BIOS611 Project1

Minxin Lu

9/23/2018

Dataset 1 catsM

```
mean(catM_tb$Bwt)
```

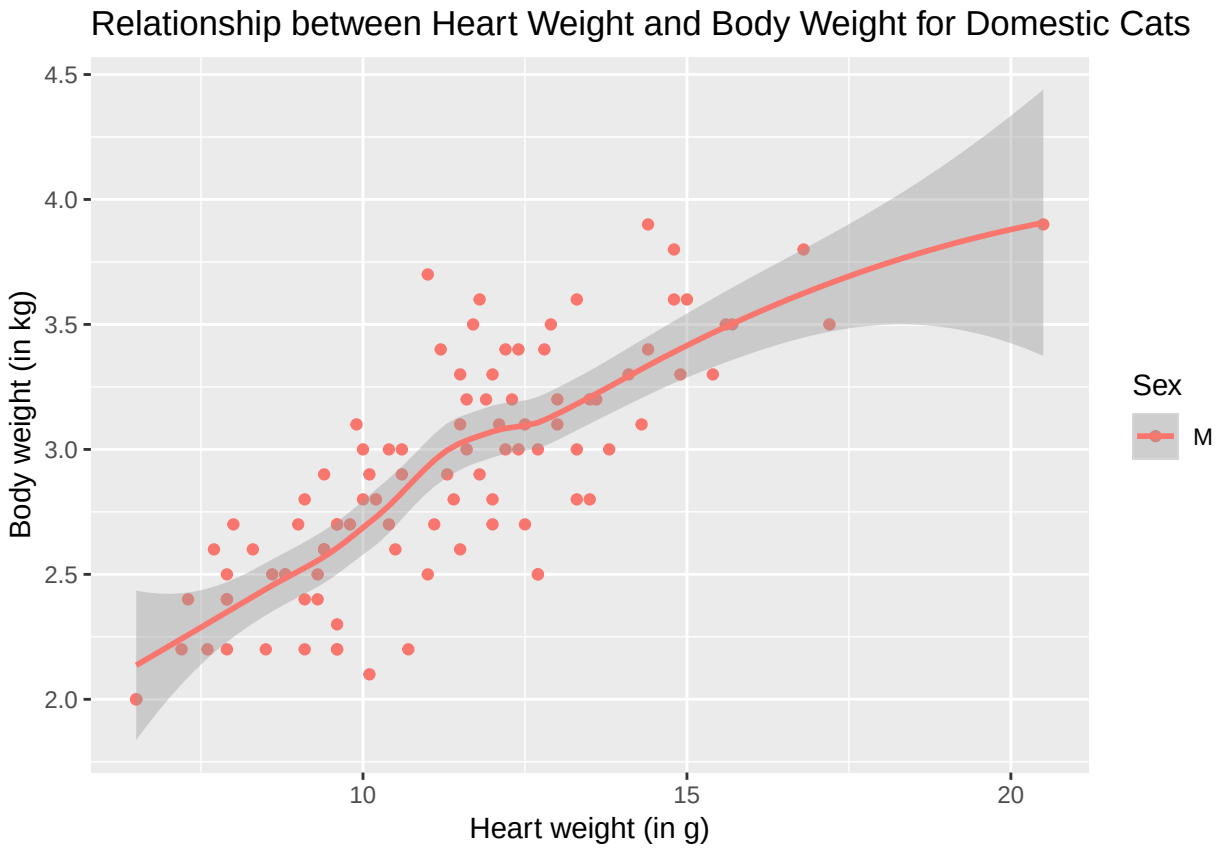
```
## [1] 2.9
```

```
mean(catM_tb$Hwt)
```

```
## [1] 11.32268
```

```
ggplot(data = catM_tb, mapping = aes(x = Hwt,y = Bwt,color = Sex)) +
  geom_point()+
  geom_smooth()+
  xlab ("Heart weight (in g)")+
  ylab ("Body weight (in kg)")+
  ggtitle("Relationship between Heart Weight and Body Weight for Domestic Cats")
```

