MATH 154 - HW2 - Data Viz

Ra-Zakee Muhammad

due: Thursday, Sepemper 16, 2021

summary

In this assignment, you will think carefully about graphics and visualization. You will work with some of R's graphical tools (`ggplot()`) as well as working to create a new and improved visualization from a graphic you have found elsewhere.

The assignment will be to collect (and clean?) a dataset, and then use that dataset to construct a graphic using R's **ggplot2** package. You will apply ideas from class on deconstructing a plot as well as coding with R to construct a graphic.

Additionally, we will focus on graphics that are not particularly good or convincing. By trying to communicate details of and then re-plot information that has not been conveyed well, you will practice the art of good graphics.

requisites

- There are many **ggplot2** tutorials. One that I particularly like is https://rstudio.cloud/learn/primers/3.
- If you are so inclined (in order to create an interactive graphic), install the package **shiny**. As much as needed, walk through the shiny 'Learn Shiny' tutorial: http://shiny.rstudio.com/tutorial/.
- My website includes links to many datasets and compilations of datasets. https://hardin47.netlify.app/courses/data/
- TidyTuesday has a plethora of great datasets, and they are generally quite clean to start with. https://github.com/rfordatascience/tidytuesday/tree/master/data

assignment

1. **Pod Q** Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, something fun, etc.) 1-3 sentences.

Danny taught me that the Bob Ross data set counts the number of trees in the painting and tells us the probability of a painting possessing trees in it. he also taught me that gs4_deauth() was a necessary line of code for problem 4 of this assignment.

2. **Find a Dataset** Using any dataset you find online, create a figure using `ggplot()`. Ideally, you should be able to link directly to the dataset, and not have to download the data. If you do need to download the data, make sure it is saved into your GitHub repo so that your R Markdown file can knit reproducibly for a collaborator.

- The plot must have at least 2 `aes`thetics.
- The plot must use at least two different `geom`s.
- You might consider `facet`ing the plot according to a categorical variable.
- Be sure to provide the source of your data and information on the variables (a few sentences).

There is data here: https://hardin47.netlify.app/courses/data/

hint: The package `readr` (which lives inside the package `tidyverse`) can seamlessly read in most files / URLs. Hint: click on `Import Dataset` at the top right of the RStudio screen and use the GUI to help you figure out the correct code. [I'm serious about using the `Import Dataset` button to help you figure out what code goes into the .Rmd file, ask me if it isn't obvious after you try it out.]

For example, if you go here: https://dasl.datadescription.com/, you can download the data as:

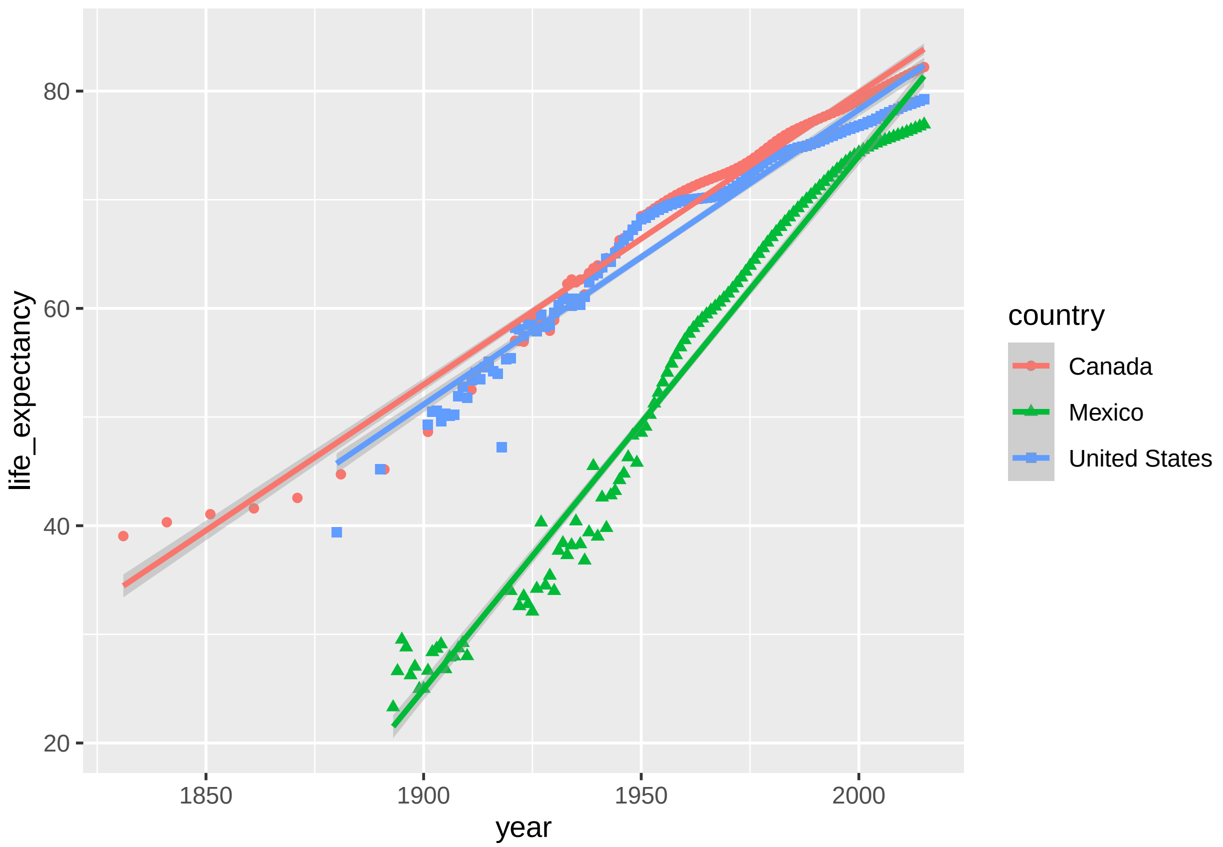
```
amazon <- readr::read_delim("https://dasl.datadescription.com/download/data/3052/amazon-books.txt",
                            delim = "\t", escape_double = FALSE, trim_ws = TRUE)
glimpse(amazon)
```

```
## Rows: 325
## Columns: 13
## $ Title          <chr> "1,001 Facts that Will Scare the S#*t Out of You: The U~
## $ Author         <chr> "Cary McNeal", "Ben Mezrich", "Smith", "Gavin Menzies",~
## $ ‘List Price‘   <dbl> 12.95, 15.00, 1.50, 15.99, 30.50, 28.95, 20.00, 15.00, ~
## $ ‘Amazon Price‘ <dbl> 5.18, 10.20, 1.50, 10.87, 16.77, 16.44, 13.46, 8.44, 18~
## $ ‘Hard/ Paper‘  <chr> "P", "P", "P", "P", "P", "H", "H", "P", "H", "H", "P", ~
## $ NumPages       <dbl> 304, 273, 96, 672, 720, 460, 336, 405, NA, 304, 624, 72~
## $ Publisher      <chr> "Adams Media", "Free Press", "Dover Publications", "Har~
## $ ‘Pub year‘     <dbl> 2010, 2008, 1995, 2008, 2011, 2011, 2010, 1987, 2011, 1~
## $ ‘ISBN-10‘      <chr> "1605506249", "1416564195", "486285537", "0061564893", ~
## $ Height         <dbl> 7.8, 8.4, 8.3, 8.8, 8.0, 8.9, 7.8, 8.2, 9.6, 9.6, 7.7, ~
## $ Width          <dbl> 5.5, 5.5, 5.2, 6.0, 5.2, 6.3, 5.3, 5.3, 6.5, 6.4, 5.1, ~
## $ Thick          <dbl> 0.8, 0.7, 0.3, 1.6, 1.4, 1.7, 1.2, 0.8, 2.1, 1.1, 1.7, ~
## $ ‘Weight (oz)‘  <dbl> 11.2, 7.2, 4.0, 28.8, 22.4, 32.0, 15.5, 11.2, NA, 19.2,~
```

```
life_expectancy_dat <- read.csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
glimpse(life_expectancy_dat)
```

```
## Rows: 17,894
## Columns: 4
## $ country         <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanis~
## $ code            <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG"~
## $ year            <int> 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, ~
## $ life_expectancy <dbl> 27.537, 27.810, 28.350, 28.880, 29.399, 29.907, 30.404~
```

```
North_American_Nations <- filter(life_expectancy_dat, country == "United States" | country == "Mexico"
ggplot(North_American_Nations,aes(x = year, y = life_expectancy, color = country,  shape = country)) +
```



3. **Lemurs** Earlier this summer, TidyTuesday posted a dataset on lemurs, there are several quantitative and categorical variables available in the dataset.

