Final Project

Dominic Colter

4/27/23

Table of Contents

Background

Aging has been of scientific interest for many years. With the advent of new technology and breakthrough scientific discoveries in biotechnology, new approaches are being examined to combat human aging. Over the past several decades, lifespan has significantly increased. Scientific successes in aging biology has created new technological avenues that are being exploited among scientific researchers. As a result, science is trying to extend lifespan even more.

In the scientific literature, model organism are being utilized to discover new insights about the genetics of aging. Various chemical compounds have been synthesized and tested in model organisms to observe any changes in longevity. Discovering effective drugs for delaying the aging process in model organisms is of considerable interest within the scientific community. The data from these various studies needs to be analyzed and summarized to potentially discover new approaches in human beings.

To analyze lifespan change with different chemical compounds in various model organisms, I used the DrugAge database provided by the The Human Ageing Genomic Resources (HAGR). The DrugAge database aims to provide “high-quality summary data on lifespan-extending drugs and compounds in model organisms” (Barardo et al., 2017). This

database contains 324 research articles from the scientific community. From this data, statistically significant lifespan extension data can be examined.

Problem Statement

Since the DrugAge database originally contains over 1000 entries, my goal is to cleanup and wrangle the data to focus on larger correlations. While the DrugAge database contains a variety of compounds that have been shown to increase the lifespan in various model organisms, which compounds should be prioritized and studied more to potentially combat human mortality? I would also like to discover which compounds showed the most success in what organisms and possibly at what dosages.

Data Definitions

A "readme" style was adopted from Cornell University Library for data definitions. Additional information about the DrugAge database including methods of data collection and variable parameters are included.

Methods

Exploring and Summarizing Data

The database was imported and wrangled to eliminate non-essential data entries for the analysis. Then, I converted all the variables to factors and renamed redundant entries. Lastly, I performed basic summary functions to take a peak at the data and some of its variables.

```
```{r}
library(tidyverse)
library(dplyr)
library(RColorBrewer)
library(ggtext)

#Read in raw data.
drugage_raw <- read_csv("./raw-data/drugage.csv")

#View raw data.
view(drugage_raw)

#Eliminate empty columns and PubMed ID column.
drugage_raw <- drugage_raw[,-c(10,11,12)]

#Omit all NA values.
drugage_raw <- na.omit(drugage_raw)

#Select for only significant data entries.
drugage_raw <- drugage_raw[drugage_raw$significance == "S",]
```
```

```
#Only 1 significant result, decided to eliminate it due to lack of data.
drugage_raw <- drugage_raw[drugage_raw$species != "Caenorhabditis elegans",]

#Single entries for these dosage amounts, decided to omit them due to lack of
data.
drugage_raw <- drugage_raw[drugage_raw$dosage != "0.01",]
drugage_raw <- drugage_raw[drugage_raw$dosage != "0.002",]
```

```{r}
#Assign wrangled data to a new variable and convert variables to factors.
drugage_final <- drugage_raw %>%
  mutate(compound_name = as.factor(compound_name),
         cas_number = as.factor(cas_number),
         species = as.factor(species),
         dosage = as.factor(dosage),
         gender = as.factor(gender),
         significance = as.factor(significance),
         strain = as.factor(strain))

#Rename redudant entries for the gender variable.
drugage_final <- drugage_final %>%
  mutate(gender=recode(gender,
                       "Male" = "MALE",
                       "male" = "MALE",
                       "MALE" = "MALE",
                       "Female" = "FEMALE",
                       "FEMALE" = "FEMALE",
                       "female" = "FEMALE",
                       "Hermaphrodite" = "HERM",
                       "HERMAPHRODITE" = "HERM",
                       "Mixed" = "BOTH",
                       "Pooled" = "BOTH"))

#Shorten this strain name for visulizations
drugage_final <- drugage_final %>%
  mutate(strain=recode(strain, "cross between Canton S females with Oregon RC
males" = "Canton-Oregon"))

#View final dataset and perform summary methods
view(drugage_final)
summary(drugage_final)
names(drugage_final)

str(drugage_final)
```

```
                                  compound_name      cas_number
 Curcumin                                :13    458-37-7  :13
```

```
 Rapamycin                                    :11    53123-88-9:11
 Genistein                                    :10    446-72-0  :10
 2-Ethyl-6-methylpyridin-3-ol hydrochloride: 6    616-91-1  : 7
 beta-Aminopropionitrile fumarate             : 5    13258-59-8: 6
 Hydrocortisone acetate                       : 4    352-96-5  : 5
 (Other)                                      :71    (Other)   :68
                     species         strain       dosage
avg_lifespan_change
 Aedes aegypti           : 8   Canton S :13   100 µM : 9   Min.   :-42.860
 Drosophila melanogaster :70   Oregon-R :11   100 mM : 9   1st Qu.:  4.742
 Mus musculus            :35   Ra       :11   5 ppm  : 6   Median : 13.000
 Nothobranchius guentheri: 2   UM-HET3  :11   0.3µM  : 4   Mean   : 12.129
 Rattus norvegicus       : 5   Canton-S :10   0.01%  : 3   3rd Qu.: 23.075
                               Liverpool: 8   0.10%  : 3   Max.   : 64.790
                               (Other)  :56   (Other):86
 max_lifespan_change    gender   significance
 Min.   :-40.000     FEMALE:51   S:120
 1st Qu.: -1.312     MALE  :69
 Median :  6.565
 Mean   :  7.606
 3rd Qu.: 16.035
 Max.   : 50.000

[1] "compound_name"       "cas_number"          "species"
[4] "strain"              "dosage"              "avg_lifespan_change"
[7] "max_lifespan_change" "gender"              "significance"
tibble [120 × 9] (S3: tbl_df/tbl/data.frame)
 $ compound_name      : Factor w/ 53 levels "17-α-estradiol",..: 2 2 2 2 2 2
37 12 12 12 ...
 $ cas_number         : Factor w/ 50 levels "103-90-2","107-35-7",..: 7 7 7 7
7 7 6 19 19 19 ...
 $ species            : Factor w/ 5 levels "Aedes aegypti",..: 2 2 2 2 2 2 2
3 3 3 ...
 $ strain             : Factor w/ 30 levels "(w1118/+)","129/Sv",..: 12 12 12
12 23 23 12 15 15 15 ...
 $ dosage             : Factor w/ 71 levels "0.00%","0.01%",..: 7 2 7 2 7 2 1
21 21 21 ...
 $ avg_lifespan_change: num [1:120] 23 18 14.3 20.4 15 22 -37 4.57 5.79 5.79
...
 $ max_lifespan_change: num [1:120] 21 2.32 -2.73 12.4 10 17 -33 -1.29 -0.77
6.73 ...
 $ gender             : Factor w/ 2 levels "FEMALE","MALE": 1 1 2 2 2 2 2 2 2
2 ...
 $ significance       : Factor w/ 1 level "S": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:3050] 1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "names")= chr [1:3050] "1" "2" "3" "4" ...
```

