Data Visualization

`geom_hist and geom_density`	distribution of numerical columns
`geom_bar`	number of occurences in a categorical col
`geom_boxplot`	shape & distribution of numerical vars
`geom_scatter + geom_line*`	numerical vs. numerical
`geom_bar`	bar plot for count of categorical vars
`geom_hline(yintercept)`	horizontal line
`geom_vline(xintercept)`	vertical line
`geom_abline(slope, intercept)`	linear function, requires
`geom_segment`	straight line between (x, y) and (xend, yend)
`geom_smooth`	plots a line/curve of best fit

*`geom_line` only makes sense with an ordering (e.g. the x-axis is year and observations connect together)

Data Manipulation

`arrange(asc(col))`	arranges *col* by ascending order
`arrange(desc(col))`	arranges *col* by descending order
`relocate(data, col, .before, .after)`	relocates a column relative to its neighbors*
`arrange(desc(col))`	arranges *col* by descending order
`slice(data, pos)`	indexes rows
`bind_rows(df1, df2, ...)`	dfs w/ same columns, concats rows
`bind_cols(df1, df2, ...)`	dfs w/ same # rows, concats cols, renames repeated cols
`semi_join(x, y, by)`	returns rows from x w/ matching val for by in y
`anti_join(x, y, by)`	returns rows from x w/o a match in y
`full_join(x, y, by)`	standard outer join
`left_join(x, y, by)`	standard left join, x is the left df
`right_join(x, y, by)`	standard right join, y is the right df

*specifying no neighbors moves *col* to leftmost col, specifyfing both is error

Suppose we have the following table `fish_encounters`

fish	station	seen
4842	Release	1
4842	I80_1	1
4842	Lisbon	1
4842	Rstr	1
4842	Base_TD	1
4842	BCE	1
4842	BCW	1
4842	BCE2	1
4842	BCW2	1
4842	MAE	1
4845	BCE	0

pivot_wider(fish_encounters, names_from = station, values_from = seen, values_fill = 0)

Fish	Release	I80_1	Lisbon	Rstr	Base_TD	BCE	BCW	BCE2	BCW2	MAE
1	4842	1	1	1	1	1	1	1	1	1
2	4843	1	1	1	1	1	1	1	1	1
3	4844	1	1	1	1	1	1	1	1	1
4	4845	1	1	1	1	0	0	0	0	0

Suppose we have the following table `billboard`

artist	track	date.entered	wk1	wk2	wk3	wk4	wk5	wk6	wk7
2 Pac	Baby...	2000-02-26	87	82	72	77	87	94	99
2Ge+her	The ...	2000-09-02	91	87	92	NA	NA	NA	NA
3 Doors D...	Kryp...	2000-04-08	81	70	68	67	66	57	54
3 Doors D...	Loser	2000-10-21	76	76	72	69	67	65	55
504 Boyz	Wobb...	2000-04-15	57	34	25	17	17	31	36

pivot_longer(billboard, cols = starts_with("wk"), names_to = "week", names_prefix = "wk", values_to = "rank", values_drop_na = TRUE)

artist	track	date.entered	week	rank
2 Pac	Baby Don't Cry (Keep...	2000-02-26	1	87
2 Pac	Baby Don't Cry (Keep...	2000-02-26	2	82
2 Pac	Baby Don't Cry (Keep...	2000-02-26	3	72
2 Pac	Baby Don't Cry (Keep...	2000-02-26	4	77
2 Pac	Baby Don't Cry (Keep...	2000-02-26	5	87
2 Pac	Baby Don't Cry (Keep...	2000-02-26	6	94
2 Pac	Baby Don't Cry (Keep...	2000-02-26	7	99
2Ge+her	The Hardest Part Of ...	2000-09-02	1	91
2Ge+her	The Hardest Part Of ...	2000-09-02	2	87
2Ge+her	The Hardest Part Of ...	2000-09-02	3	92

