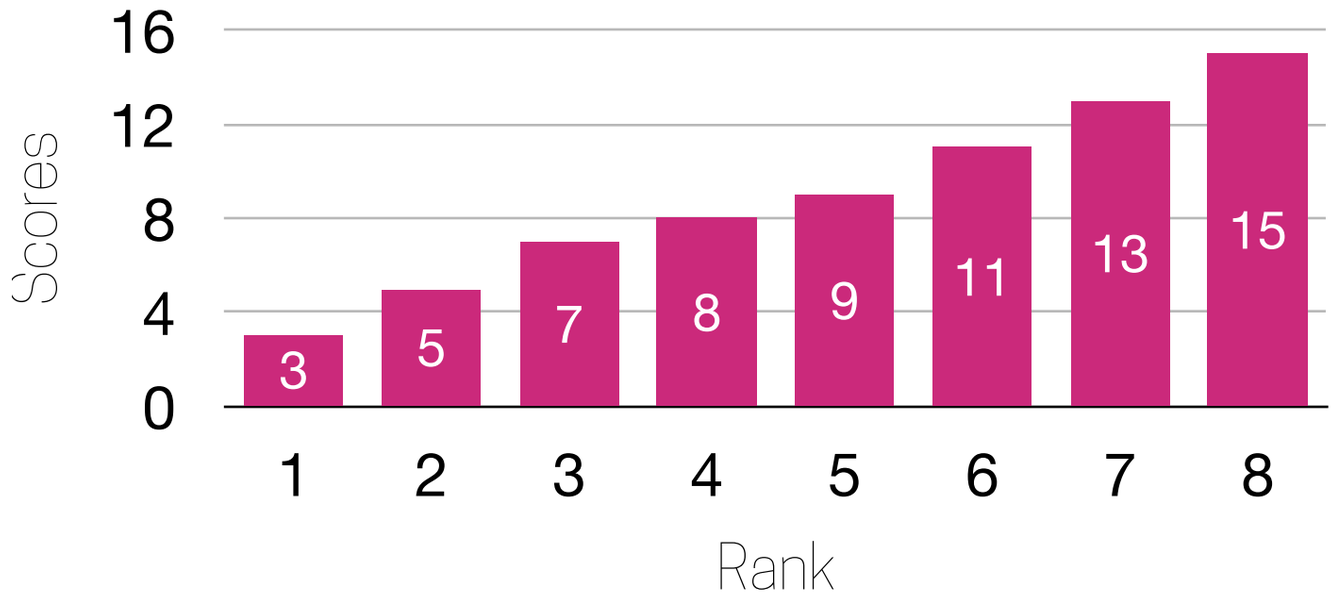
Which type of sampling would you choose for this case?

Overview

- Introduction
- Descriptive statistics
- Inferential statistics
- **Percentiles**
- Distributions
- Significance
- Mean and variance

Percentiles

Test scores



Percentile: the 65th percentile can be defined as the lowest score that is greater than 65% of the scores

How to compute the 25th percentile?

1. Compute the rank: $R = \frac{P}{100} \cdot (N+1) \rightarrow R = \frac{25}{100} \cdot (8+1) = 2.25$
2. R is integer —> percentile is the number with that rank

   R is not an integer —>

$IR = 2$ $FR = 0.25$

25th percentile

$(0.25) \cdot (7-5) + 5 = 5.5$

Overview

- Introduction
- Descriptive statistics
- Inferential statistics
- Percentiles
- **Distributions**
- Significance
- Mean and variance

Distributions

Continuous distributions: we focus in them because they are similar to geo-spatial data. For example a list of response times to perform an activity (table 2)

Table 2. Response Times

568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

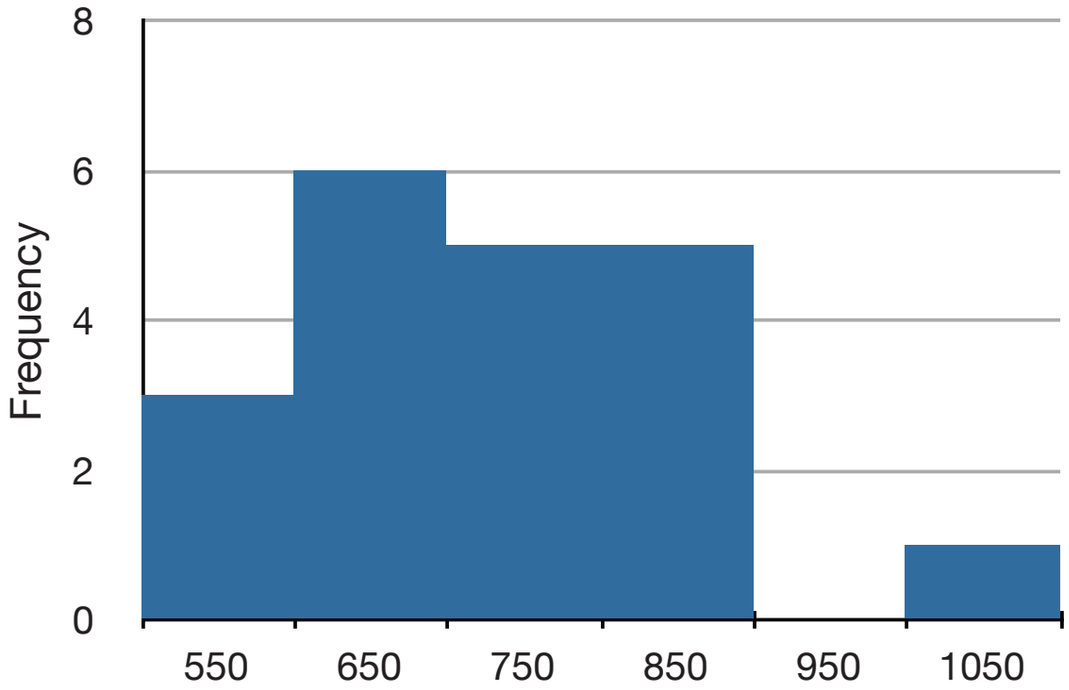


Figure 3. A histogram of the grouped frequency distribution shown in Table 3. The labels on the X-axis are the middle values of the range they represent.

Table 3. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Distributions

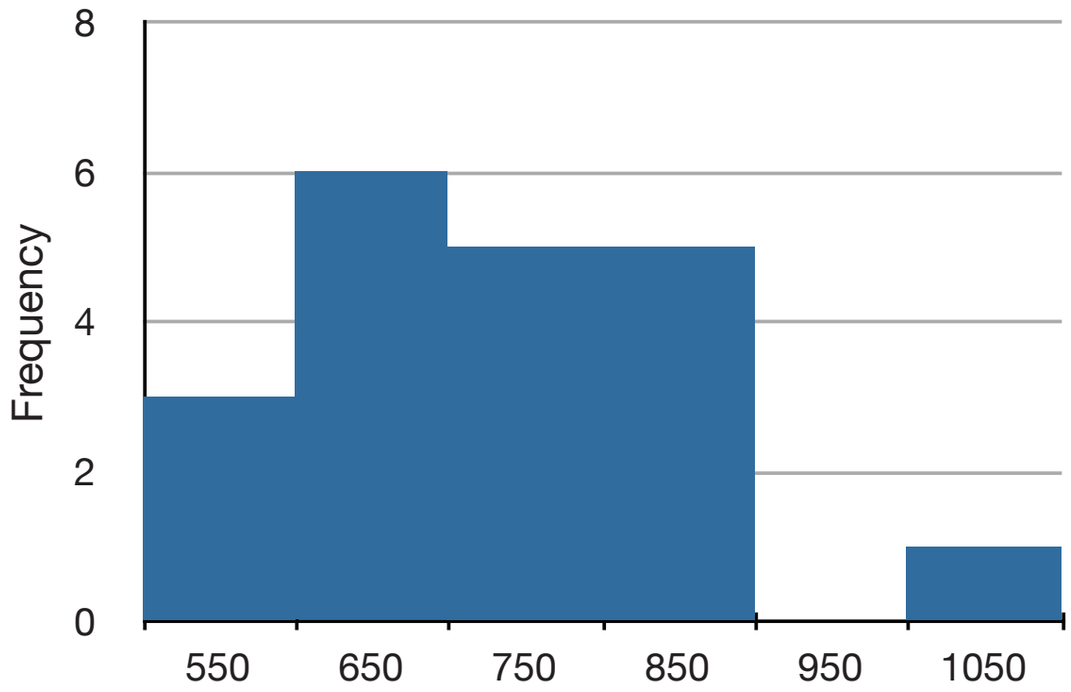


Figure 3. A histogram of the grouped frequency distribution shown in Table 3. The labels on the X-axis are the middle values of the range they represent.

- This is a distribution of continuous variable, which is also called: **“probability density (function) (pdf)”**
- Some pdfs have particular importance in statistics, such as the **normal distribution** —> because many naturally-occurring phenomena can be approximated surprisingly well by this distribution.
- What is the area under the curve?

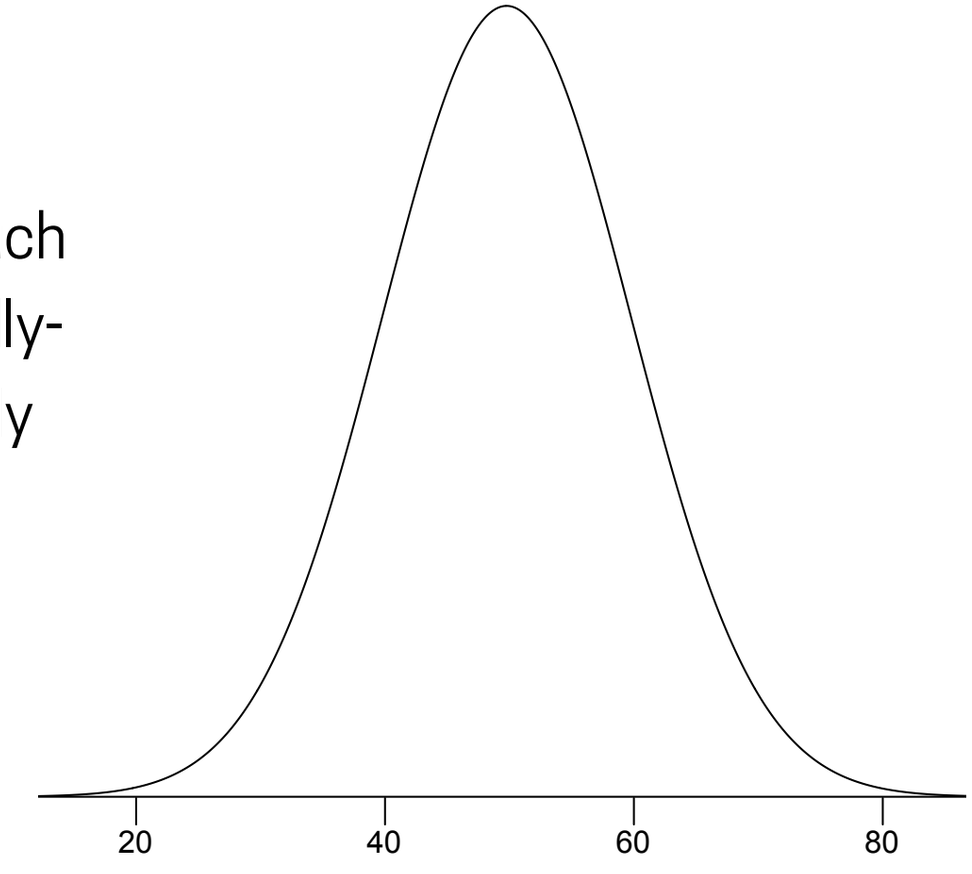


Figure 4. A normal distribution.

Distributions

- The distribution that your data follows gives you already a lot of information regarding the data you are dealing with.

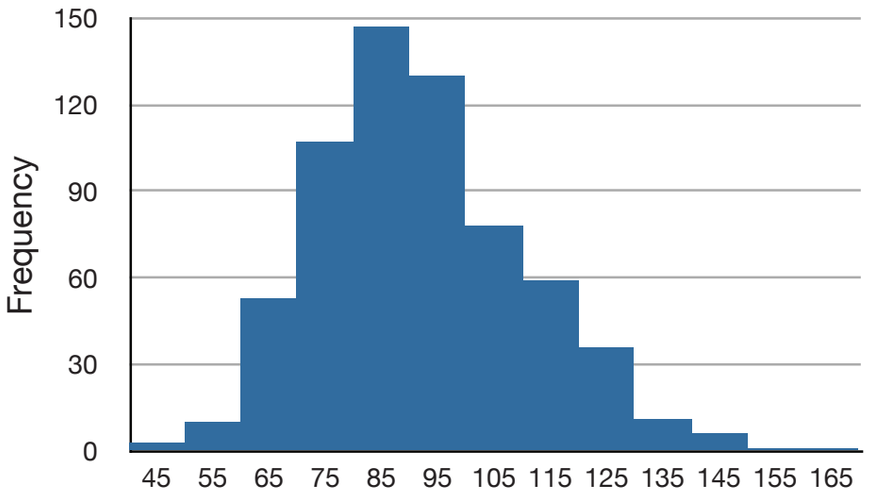


Figure 5. A distribution with a positive skew.

