Intro to dplyr and ggplot

Guest lecture at Biostatistics course (Dr. Itamar Giladi) @ Blaustein Institues for Desert Research, Ben Gurion University (January, 2010) License: CC BY-NC-SA 4.0

Dr. Shai Pilosof, Ecological Complexity Lab, Department of Life Scienecs,Ben Gurion University www.bgu.ac.il/ ecomplab

What is this file?

This class is a very brief introduction to `dplyr` and `ggplot`, which are part of the `tidyverse` world (or universe?). During the lecture I will give examples and the purpose of this file is to have the basic examples written down so students can copy-paste the text or be reminded later. We expand on these examples during the class.

This class follows the excellent tutorial Data Analysis and Visualization in R for Ecologists: https://datacarpentry.org/R-ecology-lesson/index.html.

Load functions

```
library(tidyverse)
library(lubridate)
library(magrittr)
```

Looking at data

```
surveys <- read_csv('portal_data_joined.csv')
is_tibble(surveys)
glimpse(surveys)
```

Basic dplyr capabilities

```
# Filtering
surveys %>% filter(year==1977)
surveys %<>% drop_na

# Selecting
surveys %>% select(record_id, sex)
surveys %>% select(-sex) # deselect a column

# Mutating
surveys %<>%
  mutate(Date=as_date(paste(year,month,day,sep='-'))) %>%
  drop_na

# Summarizing
```

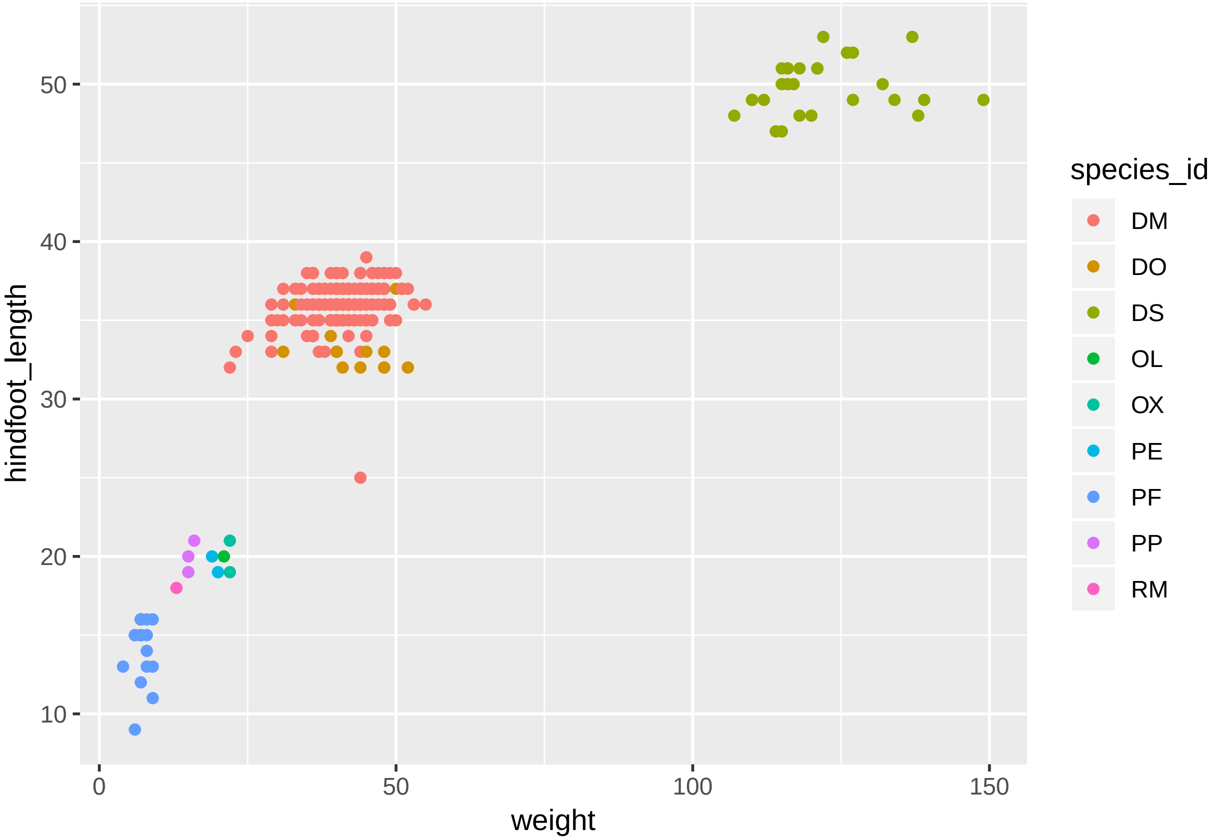
```r
surveys %>%
  group_by(species_id, sex) %>%
  summarise(mean_weight=mean(weight, na.rm=T))

surveys %>%
  group_by(species_id,sex) %>%
  summarise(n=n_distinct(record_id))

surveys %>%
  group_by(sex, species_id) %>%
  select(hindfoot_length, weight) %>%
  summarise_all(list(m=mean, mx=max))

# Grouping can be useful without a summary
surveys %>%
  group_by(species_id) %>%
  select(year) %>% table()

# Window functions
surveys %>%
  group_by(species_id) %>%
  select(weight) %>%
  top_n(3) %>% arrange(species_id, desc(weight))
```