Dates & Strings

`ymd(), dmy(), ...`	converts string to datetime according to order of y-m-d
`wdate(date)`	gets the day of the week for a given date
`strc(str1, str2, ...)`	concatenates strings/vectors of strings
`str_detect(str, pattern)`	TRUE if ∃ a substring of `str` that matches `pattern`
`str_extract(str, pat, group)`	finds 1st match in `str` for `pat`, `group` takes matched pattern, returns text matching `group`
`str_extract_all(string, pattern)`	returns all matches to `pattern`
`str_sub(string, start, end)`	indexes into `string`
`str_count(string, pattern)`	count # of matches to `pattern` in `string`

str_replace(string, pattern, replacement), str_replace_all(string, pattern, replacement) - these exist

putting color, fill, alpha, etc. outside of aes(), i.e. typically inside of geom_x() functions will set it as a constant for the whole graph

putting color, fill, alpha, etc. inside of aes() typically implies you have a column in your df (like year) that sets the groups appropriately

every geom_x() function inherits the aes() from ggplot, unless they have their own aes() which overrides the ggplot

R always prints dates as YYYY-MM-DD

Regex

`\d`	digits
`\s`	whitespace
`\w`	alphabetic and numeral
`^`	matches the start of each line
`$`	matches the end of each line
`?`	0 or 1
`+`	1 or more
`*`	0 or more
`{n}`	exactly n
`{n, }`	n or more
`{n, m}`	between n and m

Capitalizing any of the above is the complement

You can also create your own character classes using []:

`[abc]`	matches a, b, or c
`[a-z]`	matches every character between a and z
`[^abc]`	matches anything except a, b, or c
`[\^\-]`	matches ^or -

Parenthesis make groups which can be backreferenced

pattern <- "(..)\\1" #(..) is some pair of anything, and 1 takes that same pair

fruit %>% str_subset(pattern)

"banana" "coconut" "cucumber" "jujube" "papaya" "salal berry"

Probability Theory

For some random variable X, $E(X) = \sum_{x=0}^{n} x * P(X = x)$.

The expected value is just the sum of each outcome multiplied by its porbability.

$Var(X) = E((X - \mu)^2)$, $\mu = E(X)$

Again, this is just multiplying the squared difference of the mean from each observation with each observation's respective probability,

$sum((x - mu)^2 * p)$.

Suppose that the distribution of X is proportional with the function $g(x) = 6 - |x - 5|$.

Say that we have outcomes $1, 2, \cdots, 10$, this means $P(X = x) = a(6 - |x - 5|)$.

We know that the total number of outcomes and number of current outcomes must be proportional to the function.

The way to make the number of outcomes proportional is to find $\sum_{i=1}^{10} 6 - |i - 5|$.

To keep the possible values proportional, each probability is $\frac{j}{\sum_{i=1}^{10} 6-|i-5|}$, $j \in g(1), g(2), \cdots, g(10)$.

Binomial Distributions

Properties of Binomials

`b`	binary outcomes
`i`	independence
`n`	fixed sample size
`s`	same probability

Binomial Formulas

$\mu = np$ $\quad \sigma = np(1-p)$

binom prob $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

R Binomial Functions

`rbinom(n, size, prob)`	random binomial samples
`dbinom(x, size, prob)`	density fcn at `x`
`qbinom(p, size, prob)`	get the smallest value in the qth quantile
`pbinom(q, size, prob)`	$P(X <= q)$
`pbinom(q, size, prob, lower.tail = T)`	$1 - P(X <= q) = p(X > q)$

Note that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Normal Distributions

R Normal Functions*

`pnorm(q, size, prob)`	$P(X < q)$

At any given x, $X \sim N(\mu, \sigma)$, $P(X = x) = 0$.

The standard normal is $X \sim N(0, 1)$.

Any normal has its z-scores as equivalent observations in the standard normal.

In other words, $X \sim N(\mu, \sigma) \implies Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

*includes `rnorm()`, `qnorm()` which have same functionality as the binom fcns

Suppose that $\exists X \sim Binom(n, p)$, with $np(1-p) \geq 10$

Note the conditions this tests, p can't be too close to 0 or 1 (causes skew), and n must be sufficiently large (reduces variance).

We can approximate that binomial with $X \sim N(np, \sqrt{np(1-p)})$.

Recall that this appoximation isn't perfect, the normal has an effect of "cutting off" the binomial distribution.