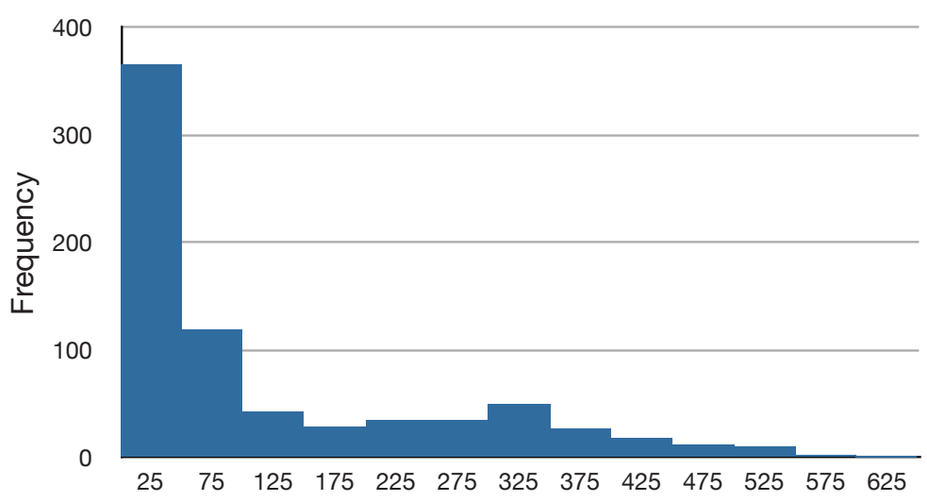
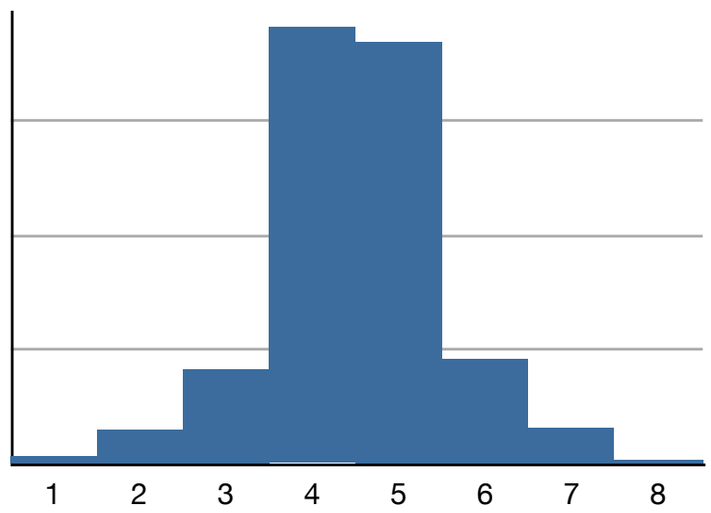
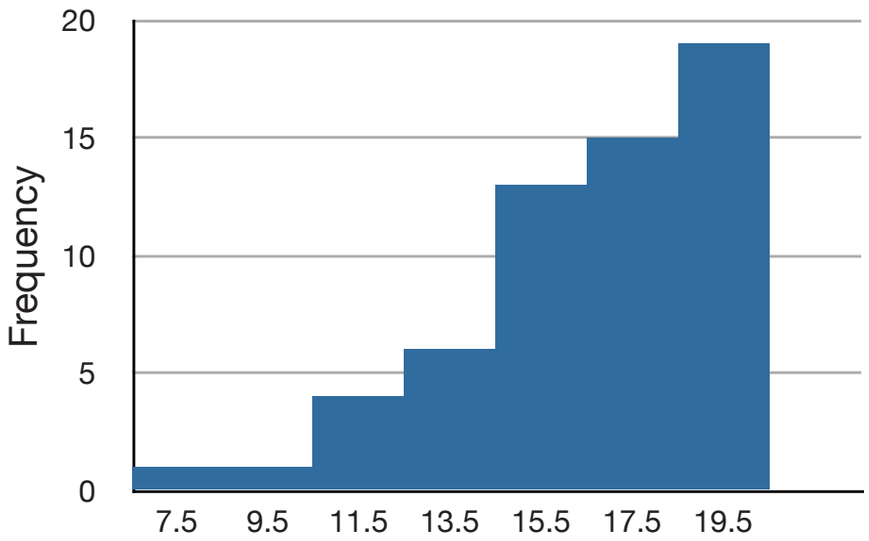
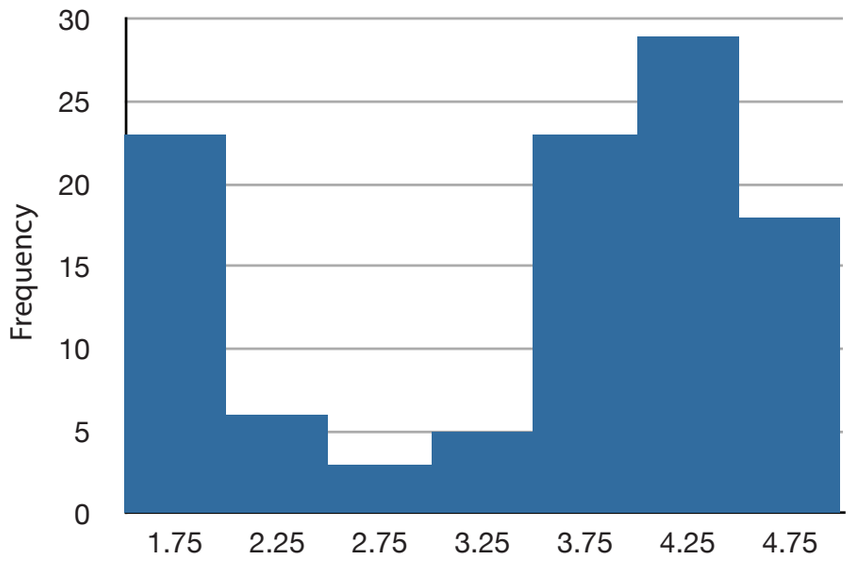


Figure 6. A distribution with a very large positive skew.







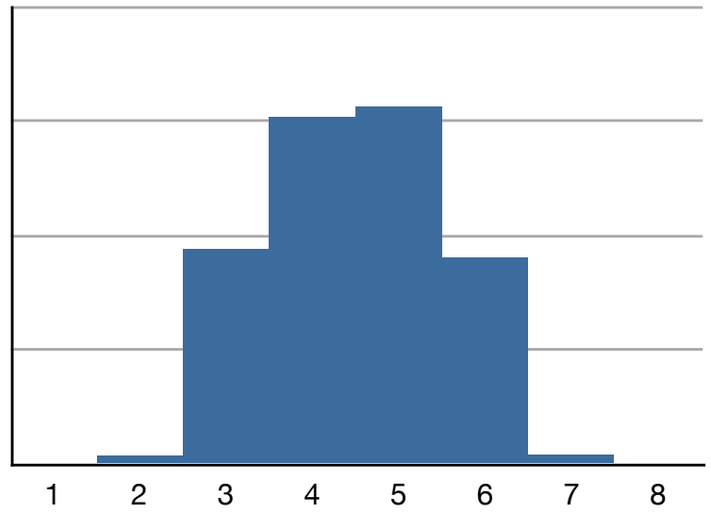


Figure 8. A distribution with negative skew. This histogram shows the frequencies of various scores on a 20-point question on a statistics test.

Overview

- Introduction
- Descriptive statistics
- Inferential statistics
- Percentiles
- Distributions
- **Significance**
- Mean and variance

Significance

Plenty of times in statistics we compare two sets of data or distributions looking for significant differences. Some of the questions we ask are:

- If the two groups have different means, what about other **summary statistics**, like median and variance? Can we be more precise about how the groups differ?

- Is it possible that the difference we saw could occur by chance, even if the groups we compared were actually the same? If so, we would conclude that the effect was not **statistically significant**.

- Is it possible that the apparent effect is due to selection bias or some other error in the experimental setup? If so, then we might conclude that the effect is an **artefact**; that is, something we created (by accident) rather than found.

Overview

- Introduction
- Descriptive statistics
- Inferential statistics
- Percentiles
- Distributions
- Significance
- **Mean and variance**

Mean and Variance

- **Mean:** if you have a sample of n values, x_i, the mean, μ, is the sum of the values divided by the number of values:

$$\mu = \frac{1}{n} \sum_i x_i$$

- **Variance:** in the same way that the mean describes the central tendency, the variance describes the spread. The variance can be calculated as:

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

- **Standard deviation:** is the square root fo the variance:

$$\sigma = \sqrt{\sigma^2}$$

Practice

Go to Gitlab if you haven't already download the materials.

Download the scripts and data inside folder "lectureA1" and the folder "data".

Keep these, we will use them in future lectures.

Compute mean, variance and standard deviation using standard libraries from python for wind direction and wind speed!

Practice

The best way to learn statistics is to find a set of data that are interesting for you.

For next Tuesday, I would like you to:

1) Think about and find a set of data that you think might be interesting to analyse and try to pose a question you are curious about.

   An example from thinkStats:

   > "Do first babies arrive late" —> National Survey of Family Growth (NSFG)

   You could think in sports results, COVID data, meteorological data, look at the 4TU repository…

2) Use the data to compute mean, variance and standard deviation with python for some of the dataset variables (not using predefined functions) and put it on Git!

Practice

The best way to learn statistics is to find a set of data that are interesting for you.

Formative!!! (but beneficial for final assignment)

2) Use the data to compute mean, variance and standard deviation with python for some of the dataset variables (not using predefined functions) and put it on Git!

Lesson A1

Visualisation

Overview

- Histograms
- Frequency Polygons
- Box Plots
- Line Graphs
- Scatter/dot plots

Overview

- **Histograms**
- Frequency Polygons
- Box Plots
- Line Graphs
- Scatter/dot plots

Histograms

It is a graphical method for displaying the shape of a distribution. This type of visualisation is particularly useful with a lot of data/ observations.

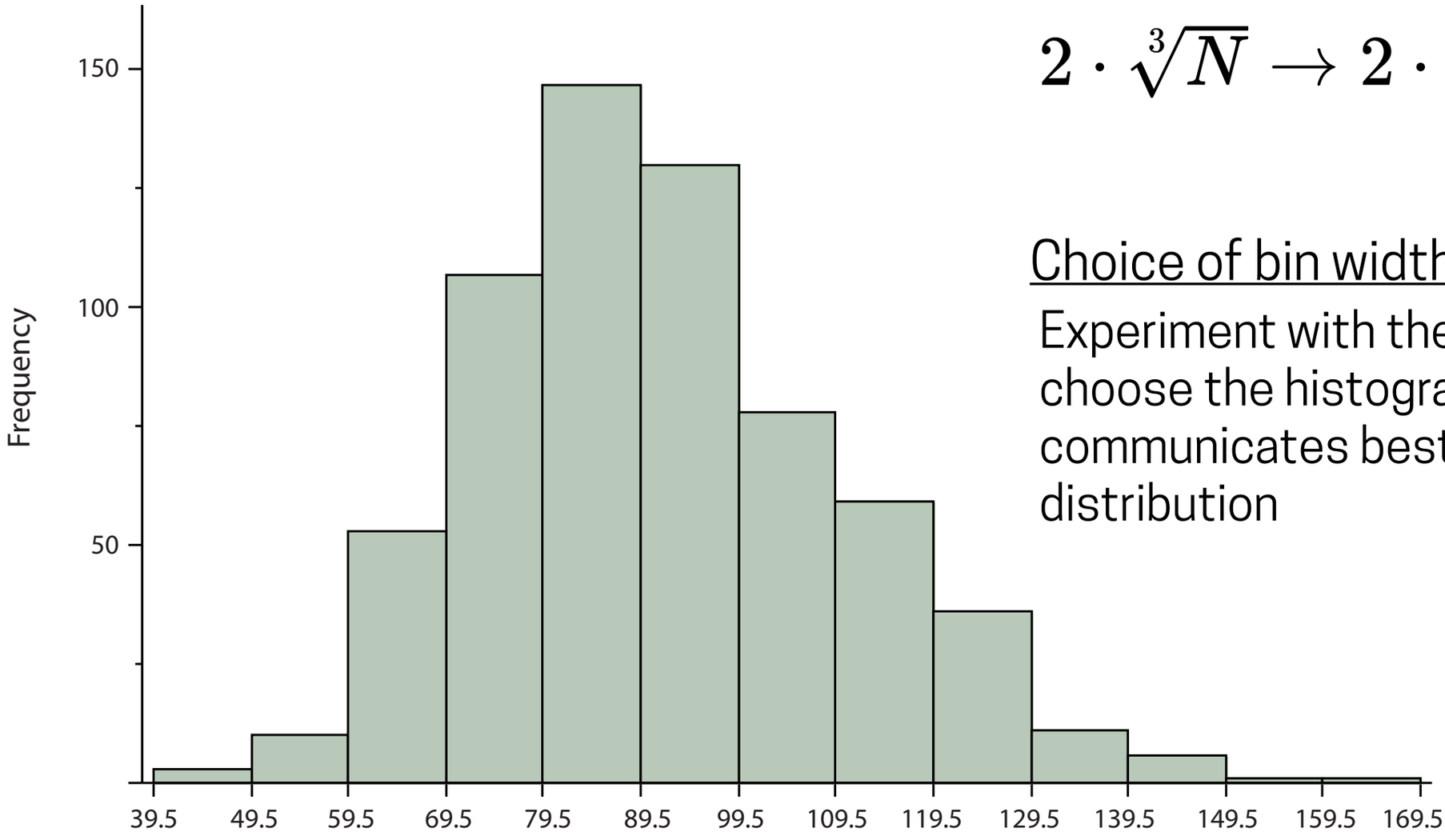


Figure 1. Histogram of scores on a psychology test.

Choice of bin number (Rice):

$$2 \cdot \sqrt[3]{N} \rightarrow 2 \cdot \sqrt[3]{1000} = 20$$

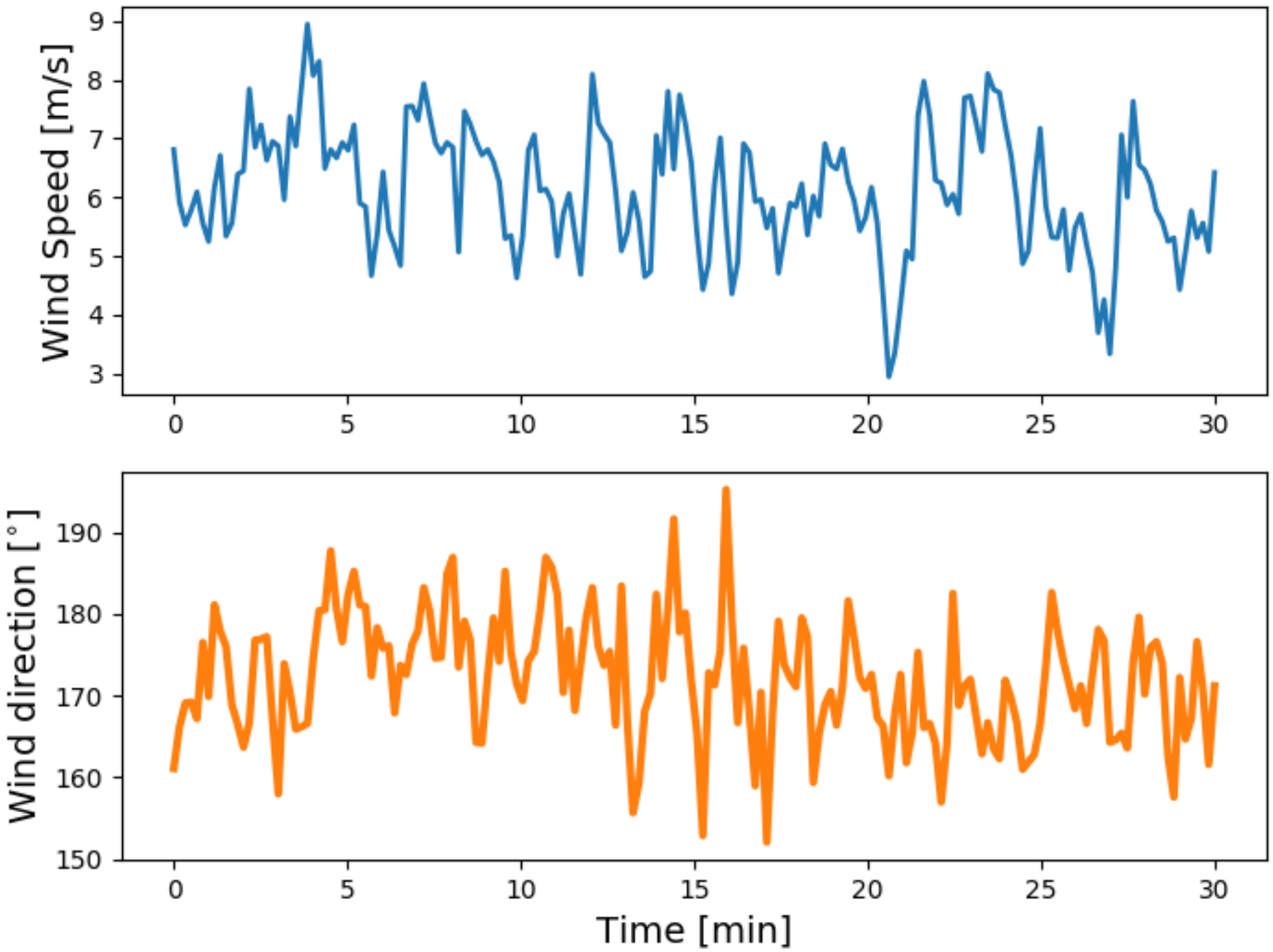
Choice of bin width:

Experiment with the data and width to choose the histogram that communicates best the shape of the distribution

Histograms

Now it is your turn to experiment. These are two wind measurement series with a frequency of 10s at 15m height.

What can your say already about the time series?



Practice

Go to Gitlab if you haven't already download the materials.

