Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

Welcome!
ATTENDANCE DAY 1:
Name, Unit, Favorite Emoji https://emojipedia.org/
Mike Trizna, OCIO,
Jorge A. Santiago-Blay, NMNH-Paleobiology,
Rob Fleischer, CCG-SCBI-NZP, ☺
Joseph A. Campbell, NMAAHC,
Pietro Tardelli Canedo, NMNH-LAB,
Mike O'Mahoney, Invert Zoo,
Sonia C. Haro, OP&A, 😊
Anna Kearns, SCBI CCG,
Chelsea Robertson, OA,
Bryton A. Smith, NMNH-Paleobiology,
Mohammad Vatanparast, RA-Botany, NMNH 😂
Carrie Craig, NMNH-LAB, 😎
Matthew Kweskin, NMNH, LAB, 😸
Bamila Cardoso, GWU, Orti LAB

Data Organization in Spreadsheets: https://datacarpentry.org/spreadsheet-ecology-lesson/

Please describe something you have accidentally done in a spreadsheet that caused you frustration (if anything)
can never remember how to split or unsplit a screen in Excel.

Remove/add needed stuff or not doing what I thought I wanted to do. In general, I look forward to live and very well to actually use a machine that would read my mind to the computer and have it do whatever I want. In other words, no learning the commands another mortal designed. For the time being, we have to learn what someone else decided is the way to tell the machine what to do.

Sorted only one column rather than the whole table
accidentally changing set formula and calculated values but not realizing
Sorting spreadsheet with a missing column in the middle, so data all jumbled
Accidentaly losing data by writing the wrong formula
Many times just simply hitting enter get rid of the cells content, happend a lot!
Replaceing every instance of a value instead of the ones in the target column and many more annoying little things like that

Tried to sort by a column, and only that column got sorted

Not having the correct data type for the formula I want to use
Can't get Pivot tables to stop breaking down dates to days, quarters and years
lost data replacing or filtering

Cardinal rule: keep it “tidy”

1. Put all your variables (weight, temperature) in columns
2. Put each observation on its own row
3. Don’t combine multiple pieces of information in one cell
4. Leave raw data raw—don’t change it!
5. Export the cleaned data to text based format like CSV (comma separated values) format

Keep track of steps in a readMe file

Exercise:
We’re going to take a messy version of the survey data and describe how we would clean it up.

- Download the data by clicking here https://ndownloader.figshare.com/files/2252083 to get it from FigShare.

1. Open up the data in a spreadsheet program.
2. You can see that there are two tabs. Two field assistants conducted the surveys, one in 2013 and one in 2014, and they both kept track of the data in their own way. Now you’re the person in charge of this project and you want to be able to start analyzing the data.
3. Identify what is wrong with this spreadsheet and the steps you would need to take to clean up the 2013 and 2014 tabs, and to put them all together in one spreadsheet.

After you go through this exercise, we’ll discuss as a group what was wrong with this data and how you would fix it.

Meet back around 9:58 am

BREAK: meet back at 10:31

OpenRefine Lesson: https://datacarpentry.org/OpenRefine-ecology-lesson/

Data File we will be using: https://ndownloader.figshare.com/files/7823341

- Using faceting, find out how many years are represented in the census?

26 26 2626
26262626

- Is the column formatted as Number, Date, or Text? How does changing the format change the faceting display?

TextTextTextTextText Text
Facet by number is displayed as histogram+1

- Which years have the most and least observations?

Most 1997 1997199719971997 1997
Least 1977 1977197719771977 1977

If you have any problems with installing DB Browser for SQLite, email me (Matt) at kweskinm@si.edu
See you tomorrow!

Welcome back!
Download DB Browser for SQLite

- Mac: https://download.sqlitebrowser.org/DB.Browser.for.SQLite-3.11.2.dmg
- Windows: https://download.sqlitebrowser.org/DB.Browser.for.SQLite-3.11.2-win64.zip

Dataset
https://ndownloader.figshare.com/articles/1314459/versions/9
Download then unzip. We'll be using surveys.csv, species.csv, plots.csv

Lesson
https://datacarpentry.org/sql-ecology-lesson/00-sql-introduction/index.html

Amanda Devine's SQL cheat sheet:
https://gist.github.com/amdevine/9460baa6f0fc6525b16724947cfb20b4

Day two attendance: *Name and something you've enjoyed doing during this stay at home time*
Matt Kweskin, baking treats
Carrie Craig, watching birds at our bird feeder
Bryton A. Smith, more relaxed mornings and evenings due to not having to go on long commutes
Rob Fleischer - Long walks with our dog. And baking bread.
Pietro Tardelli Canedo - puzzling and reading
Mike O'Mahoney, building a media server
JorgeA. Santiago-Blay - fewer drives to DC (= wake up time = 3AM) from York, PA
Joseph Campbell, Working on 3D and playing with the dog.
Mike Trizna, training a new dog
Sonia Haro, reading and baking.
Chelsea Robertson - learning card tricks
Anna Kearns, exploring different neighborhood trails with my dog
Mohammad, doing some meditation!
Vanessa Gonzalez, hanging out with my new dog.

