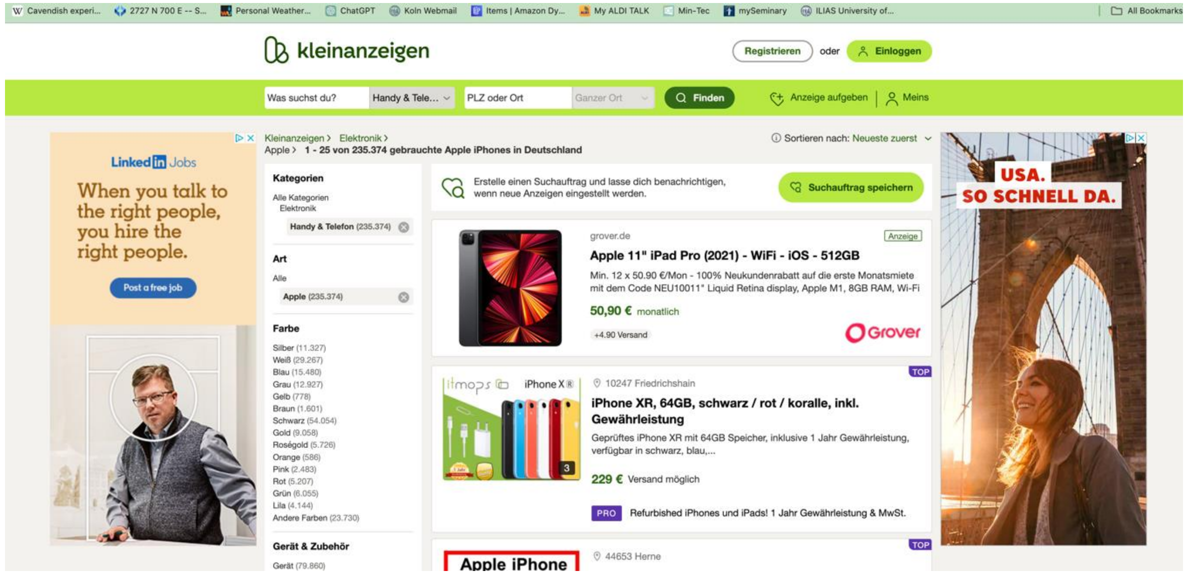
In Class Exercises

Cleaning

Download:

rawdata.csv, and kickstarter-campaigns.csv



Kleinanzeigen › Elektronik ›
Apple › 1 - 25 von 235.374 gebrauchte Apple iPhones in Deutschland

Sortieren nach: Neueste zuerst

Kategorien

Alle Kategorien
Elektronik
Handy & Telefon (235.374)

Art

Alle
Apple (235.374)

Farbe

Silber (11.327)
Weiß (29.267)
Blau (15.480)
Grau (12.927)
Gelb (778)
Braun (1.601)
Schwarz (54.054)
Gold (9.058)
Roségold (5.726)
Orange (586)
Pink (2.483)
Rot (5.207)
Grün (6.055)
Lila (4.144)
Andere Farben (23.730)

Gerät & Zubehör

Gerät (79.860)

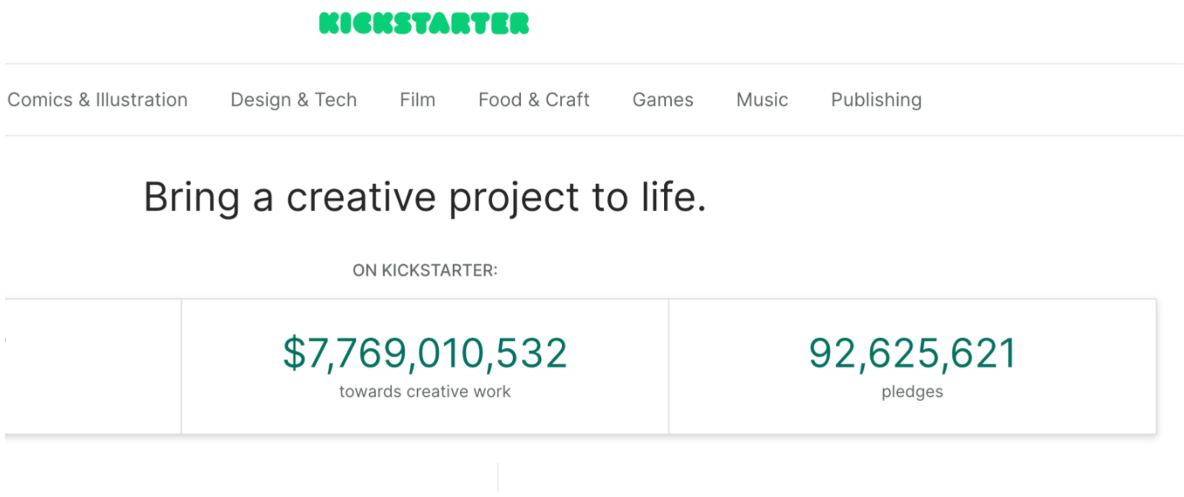
Erstelle einen Suchauftrag und lasse dich benachrichtigen, wenn neue Anzeigen eingestellt werden.