Download the scripts and data inside folder "lectureA1"

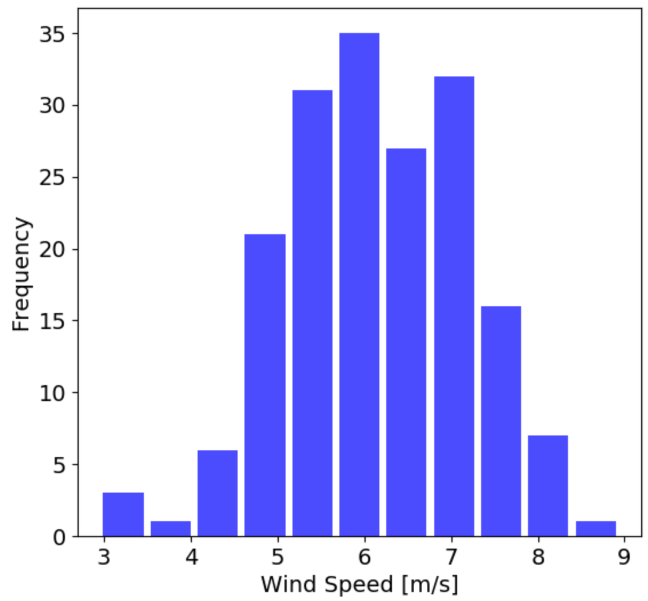
Try changing the number of bins in the histogram, use Rice's rule and then try to increase it or reduce it.

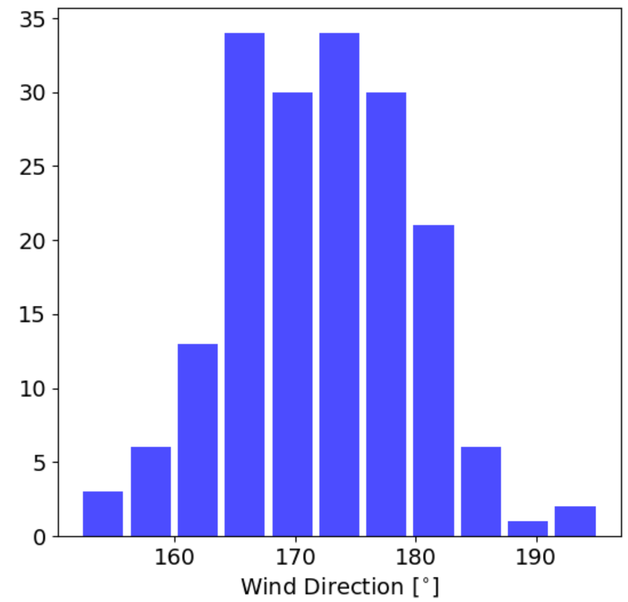
Comment with the class, what do you see? How many bins are necessary to extract conclusions about the data?

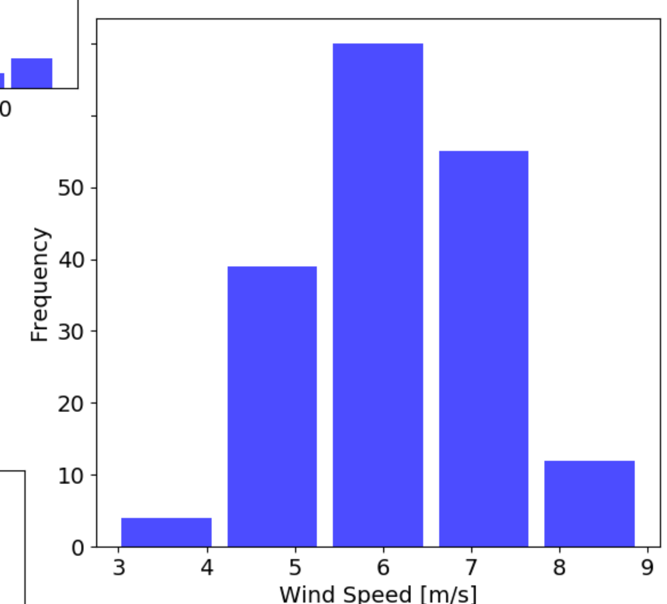
How can you normalise the frequency?

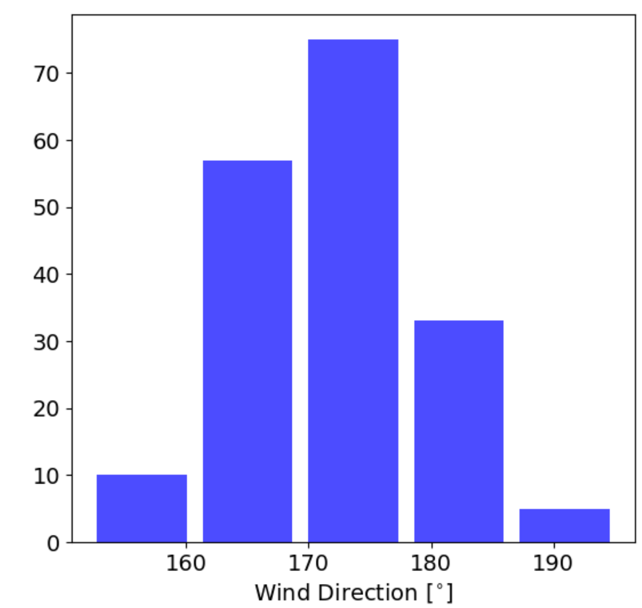
Histograms

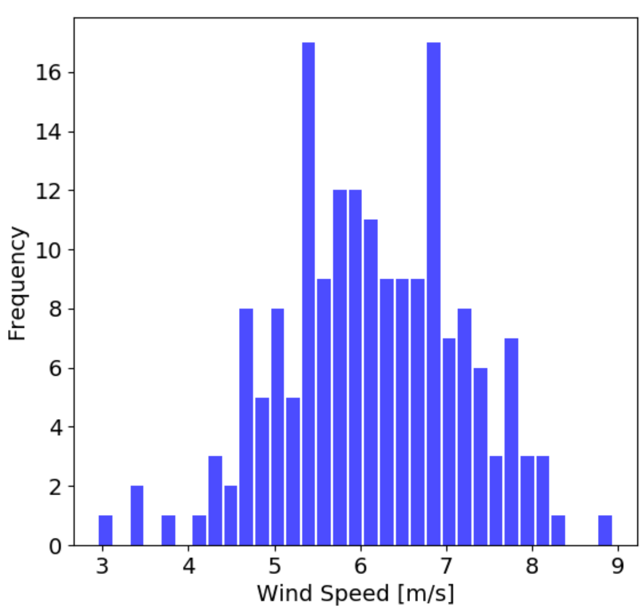
Now it is your turn to experiment. These are two wind measurement series with a frequency of 10s at 15m height.

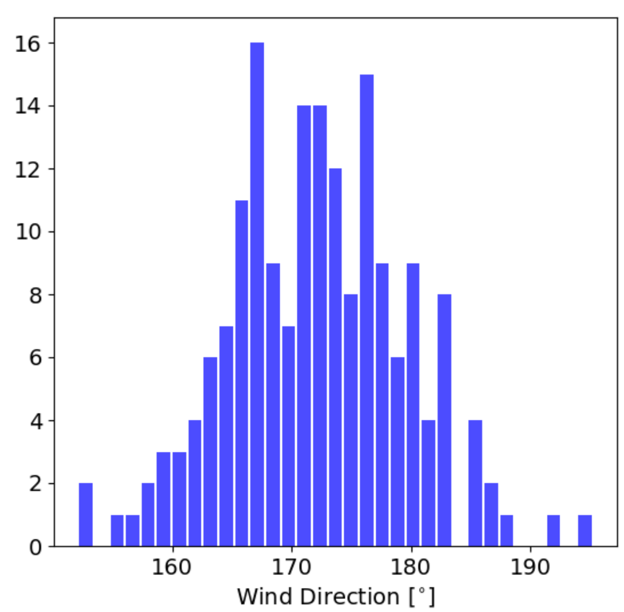








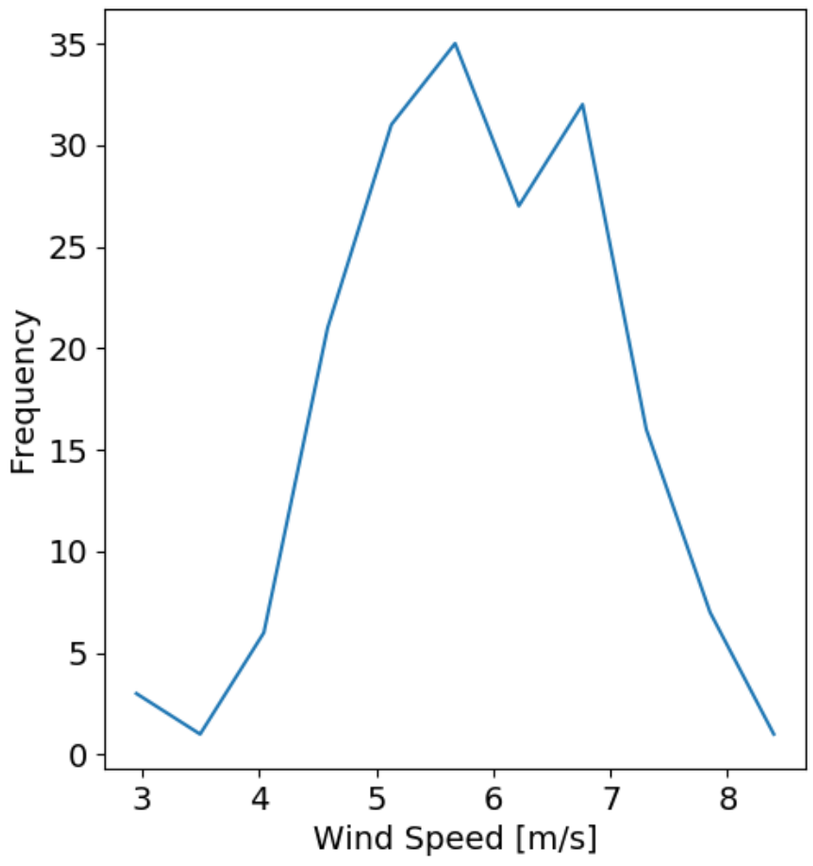


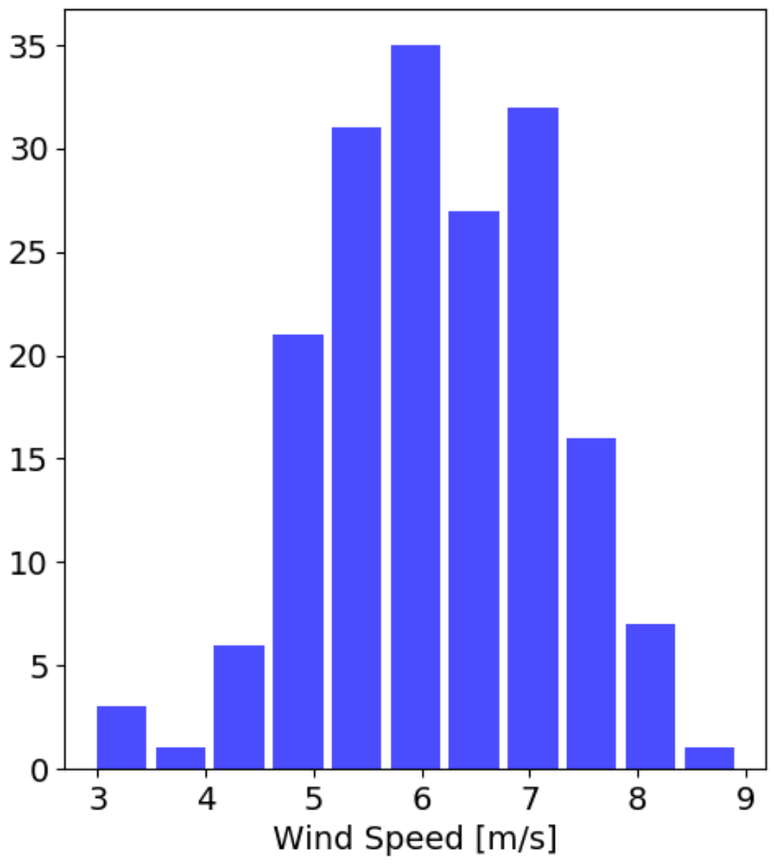


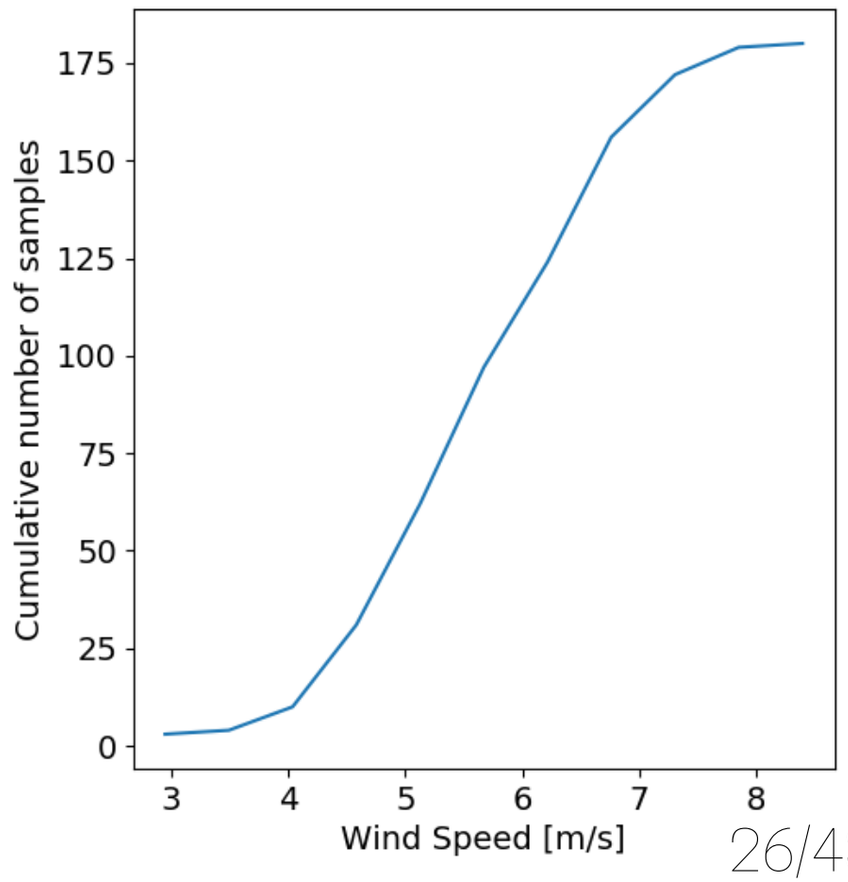
Histograms

We can also plot the envelope of the histogram as you can see in the leftmost plot

Cumulative density functions (cdf, rightmost plot) are very useful to determine locations for confidence intervals in statistics.