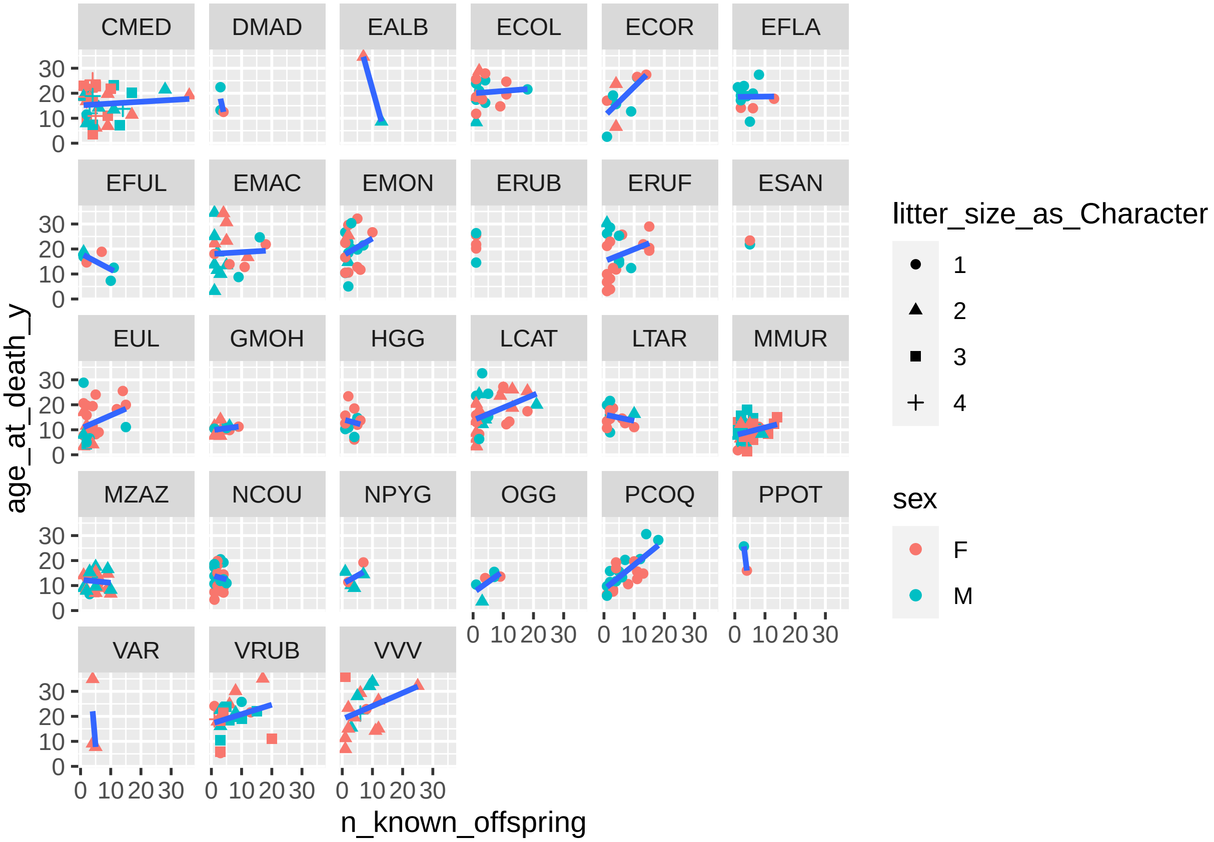
note: I selected only the most recent observation for each lemur (making the dataset ~2000 observations instead of ~8000 observations). If you are an advanced data wrangler, you are welcome to filter more or filter less depending on what you are hoping to see in the data.