Correct for this with $P(X \leq x + .5)$ wheing finding $P(X \leq x)$, $P(X \geq x - .5)$ when finding $P(X \geq x)$

As a general rule,

65% of data 1 SD from the mean

95% of data 2 SD from the mean

99% of data 3 SD from the mean

Inference on Proportions

`\begin{enumerate}`	Numbered list.
`\begin{itemize}`	Bulleted list.
`\begin{description}`	Description list.
`\item` *text*	Add an item.
`\item[`*x*`]` *text*	Use *x* instead of normal bullet or number. Required for descriptions.

References

`\label{`*marker*`}`	Set a marker for cross-reference, often of the form `\label{sec:item}`.
`\ref{`*marker*`}`	Give section/body number of marker.
`\pageref{`*marker*`}`	Give page number of marker.
`\footnote{`*text*`}`	Print footnote at bottom of page.

Floating bodies

`\begin{table}[`*place*`]`	Add numbered table.
`\begin{figure}[`*place*`]`	Add numbered figure.
`\begin{equation}[`*place*`]`	Add numbered equation.
`\caption{`*text*`}`	Caption for the body.

The *place* is a list valid placements for the body. `t`=top, `h`=here, `b`=bottom, `p`=separate page, `!`=place even if ugly. Captions and label markers should be within the environment.

Text properties

Font face

Command	*Declaration*	*Effect*
`\textrm{`*text*`}`	`{\rmfamily` *text*`}`	Roman family
`\textsf{`*text*`}`	`{\sffamily` *text*`}`	Sans serif family
`\texttt{`*text*`}`	`{\ttfamily` *text*`}`	Typewriter family
`\textmd{`*text*`}`	`{\mdseries` *text*`}`	Medium series
`\textbf{`*text*`}`	`{\bfseries` *text*`}`	**Bold series**
`\textup{`*text*`}`	`{\upshape` *text*`}`	Upright shape
`\textit{`*text*`}`	`{\itshape` *text*`}`	*Italic shape*
`\textsl{`*text*`}`	`{\slshape` *text*`}`	*Slanted shape*
`\textsc{`*text*`}`	`{\scshape` *text*`}`	Small Caps shape
`\emph{`*text*`}`	`{\em` *text*`}`	*Emphasized*
`\textnormal{`*text*`}`	`{\normalfont` *text*`}`	Document font
`\underline{`*text*`}`		Underline

The command (t*ttt*) form handles spacing better than the declaration (t*ttt*) form.

Font size

`\tiny`	tiny
`\scriptsize`	scriptsize
`\footnotesize`	footnotesize
`\small`	small
`\normalsize`	normalsize
`\large`	large
`\Large`	Large
`\LARGE`	LARGE
`\huge`	huge
`\Huge`	Huge

These are declarations and should be used in the form `{\small ...}`, or without braces to affect the entire document.

Verbatim text

`\begin{verbatim}`	Verbatim environment.
`\begin{verbatim*}`	Spaces are shown as ␣.
`\verb!text!`	Text between the delimiting characters (in this case '!') is verbatim.

Justification

Environment	*Declaration*
`\begin{center}`	`\centering`
`\begin{flushleft}`	`\raggedright`
`\begin{flushright}`	`\raggedleft`

Miscellaneous

`\linespread{`*x*`}` changes the line spacing by the multiplier x.

Text-mode symbols

Symbols

&	`\&`	_	`\_`	...	`\ldots`	•	`\textbullet`
$	`\$`	^	`\^{}`	|	`\textbar`	\	`\textbackslash`
%	`\%`	~	`\~{}`	#	`\#`	§	`\S`

Accents

ò	`\`o`	ó	`\'o`	ô	`\^o`	õ	`\~o`	ō	`\=o`
ȯ	`\.o`	ö	`\"o`	o͡	`\c o`	ō	`\v o`	ő	`\H o`
ç	`\c c`	o̤	`\d o`	o͡o	`\t oo`	o͡o	`\t oo`	œ	`\oe`
Œ	`\OE`	æ	`\ae`	Æ	`\AE`	å	`\aa`	Å	`\AA`
ø	`\o`	Ø	`\O`	ł	`\l`	Ł	`\L`	ı	`\i`
ȷ	`\j`	¡	`!‘`	¿	`?‘`				

Delimiters

‘ ‘	“ ‘‘	{ \{	[[	((	< \textless
’ ’	” ’’	} \}	]]	))	> \textgreater

Dashes

Name	*Source*	*Example*	*Usage*
hyphen	-	X-ray	In words.
en-dash	--	1–5	Between numbers.
em-dash	---	Yes—or no?	Punctuation.