Results

The questions I would like to answer are as follows:

1. Based on their success in model organisms, which compounds should be prioritized and studied more?
2. Which compounds showed the most success and in what organisms? Is there a correlation with the studied genders?
3. What dosages are being used the most and which dosages are significant?

To answer these questions I decided to first visualize the proportion of species in the dataset. I created CSV files to save the proportions and their numeric quantities. I then wanted to see if there were trends between the average lifespan change, species, and gender.

```{r}
#Calculate the proportion of dosages in the dataset.
drugage_final_dosagepro <- drugage_final %>%
 group_by(dosage) %>%
 summarize(total = n()) %>%
 mutate(proportion = total / sum(total))

#Calculate the proportion of species in the dataset.
drugage_final_speciespro <- drugage_final %>%
 group_by(species) %>%
 summarize(total = n()) %>%
 mutate(proportion = total / sum(total))

#Save proportions as CSV files.
write.csv(drugage_final_dosagepro, "./data-output/drugprop.csv")
write.csv(drugage_final_speciespro, "./data-output/speciesprop.csv")

#Create a figure to visulize species proportion. Figure 1.
species_proportion_plot <- drugage_final_speciespro %>%
  ggplot(mapping = aes(species, proportion)) +
  geom_col(fill = "#6666ff") +
  scale_x_discrete(guide = guide_axis(check.overlap = TRUE)) +
  theme_light() +
  labs(
    title = "Proportion of Species",
    subtitle = "Selected from significant entries in DrugAge Database",
    x = "Species Name",
    y = "Proportion"
  ) +
  theme(
    text = element_text(family = "serif", color = "#063970"),
    axis.text.x = element_text(size = 9, color = "black", face = "italic"),
    axis.text.y = element_text(size = 9, color = "black"),
    plot.title = element_text(face = "bold", hjust = 0.5),
```

```
    plot.subtitle = element_text(hjust = 0.5, size = 8),
    panel.border = element_rect(colour = "#063970", size = 1)
  )

#Display figure.
species_proportion_plot

#Save figure as a png file.
ggsave("./fig-output/SpeciesProportion.png", species_proportion_plot, dpi =
300, width = 11, height = 8.5)
```