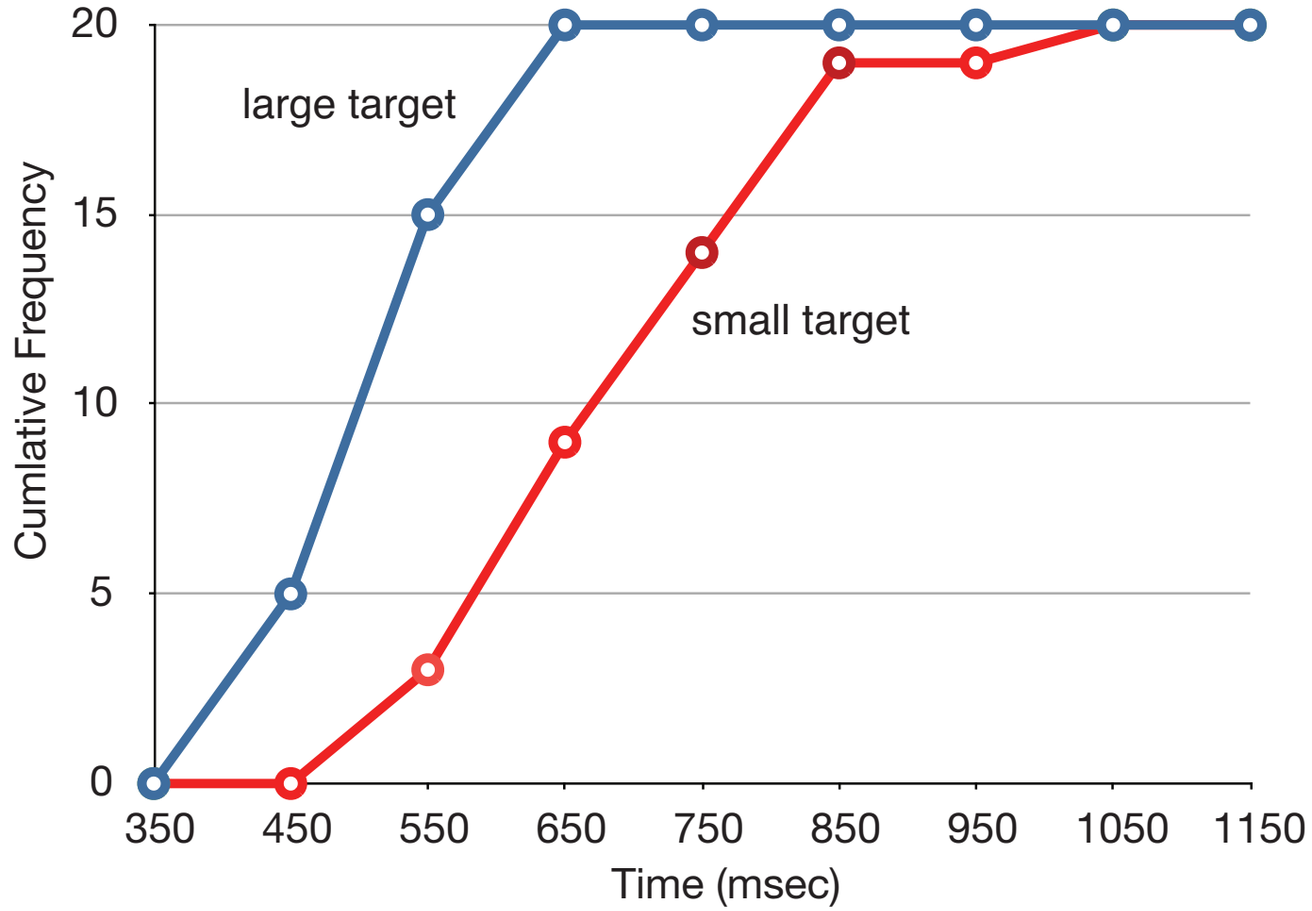
Overview

- Histograms
- **Frequency Polygons**
- Box Plots
- Line Graphs
- Scatter/dot plots

Frequency polygons

Graphical device for understanding the shapes of distributions, specially useful to compare diverse distributions.

They normally provide the same kind of information as histogram plots.



Overview

- Histograms
- Frequency Polygons
- **Box Plots**
- Line Graphs
- Scatter/dot plots

Box plots

Box plots are useful for **identifying outliers** and for **comparing distributions**

Steps to construct a box plot:

1. Compute 25th (Q1), 50th (Q2) and 75th (Q3) percentiles in the distribution scores.

2. Compute distance between Q3 and Q1 (Interquartile Range, IQR).

3. Compute the whiskers:

    3.1. Q1-1.5IQR and Q3+1.5IQR

    3.2. Find the largest value below upper whisker and smallest value above lower whisker

4. Compute outliers, values outside the whiskers

5. Add means in the plot

Box plots

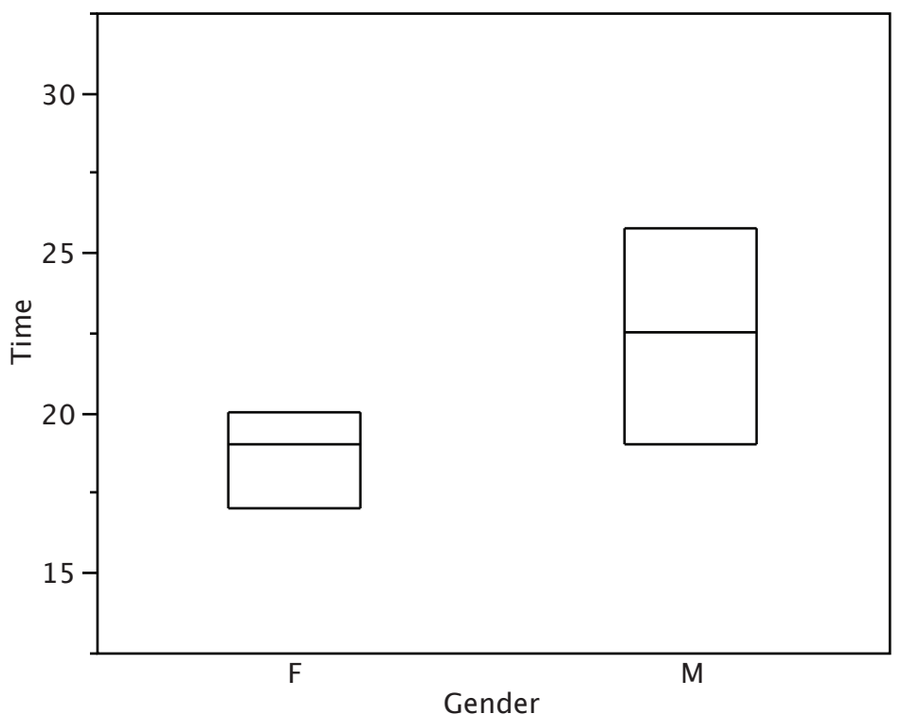
Table 1. Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

Times in seconds to identify distribution of colours in plot.

1. Compute 25th (Q1), 50th (Q2) and 75th (Q3) percentiles in the distribution scores.

$$R = \frac{P}{100} \cdot (N+1) \longrightarrow R(25) = 8, R(75) = 24 \longrightarrow Q1 = 17, Q3 = 20$$
$$Q2 = 19$$



Box plots

Table 1. Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

Times in seconds to identify distribution of colours in plot.

2. Compute distance between Q3 and Q1 (Interquartile Range, IQR).

$$IQR = 20 - 17 = 3$$

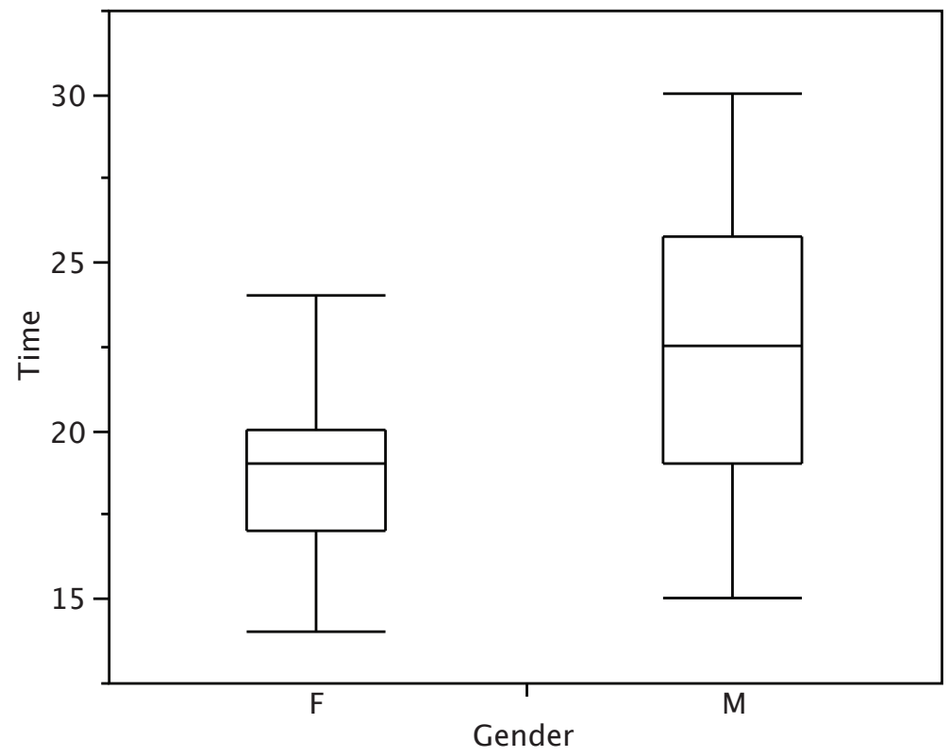
3. Compute the whiskers:

$$Q1 - 1.5IQR = 17 - 4.5 = 12.5$$

$$Q3 + 1.5IQR = 20 + 4.5 = 24.5$$

$$W1 = 14$$

$$W2 = 24$$



Box plots

Table 1. Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

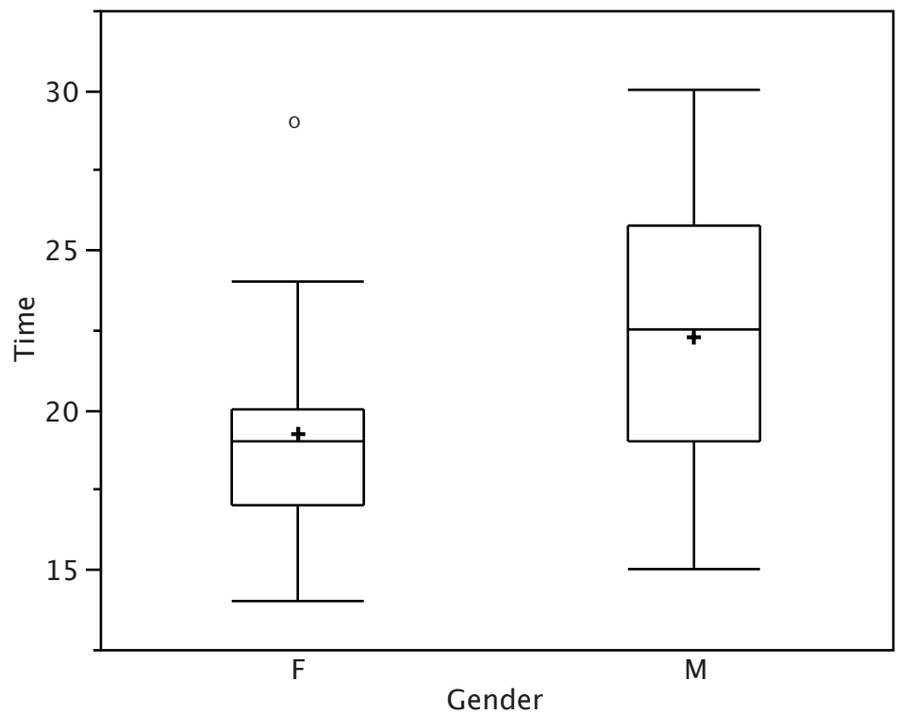
Times in seconds to identify distribution of colours in plot.