Joining data

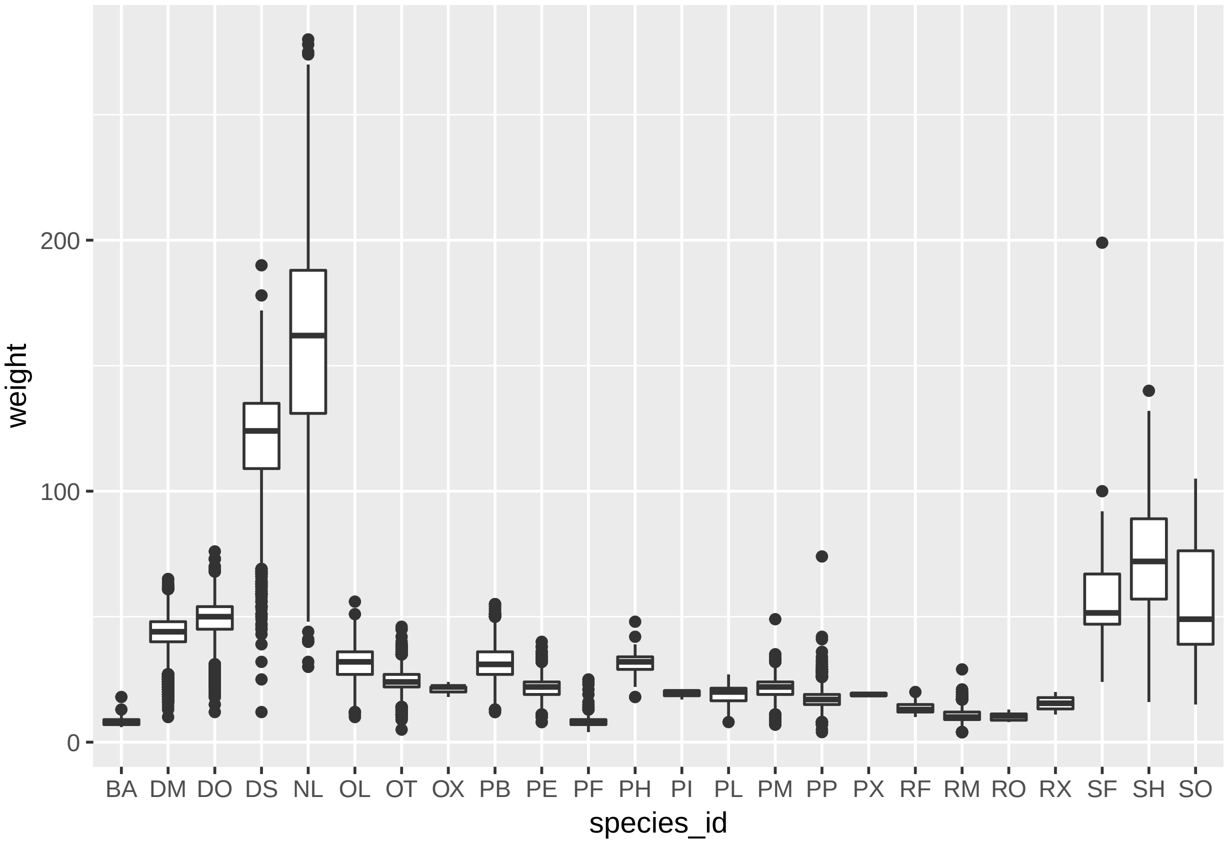
```r
# First prepare two data sets:
data_plots <- surveys %>%
  select(plot_id, month, day, year, plot_type) %>%
  distinct_all() %>%
  mutate(Date=as_date(paste(year,month,day,sep='-'))) %>%
  arrange(Date, plot_id) %>%
  tibble::rowid_to_column() %>%
  rename(visit_id=rowid) %>%
  filter(visit_id%in%1:20)

data_2 <- surveys %>%
  mutate(Date=as_date(paste(year,month,day,sep='-'))) %>%
  arrange(Date, plot_id) %>%
  group_by(Date, plot_id) %>%
  mutate(visit_id=group_indices()) %>%
  ungroup() %>%
  select(visit_id, everything(), -Date, -month, -day, -year, -plot_id, -plot_type) %>%
  filter(visit_id%in%1:15)

# Joining:
joined_left <- data_2 %>% left_join(data_plots)
joined_full <- data_2 %>% full_join(data_plots)
joined_right <- data_2 %>% right_join(data_plots)
```