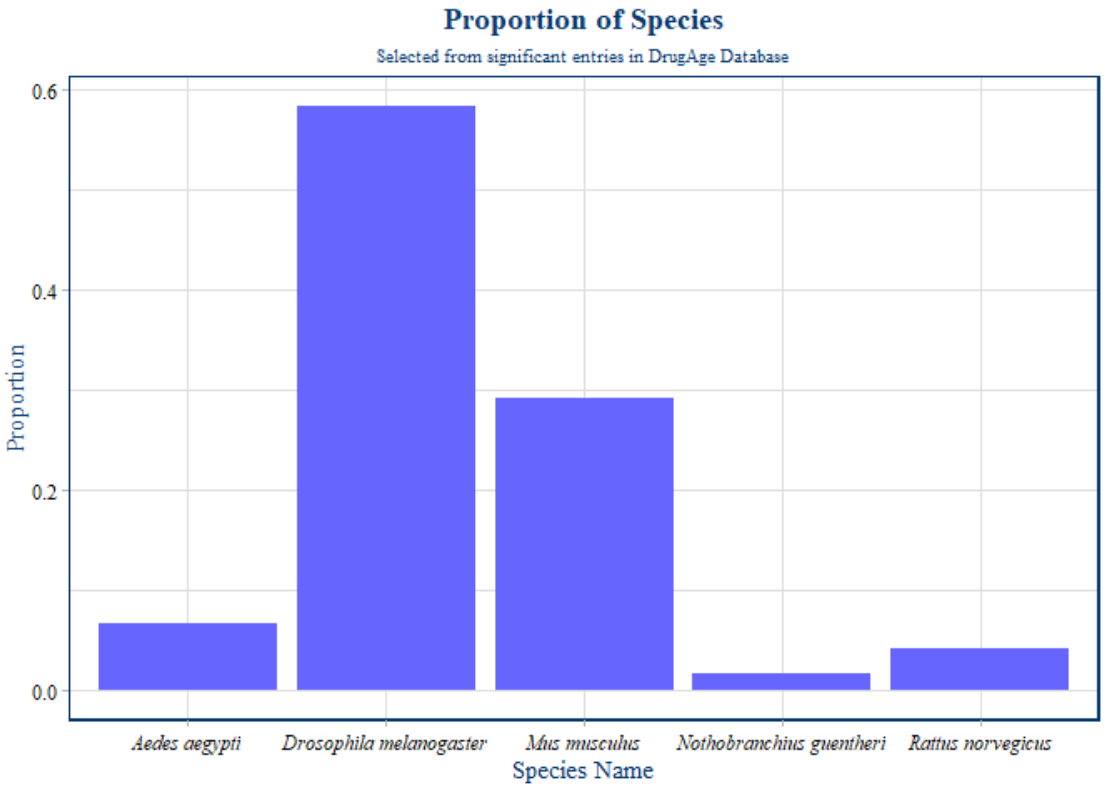

Figure 1. Proportion of Species

The goal of this figure is to visualize how the different species are distributed within the data after selecting for only significant data entries (121 entries). As shown by Figure 1, almost 60% of the data involves *Drosophila melanogaster* while *Mus musculus* comprises about 30% of the data. From these results, I can observe that most of the significant research findings included in the DrugAge database were using *Drosophila melanogaster* as their model organism. I may conclude that *Drosophila melanogaster* is popular among scientific researchers studying the aging process, highlighting its usefulness for scientific study. I decided to focus on *Drosophila melanogaster* and *Mus musculus* in my report to attempt to answer the questions presented by my problem statement.

```
```{r}
#Figure 2.
```
```

```
avgbyspecies_plot <- drugage_final %>%
  ggplot(mapping = aes(x = species, y = avg_lifespan_change)) +
  geom_violin(fill = "#6666ff") +
  theme_bw() +
  labs(
    title = "Average Lifespan Change by Species",
    subtitle = "Selected from significant entries in DrugAge Database",
    x = "Species Name",
    y = "Average Lifespan Change (%)"
  ) +
  theme(
    text = element_text(family = "serif", color = "#063970"),
    axis.text.x = element_text(size = 9, color = "black", face = "italic"),
    axis.text.y = element_text(size = 9, color = "black"),
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, size = 8),
    panel.border = element_rect(colour = "#063970", size = 1)
  ) +
  geom_boxplot(width=0.1)

#View the figure.
avgbyspecies_plot

#Save the figure as a png.
ggsave("./fig-output/AvgbySpecies.png", avgbyspecies_plot, dpi = 300, width =
11, height = 8.5)
```