4. Compute outliers, values outside the whiskers

$$outlier = 29$$

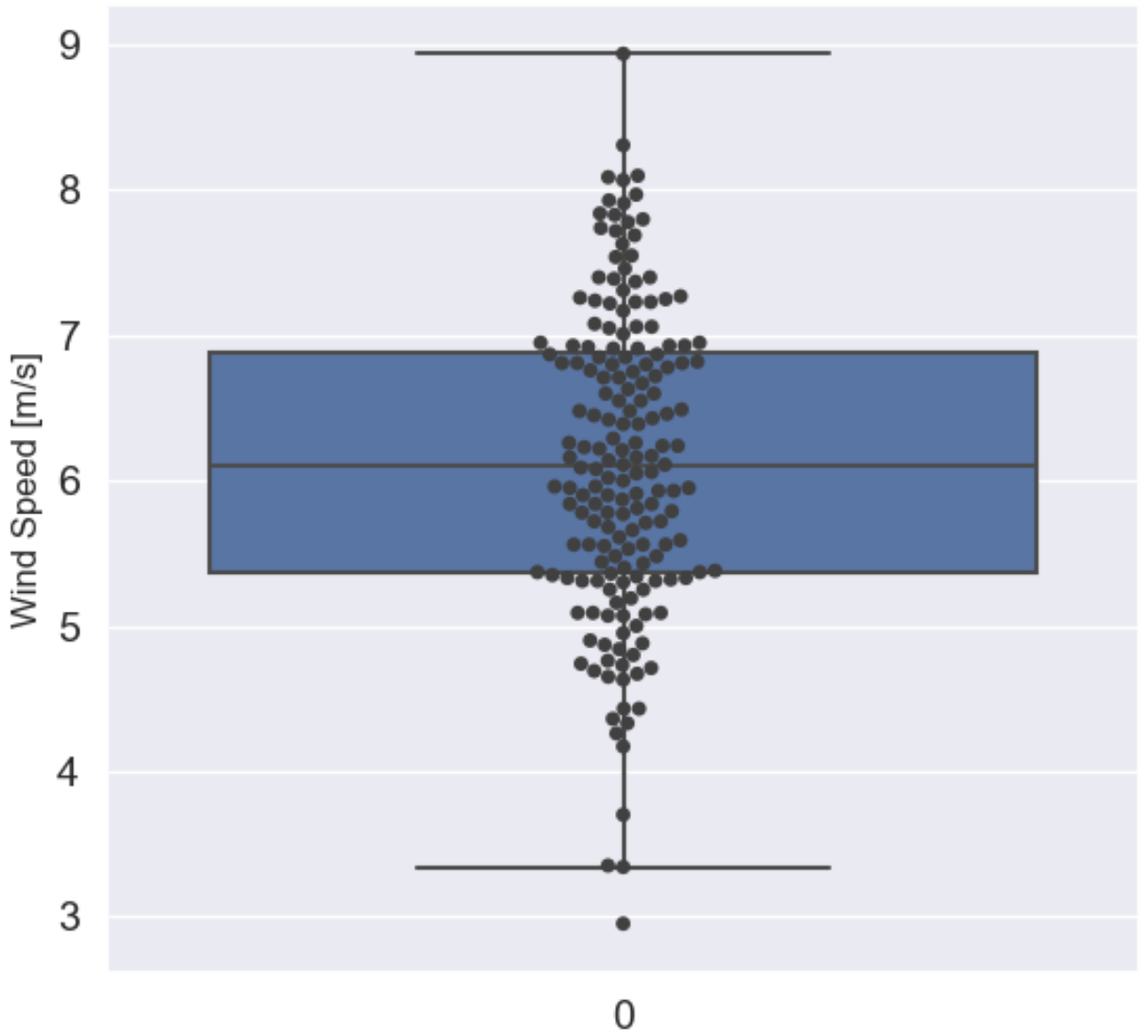
5. Add means in the plot

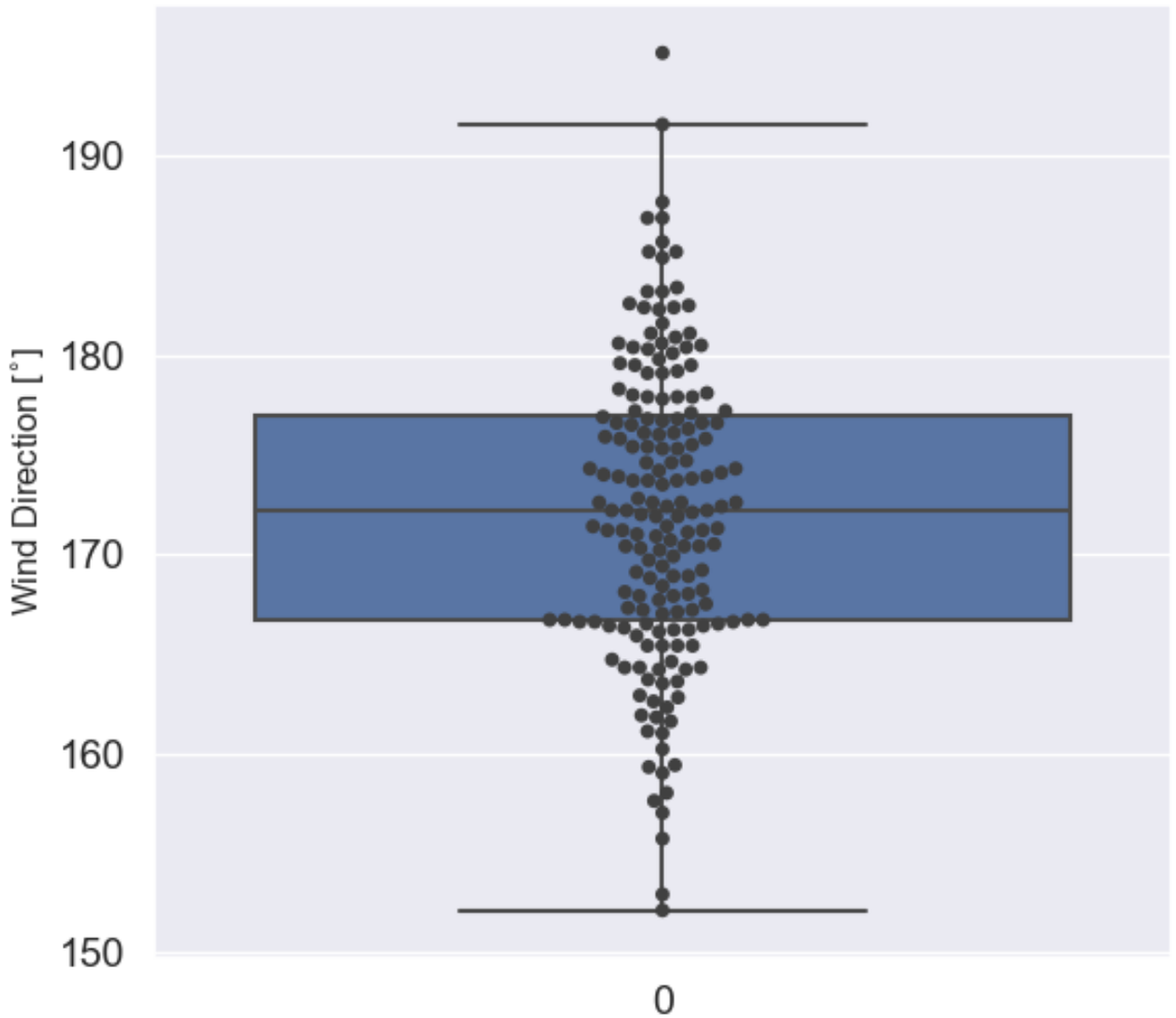
$$mean = \mu = 19.19$$



Practice

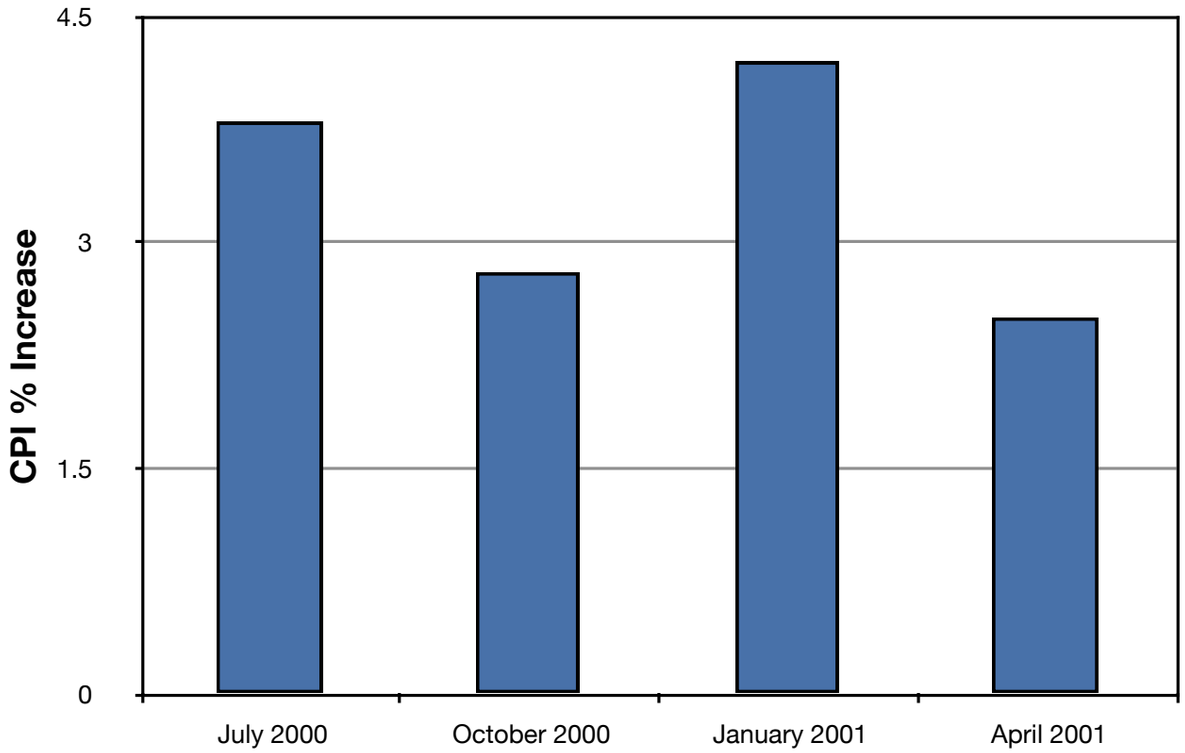
Using the wind data provided in the previous exercise, construct the box plots for the wind speed and wind direction variables.

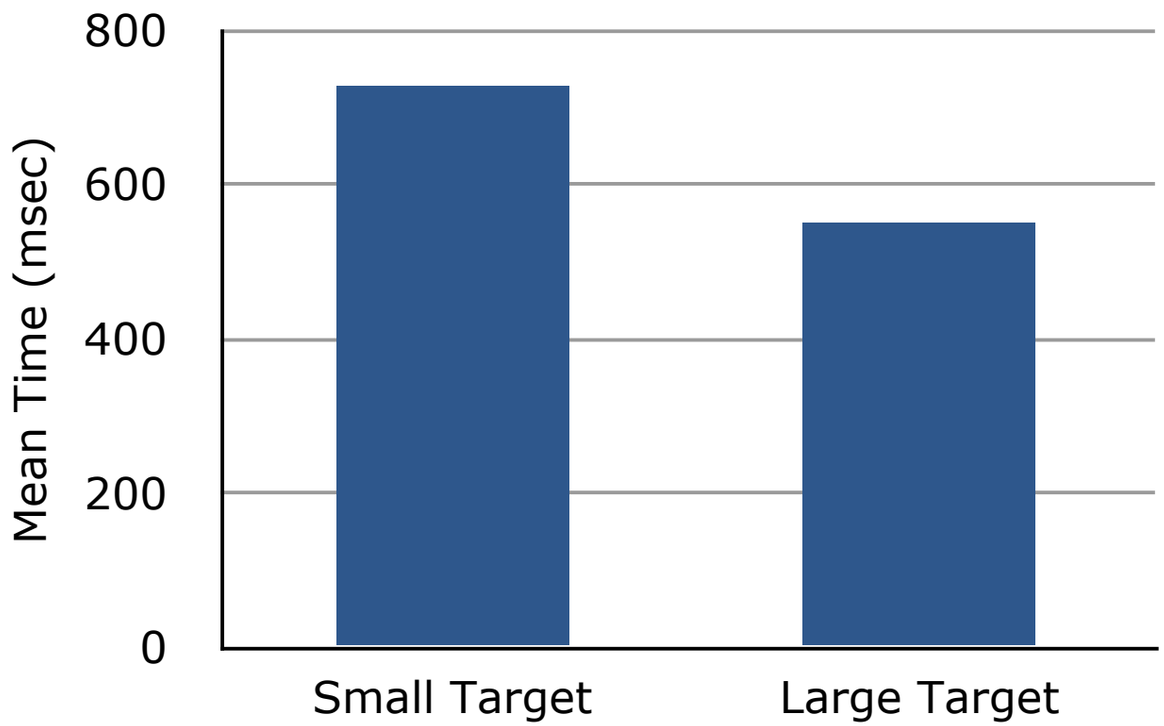




Bar charts

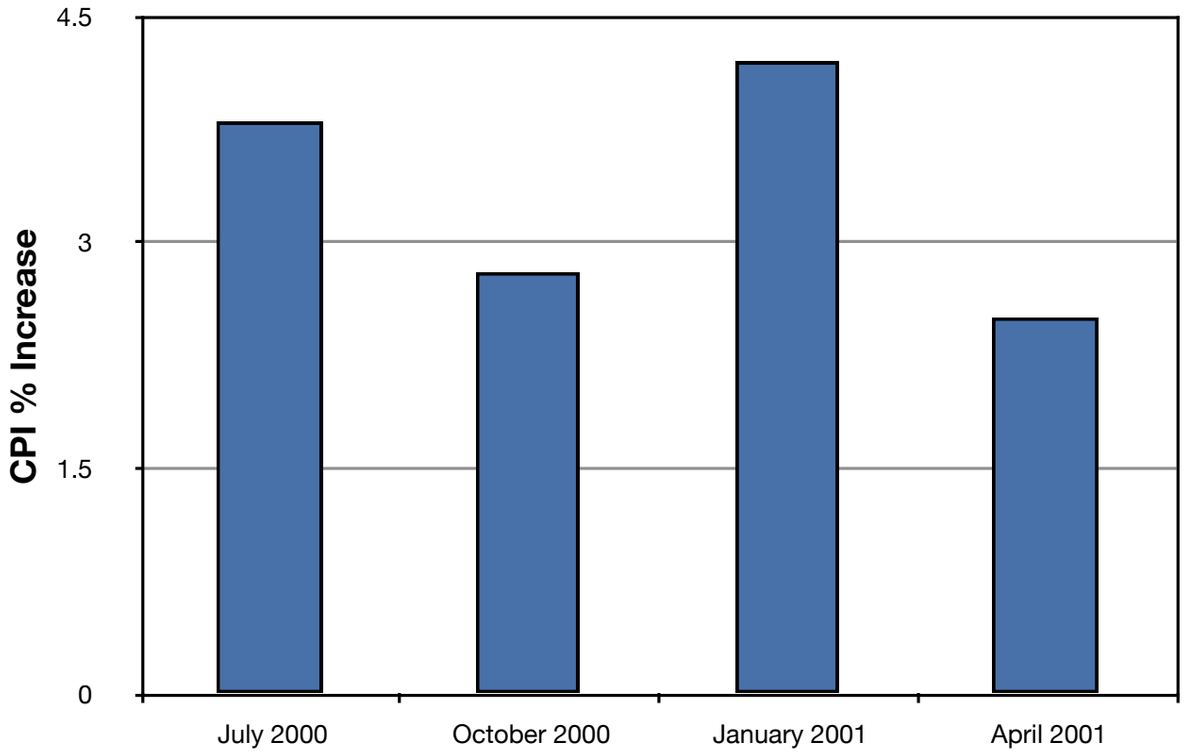
Bar charts are particularly effective for showing change over time, specially over long times.

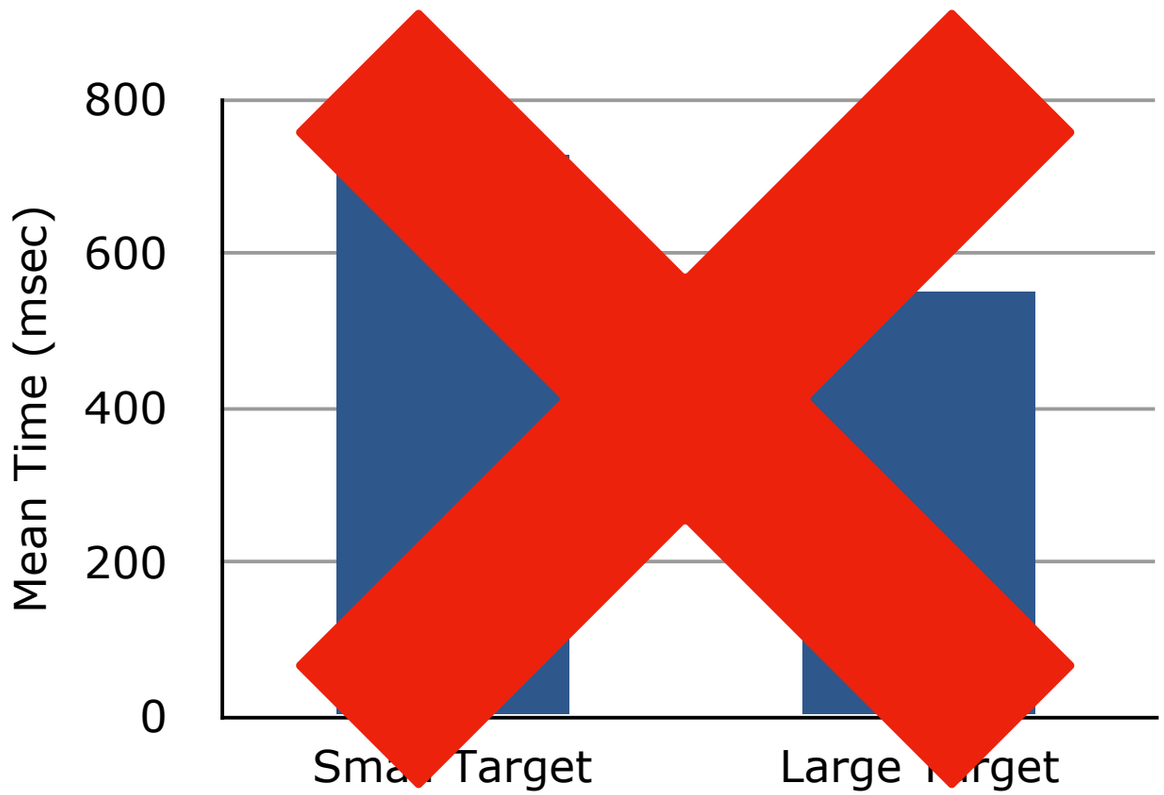




Bar charts

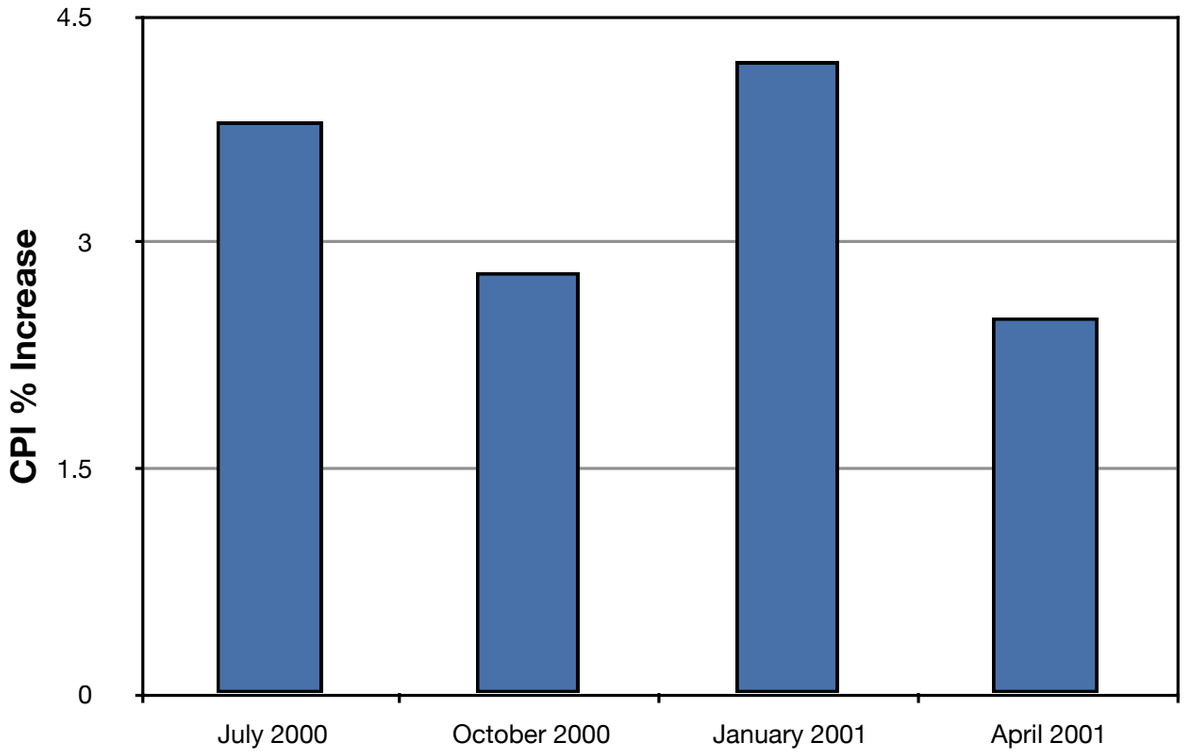
Bar charts are particularly effective for showing change over time, specially over long times.