Plotting

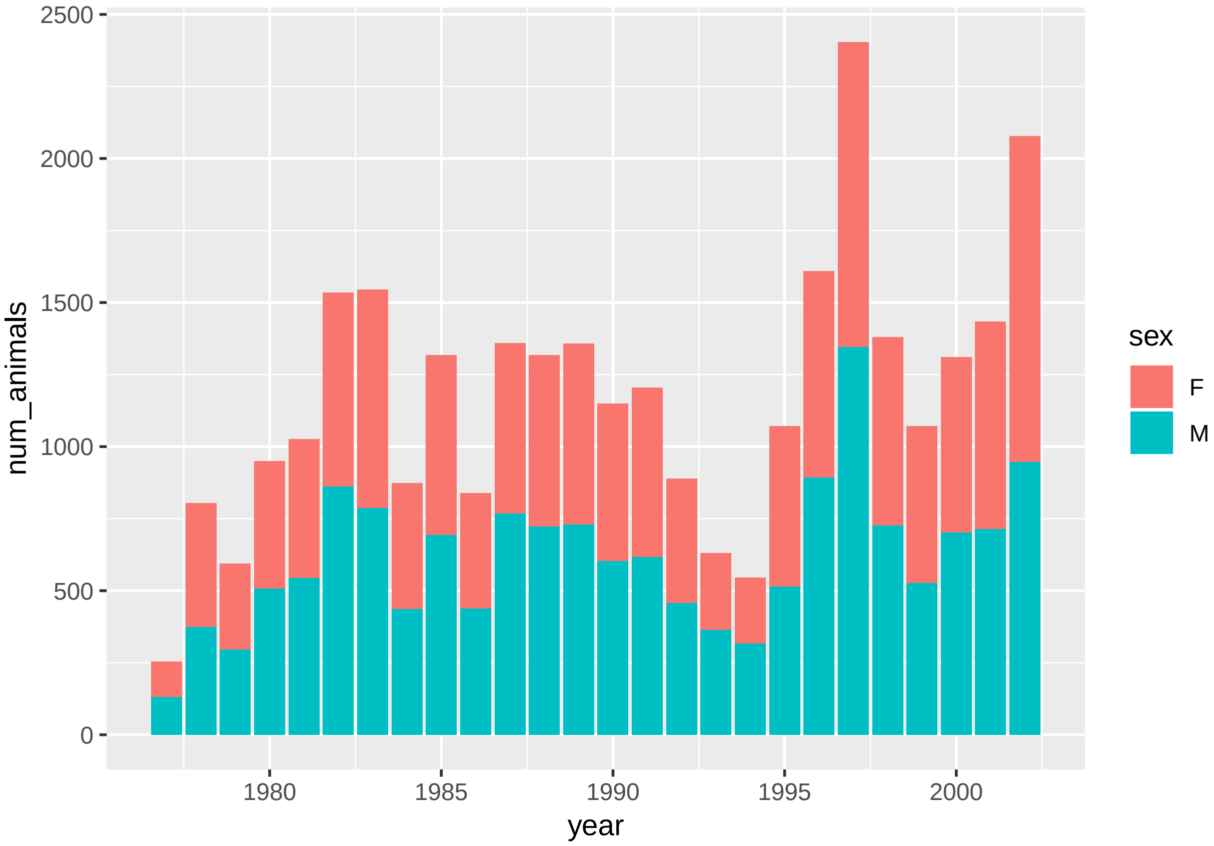
```
# Scatter plot
surveys %>%
  filter(year==1977) %>%
  filter(!is.na(weight)) %>%
  filter(!is.na(hindfoot_length)) %>%
  ggplot()+geom_point(aes(x=weight, y=hindfoot_length, color=species_id))
```