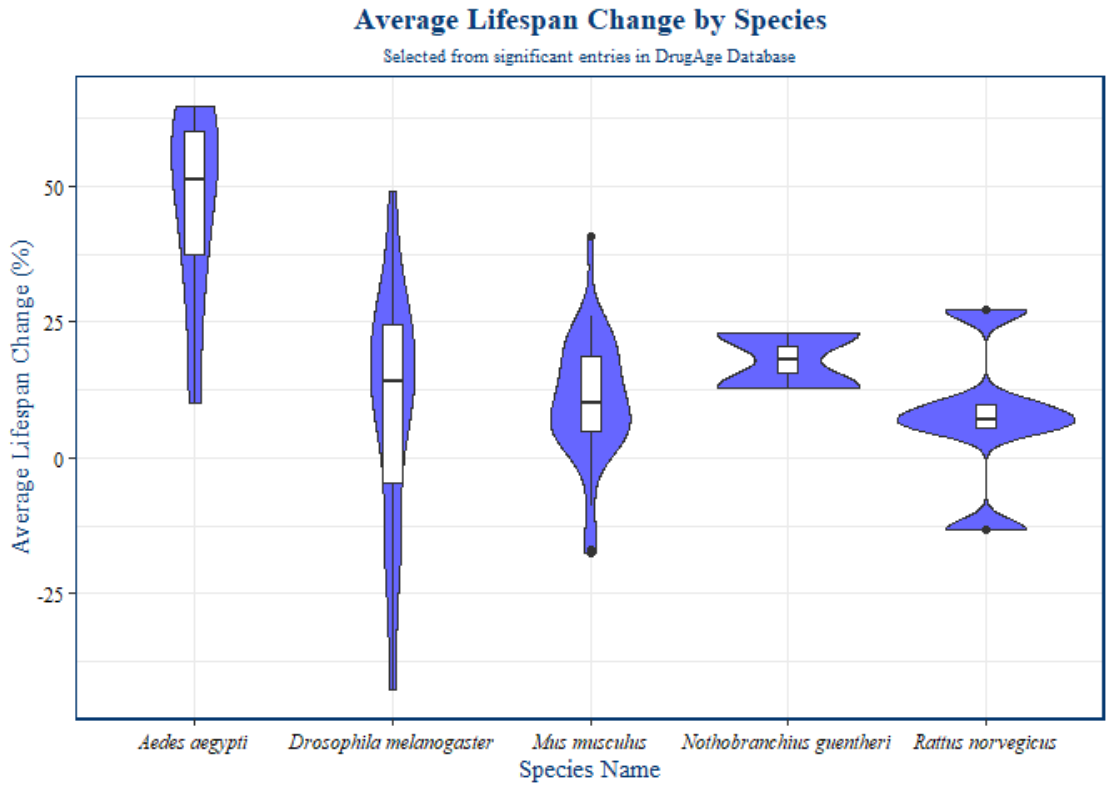

Figure 2. Average Lifespan Change by Species

Figure 2 illustrates the lifespan change based on species. The purpose of this figure is to identify which species has seen the most and least amount of change in lifespan. I decided to visualize the distribution of the data with a violin plot and a box plots. Interestingly, *Aedes aegypti* saw the highest average change in lifespan even though it posses a small proportion of the data (less than 10%). *Drosophila melanogaster* saw a wide spread in average lifespan change which may be due to the high amount of studies with this species in the dataset.

```
```{r}
#Figure 3.
avgbygender_plot <- drugage_final %>%
  filter(avg_lifespan_change > 0) %>%
  ggplot(mapping = aes(x = gender, y = avg_lifespan_change, fill = species))
+
  geom_col(position = "dodge") +
  scale_fill_brewer(palette = "Set1") +
  labs(
    title = "Average Lifespan Change by Gender",
    subtitle = "For each species of significance in the DrugAge database",
    x = "Gender",
    y = "Average Lifespan Change (%)"
  ) +
  theme_light() +
  theme(
```
```

```
      text = element_text(family = "serif", color = "#063970"),
      plot.title = element_text(face = "bold", hjust = 0.5),
      plot.subtitle = element_text(hjust = 0.5, size = 7),
      axis.text.x = element_text(size = 9, color = "black"),
      axis.text.y = element_text(size = 9, color = "black"),
      legend.title = element_blank(),
      legend.position = "top",
      legend.text = element_text(size = 8, face = "italic"),
      panel.border = element_rect(colour = "#063970", size = 1)
  )

#Show figure.
avgbygender_plot

#Save figure.
ggsave("./fig-output/AvgbyGender.png", avgbygender_plot, dpi = 300, width =
11, height = 8.5)
```