Line and page breaks

`\\`	Begin new line without new paragraph.
`\\*`	Prohibit pagebreak after linebreak.
`\kill`	Don't print current line.
`\pagebreak`	Start new page.
`\noindent`	Do not indent current line.

Miscellaneous

`\today` May 6, 2024.
`$\sim$` Prints $\sim$ instead of `\~{}`, which makes ~.
`~` Space, disallow linebreak (`W.J.~Clinton`).
`\@.` Indicate that the . ends a sentence when following an uppercase letter.
`\hspace{`*l*`}` Horizontal space of length *l* (Ex: *l* = 20pt).
`\vspace{`*l*`}` Vertical space of length *l*.
`\rule{`*w*`}{`*h*`}` Line of width *w* and height *h*.

Tabular environments

`tabbing` environment

`\=` Set tab stop. `\>` Go to tab stop.
Tab stops can be set on "invisible" lines with `\kill` at the end of the line. Normally `\\` is used to separate lines.

`tabular` environment

`\begin{array}[`*pos*`]{`*cols*`}`
`\begin{tabular}[`*pos*`]{`*cols*`}`
`\begin{tabular*}{`*width*`}[`*pos*`]{`*cols*`}`

`tabular` column specification

`l` Left-justified column.
`c` Centered column.
`r` Right-justified column.
`p{`*width*`}` Same as `\parbox[t]{`*width*`}`.
`@{`*decl*`}` Insert *decl* instead of inter-column space.
`|` Inserts a vertical line between columns.

`tabular` elements

`\hline` Horizontal line between rows.
`\cline{`*x*`-`*y*`}` Horizontal line across columns *x* through *y*.
`\multicolumn{`*n*`}{`*cols*`}{`*text*`}` A cell that spans *n* columns, with *cols* column specification.

Math mode

For inline math, use `\(...\)` or `$...$`. For displayed math, use `\[...\]` or `\begin{equation}`.

Superscript x^y	`^{x}`	Subscript x_y	`_{x}`
$\frac{x}{y}$	`\frac{x}{y}`	$\sum_{k=1}^{n}$	`\sum_{k=1}^n`
$\sqrt[n]{x}$	`\sqrt[n]{x}`	$\prod_{k=1}^{n}$	`\prod_{k=1}^n`

Math-mode symbols

$\leq$	`\leq`	$\geq$	`\geq`	$\neq$	`\neq`	$\approx$	`\approx`
$\times$	`\times`	$\div$	`\div`	$\pm$	`\pm`	$\cdot$	`\cdot`
$\circ$	`^{\circ}`	$\circ$	`\circ`	$\prime$	`\prime`	$\cdots$	`\cdots`
∞	`\infty`	$\neg$	`\neg`	$\wedge$	`\wedge`	$\vee$	`\vee`
$\supset$	`\supset`	$\forall$	`\forall`	$\in$	`\in`	$\rightarrow$	`\rightarrow`
$\subset$	`\subset`	$\exists$	`\exists`	$\notin$	`\notin`	$\Rightarrow$	`\Rightarrow`
$\cup$	`\cup`	$\cap$	`\cap`	$\mid$	`\mid`	$\Leftrightarrow$	`\Leftrightarrow`
$\dot{a}$	`\dot a`	$\hat{a}$	`\hat a`	$\bar{a}$	`\bar a`	$\tilde{a}$	`\tilde a`
α	`\alpha`	β	`\beta`	γ	`\gamma`	δ	`\delta`
ϵ	`\epsilon`	ζ	`\zeta`	η	`\eta`	ε	`\varepsilon`
θ	`\theta`	ι	`\iota`	κ	`\kappa`	ϑ	`\vartheta`
λ	`\lambda`	μ	`\mu`	ν	`\nu`	ξ	`\xi`
π	`\pi`	ρ	`\rho`	σ	`\sigma`	τ	`\tau`
υ	`\upsilon`	ϕ	`\phi`	χ	`\chi`	ψ	`\psi`
ω	`\omega`	Γ	`\Gamma`	Δ	`\Delta`	Θ	`\Theta`
Λ	`\Lambda`	Ξ	`\Xi`	Π	`\Pi`	Σ	`\Sigma`
Υ	`\Upsilon`	Φ	`\Phi`	Ψ	`\Psi`	Ω	`\Omega`