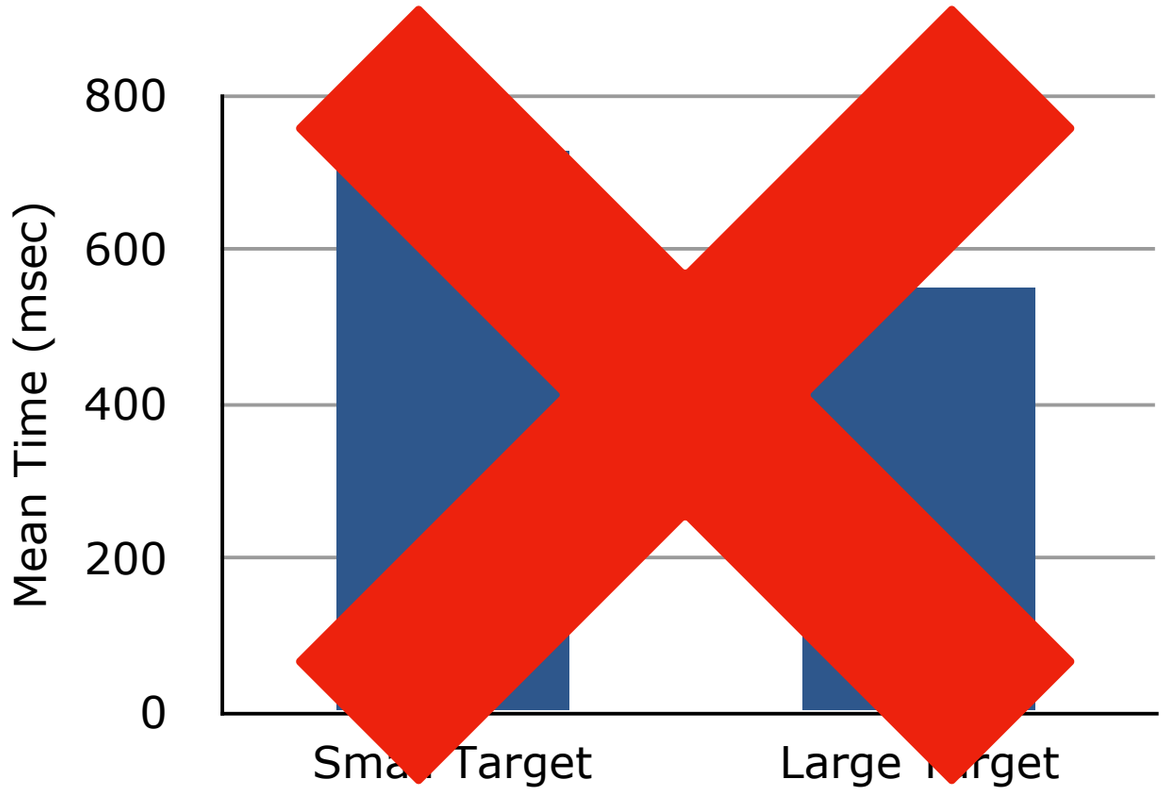


Bar charts

Bar charts are particularly effective for showing change over time, specially over long times.



OTHER PLOTS BETTER (BOX PLOTS)

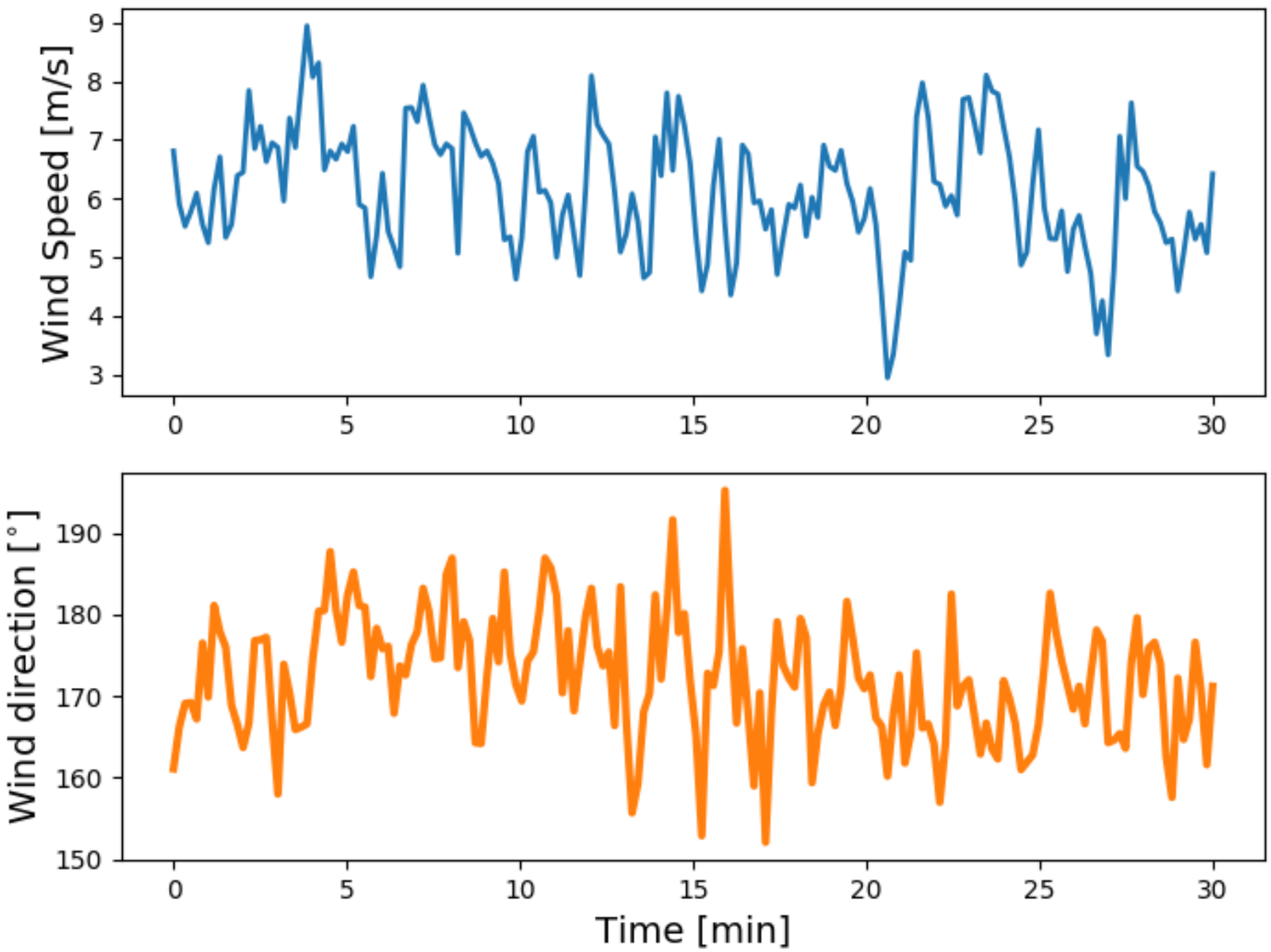


Overview

- Histograms
- Frequency Polygons
- Box Plots
- **Line Graphs**
- Scatter/dot plots

Line graphs

Line graphs are particularly effective for showing change over time as well, and they only make sense when both X and Y axes display ordered. For example for time series of measurements:



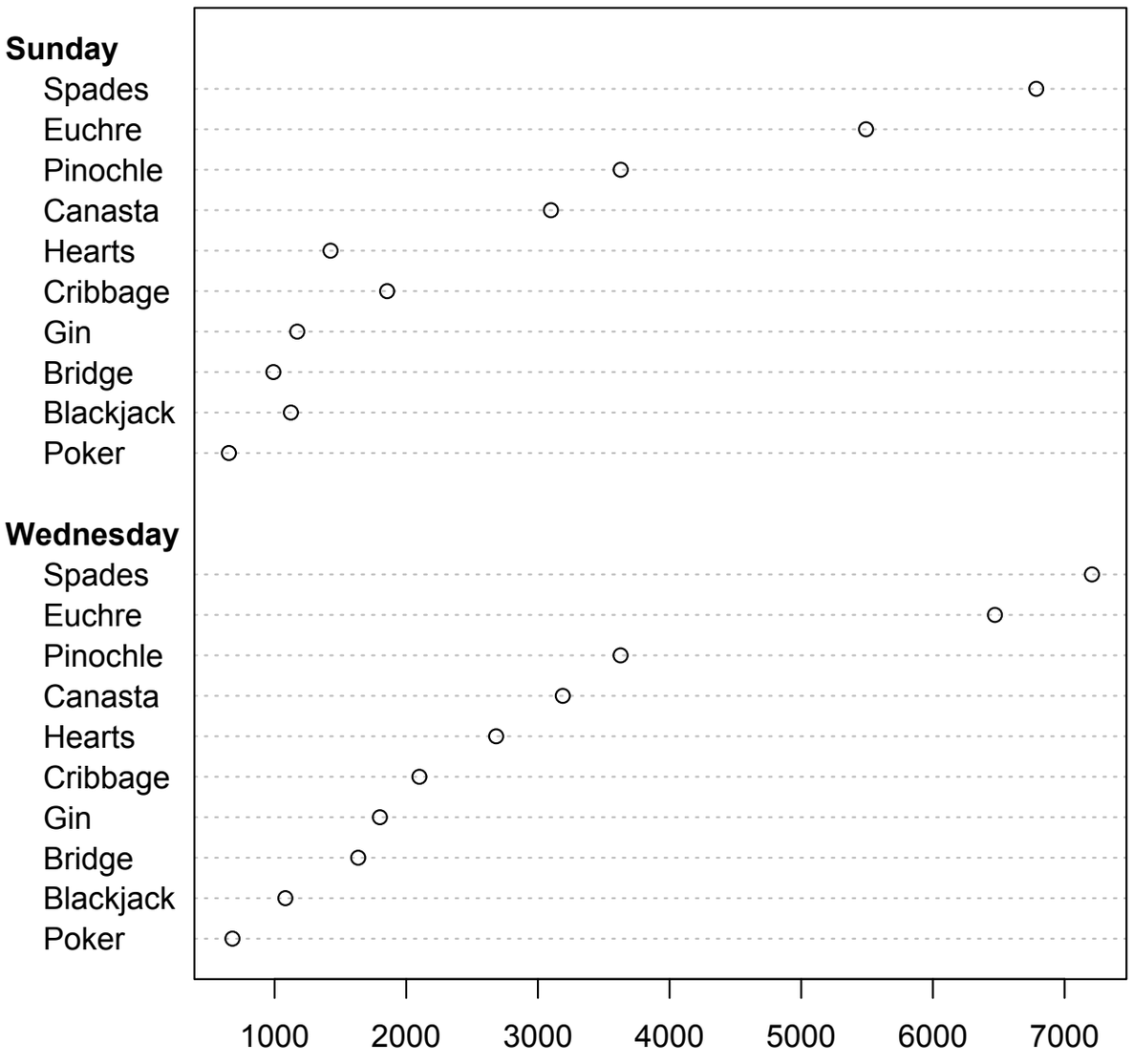
Overview

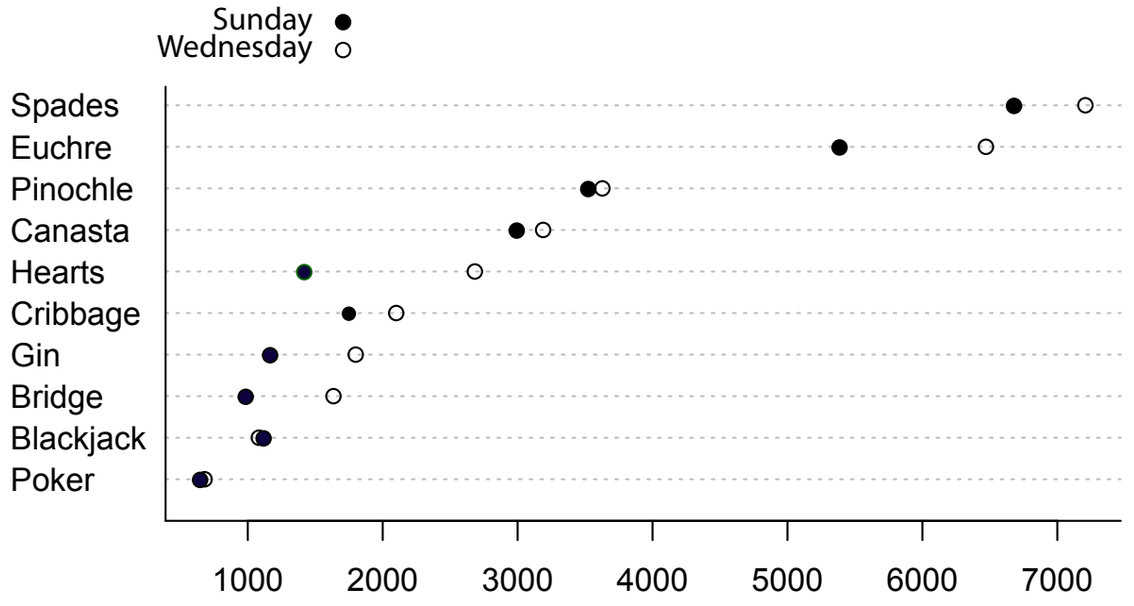
- Histograms
- Frequency Polygons
- Box Plots
- Line Graphs
- **Scatter/dot plots**

Scatter/dot plots

Scatter plots can be used in the variety of ways. When presenting experiments, the first plot normally used is typically a scatter point, since we perform discrete measurements most of the time.