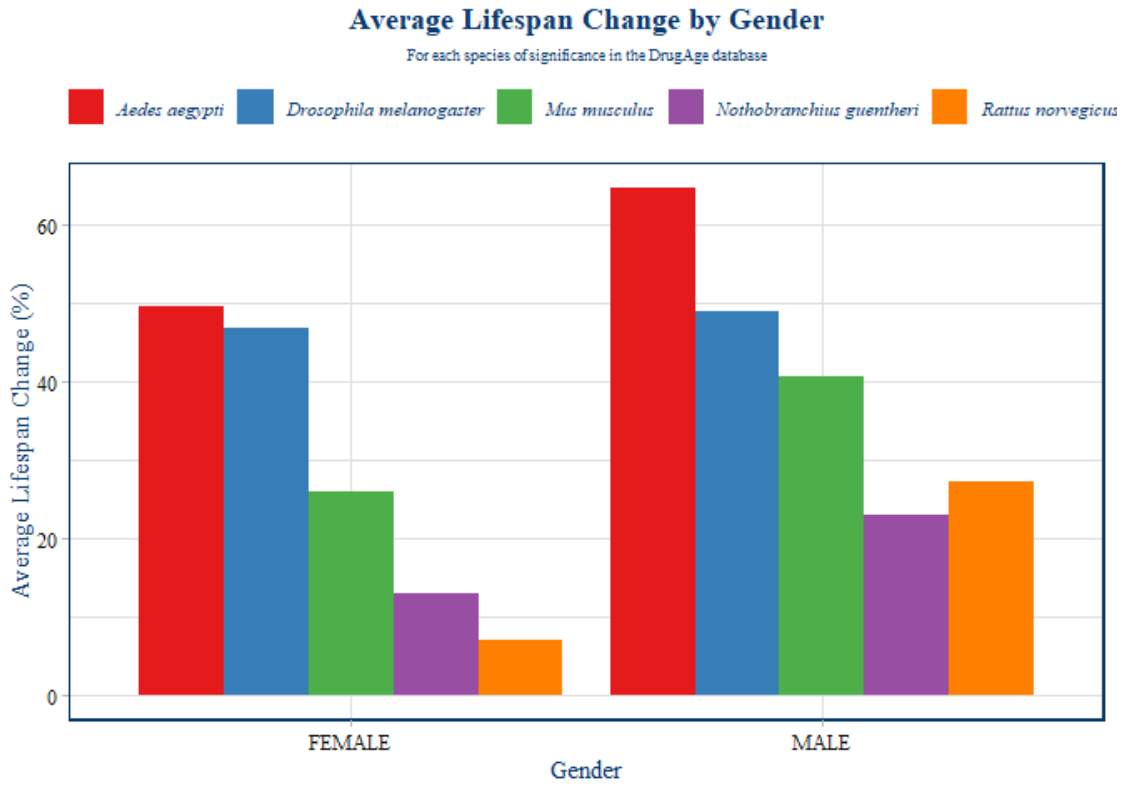

Figure 3. Average Lifespan Change by Gender

I then decided to see the distribution of gender among the species and see if there's a correlation with the average lifespan change. I wanted to see if there's any influence of gender on the research studies within the data. I also filtered the data to only include positive lifespan changes so that results that extended lifespan can be focused on. I decided to set the position of the column chart to dodge to allow visualization of both female and

male data for each species. Interestingly, only male and female data was preserved after I wrangled the data. As a result, male and female model organisms may be the most popular for studying aging. *Aedes aegypti*, *Drosophila melanogaster*, and *Mus musculus* showed the highest average lifespan change among both males and females with *Aedes aegypti* being the highest. *Rattus norvegicus* saw the lowest average percent change out of significant entries in the data. Male organisms seem to experience an increased average lifespan change across the species. Ultimately, the choice of male or female may depend on what organism and drug compound a researcher decides to study.

```{r}
#Count the frequency of dosages and compound names in the data.
raw_dosage_compound_count <- drugage_final %>%
  count(dosage, compound_name, sort = TRUE)

#View raw frequency counts.
raw_dosage_compound_count

#Filter raw counts to include only enteries with more than 2 counts in the
data.
dosage_compound_count <- raw_dosage_compound_count %>%
 filter(n>2)

#View final frequency counts.
dosage_compound_count

#Figure 4. Visualize counts and fill by dosage.
dosage_compound_count_plot <- dosage_compound_count %>%
  ggplot(mapping = aes(x = compound_name, y = n, fill = dosage)) +
  geom_col() +
  scale_x_discrete(guide = guide_axis(n.dodge = 2))+
  theme_light() +
  labs(
    title = "Selected Compounds by Count",
    subtitle = "Compounds with counts > 2",
    y = "Count",
    x = "Compound Name"
  ) +
  theme(
    text = element_text(family = "serif", color = "#063970"),
    axis.text.y = element_text(size = 12, color = "black"),
    axis.text.x = element_text(size = 12, colour = "black"),
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, size = 10),
    panel.border = element_rect(colour = "#063970", size = 1),
    legend.position = "top",
    legend.title = element_blank(),
    legend.text = element_text(size = 11)
  )
```

```
#View figure.
dosage_compound_count_plot

#Save figure
ggsave("./fig-
output/Selected_Dosage_Compound_Counts.png",dosage_compound_count_plot, dpi =
300, width = 11, height = 8.5)
```

```
# A tibble: 84 × 3
   dosage      compound_name                                n
   <fct>       <fct>                                    <int>
 1 100 mM      Curcumin                                     9
 2 0.01%       2-Ethyl-6-methylpyridin-3-ol hydrochloride     3
 3 0.10%       2-Ethyl-6-methylpyridin-3-ol hydrochloride     3
 4 1 mg/mL     beta-Aminopropionitrile fumarate               3
 5 0.3µM       Epicatechin                                    2
 6 0.3µM       Quercetin                                      2
 7 0.5% w/w    Ethoxyquin                                     2
 8 1 mg/g      Curcumin                                       2
 9 1 µL/100 mL Genistein                                      2
10 10 µM       Trichostatin A                                 2
# … with 74 more rows
# A tibble: 4 × 3
  dosage  compound_name                                n
  <fct>   <fct>                                    <int>
1 100 mM  Curcumin                                     9
2 0.01%   2-Ethyl-6-methylpyridin-3-ol hydrochloride     3
3 0.10%   2-Ethyl-6-methylpyridin-3-ol hydrochloride     3
4 1 mg/mL beta-Aminopropionitrile fumarate               3
```