Suchauftrag speichern

grover.de
Anzeige
Apple 11" iPad Pro (2021) - WiFi - iOS - 512GB
Min. 12 x 50.90 €/Mon - 100% Neukundenrabatt auf die erste Monatsmiete mit dem Code NEU10011" Liquid Retina display, Apple M1, 8GB RAM, Wi-Fi
50,90 € monatlich
+4.90 Versand

TOP
10247 Friedrichshain
iPhone XR, 64GB, schwarz / rot / koralle, inkl. Gewährleistung
Geprüftes iPhone XR mit 64GB Speicher, inklusive 1 Jahr Gewährleistung, verfügbar in schwarz, blau,...
229 € Versand möglich
PRO Refurbished iPhones und iPads! 1 Jahr Gewährleistung & MwSt.

TOP
Apple iPhone
44653 Herne



Bring a creative project to life.

ON KICKSTARTER:

$7,769,010,532
towards creative work

92,625,621
pledges

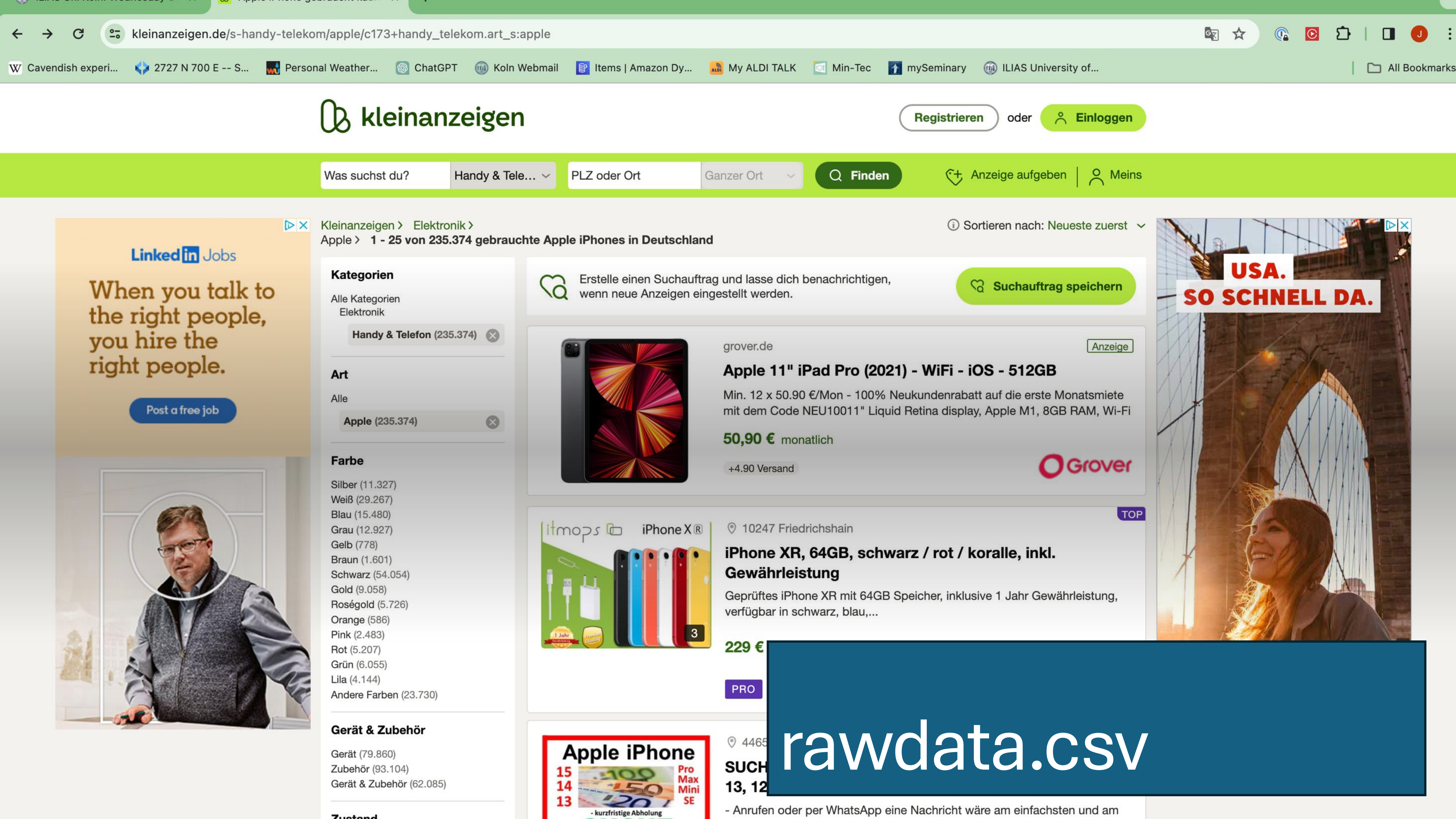
Brief review of gsub

```
gsub("pattern", "what to replace with", "the string to analyze")
gsub(" . ", "*", "data$title")
```

Code	Description
.	Any character
\\w	Any alpha numeric character
\\s	Any white space (including new line: \n)
\\d	Any digit

Code	Description
[a-z]	Lower case letters
[A-Z]	Upper case letters
[a-zA-Z]	Lower or upper case letters
Any word	Any word

