

HW#3

Lucía Carrera

10/1/2021

```
knitr::opts_chunk$set(echo = TRUE)

# load the tidyverse here -- OR-- load dplyr and ggplot2
library(tidyverse)
```

Question 1: Mapping Aesthetics – 4 pts

Variable – Aesthetic

- Year -> x
- Totalprop -> y
- Letter -> facet
- Sex -> fill

```
ggplot(data = babies, mapping = aes(x=year, y=totalprop, color=sex)) + geom_line() + facet_wrap(~ letter)
```

Question 2: dplyr

part a – lists in order 3 pts

```
# View and Help
# View(msleep)
# help(msleep)

# Method 1
msleep %>% filter(!is.na(awake), !is.na(name)) %>% arrange(desc(awake)) %>% select
(name, awake) %>% head(6)
```

```
## # A tibble: 6 x 2
##   name                awake
##   <chr>              <dbl>
## 1 Giraffe            22.1
## 2 Pilot whale       21.4
## 3 Horse             21.1
## 4 Roe deer          21
## 5 Donkey            20.9
## 6 African elephant  20.7
```

```
# Method 2
# msleep_nameAwake <- msleep %>% select(name, awake)
# msleep_nameAwake %>% arrange(desc(awake)) %>% head(6)
```

```
msleep %>% filter(!is.na(awake), !is.na(name)) %>% arrange(awake) %>% select(name, awake) %>% head(6)
```

```
## # A tibble: 6 x 2
##   name                awake
##   <chr>              <dbl>
## 1 Little brown bat    4.1
## 2 Big brown bat      4.3
## 3 Thick-tailed opossum 4.6
## 4 Giant armadillo    5.9
## 5 North American Opossum 6
## 6 Long-nosed armadillo 6.6
```

part b – group_by – 3 pts

```
mpgby <- msleep %>% group_by(sleep_rem)
msleep %>% group_by(vore) %>% filter(!is.na(vore)) %>%
  summarise(meanawake = mean(awake),
            sdawake = sd(awake),
            n_vore = n())
```

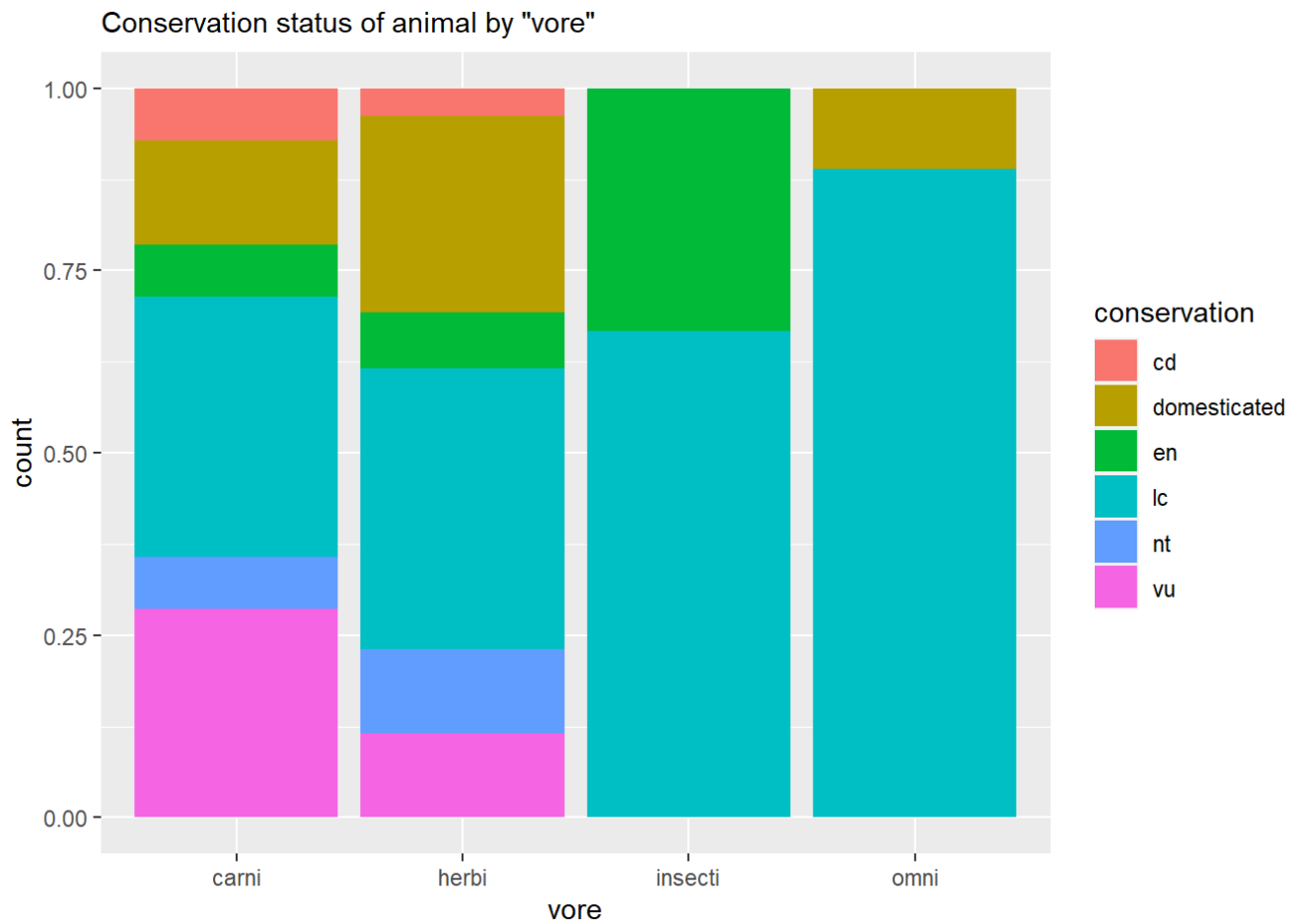
```
## # A tibble: 4 x 4
##   vore      meanawake sdawake n_vore
##   <chr>      <dbl>    <dbl> <int>
## 1 carni      13.6      4.68    19
## 2 herbi      14.5      4.88    32
## 3 insecti     9.06      5.92     5
## 4 omni       13.1      2.95    20
```