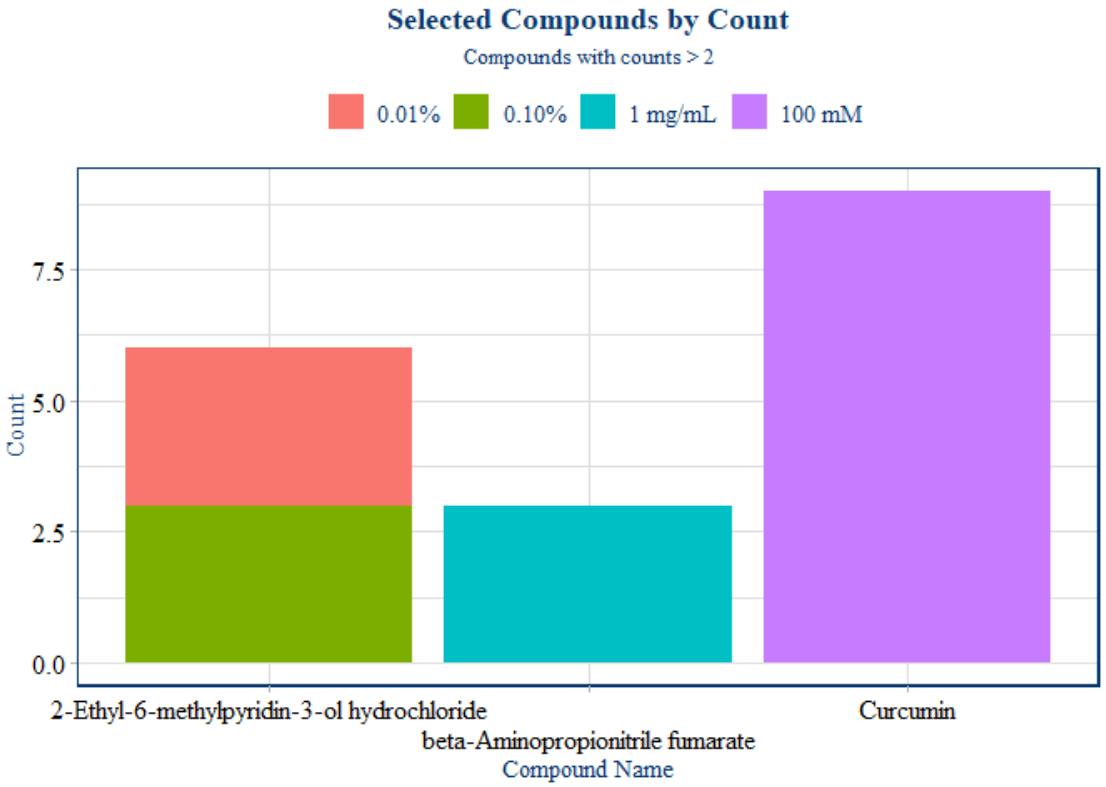

Figure 4. Selected Compounds by Count

In this figure, I wanted to see what compounds and dosages were used the most within the data. I discovered that the dataset had compounds and dosages with several single count entries. These single entries caused problems when attempting to create visualizations. To solve this, I decided to include only counts of 2 or higher. Interestingly, filtering the data in this way significantly reduced the number of entries. Therefore, the three compounds in the figure are of the highest counts in the data. Of these compounds, curcumin at a dosage of 100 mM was used the most in *Drosophila melanogaster*. However, this does not suggest that this compound at this dosage was the most effective at extending lifespan.

```{r}
#Select for Drosophila melanogaster from the dataset.
dm <- drugage_final %>%
  filter(str_detect(drugage_final$species, "Drosophila melanogaster"))

#Figure 5.
dm_lifespan_compound <- dm %>%
  filter(avg_lifespan_change > 25) %>%
  ggplot(mapping = aes(x = compound_name, y = avg_lifespan_change, label =
avg_lifespan_change)) +
  geom_text(nudge_x = 0.40, check_overlap = TRUE, size = 3)+
  geom_point(aes(color = factor(dosage)), size = 4) +
  coord_flip()+
  theme_light()+
  labs(
```

```
    title = "**Compounds by Lifespan Change in *Drosophila melanogaster***",
    subtitle = "Average lifespan change > 25%",
    x = "Compound Name",
    y = "Average Lifespan Change (%)"
  ) +
  theme(
    text = element_text(family = "serif", color = "#063970"),
    axis.text.y = element_text(size = 10, color = "black"),
    axis.text.x = element_text(size = 10, colour = "black"),
    plot.title = element_markdown(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, size = 10),
    legend.position = "right",
    legend.title = element_blank(),
    legend.text = element_text(size = 10)
  )

#View the figure.
dm_lifespan_compound

#Save the figure.
ggsave("./fig-output/Dm_lifespan_compound.png",dm_lifespan_compound, dpi =
300, width = 11, height = 8.5)
```