Code	Description
^	Match beginning of string
$	Match end of string



kleinanzeigen.de/s-handy-telekom/apple/c173+handy_telekom.art_s:apple

Cavendish experi... | 2727 N 700 E -- S... | Personal Weather... | ChatGPT | Koln Webmail | Items | Amazon Dy... | My ALDI TALK | Min-Tec | mySeminary | ILIAS University of... | All Bookmarks

kleinanzeigen

Registrieren oder Einloggen

Was suchst du? | Handy & Tele... | PLZ oder Ort | Ganzer Ort | Finden | Anzeige aufgeben | Meins

Kleinanzeigen › Elektronik › Apple › 1 - 25 von 235.374 gebrauchte Apple iPhones in Deutschland

Sortieren nach: Neueste zuerst

Kategorien

Alle Kategorien
Elektronik
Handy & Telefon (235.374)

Art

Alle
Apple (235.374)

Farbe

Silber (11.327)
Weiß (29.267)
Blau (15.480)
Grau (12.927)
Gelb (778)
Braun (1.601)
Schwarz (54.054)
Gold (9.058)
Roségold (5.726)
Orange (586)
Pink (2.483)
Rot (5.207)
Grün (6.055)
Lila (4.144)
Andere Farben (23.730)

Gerät & Zubehör

Gerät (79.860)
Zubehör (93.104)
Gerät & Zubehör (62.085)

Zustand

Erstelle einen Suchauftrag und lasse dich benachrichtigen, wenn neue Anzeigen eingestellt werden.

Suchauftrag speichern

grover.de — Anzeige

Apple 11" iPad Pro (2021) - WiFi - iOS - 512GB

Min. 12 x 50.90 €/Mon - 100% Neukundenrabatt auf die erste Monatsmiete mit dem Code NEU10011" Liquid Retina display, Apple M1, 8GB RAM, Wi-Fi

50,90 € monatlich

+4.90 Versand

Grover

TOP

10247 Friedrichshain

iPhone XR, 64GB, schwarz / rot / koralle, inkl. Gewährleistung

Geprüftes iPhone XR mit 64GB Speicher, inklusive 1 Jahr Gewährleistung, verfügbar in schwarz, blau,...

229 €

PRO

4465

SUCH

13, 12

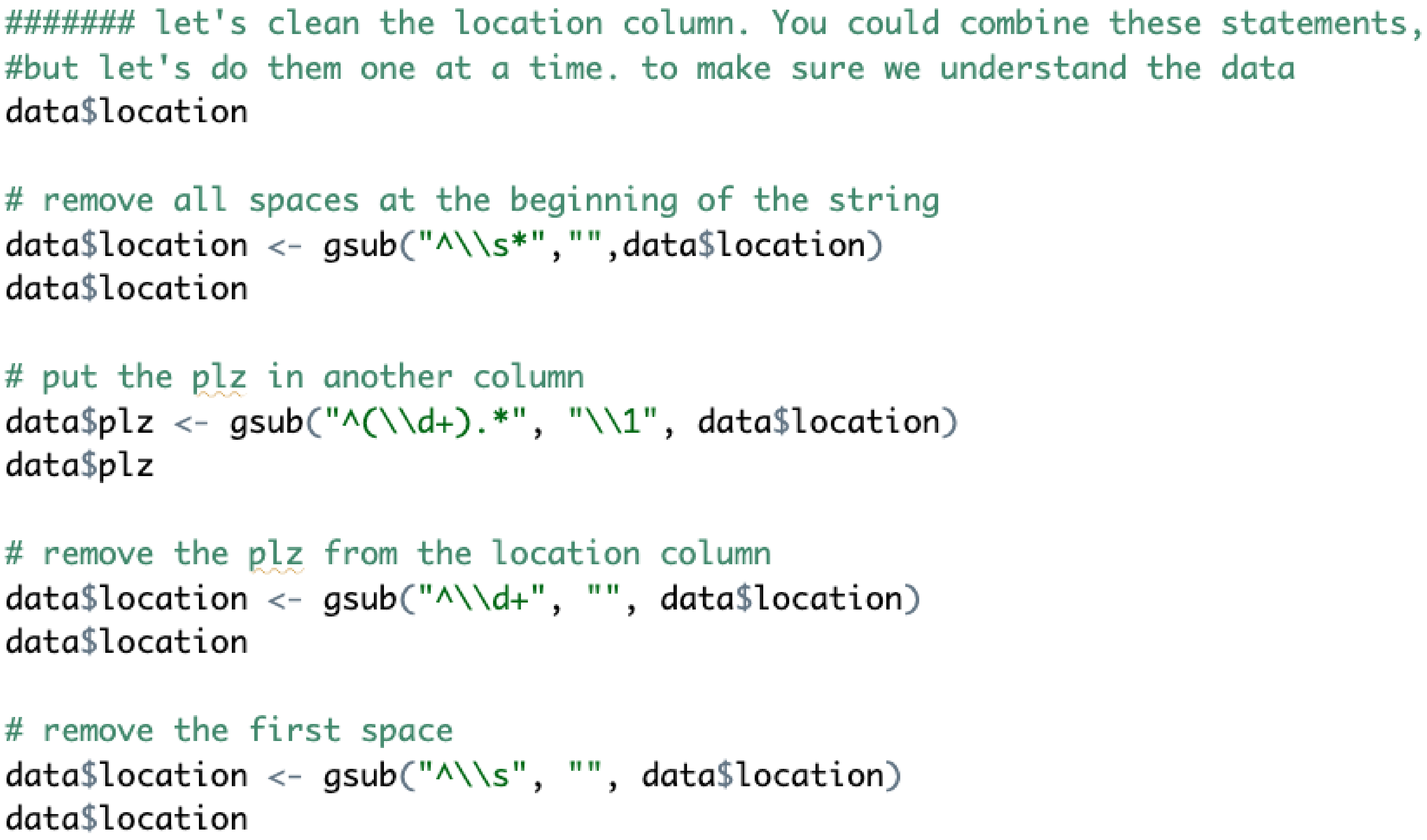
- Anrufen oder per WhatsApp eine Nachricht wäre am einfachsten und am

rawdata.csv

Let´s clean the ”location” column:

1. remove all spaces at the beginning of the string
2. put the plz (postal code) in another column
3. remove the plz from the location column
4. remove the first space

Let´s clean the ”location” column:



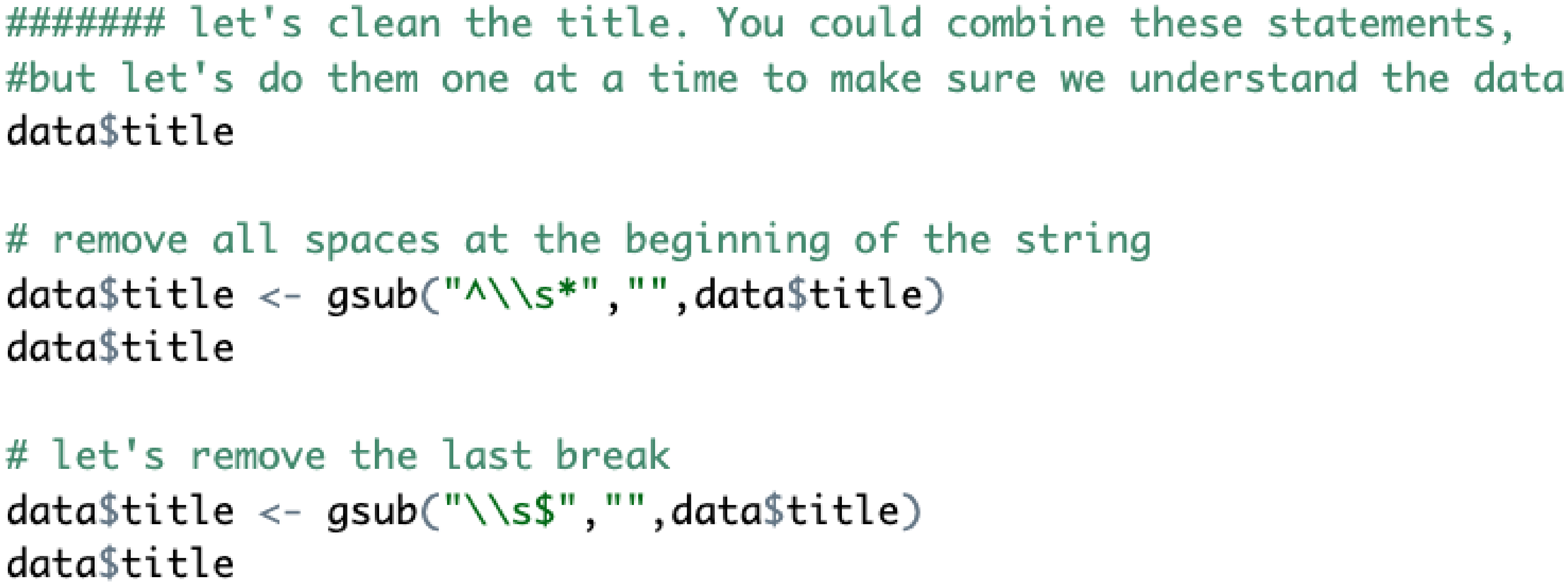
```
####### let's clean the location column. You could combine these statements,
#but let's do them one at a time. to make sure we understand the data
data$location

# remove all spaces at the beginning of the string
data$location <- gsub("^\\s*","",data$location)
data$location

# put the plz in another column
data$plz <- gsub("^(\\d+).*", "\\1", data$location)
data$plz

# remove the plz from the location column
data$location <- gsub("^\\d+", "", data$location)
data$location

# remove the first space
data$location <- gsub("^\\s", "", data$location)
data$location
```

Let's clean the “title” column:

1. remove all spaces at the beginning of the string

2. remove the last break

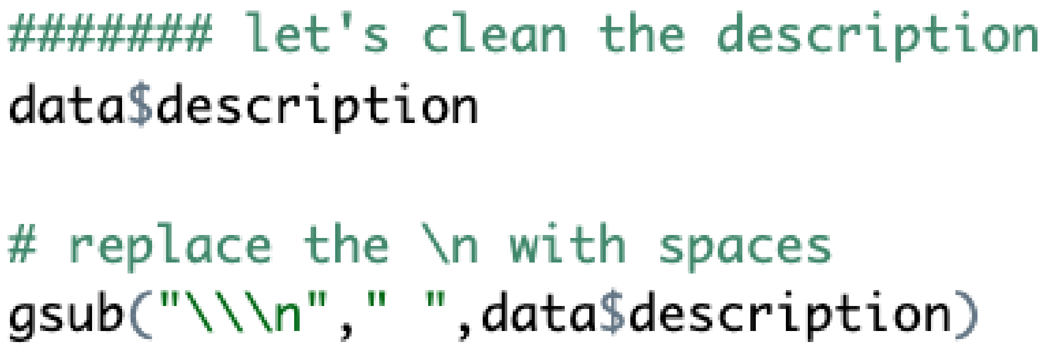
Let's clean the “title” column:



```
####### let's clean the title. You could combine these statements,
#but let's do them one at a time to make sure we understand the data
data$title

# remove all spaces at the beginning of the string
data$title <- gsub("^\\s*","",data$title)
data$title

# let's remove the last break
data$title <- gsub("\\s$","",data$title)
data$title
```