Import species.csv and plots.csv into the database.

Return from break at 10:30

Write a query that returns the plot_id, species_id, sex, and weight in mg.

```
SELECT plot_id, species_id, sex, weight*1000 AS weight_mg
FROM surveys;
SELECT plot_id,species_id, sex, weight*1000
FROM surveys;
SELECT plot_id, species_id, sex, ROUND(weight*1000,2) AS weight_mg
FROM surveys;
SELECT plot_id, species_id, sex, weight*1000
FROM surveys;
SELECT plot_id, species_id, sex, weight*1000 AS weight_mg
FROM surveys;
SELECT plot_id, species_id, sex, weight*1000 AS weight_mg
FROM surveys;
```

Convert to integer: CAST(weight AS INTEGER) Thanks Mike

Write a query that returns the total weight, average weight, minimum and maximum weights for all animals caught over the duration of the survey. Can you modify it so that it calculates these values only for animals >= 5 and <10 grams?

```
SELECT CAST(sum(weight) AS INT) AS total_weight_g, round(avg(weight), 2) AS avg_weight_g,
min(weight) AS min_weight_g, max(weight) AS max_weight_g
FROM surveys
WHERE weight >= 5 AND weight < 10;
```

```
select SUM(weight) total_weight, AVG(weight) avg_weight, MIN(weight) min_weight
FROM surveys
WHERE weight BETWEEN 5 and 10
```
(I changed data type for weight column to integer by modifying the table)

```
SELECT SUM(weight), AVG(weight), MIN(weight), MAX(weight)
FROM surveys
WHERE weight >= 5 AND weight < 10;
```

```
SELECT sum(weight), avg(weight), min(weight), max(weight)
FROM surveys
WHERE weight >= 5 AND weight <= 10;
SELECT SUM(weight), AVG(weight), max(weight), min(weight)
FROM surveys
WHERE weight >=5 AND weight <10;
SELECT sum(weight), avg(weight), min(weight), max(weight)
FROM surveys
WHERE (weight >=5 )AND( weight <=10);
```

```
-- Write a query that returns the total weight, average weight, minimum and maximum weights for all animals caught over the duration of the survey
SELECT total(weight), avg(weight), min(weight), max(weight)
FROM surveys;
```

```
-- Modify above query so that it calculates these values only for animals >= 5 and <10 grams
SELECT total(weight), avg(weight), min(weight), max(weight)
FROM surveys
WHERE weight >= 5 AND weight <10;
SELECT SUM(weight), AVG(weight), MIN(weight), MAX(weight)
FROM surveys
WHERE weight >= 5 AND weight < 10;
```

```
--Write a query that returns the total weight, average weight, minimum and maximum weights for all animals caught over the duration of the survey. Can you modify it so that it calculates these values only for animals >= 5 and <10 grams?
```

```
--query practice 1
SELECT SUM(weight), AVG(weight), MIN(weight), MAX(weight)
FROM surveys;
```

```
--query practice 2
SELECT SUM(weight), AVG(weight) AS avg_weight_g, MIN(weight), MAX(weight)
FROM surveys
WHERE weight>=5 AND weight<10;
```

Break until 11:30

SELECT *
FROM surveys
Where * IS NOT NULL; (this won't work, you have to list each field name in the WHERE statement)

Write a query that returns the genus, the species name, and the weight of every individual captured.