Question 3

part a – bar graph – 3 pts

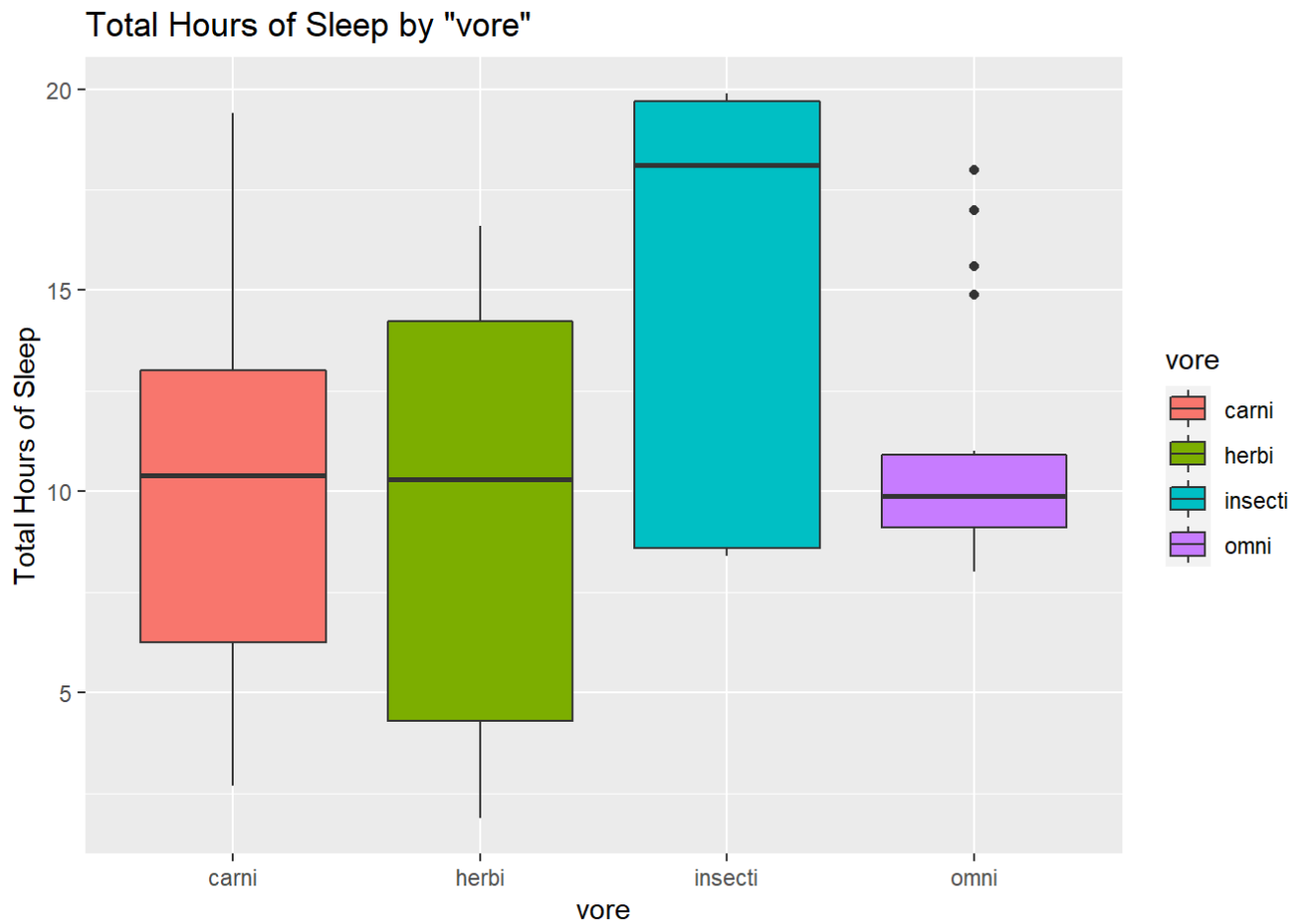
```
vores_nona <- msleep %>% filter(!is.na(conservation), !is.na(vore))

ggplot(data = vores_nona,
       mapping = aes( x = vore, fill = conservation)) +
  geom_bar(position = 'fill') +
  labs(subtitle = 'Conservation status of animal by "vore"')
```