It is very easy to combine data together through legends.



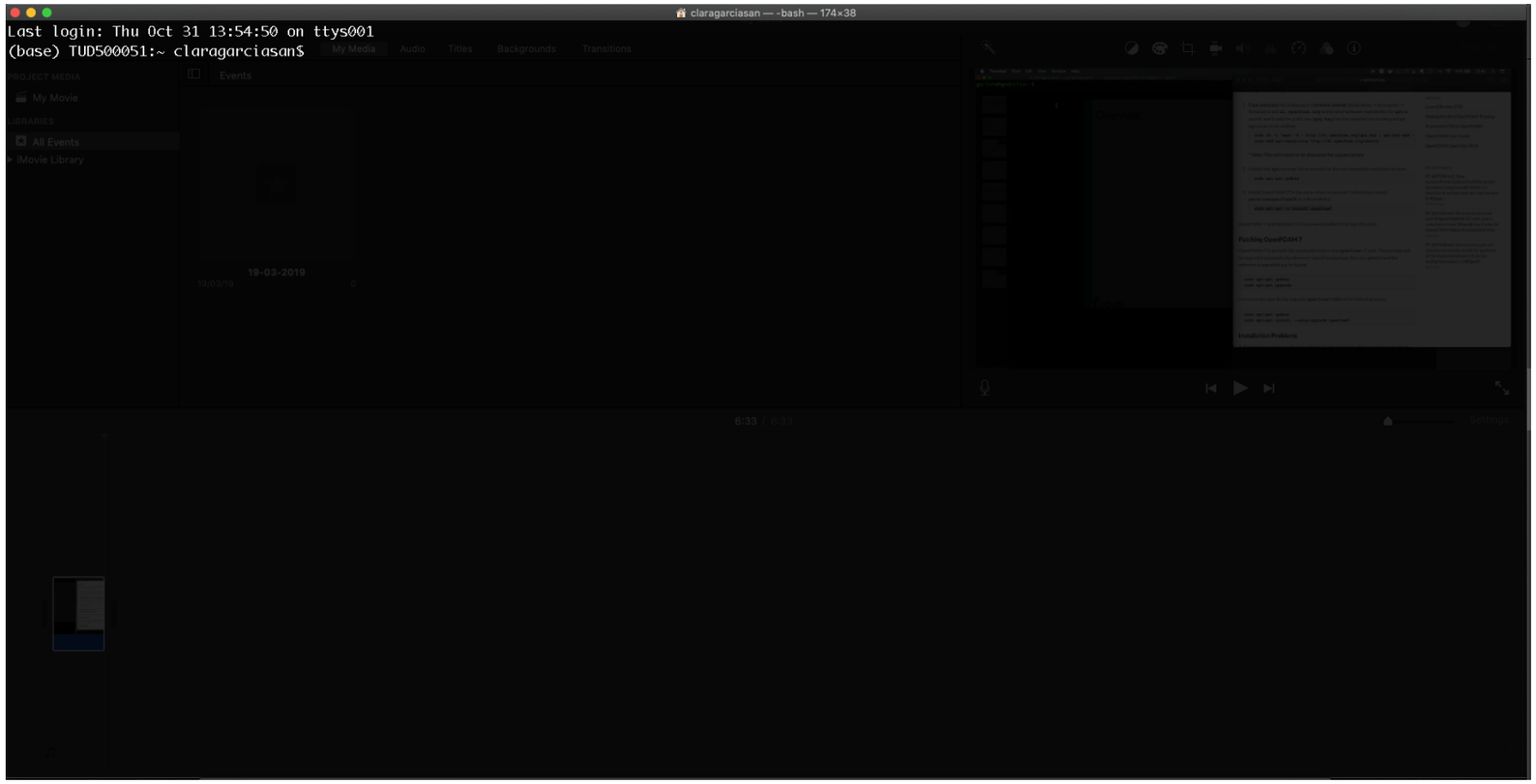


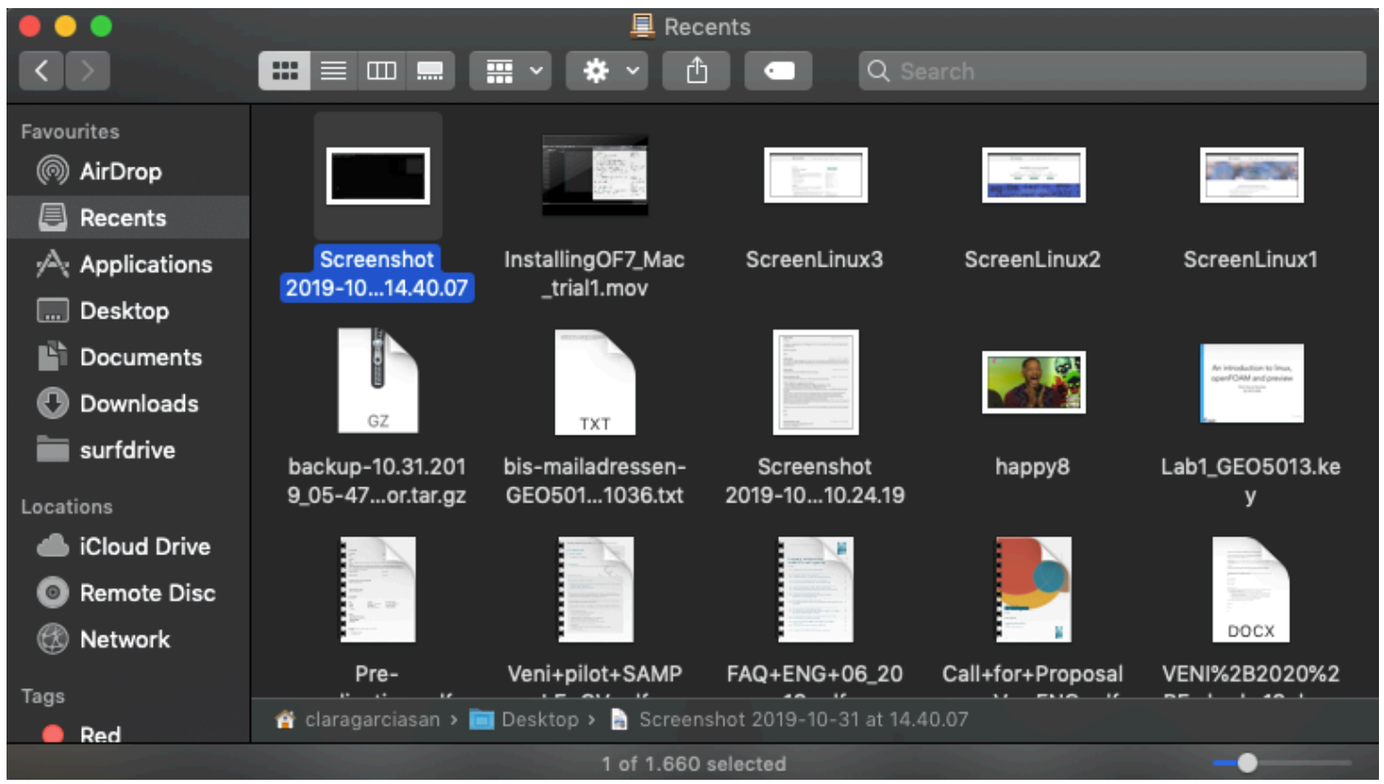
Lesson Z1

Linux basics & console

Does anyone worked with Linux before?

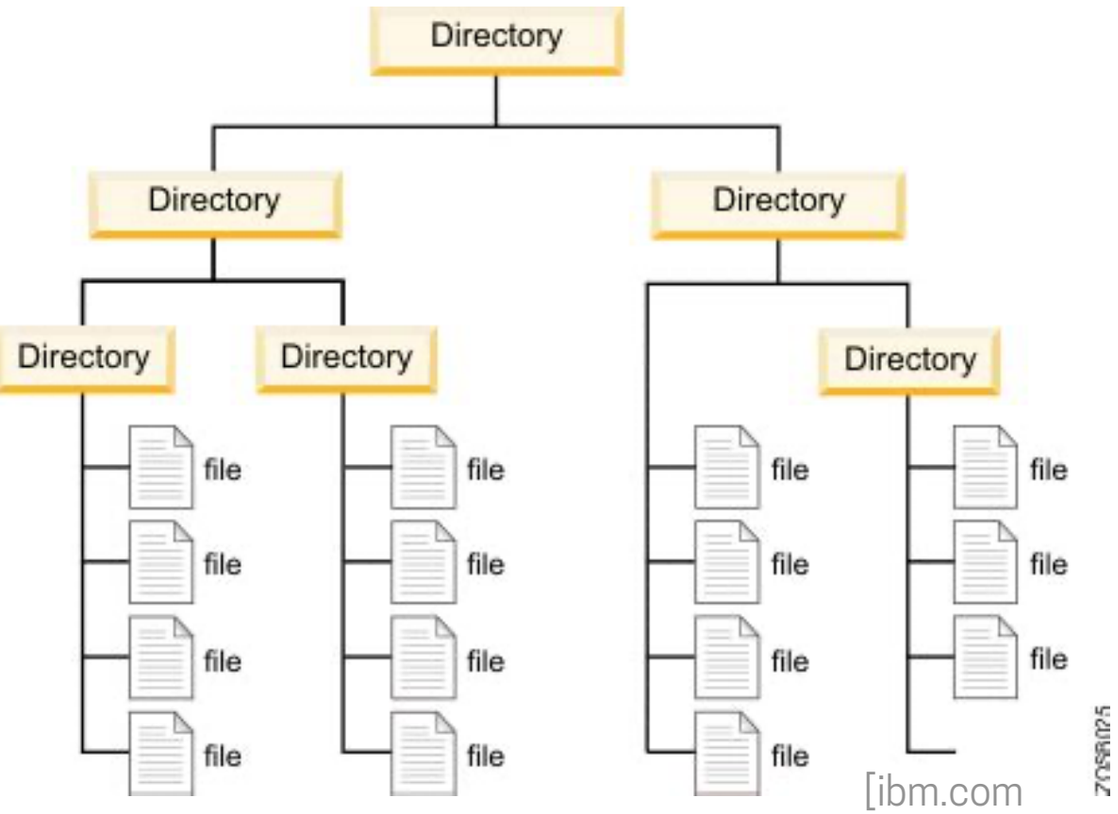
Linux basics & console



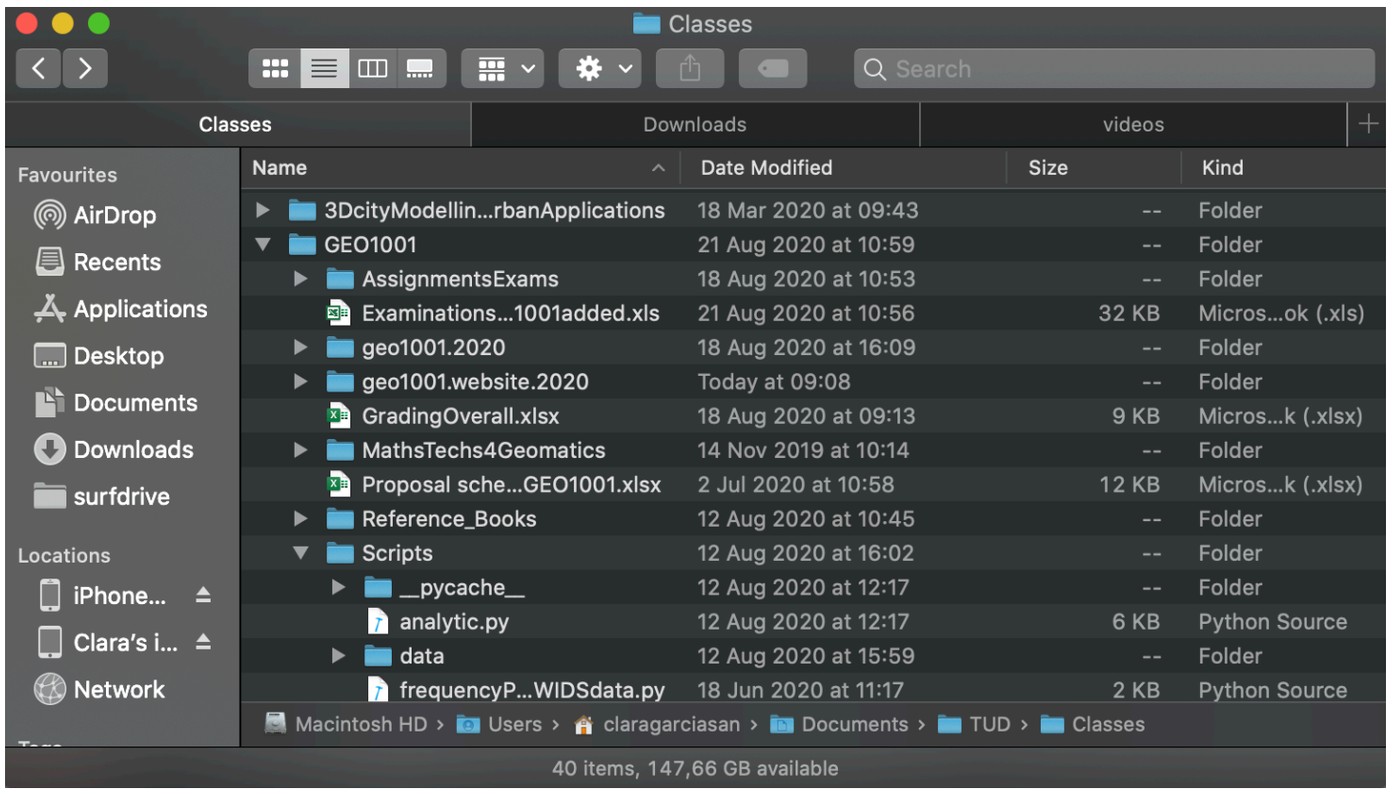


Linux basics & console

Files system is a tree:



[ibm.com



Linux basics & console

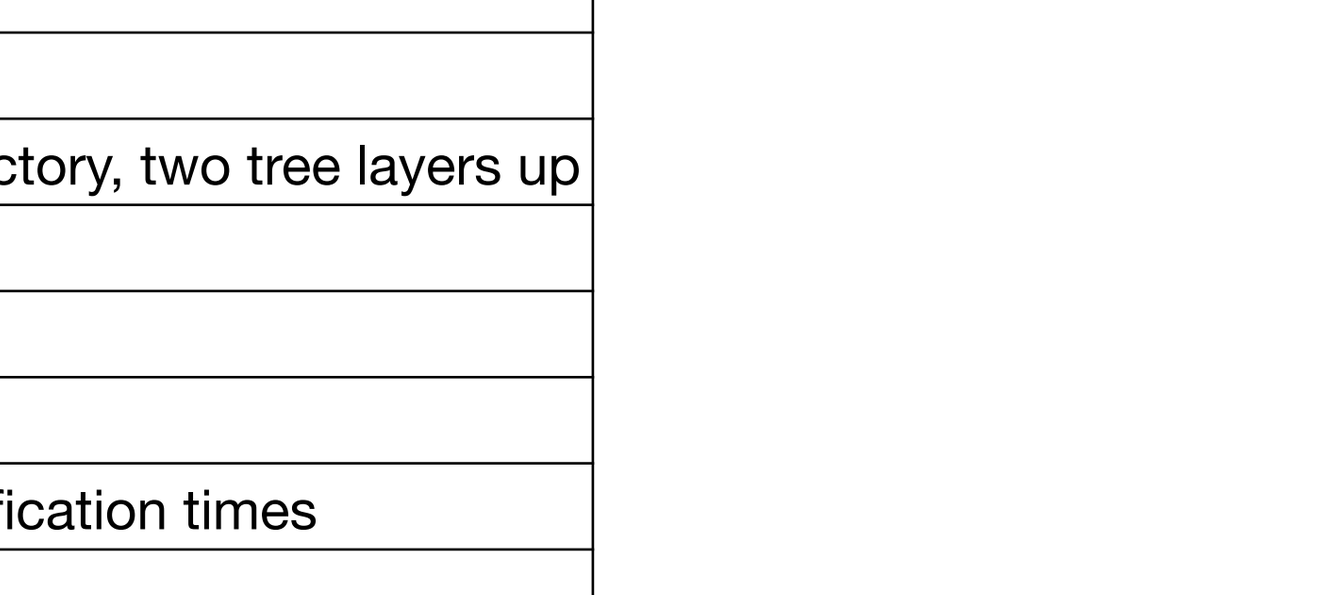
pwd	Returning working directory
ls	List directory contents
cd	Current directory
mkdir	Make directories
cd ../../	Going out from the current directory, two tree layers up
cat ./myfile.txt	Concatenate and print files
exit	Exit the shell
nano/vim	Command line text editors
touch myfile.txt	Change a file access and modification times
mv	Move files
rm	Remove directory entries
tldr	Shorter man
apt	Annotation processing tool