Let's clean the “description” column:

1. remove the breaks



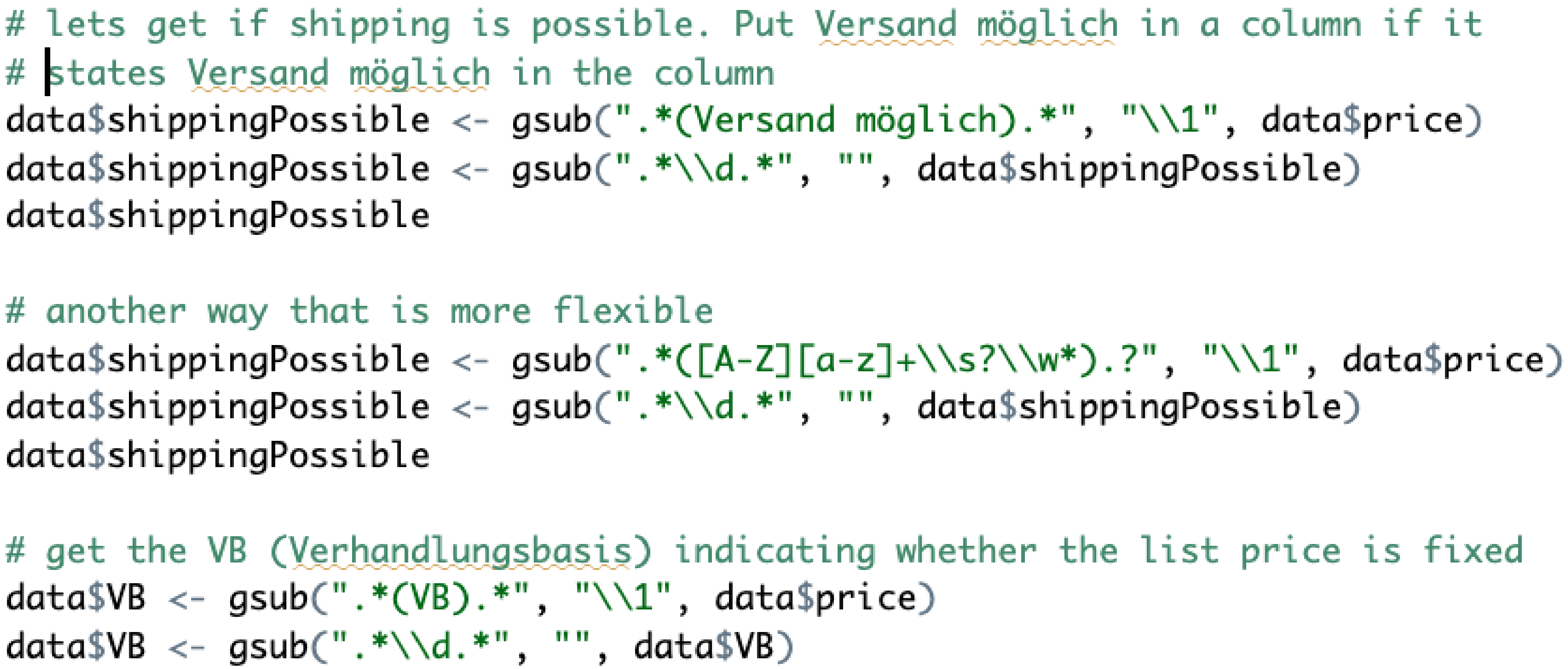
```
####### let's clean the description
data$description

# replace the \n with spaces
gsub("\\\n"," ",data$description)
```

Let’s clean the price “price” column:

1. Put shipping possible (Versand möglich) in a cloumn if shipping is possible

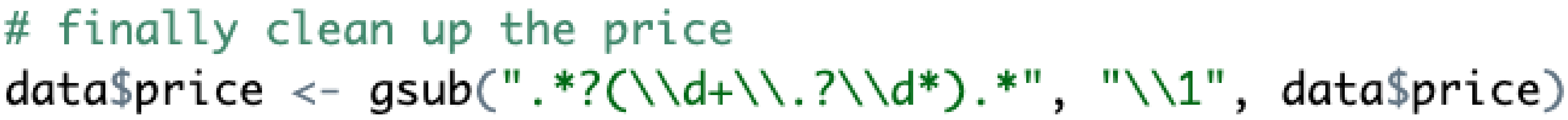
2. Put VB (Verhandlungsbasis) in a column when the list price is indicated as basis for negotiation