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

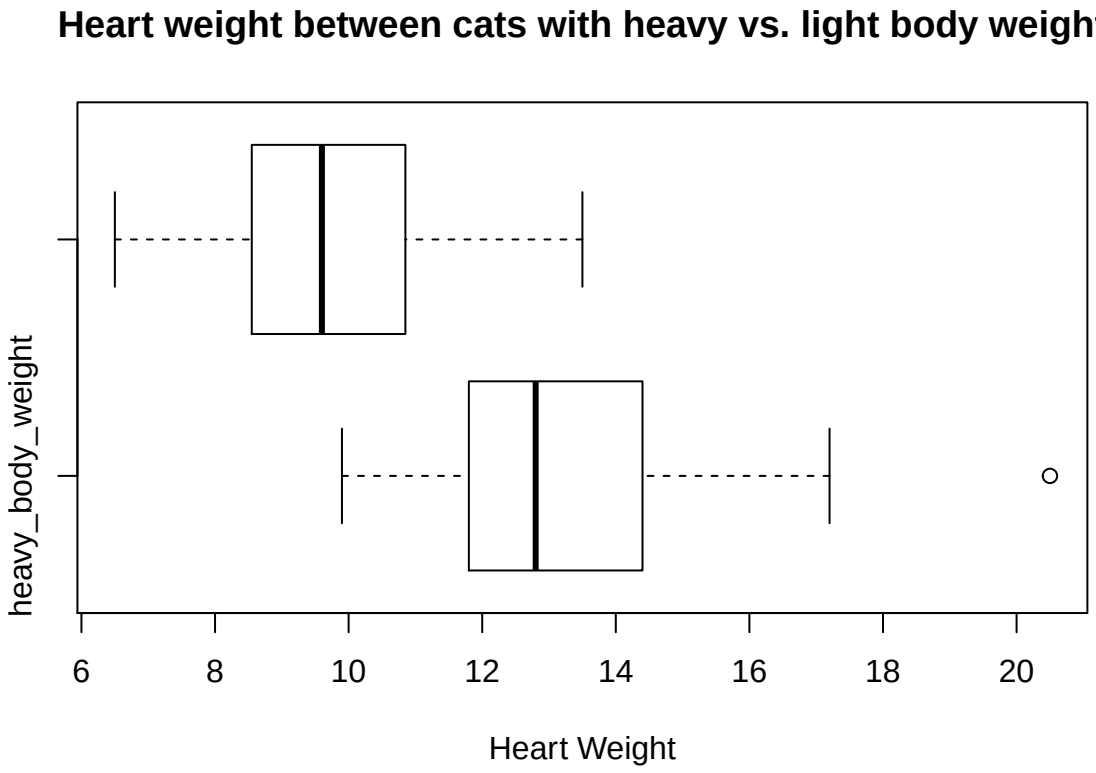



```
heavy_body_weight = catM_tb %>%
  filter(Bwt > mean(catM_tb$Bwt))
```

```
light_body_weight = catM_tb %>%
  filter(Bwt <= mean(catM_tb$Bwt))

heavy_Hwt <- heavy_body_weight$Hwt
light_Hwt <- light_body_weight$Hwt

Bodyweight<- c("heavy_body_weight", "light_body_weight")
boxplot(heavy_Hwt, light_Hwt, names=Bodyweight, horizontal = TRUE,
        main = "Heart weight between cats with heavy vs. light body weight",
        xlab = "Heart Weight")
```