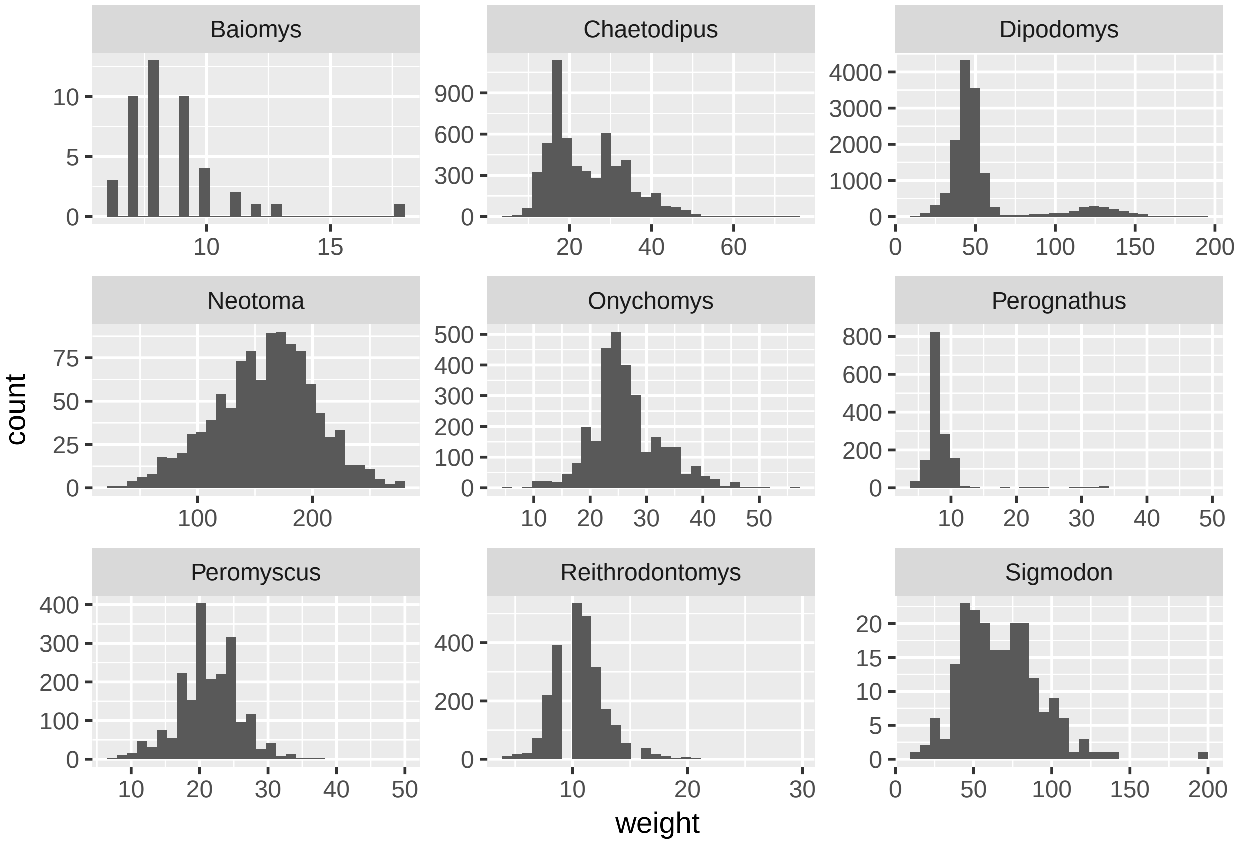
```
# Box plots
surveys %>%
  ggplot()+geom_boxplot(aes(x=species_id, y=weight))
```