Bibliography and citations

When using BibTeX, you need to run `latex`, `bibtex`, and `latex` twice more to resolve dependencies.

Citation types

`\cite{`*key*`}` Full author list and year. (Watson and Crick 1953)
`\citeA{`*key*`}` Full author list. (Watson and Crick)
`\citeN{`*key*`}` Full author list and year. Watson and Crick (1953)
`\shortcite{`*key*`}` Abbreviated author list and year. ?
`\shortciteA{`*key*`}` Abbreviated author list. ?
`\shortciteN{`*key*`}` Abbreviated author list and year. ?
`\citeyear{`*key*`}` Cite year only. (1953)
All the above have an `NP` variant without parentheses; Ex. `\citeNP`.

BibTeX entry types

`@article` Journal or magazine article.
`@book` Book with publisher.
`@booklet` Book without publisher.
`@conference` Article in conference proceedings.
`@inbook` A part of a book and/or range of pages.
`@incollection` A part of book with its own title.
`@misc` If nothing else fits.
`@phdthesis` PhD. thesis.
`@proceedings` Proceedings of a conference.
`@techreport` Tech report, usually numbered in series.
`@unpublished` Unpublished.

BibTeX fields

`address` Address of publisher. Not necessary for major publishers.
`author` Names of authors, of format ....
`booktitle` Title of book when part of it is cited.
`chapter` Chapter or section number.
`edition` Edition of a book.
`editor` Names of editors.
`institution` Sponsoring institution of tech. report.
`journal` Journal name.
`key` Used for cross ref. when no author.
`month` Month published. Use 3-letter abbreviation.
`note` Any additional information.
`number` Number of journal or magazine.
`organization` Organization that sponsors a conference.
`pages` Page range (`2,6,9--12`).
`publisher` Publisher's name.
`school` Name of school (for thesis).
`series` Name of series of books.
`title` Title of work.
`type` Type of tech. report, ex. "Research Note".
`volume` Volume of a journal or book.
`year` Year of publication.
Not all fields need to be filled. See example below.

Common BibTeX style files

`abbrv`	Standard	`abstract`	`alpha` with abstract
`alpha`	Standard	`apa`	APA
`plain`	Standard	`unsrt`	Unsorted

The LaTeX document should have the following two lines just before `\end{document}`, where `bibfile.bib` is the name of the BibTeX file.

```
\bibliographystyle{plain}
\bibliography{bibfile}
```

BibTeX example

The BibTeX database goes in a file called *file*`.bib`, which is processed with `bibtex file`.

```
@String{N = {Na\~ture}}
@Article{WC:1953,
  author  = {James Watson and Francis Crick},
  title   = {A structure for Deoxyribose Nucleic Acid},
  journal = N,
  volume  = {171},
  pages   = {737},
  year    = 1953
}
```

Sample LaTeX document

```
\documentclass[11pt]{article}
\usepackage{fullpage}
\title{Template}
\author{Name}
\begin{document}
\maketitle

\section{section}
\subsection*{subsection without number}
text \textbf{bold text} text. Some math: $2+2=5$
\subsection{subsection}
text \emph{emphasized text} text. \cite{WC:1953}
discovered the structure of DNA.

A table:
\begin{table}[!th]
\begin{tabular}{|l|c|r|}
\hline
first  &  row  &  data \\
second &  row  &  data \\
\hline
\end{tabular}
\caption{This is the caption}
\label{ex:table}
\end{table}

The table is numbered \ref{ex:table}.
\end{document}
```


http://wch.github.io/latexsheet/