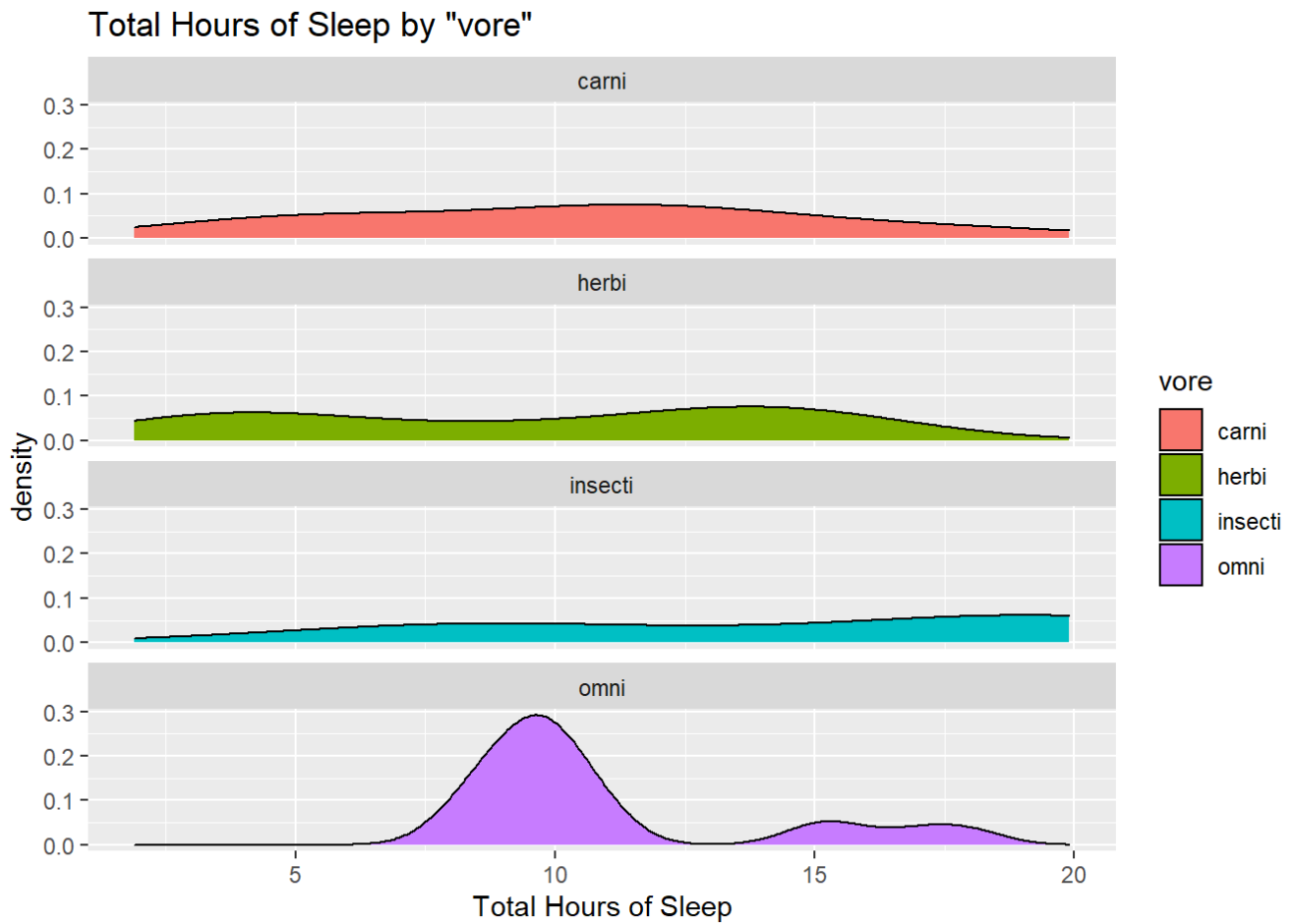
part b – box plot – 3 pts

```
msleep %>% filter(!is.na(vore)) %>%  
  ggplot(mapping = aes(x = vore, y = sleep_total, fill = vore)) +  
  geom_boxplot() +  
  labs(title = 'Total Hours of Sleep by "vore"', y = 'Total Hours of Sleep')
```