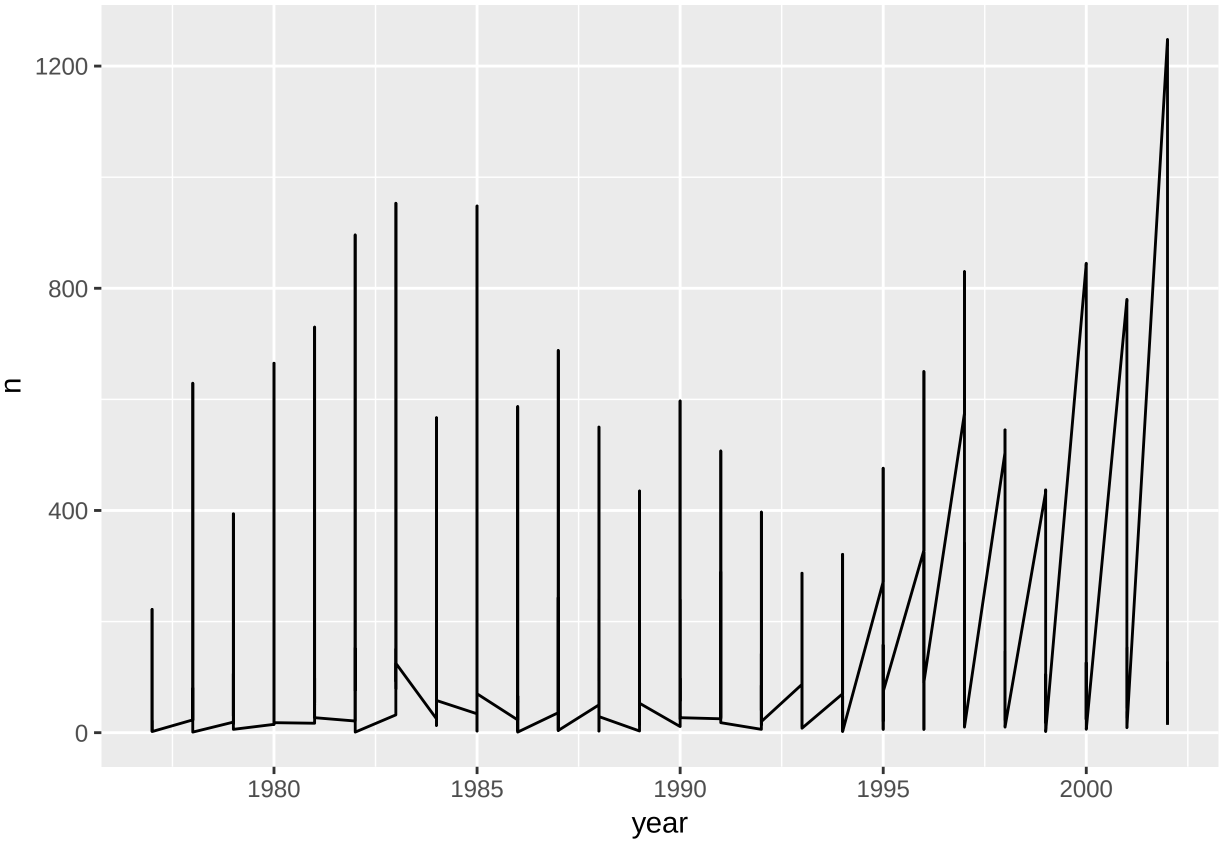
```
# Columns
surveys %>%
  group_by(year, sex) %>%
  summarise(num_animals=n()) %>%
  ggplot()+
  geom_col(aes(x=year, y=num_animals, fill=sex))
```