Let’s clean the price “price” column:



```
# lets get if shipping is possible. Put Versand möglich in a column if it
# states Versand möglich in the column
data$shippingPossible <- gsub(".*(Versand möglich).*", "\\1", data$price)
data$shippingPossible <- gsub(".*\\d.*", "", data$shippingPossible)
data$shippingPossible

# another way that is more flexible
data$shippingPossible <- gsub(".*([A-Z][a-z]+\\s?\\w*).?", "\\1", data$price)
data$shippingPossible <- gsub(".*\\d.*", "", data$shippingPossible)
data$shippingPossible

# get the VB (Verhandlungsbasis) indicating whether the list price is fixed
data$VB <- gsub(".*(VB).*", "\\1", data$price)
data$VB <- gsub(".*\\d.*", "", data$VB)
```

Finally clean up the price:



```
# finally clean up the price
data$price <- gsub(".*?(\\d+\\.?\\d*).*", "\\1", data$price)
```

In Class Exercises

Wrangling

`ggplot2`
Visualization for exploration and reporting

`dplyr`
Data organization, transformation, and summaries

`stringr`
Text analysis using regular expressions

`readr`
Reading rectangular text files (.txt, .csv)

`tibble`
Lean version of the data frame

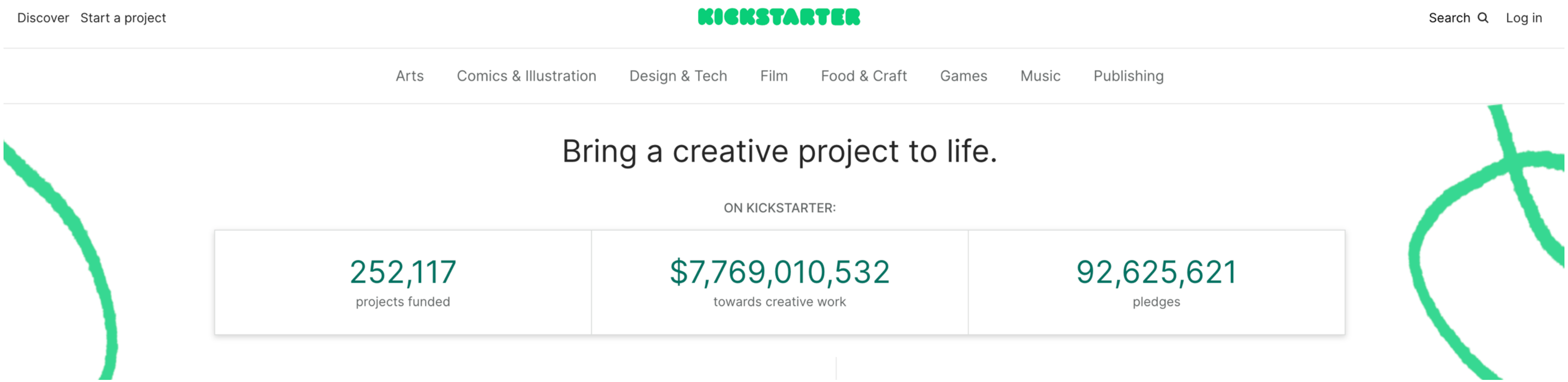
`tidyr`
dplyr-helper package for data pivoting

`purrr`
Functional programming

`forcats`
Handling factors

Key verbs	**Purpose**
Transformation	
`rename()`	Rename column names
`mutate()`	Create/change columns
Organization	
`arrange()`	Sort
`select()`	Select variables
`slice()`, `filter()`	Select rows
`left_join()`, `inner_join()`, etc.	Join data sets
Aggregation	
`summarize()`	Calculate statistics
`group()`	Summarize group-wise

Introduction to the Data



Bring a creative project to life.

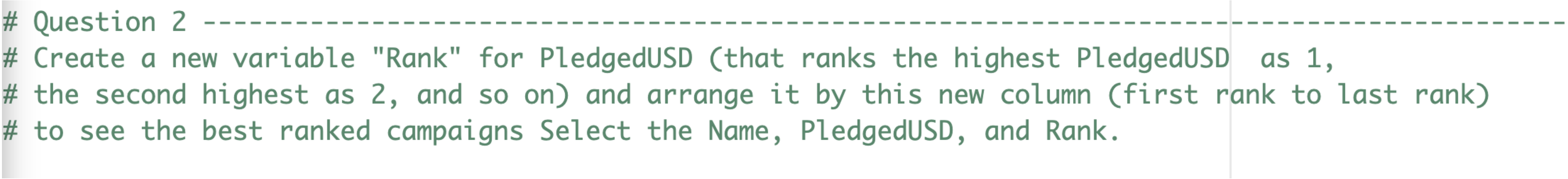
ON KICKSTARTER:

252,117 projects funded

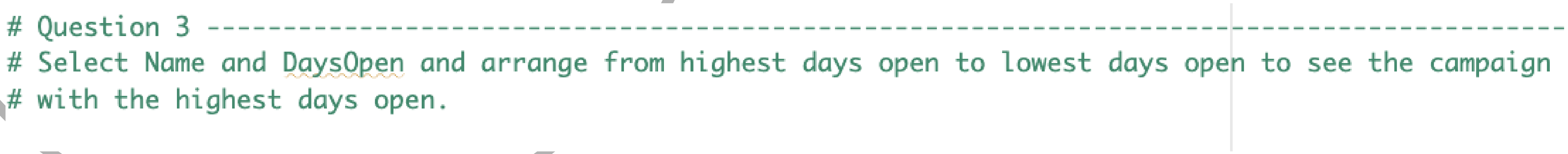
$7,769,010,532 towards creative work

92,625,621 pledges

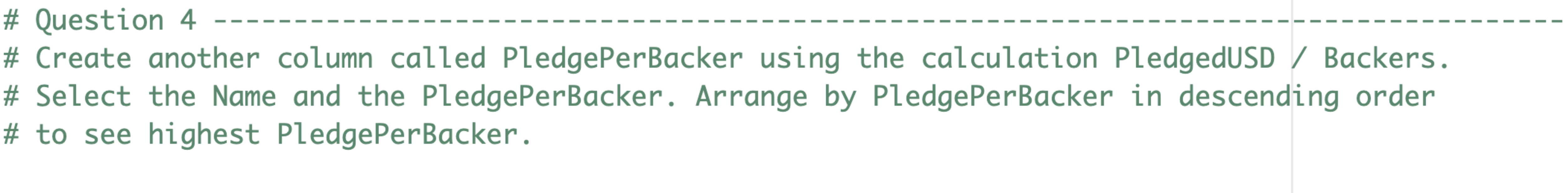
```
# Question 1 -----------------------------------------------------------
# Select only the name, country, state, and goal. Arrange name from A - Z.
```



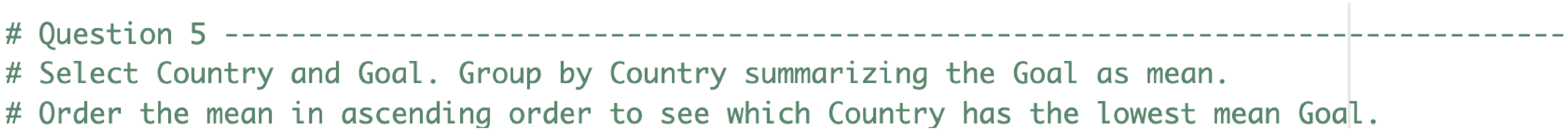
```
# Question 2 -----------------------------------------------------------------------------------------
# Create a new variable "Rank" for PledgedUSD (that ranks the highest PledgedUSD  as 1,
# the second highest as 2, and so on) and arrange it by this new column (first rank to last rank)
# to see the best ranked campaigns Select the Name, PledgedUSD, and Rank.
```



```
# Question 3 -----------------------------------------------------------------------------------------
# Select Name and DaysOpen and arrange from highest days open to lowest days open to see the campaign
# with the highest days open.
```

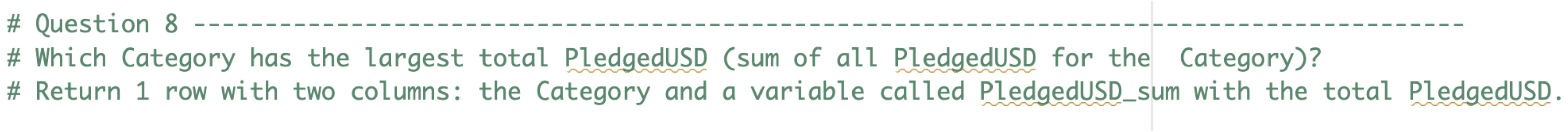


```
# Question 4 ---------------------------------------------------------------------------------------------
# Create another column called PledgePerBacker using the calculation PledgedUSD / Backers.
# Select the Name and the PledgePerBacker. Arrange by PledgePerBacker in descending order
# to see highest PledgePerBacker.
```

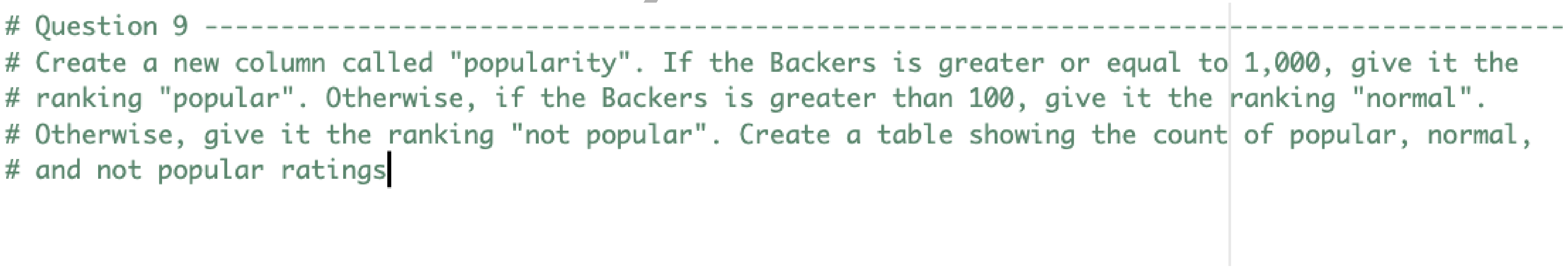


```
# Question 5 -------------------------------------------------------------------------------
# Select Country and Goal. Group by Country summarizing the Goal as mean.
# Order the mean in ascending order to see which Country has the lowest mean Goal.
```