See GitHub repo here: https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-08-24

```
lemurs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
  group_by(dlc_id) %>%
  arrange(weight_date) %>%
  slice(tail(row_number(), 1))
```

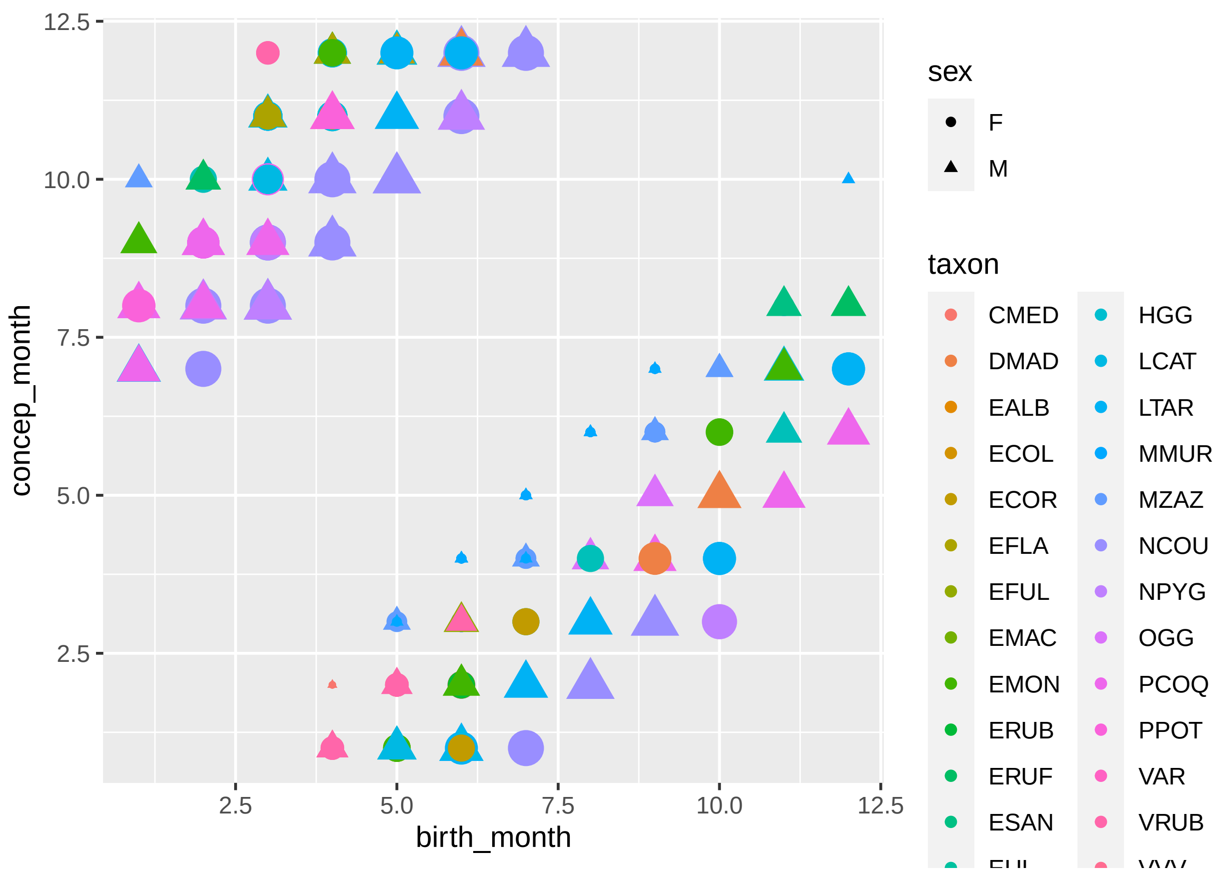
a. Create a single plot using as many variables as you can. [Note: this is *not* good graphical practice, you are likely to create a terrible graph! The assignment is merely an exercise to help you understand how to use visual cues and aesthetics!]

```
lemurs %>%
  filter(!is.na(age_at_death_y), !is.na(n_known_offspring), !is.na(birth_month), !is.na(litter_size))%>
  mutate(litter_size_as_Character = toString(litter_size))%>%
  ggplot(mapping = aes(x = n_known_offspring, y = age_at_death_y)) +
  geom_point(aes(color = sex, shape = litter_size_as_Character )) +
  geom_smooth(method = "lm", se =FALSE)+
  facet_wrap(~taxon)
```



b. Again, using the lemur data, create a single plot which is terrible. The worse the plot, the better. Include the worst plot you create. Explain why you think the graph is terrible.

```
lemurs %>%
  filter(!is.na(age_at_death_y), !is.na(n_known_offspring), !is.na(birth_month), !is.na(litter_size) ,
  ggplot(mapping = aes(x = birth_month, y = concep_month, size = expected_gestation, color = taxon, sha
geom_point()
```