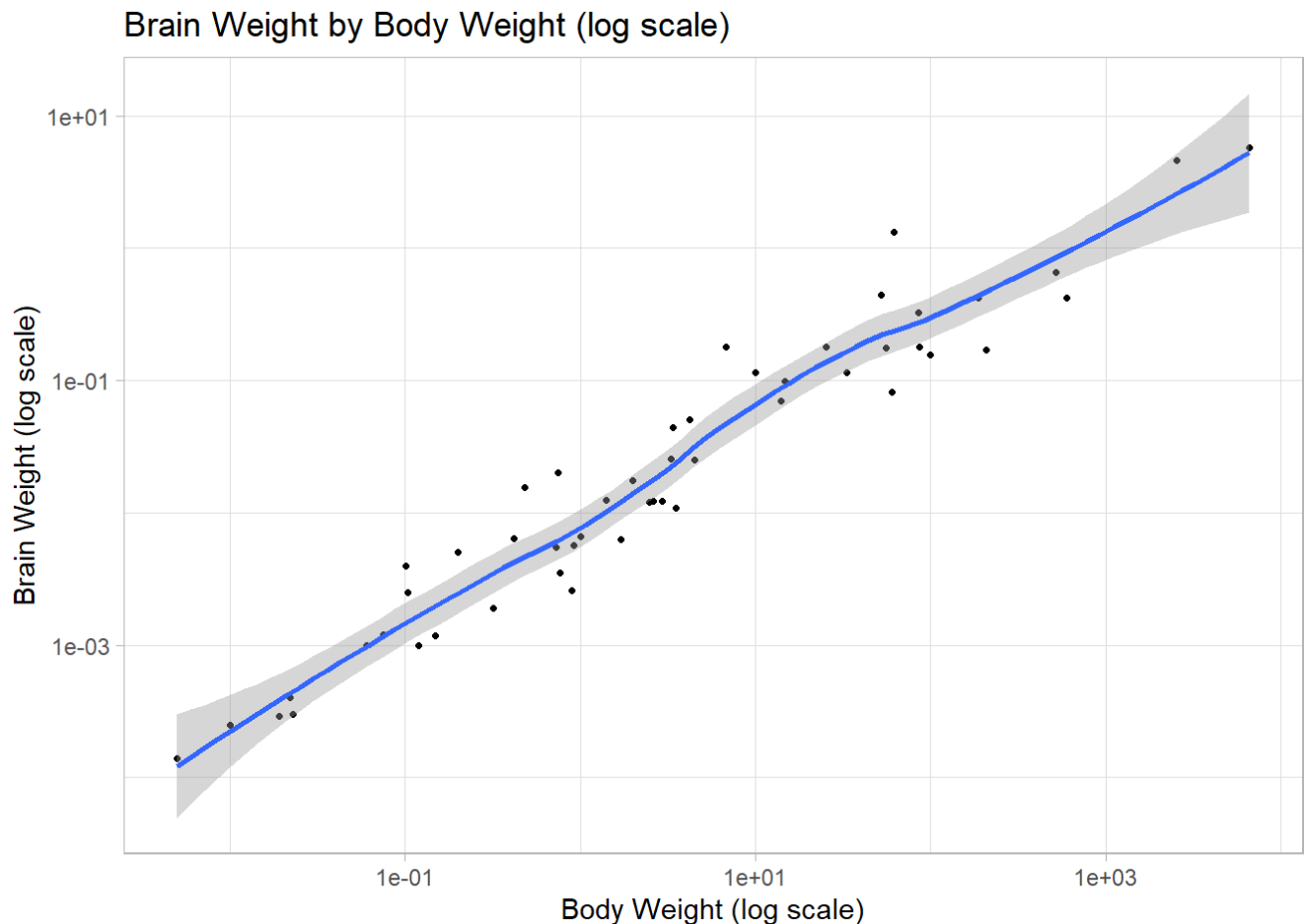
part c – facets – 3 pts

```
msleep %>% filter(!is.na(vore)) %>%  
  ggplot(mapping = aes(x = sleep_total, fill = vore)) +  
  geom_density() +  
  labs( title = 'Total Hours of Sleep by "vore"', x = 'Total Hours of Sleep') +  
  facet_wrap(~ vore, ncol = 1)
```