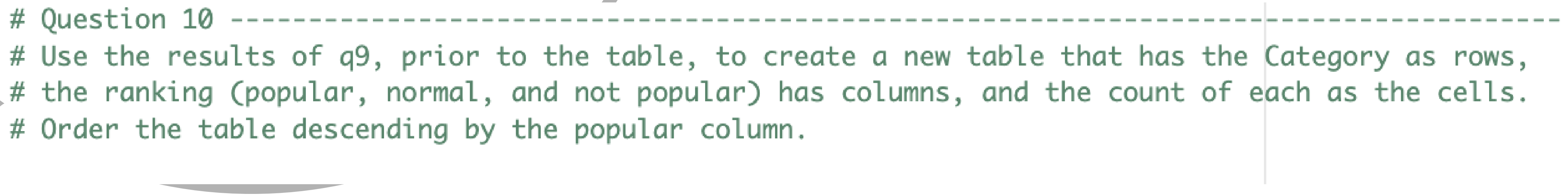
```
# Question 6 ----------------------------------------------------------------------------------------
# Repeat the previous command but limit it to countries that have at least ten campaigns
# Save the result to a tibble called q6.
# Question 7 -------------------------------------------------------------------
# Which campaigns are Staff Picks? Select only those titles:
```



```
# Question 8 -----------------------------------------------------------------------------------------
# Which Category has the largest total PledgedUSD (sum of all PledgedUSD for the  Category)?
# Return 1 row with two columns: the Category and a variable called PledgedUSD_sum with the total PledgedUSD.
```



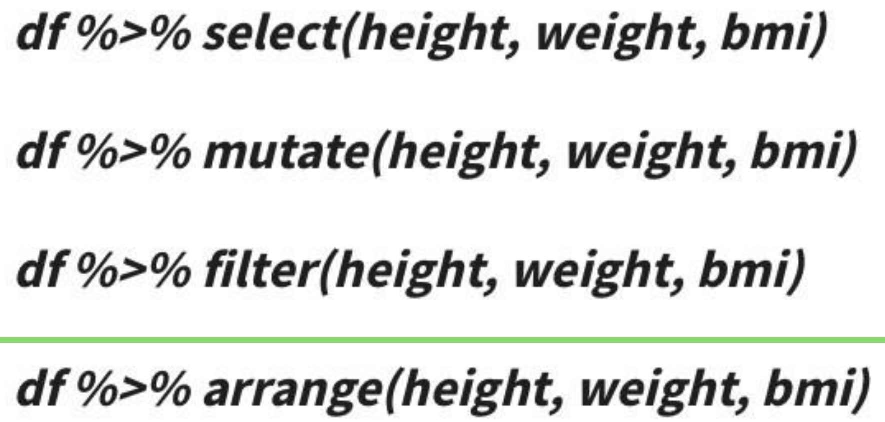
```
# Question 9 -----------------------------------------------------------------------------------------
# Create a new column called "popularity". If the Backers is greater or equal to 1,000, give it the
# ranking "popular". Otherwise, if the Backers is greater than 100, give it the ranking "normal".
# Otherwise, give it the ranking "not popular". Create a table showing the count of popular, normal,
# and not popular ratings
```



```
# Question 10 -------------------------------------------------------------------------------------
# Use the results of q9, prior to the table, to create a new table that has the Category as rows,
# the ranking (popular, normal, and not popular) has columns, and the count of each as the cells.
# Order the table descending by the popular column.
```

Data Wrangling Quiz

If you had six different variables in the columns of a dataset (**df**) and wanted to order the data frame by **height**, then by **weight**, and then by **bmi**, which of these would you use?



```
df %>% select(height, weight, bmi)

df %>% mutate(height, weight, bmi)

df %>% filter(height, weight, bmi)

df %>% arrange(height, weight, bmi)
```

Data Wrangling Quiz

If you had six different variables in the columns of a dataset (df) and wanted to **select three** of them (**height, weight, bmi**) and display them in **alphabetical order**, which of these would achieve that?

1. filter(bmi, height, weight)

2. df %>% mutate(bmi, height, weight)

3. df %>% filter(height, weight, bmi)

4. arrange(bmi, height, weight)

[✓ ***5. df %>% select(bmi, height, weight)***]

Data Wrangling Quiz

What function of what package should you use to create a new column or variable?

[✓ ***1. mutate()*** *function of the* ***dplyr package***]

2. new_col() *function of the* ***dplyr package***

3. new_col() *function of the* ***tidyr package***

4. new_var() *function of the* ***dplyr package***

5. new_var() *function of the* ***tidyr package***

6. mutate() *function of the* ***tidyr package***

Data Wrangling Quiz

What package and function should you use to change the order of the values of a variable?

1. sort() *function of the* ***tidyr package***

[✓ ***2. arrange()*** *function of the* ***dplyr package***]

3. reorder() *function of the* ***dplyr package***

4. order() *function of the* ***reorder package***

Assignment Week 2

NETFLIX

R file: NetflixAssignment.R
Data: NetflixOriginals.csv

Sample Final Exam Questions

- What do common regex / gsub commands do (\\w \\d \\s . * + ^ $?)
- How do you get data versus replace data in gsub?
- What do common commands do in tidyverse (select, summarize, filter, group by, arrange desc, mutate, table, pivot_longer, pivot_winder)
- What is the difference between long versus wide data
- Provide answers in response to code snippets