SELECT species.genus, species.species, surveys.weight
FROM species
LEFT JOIN surveys
USING (species_id);
SELECT species.genus, species.species,surveys.weight
FROM surveys
LEFT JOIN species
USING (species_id);

Feedback on today's section:
https://forms.gle/Wb2Sm7C98aqZF73u8

Next session in next Tuesday at 9am, same Zoom link!

pad.carpentries.org/2020-05-19-smithsonian

DAY THREE ATTENDANCE
Please add your name, and 1 tool or concept you learned last week that you will use again in the future
Bryton A. Smith-joining in SQL
Sonia Haro, OpenRefine for upcoming large datasets.
Rob Fleischer, learned better ways to handle very large databases than in Excel.
Joseph Campbell - OpenRefine and parsing data sets
Chelsea Robertson, OpenRefine was great, I'll be using that to parse large datasets rather than pivots in Excel where possible
Anna Kearns - Both openrefine and SQL. Joining!
Pietro Tardelli Canedo - faceting in OpenRefine

Mohammad Vatanparast: Openrefine, in particular removing spaces!

Jorge Santiago-Blay - Both Open Refine and SQL. The database I wish to use to practice with is relatively small that I can do cleaning by hand. Will we have time to play with our data using R and ask help if we get stuck? :)

Is it possible to control OpenRefine or DB Browser via apple script or bash? Thinking about automation for data sets with millions+ objects -- if so are there any example projects?

R Lesson link: https://datacarpentry.org/R-ecology-lesson/01-intro-to-r.html
RStudio Binder back-up: https://mybinder.org/v2/gh/SmithsonianWorkshops/binders/rstudio?urlpath=rstudio

downloaded a lot!

URL: **download.file**(url="https://ndownloader.figshare.com/files/2292169",
`destfile = "data_raw/portal_data_joined.csv")`

CHALLENGE:
Select a subset of the surveys data that only comes from plot id 2, and save that to a variable called plot2.

plot2 <- filter(surveys, plot_id == 2)

CHALLENGE:

Using pipes, subset the surveys data to include animals collected before 1995 and retain only the columns year, sex, and weight.

```
filter(surveys, year < 1995) %>%
  select(year, sex, weight)
```

```
surveys %>%
  filter(year < 1995) %>%
  select(year, sex, weight)
```

CHALLENGE:
How many animals were caught in each plot_type surveyed?

```
surveys %>%
  count(plot_type)
```

What was the heaviest animal measured in each year? Return the columns year, genus, species_id, and weight.

```
fatrats <- surveys %>%
  filter(!is.na(weight)) %>%
  group_by(year) %>%
  filter(weight == max(weight)) %>%
  select(year, genus, species_id, weight) %>%
  arrange(year)
```

Feedback on today's section:
https://forms.gle/oPY2PmTTyyaNfHEb7

DAY 4 SIGN-IN
Name | Question from yesterday, or thing you wish we had spent more time on

Bryton A. Smith-just a little more elaboration on how Python is related to R and the different uses each is ideal for

Jorge - A little slower pace for all to follow if they so wish in their hearts :)

Joseph Campbell: I was looking forward to todays session of plotting data. yesterday was great intro.

Mohammad Vatanparast: It was great! It would be great to have more session in R, since we can do a lot!

Mike O'Mahoney: How to write out a list that contains multiple datatypes (dataframes, lists, etc.) into a dataframe

Rob Fleischer: why do we also have R console as well as RStudio? Do we need R if we use RStudio?

Pietro Tardelli Canedo: what are the possible summarize commands?

Sonia Haro-great introduction to R, overly excited for data visualization !

Chelsea - did we cover all of the most commonly used commands yesterday?

Anna - ready to learn more about data visualization in R!

R for Data Science: https://r4ds.had.co.nz/

no font could be found for family "Arial". And no axis or numbers are on the plot. Any idea how to fix?

Error in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : polygon edge not found

```{r}
ggplot(data = surveys_complete,
      mapping = aes(x = weight,
                    y = hindfoot_length))
```

https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html

ggplot(data.frame(), aes(1, 1)) + theme(text = element_text(family = 0))

ggplot(data.frame(), aes(1, 1)) + theme(text = element_text(family = "Comic Sans MS"))

Yes. Also, I tried the graph by sex.

Challenge

Use what you just learned to create a scatter plot of weight over species_id with the plot types showing in different colors. Is this a good way to show this type of data?

Come back at 10:38

To get keyboard shortcut to %in% we need to install Rstudio addins (a lot of downloads)
https://rstudio.github.io/rstudioaddins/
Then add shortcut manually!

```
ggplot(data = yearly_counts,
      mapping = aes(x = year,
                    y = n)) +
```

```
  geom_line() +
  facet_wrap(facets = vars(genus))
```

Dipodomys are kangaroo rats!

Challenge
Use what you just learned to create a plot that depicts how the average weight of each species changes through the years.

Try to work on this until 11:40