This plot isn't a good because of the way that the data points layer on top of each other which prevents us from observing color and size differences among the points which is crucial to our understanding of the taxonomies and the length of gestation periods of the individual lemurs respectively. Additionally, we have no certainty as to whether points of different shapes are placed underneath larger points and this means that we cannot effectively observe the sex of individual data point potted on this scatter plot. We have no way of knowing how many points are in this plot because the overlapping.

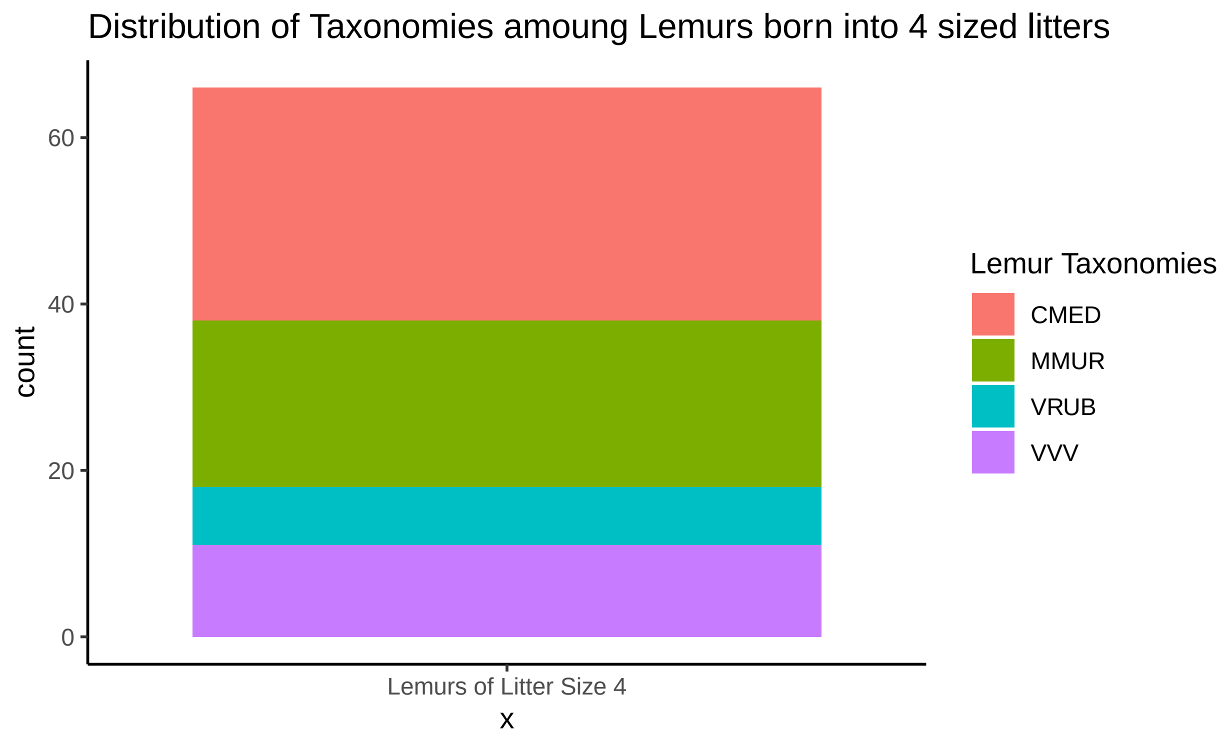
c. Again, using the lemur data, create a single plot that you like.

   - Provide a few sentences interpreting what the plot shows - as if for a figure caption.