Ownership of files:

To see: ls -l

To change: chmod -+ u/g/a/o r/w/x

Example: chmod a-rw file1



makeuseof.com

Unix/Linux Command Reference

File commands

ls	Directory listing
ls -al	Formatted listing with hidden files
cd dir	Change directory to dir
cd	Change to home
pwd	Show current directory
mkdir dir	Create a directory dir
rm file	Delete file
rm –r dir	Delete directory dir
rm –f file	Force remove fire
rm –rf dir	For remove directory dir
cp file1 file2	Copy file1 to file2
cp –r dir1 dir2	Copy dir1 to dir2; create dir2 if it doesn't exit
mv file1 file2	Rename or move file1 to file2. If filed 2 is an existing directory, moves file1 into directory file2
ln –s file link	Create symbolic link link to file
touch file	Create or update file
cat > file	Places standard input into file
more file	Output the contents of file
head file	Output the firest 10 lines of file
tail file	Output the last 10 lines of file
tail –f file	Output the contents of file as it grows, starting with the last 10 lines

Process Management

ps	display all currently active processes
top	display all running processes
kill pid	kill process id pid
killall proc	kill all processes named proc *
bg	lists stopped or background jobs; resume a stopped job in the background
fg	Brings the most recent job to the foreground
fg a	brings job a to the foreground

File Permissions

chmod octal file	change the permissions of file to octal, which can be found separately for user, group, and world by adding: • 4 – read (r) • 2 – write (w) • 1 – execute (x) **Examples:** ***chmod 777*** *– read, write, execute for all* ***chmod 755*** *– rwx for owner, rx for group and world. For more options, see* **man chmod.**

SSH

ssh user@host	connect to host as user
ssh -p port user@host	connect to host on port port as user
ssh-copy-id user@host	add your key to host tor user to enable a keyed or passwordless login

Searching

grep pattern files	search for pattern in files
grep -r pattern dir	search recursively for pattern in dir
command | grep pattern	search for pattern in the output of command
locate file	find all instances of file

System Info

date	show the current date and time
cal	show this month's calendar
uptime	show current uptime
w	display who is online
whoami	who you are logged in as
finger user	display information about user
uname -a	show kernel information
cat /proc /cpuinfo	cpu information
cat /proc /meminfo	memory information
man command	show the manual for command
df	show disk usage
du	show directory space usage
free	show memory and swap usage
whereis app	show possible locations of app
which app	show which app will be run by default

Compression

tar cf file.tar files	create a tar named file.tar containing files
tar xf file.tar	extract the files from file.tar
tar czf file.tar.gz files	create a tar with Gzip compression
tar xzf file.tar.gz	extract a tar using Gzip
tar cjf file.tar.bz2	create a tar with Bzip2 compression
tar xjf file.tar.bz2	extract a tar using Bzip2
gzip file	compresses file and renames it to file.gz
gzip -d file.gz	decompresses file.gz back to file

Network

ping host	ping host and output results
whois domain	get whois information for domain
dig domain	get DNS information for domain
dig -x host	reverse lookup host
wget file	download file
wget -c file	continue a stopped download

Installation

Install from source:	
./configure	
make	
make install	
dpkg -i pkg.deb	install a package (Debian)
rpm -Uvh pkg.rpm	install a package (RPM)

Shortcuts

Ctrl+C	halts the current command
Ctrl+Z	stops the current command, resume with fg in the foreground or bg in the background
Ctrl+D	log out of current session, similar to exit
Ctrl+W	erases one word in the current line
Ctrl+U	erases the whole line
Ctrl+R	type to bring up a recent command
!!	repeats the last command
exit	log out of current session
*	use with extreme caution

Linux basics & console

Linux Cheat sheet
by A.Mahouachi

1 File Commands

ls [options] file

options

-a: show hidden files
-A: show hidden files except . and ..
-d: only show directories
-h: human readable size
-i: inode info
-l: long list format
-m: output as csv
-n: numeric uid and guid
-r: sort in reverse order
-S: sort by file size
-t: sort by modification time

tree [options] dir

options

-d: only directories
-f: show full paths
-P pattern: only matching pattern
-I pattern: except matching pattern
-h: print sizes in human readable format
-C: use colors
-L max: max level depth

cp [options] source dest

options

-b: backup dest before overwrite
-r: recursive
-f: force
-l: link files instead of copy
-P: dont follow sym links
-i: interactive
-u: copy only if source newer than dest

mv [options] source dest

options

-b: backup dest before overwrite
-f: force
-i: interactive
-u: move only if source newer than dest

ln [options] file link

options

-s: sym link (hard by default)
-f: overwrite link if exists
-b: backup old link before overwrite

rm [options] file

options

-f: force
-i: interactive
rm – -foo if file name is -foo

chmod [options] mode file(s)

options

-R: recursive

symbolic mode

format: [ugoa][[+-=][perms]],...
example: u+x,o-wx,g-w
u: owner
g: group
o: others
a: all
+: add mode
-: remove mode
=: exact mode
r: read
w: write
x: execute files and search for dirs
X: search for dirs
s: setuid or setguid
t: sticky bit

numeric mode

format: [0-7]1,4
example: 755
first digit: setuid(4), setguid(2)
second digit: owner perms
third digit: group perms
fourth digit: others perms
read: 4
write: 2
execute: 1

find path [options] [tests] [actions]

options:

-mindepth: start from min level in hierarchy
-maxdepth: end with max level in hierarchy

tests:

-name "xyz*": name like xyz*
-iname "xyz*": like -name but case insensitive
-type d: only directories
-type f:only files
-mtime 0: modified < 1 day
-mtime -x: modified < x days
-mtime +x: modified > x days
-mmin: like -mtime but in minutes
-size +100M: size > 100mb
-size -100M: size < 100mb (k for kb, G for gb)
-perm /o+w: writable by others
! -perm /o+r : not readable by others

actions:

-print: print matching
-delete: rm matching files
-exec cmd '{}' ; : run cmd for every match
-exec cmd '{}' + : run cmd at the end of search
-exec rm -rf '' : rm -rf matching items
-fprint /tmp/result: write matches to /tmp/result

diff [options] files

options

-r: recursive
-w: ignore whitespaces
-B: ignore blank lines
-q: only show file names
-x".sync*": exclude files with path like .sync*

grep [options] pattern files

options

-i: ignore case
-P: pattern is a perl regex
-m: stop after m matches
-n: also show matching line number
-R: recurse directories
-c: only show matching lines count
–exclude=glob : exclude these
–include=glob : only consider these

cat [options] file(s)

options

-v: non ascii chars except tab and eol
-T: show tabs
-t: equivalent to -vT
-E: show eol end of line
-e: equivalent to -vE
-A: equivalent to -vET
-s: remove repeat empty lines

tail [options] file

options

-f: show end of file live
-35: show last 35 lines
-q: be quiet

head [options] file

options

-35: show first 35 lines
-q: be quiet

tac file(s)

print files starting from last line

cut [options] file

options

-d char: use char as delimiter
-f 1,3,5: print fields 1, 3 and 5

uniq [options] input output

options

-c: prefix lines by number of occurrences
-d: only print duplicate lines
-u: only print unique lines

sort [options] file

options

-n: numeric sort
-b: ignore blank lines
-f: ignore case
-r: reverse order

tar [options] file

options

-f file: archive file
-c: create
-t: list
-x: extract
-C DIR: cd to DIR
-z: gzip
-j: bzip2

du [options] file

options:

-c: a grand total
-h: human readable
-L: dereference sym links
-P: no dereference of sym links
-s: total for each argument
–exclude=pattern
-max-depth=N: dont go deeper than N

df [options] file

options:

-h: human readable
-i: list inodes info
-P: no dereference of sym links

2 Process Commands

ps [options]

options:

-e: all processes
-f: full listing
-H: show hierarchy
-p pid: this process pid
-C cmd: this command name cmd
-w: wide output
-ww: to show long command lines
-l: long listing, including wchan
-o x,y,z: show columns x y z
-o user,pid,cmd: show columns user, pid command
-N: negation
-u user: processes owned by user
-u user -N: processes not owned by user
–sort=x,y: x y are columns in ps output
–sort=user: sort by user
–sort=+time: sort by cpu time asc
–sort=-time: sort by cpu time desc
–sort=size: sort by memory size
–sort=vsize: sort by vm size

top [options]

options

-d x: refresh every x seconds
-p pid1 -p pid2: only processes with pid1 pid2
-c : show command lines

interactive commands

space: udpdate display
n: change number of displayed processes
up and down: browse processes
k: kill a process
o: change order
T: sort by time
A: sort by age
P: sort by cpu
M: sort by memory
c: display/hide command line
m: display/hide memory
t: display/hide cpu
f: manage list of displayed columns
up and down: move between columns
d: display/hide the selected column
q: apply and quit the field mgmt screen

pgrep [options] pattern

options

-l : show pid and process name
-a : show pid and full command line
-n : if more than one show newest
-o : if more than one show oldest
-u uid : show only processes of uid
-c : count results

3 Network & Remote

ssh [options] user@host ["cmd1;cmd2"]

options:

-2: force protocol 2
-o StrictHostKeyChecking=no: ignore warnings due to remote host key change
-X: forward X11 display

wget [options] url

options:

-b: run in background
-o file: print wget output in file
-o /dev/null: suppress wget output
-q: be quiet
-d: debug
-O file: save response to file
-c: resume file download
-S: print server headers
-T N: timeout after N seconds
–user=user: basic http auth user
–password=password: basic http auth password
–save-cookies file: save cookies to file
–load-cookies file: use file as cookies
–post-data=string
–post-file=file
–no-check-certificate: ignore ssl certificate

curl [options] url

options:

-H header: like -H "Host: st.com"
-u <user:password>: basic http auth
-s: be silent
-S: show errors if silent mode
-L: follow new location in case 301
–data "field=value": x-www-form-urlencoded query
–data-binary data: post data as is without encoding
–data-binary @filename: post filename content as is
-X method: use PUT, GET, POST etc.
–request method: use PUT, GET, POST etc.

mail [options] to-address

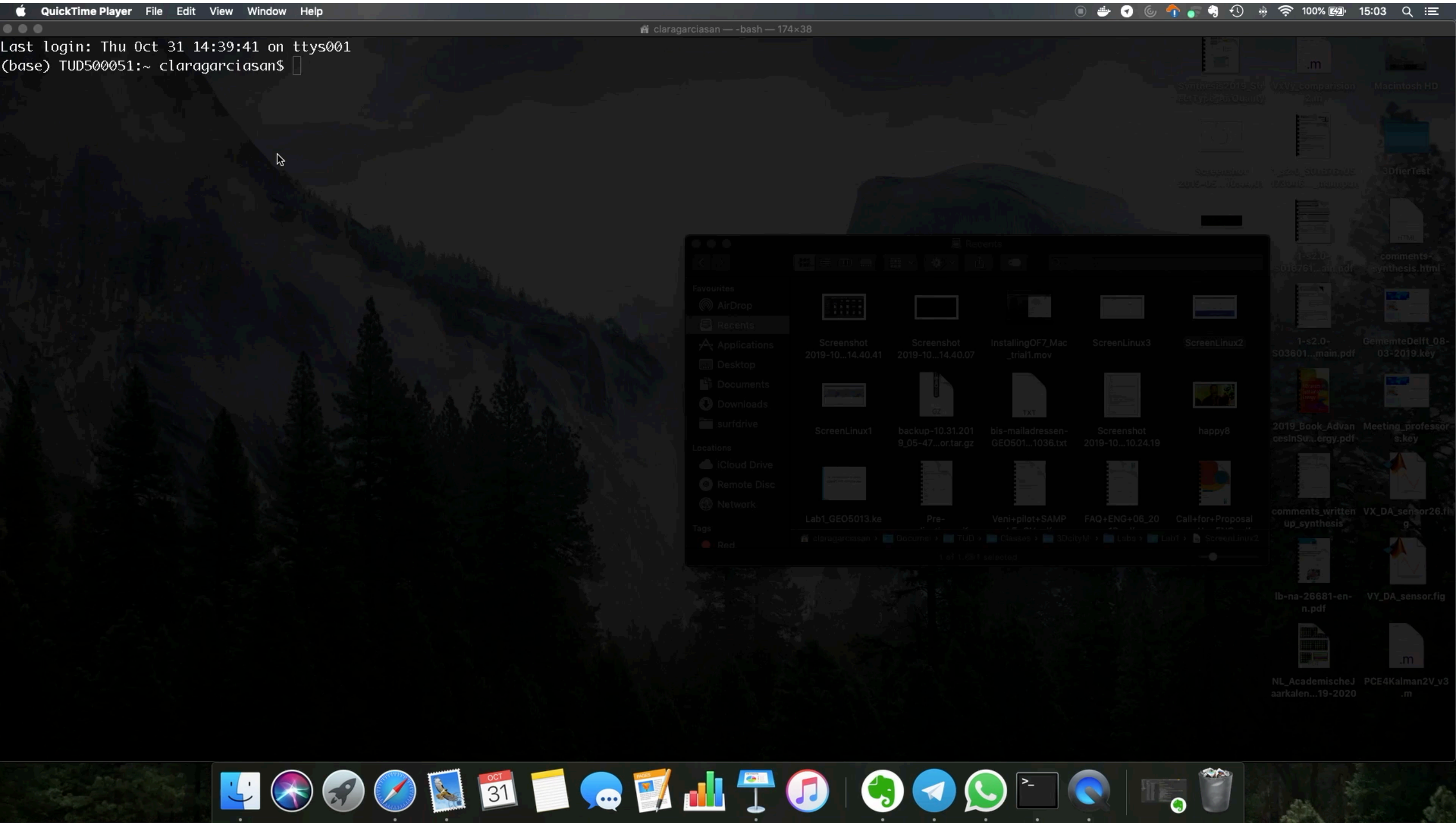
options:

-s subject: email with subject
-c address1,address2: cc copy
-b address1,address2: bcc copy
mail -s 'hello there' 'joe@st.com' < somefile

4 Terminal

Ctrl+C: halt current command
Ctrl+Z: pause current command
bg %1: resume paused command in background
fg %1: resume paused command in foreground
Ctrl+D: logout
Ctrl+W: remove a word from current line
Ctrl+U: remove current line
Ctrl+A: go to beginning of current line
Ctrl+E: go to end of current line

Linux basics & console



Linux basics & console