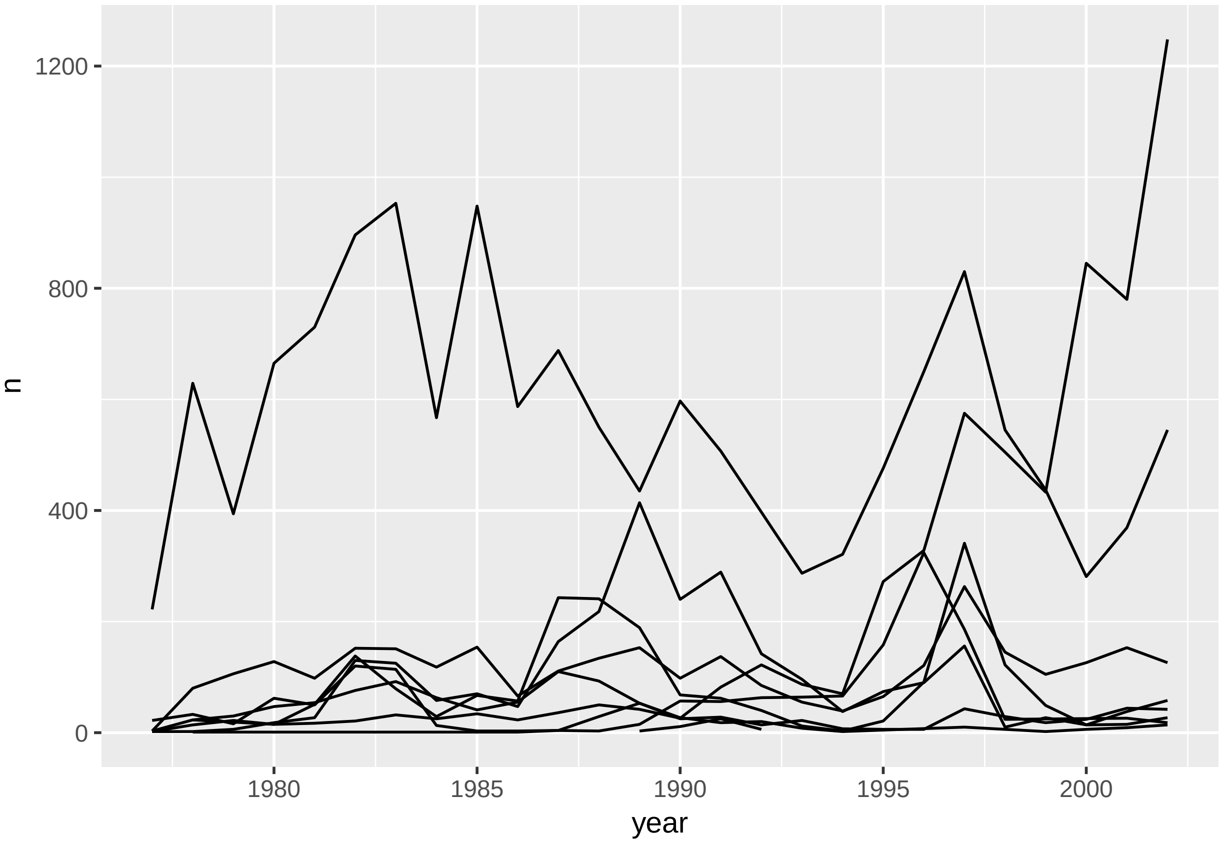
```
# Histogram
ggplot(surveys, aes(x=weight))+geom_histogram()+
  facet_wrap(~genus, scales='free')
```