   - Provide a few sentences interpreting what the plot shows - as if for a screen reader or an individual who wasn't actually looking at the image (this is called alt text).

```
alt_text = "This single bar chart  shows us how lemurs within our data set born into litters
of size 4 are distributed into taxonomies. The distribution is in the numerical count as
opposed to percentages and we get count information from the y axis. We also roughly gauge the
taxonomy distribution based on visual cues of color and proportions of the total bar."
```

```
lemurs %>%
  filter( !is.na(litter_size) )%>% filter(litter_size == 4)%>%
  ggplot(aes(x = "Lemurs of Litter Size 4", fill = taxon)) +
geom_bar() + theme_classic()  +labs(title = "Distribution of Taxonomies amoung Lemurs born into 4 sized
of size 4 are distributed into taxonomies. The distribution is in the numerical count as
opposed to percentages and we get count information from the y axis. We also roughly gauge the
taxonomy distribution based on visual cues of color and proportions of the total bar.")
```



Caption: This single bar chart shows us how lemurs within our data set born into litters of size 4 are distributed into taxonomies. The distribution is in the numerical count as o percentages and we get count information from the y axis. We also roughly gauge the taxonomy distribution based on visual cues of color and proportions of the total bar.

4. **Information is Beautiful** Check out the website Information Is Beautiful data: http://www.informationisbeautiful.net/data/ (click on the link in the google spreadsheet to see the different viz + data).

   - Find a plot on Information Is Beautiful that you can (at least somewhat) recreate using `ggplot()`. Maybe it violates at least one concept of effective data visualization discussed in class. (To upload the data, you may need to use `googlesheets4`,https://googlesheets4.tidyverse.org/), be sure to give the URL and citation associated with the plot.
   - Include the original image in your assignment (either link to it or upload the screenshot into your GitHub repo / compile into your pdf).
   - Write a few sentences about the original image, with a critique of what aspects of the plotting could be improved. Imagine you were going to correspond with the people who designed the plot, and give them guidance about how to make a more effective depiction of the data.
   - Using the data provided from Information is Beautiful, recreate the image to the best of your ability. Indeed, ideally you would also be able to improve the original image (e.g., simplify, color choices, axes location, reference line like y=x, etc.) in some way. (You may not be able to improve the plot overall.)

The original graph is very helpful in the way that it seperates percentages of the workers that are of different ethnicities into seperate graphs that are on a scale from 0 to 100. Additionally its very helpful that these percentages are literally placed on the bars of these faceted graphs. On the other hand this seperation of graphs could also be viewed as redundant provided that the different ethnicities are already color coded and can be explored more easiely using a key. Additionally, one half of the graphic is a stacked bar chart with women and men on the same chart and the sum of their percentages making a 100 length bar for each company. This could also have been done for ethnicities so as to save more space, and be more consistent.

The reason why this probably isn't how the graph is dislayed, is that the percentages of the this data frame may not add up to 100 when taking into consideration identities such as multi-ethnic, other, and undeclared.