part d – logtrans – 3 pts

```
msleep %>% filter(!is.na(vore)) %>%
  ggplot(mapping = aes(x = bodywt, y=brainwt)) +
  geom_point(size=.9) +
  geom_smooth() +
  scale_x_log10() +
  scale_y_log10() +
  labs( title = 'Brain Weight by Body Weight (log scale)', x = 'Body Weight (log scale)', y = 'Brain Weight (log scale)' ) +
  theme_light()
```

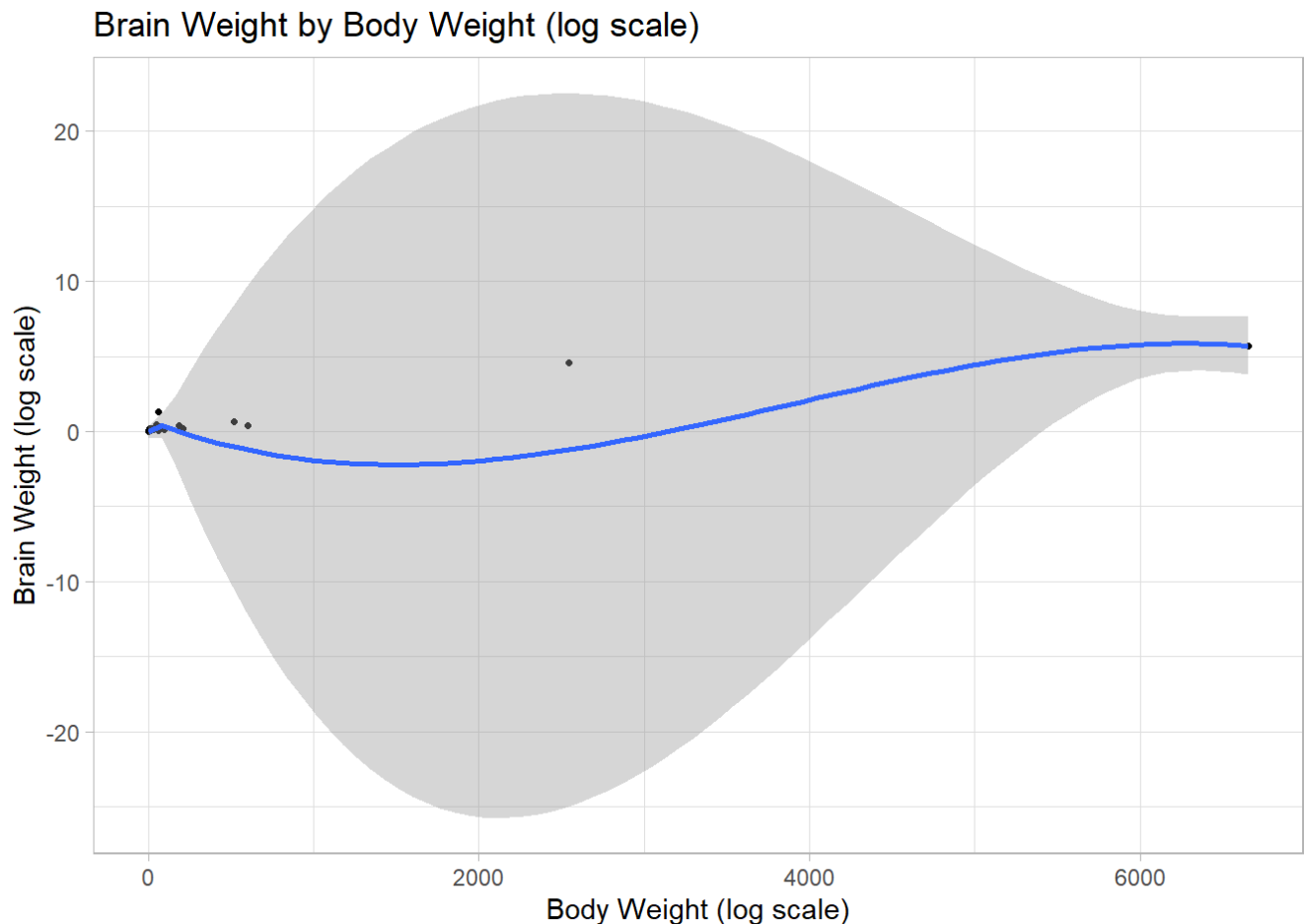


part e – whylog – 4 pts

<why is the log transformation useful? Put answer here > The log transformation is used when our original data does not follow the bell curve. With this mathematical tool we can transform the data so that its base is not as skewed. If we observe the graph below, most of the points are in the left and then the gray region occupies a large part of the graph. This is due to the data being skewed.

```
# Graph here the same scatterplot without the log transformation
msleep %>% filter(!is.na(vore)) %>%
  ggplot(mapping = aes(x = bodywt, y=brainwt)) +
  geom_point(size=.9) +
  geom_smooth() +
  labs( title = 'Brain Weight by Body Weight (log scale)', x = 'Body Weight (log scale)', y = 'Brain Weight (log scale)') +
  theme_light()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Question 4: Lie Factor – 4 pts

< Type in calculations, answer, and describe the result here > Using the following [website] (https://eleif.net/photo_measure.html) I was able to measure the exact graphical value of the bars. Effect in...

GraphicData

Barry Bonds bar	7.2	762
Babe Ruth bar	2.4	714

Formula

LIE_effect = SIZE_effect_graphic / SIZE_effect_data
 SIZE = (VALUE_one - VALUE_two) / VALUE_two

Comparison (Barry Bonds and Babe Ruth bar):

SIZE_effect_graphic = $((7.2 - 2.4) / 2.4) * 100 = 200\%$
 SIZE_effect_data = $((762 - 714) / 714) * 100 = 6.72\%$

LIE_effect = $200 / 6.72 = 29.76$

Conclusion

The Lie effect of this graph is not in range according to Tufte (0.95 to 1.05). In fact, it is very much over the desired number, meaning that the graph presented overstates the effect in the data. The bargraph is not directly proportional to the numerical data, hence we should not be making conclusions of the data collected according to it. Without making calculations, one could see that there is a potential problem by seeing that it does not commence in zero but instead at 690. If the entire y axis had been shown we would have seen that there is very little difference between each bar.