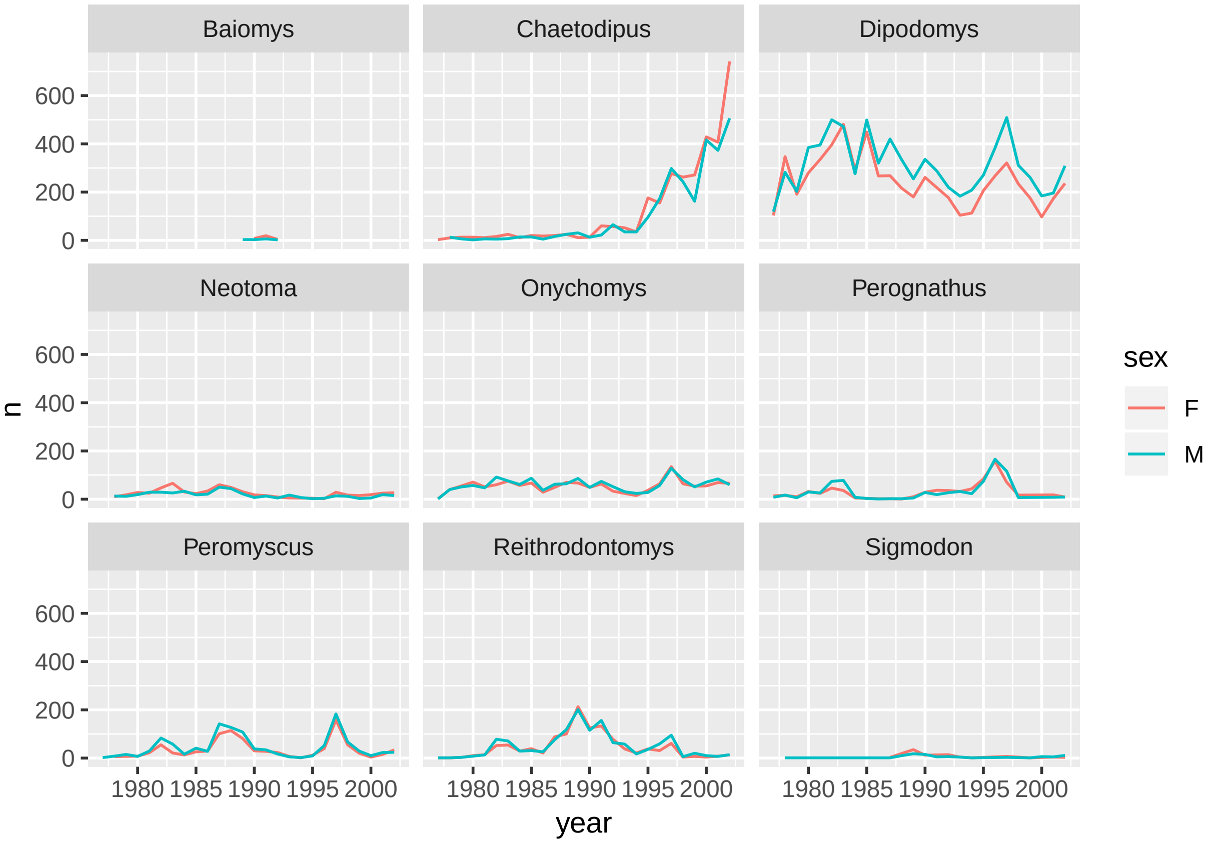
```r
# Time series
yearly_counts <- surveys %>%
  count(year, genus)

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



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



```r
yearly_sex_counts <- surveys %>%
  count(year, genus, sex)

ggplot(data = yearly_sex_counts, mapping = aes(x = year, y = n, color = sex)) +
  geom_line() +
  facet_wrap(facets =  vars(genus))
```


Baiomys
Chaetodipus
Dipodomys
Neotoma
Onychomys
Perognathus
Peromyscus
Reithrodontomys
Sigmodon
600
400
200
0
n
1980
1985
1990
1995
2000
year
sex
F
M