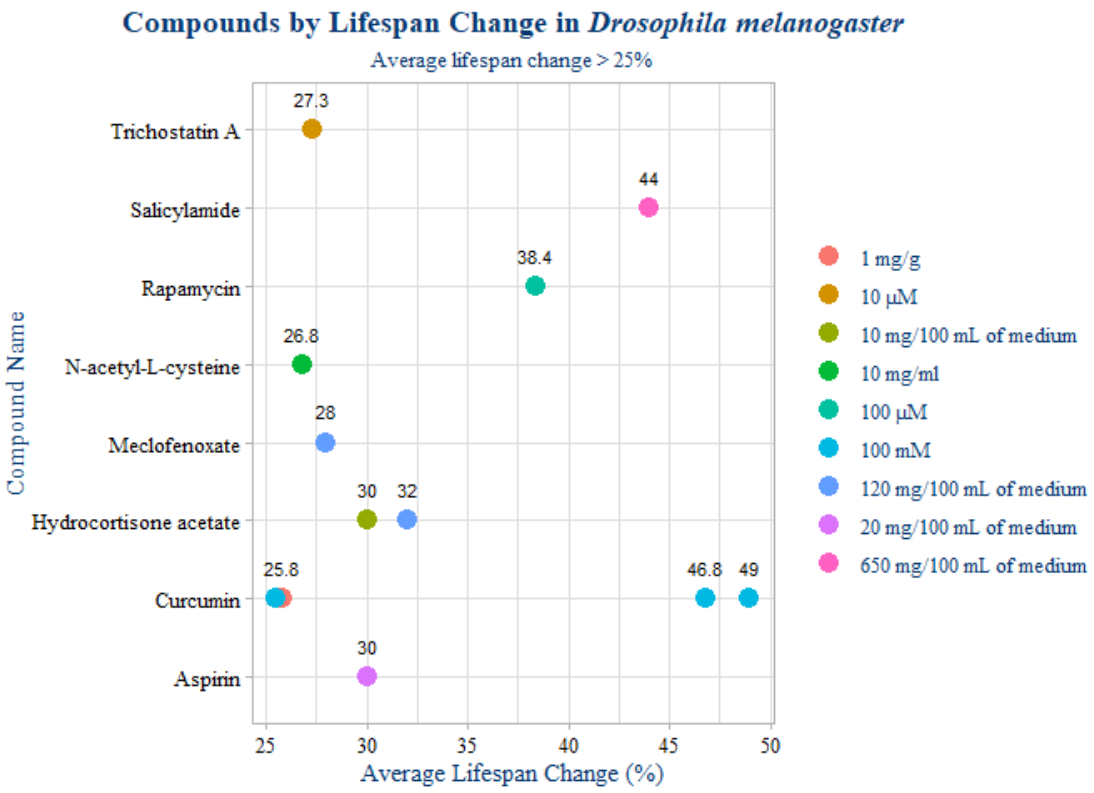

Figure 5. Compounds by Lifespan Change in Drosophila melanogaster

To see which compounds and dosages were most effective at extending lifespan, I focused on the two species that were studied the most within the data. In this figure, I am observing the average lifespan changes with compounds and dosages used in *Drosophila melanogaster*. I decided to trim the dataset to include only *Drosophila melanogaster* data that showed an average lifespan change greater than 25%. This percentage seemed to show most of the significant results within this subset of the data. As shown in this figure and in figure 4, curcumin at 100mM was the most effective at increasing the average lifespan in *Drosophila melanogaster.* 650 mg/100 mL of medium also increased the average lifespan considerably.

```{r}
#Select for Mus musculus from the dataset.
mu <- drugage_final %>%
  filter(str_detect(drugage_final$species, "Mus musculus"))

#Figure 6.
mu_lifespan_compound <- mu %>%
  filter(avg_lifespan_change > 10) %>%
  ggplot(mapping = aes(x = compound_name, y = avg_lifespan_change, label =
avg_lifespan_change)) +
  geom_text(nudge_x = 0.40, check_overlap = TRUE, size = 3)+
  geom_point(aes(color = factor(dosage)), size = 4) +
  coord_flip()+
  theme_light()+
  labs(
    title = "**Compounds by Lifespan Change in *Mus musculus***",
    subtitle = "Average lifespan change > 10%",
    x = "Compound Name",
    y = "Average Lifespan Change (%)"
  ) +
  theme(
    text = element_text(family = "serif", color = "#063970"),
    axis.text.y = element_text(size = 10, color = "black"),
    axis.text.x = element_text(size = 10, colour = "black"),
    plot.title = element_markdown(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, size = 10),
    legend.position = "right",
    legend.title = element_blank(),
    legend.text = element_text(size = 10)
  )

#Show figure.
mu_lifespan_compound

#Save figure.
ggsave("./fig-output/Mu_lifespan_compound.png",mu_lifespan_compound, dpi =
300, width = 11, height = 8.5)
```