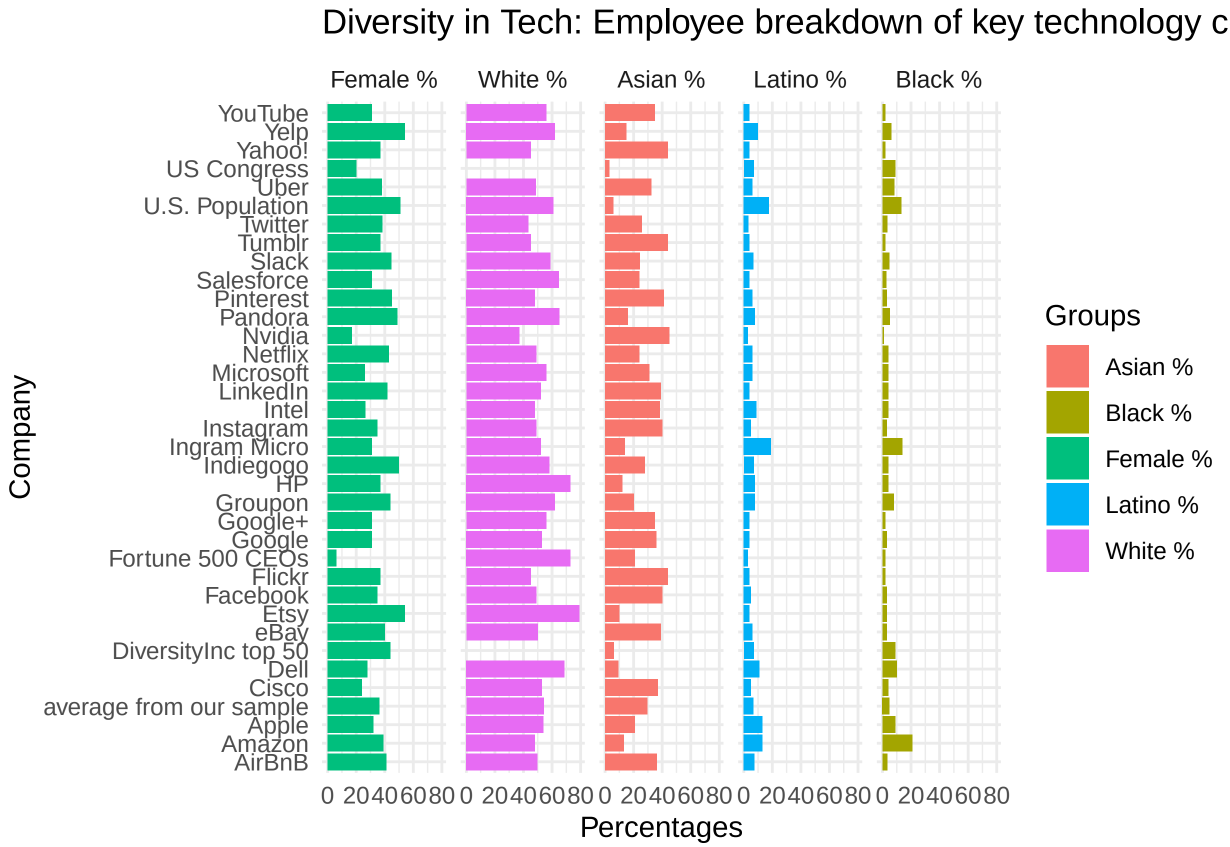
```
library(googlesheets4)
gs4_deauth()
diversity <- read_sheet("1e5jevLJTK9Aayob2msk4Ss9qIMCqfris4m_m0kXO-7s")
```

```
Cleaned_diversity <- diversity  %>% select("Company"= "...1","Female %"="% of total workforce - latest 

Cleaned_diversity <- Cleaned_diversity[-1, ] %>% pivot_longer(cols = -Company, names_to = "Groups", valu
Cleaned_diversity_1 <- Cleaned_diversity %>% mutate(graph_base = 100)
Cleaned_diversity_1
```

```
## # A tibble: 324 x 4
##    Company         Groups       Percentages graph_base
##    <chr>           <chr>        <list>           <dbl>
##  1 U.S. Population Female %     <dbl [1]>          100
##  2 U.S. Population Male %       <dbl [1]>          100
##  3 U.S. Population White %      <dbl [1]>          100
##  4 U.S. Population Asian %      <dbl [1]>          100
##  5 U.S. Population Latino %     <dbl [1]>          100
##  6 U.S. Population Black %      <dbl [1]>          100
##  7 U.S. Population multi %      <dbl [1]>          100
##  8 U.S. Population other %      <dbl [1]>          100
##  9 U.S. Population undeclared % <chr [1]>          100
## 10 Facebook        Female %     <dbl [1]>          100
## # ... with 314 more rows
```

```
Cleaned_diversity %>% filter(Groups == "Female %" | Groups == "White %" | Groups == "Asian %" | Groups =
```


Diversity in Tech: Employee breakdown of key technology cc
Female %
White %
Asian %
Latino %
Black %
Company
YouTube
Yelp
Yahoo!
US Congress
Uber
U.S. Population
Twitter
Tumblr
Slack
Salesforce
Pinterest
Pandora
Nvidia
Netflix
Microsoft
LinkedIn
Intel
Instagram
Ingram Micro
Indiegogo
HP
Groupon
Google+
Google
Fortune 500 CEOs
Flickr
Facebook
Etsy
eBay
DiversityInc top 50
Dell
Cisco
average from our sample
Apple
Amazon
AirBnB
0 20 40 60 80
Percentages
Groups
Asian %
Black %
Female %
Latino %
White %