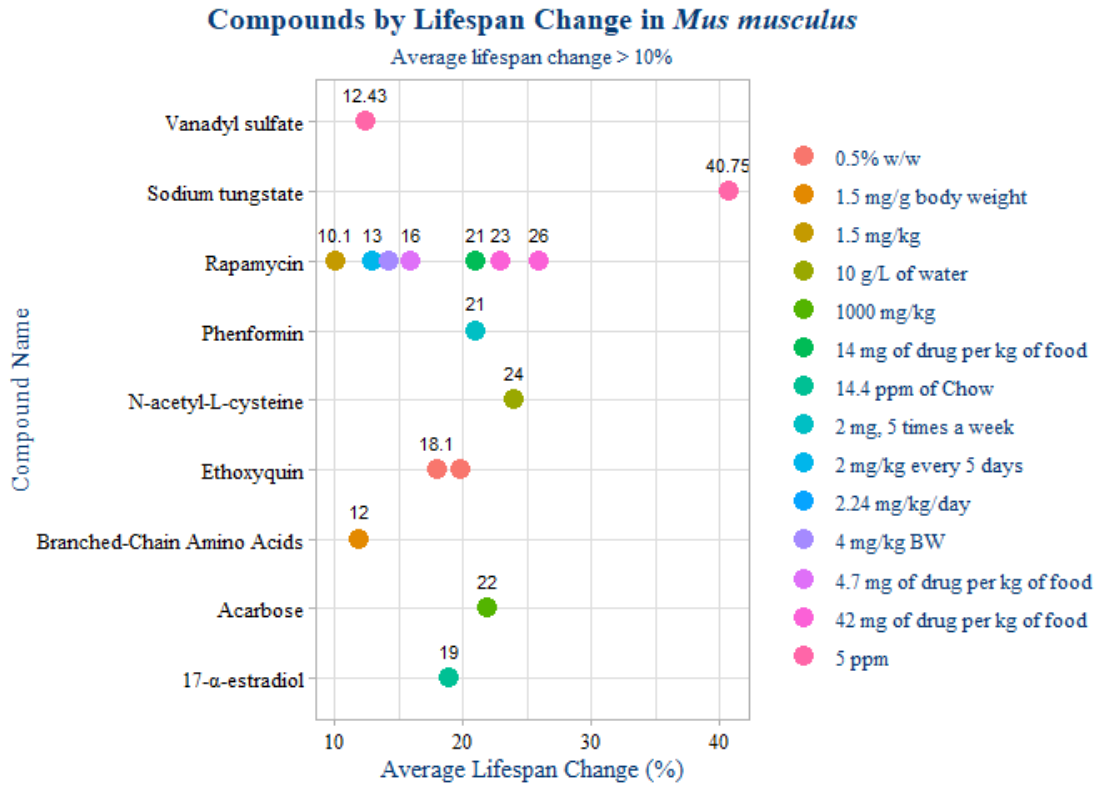

Figure 6. Compounds by Lifespan Change in Mus musculus

I also sought to see which compounds and dosages were most effective at extending lifespan in *Mus musculus* since it comprised 30% of the data. Like in figure 5, I isolated the *Mus musculus* data from the dataset. I then decided to only include compounds and dosages that increased the average lifespan by more than 10%. 10% seemed to represent an ample amount of data to analyze. This figure shows that rapamycin was used the most in this species with a dosage of 42 mg of drug per kg of food being the most effective at increasing the average lifespan. However, sodium tungstate possessed the highest increase in average lifespan change at a dosage of 5ppm.

Conclusions

As an undergraduate biology major, biological data inspires me to learn more as a student. Making design choices and observations in RStudio was relatively new to me. However, I attempted to make thoughtful decisions that could help me answer the questions I had about the DrugAge database. Initially, I observed that the database included many qualitative variables. Therefore, I knew I had to use visualizations that were best at representing relationships between qualitative variables. Violin plots and bar charts were of high interest for me in this analysis. For the quantitative variables, I decided to use boxplots and scatter plots to visualize correlations because they seemed to work best for this database.

The wrangling process did trim the raw dataset considerably. This made the dataset easier to visualize but much of the original entries were removed. The raw dataset seemed to

contain mostly non-significant data entries and a few empty columns. I wanted to only focus on significant data as determined by the original database. While most of the data was not analyzed, I think that this choice was best for making valuable conclusions from the data. Additional data can be found on the DrugAge website (*Drugage database of anti-ageing drugs,* n.d.) that can further supplement this data. In the future, I could analyze another database similar to the DrugAge database to compare and contrast with these results.

Problems Encountered

Initially, I had problems with wrangling the data and creating the criteria for trimming down the original dataset. I considered multiple options such as removing all the NA values and removing single entries. These approaches worked the best and I think I kept a decent sample size of significant data for the analysis. I also had problems with the labels on my figures and explored ways to keep the original labels without sacrificing clarity. The best approach was flipping the plot axis to position the long labels on the y-axis. This allowed for additional space and it looked more clear overall. I explored abbreviating the labels and renaming some of them manually from the raw dataset. However, these approaches were either too time consuming or sacrificed clarity in the plots.

Metacognition

During this course, I learned many new skills that I think will benefit me in my future career. I was always interested in programming and using computers to aid in analyzing biological data. This course allowed me to follow this interest and expand on it. I also gained practice in problem solving and taking initiative in discovering my own solutions for problems that arose in my code. I think that this practice was beneficial for me overall as a student and as a scientist.

I think data visualization is important to learn because it can be used in a variety of disciplines. For my interests, data visualization can be critical for conveying scientific data in journal articles and presentations. I think the experience I gained throughout this course will help me in my scientific research and will make me a more effective communicator. Communicating results and findings with others is essential for making lasting discoveries. This course allowed me to access additional avenues for communicating data and observations that will only benefit me in the future.

References

Barardo, D., Thornton, D., Thoppil, H., Walsh, M., Sharifi, S., Ferreira, S., Anžič, A., Fernandes, M., Monteiro, P., Grum, T., Cordeiro, R., De-Souza, E. A., Budovsky, A., Araujo, N., Gruber, J., Petrascheck, M., Fraifeld, V. E., Zhavoronkov, A., Moskalev, A., & de Magalhães, J. P. (2017). The DrugAge database of aging-related drugs. Aging cell, 16(3), 594–597. https://doi.org/10.1111/acel.12585

Drugage database of anti-ageing drugs. DrugAge: Database of Ageing-Related Drugs. (n.d.). Retrieved April 26, 2023, from https://genomics.senescence.info/drugs/index.php