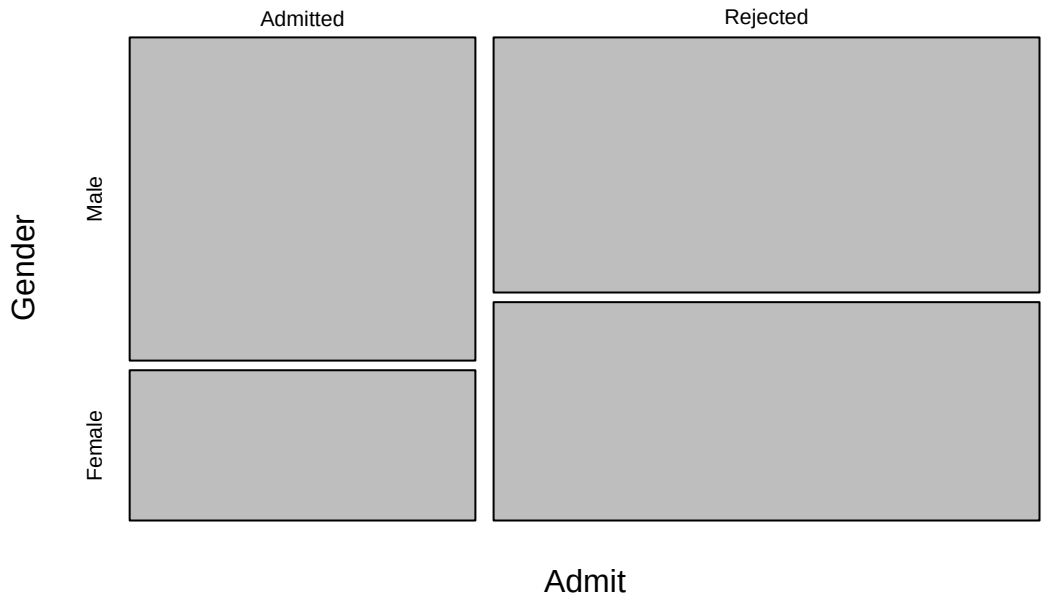
This figure was derived from the "catM" data set in R, which has 97 rows for 97 male adult (over 2 kg in weight) domestic cats. For each cat, its sex, body weight, and heart weight are recorded. The mean body weight among these 97 male cats is 2.9 kg. The mean heart weight is 11.32 g.We observe a positive relationship bewteen cat body weight and heart weight. If we group the body weight into hearvy and light, we can observe a difference in heart weight between 2 groups: The group with heavy body weight has larger median heart weight than the group with light body weight. As a next step, it will be interesting to compare female cats with male cats. We can observe whether the positive relationship between body weight and heart weight still exist, and if so, how strong is the relationship compared with the relationship we observed among male cats.

Dataset 2 UCBAdmissions

```
UCB_tb <- as_tibble(UCBAdmissions)

require(graphics)
```

```
mosaicplot(apply(UCBAdmissions, c(1, 2), sum),
           main = "Student Admissions at UC Berkeley")
```




```
by_gender_Admit <- UCB_tb %>%
  group_by(Gender, Admit) %>%
  summarise(ntotal = sum(n, na.rm = TRUE))%>%
  arrange(desc(ntotal))
by_gender_Admit
```

```
## # A tibble: 4 x 3
## # Groups:   Gender [2]
##   Gender Admit    ntotal
##   <chr>  <chr>     <dbl>
## 1 Male   Rejected   1493
## 2 Female Rejected   1278
## 3 Male   Admitted   1198
## 4 Female Admitted    557
```

```
by_gender <- UCB_tb %>%
  group_by(Gender) %>%
  summarise(ntotal = sum(n, na.rm = TRUE))
by_gender
```

```
## # A tibble: 2 x 2
##   Gender ntotal
##   <chr>   <dbl>
## 1 Female   1835
## 2 Male     2691
```

```
by_Admit <- UCB_tb %>%
  group_by(Admit) %>%
  summarise(ntotal = sum(n, na.rm = TRUE))
by_Admit
```

```
## # A tibble: 2 x 2
##   Admit    ntotal
##   <chr>     <dbl>
## 1 Admitted   1755
## 2 Rejected   2771
```

```
admission_rate <- 1755 /(1755+2771);admission_rate
```

```
## [1] 0.3877596
```

```
admission_rate_M <- 1198 /2691;admission_rate_M
```

```
## [1] 0.4451877
```

```
admission_rate_F <- 557 /1835;admission_rate_F
```

```
## [1] 0.3035422
```

Results

This figure was derived from the "UCBAdmissions" data set in R. It is a list of the aggregate data on applicants to graduate school at UC Berkeley for the six largest departments in 1973. Admission decision, sex, department code, and the number of applicants in each catagory are recorded. There are 2691 male applicants in total: 1198 of them are admitted, and the rest 1493 male applicants are rejected. Admission rate for male is around 44.5%. There are 1835 female applicants in total, 557 of them are admitted and 1278 of them are rejected. Admission rate for female is around 30.4%. In total there are 1755 admitted and 2771 rejected.The average admission rate is around 38.8%. From the figure we observe among admitted students, there are more males than females. But among rejected students, the number of male students are also larger than female students. Given that the number male applicants are greater than the number of female applicants, it is hard to decide whether the difference of admission rate between genders is significant.

As a next step, it would be interesting to construct a proper statistical test to decide whether the difference between gender is significant.

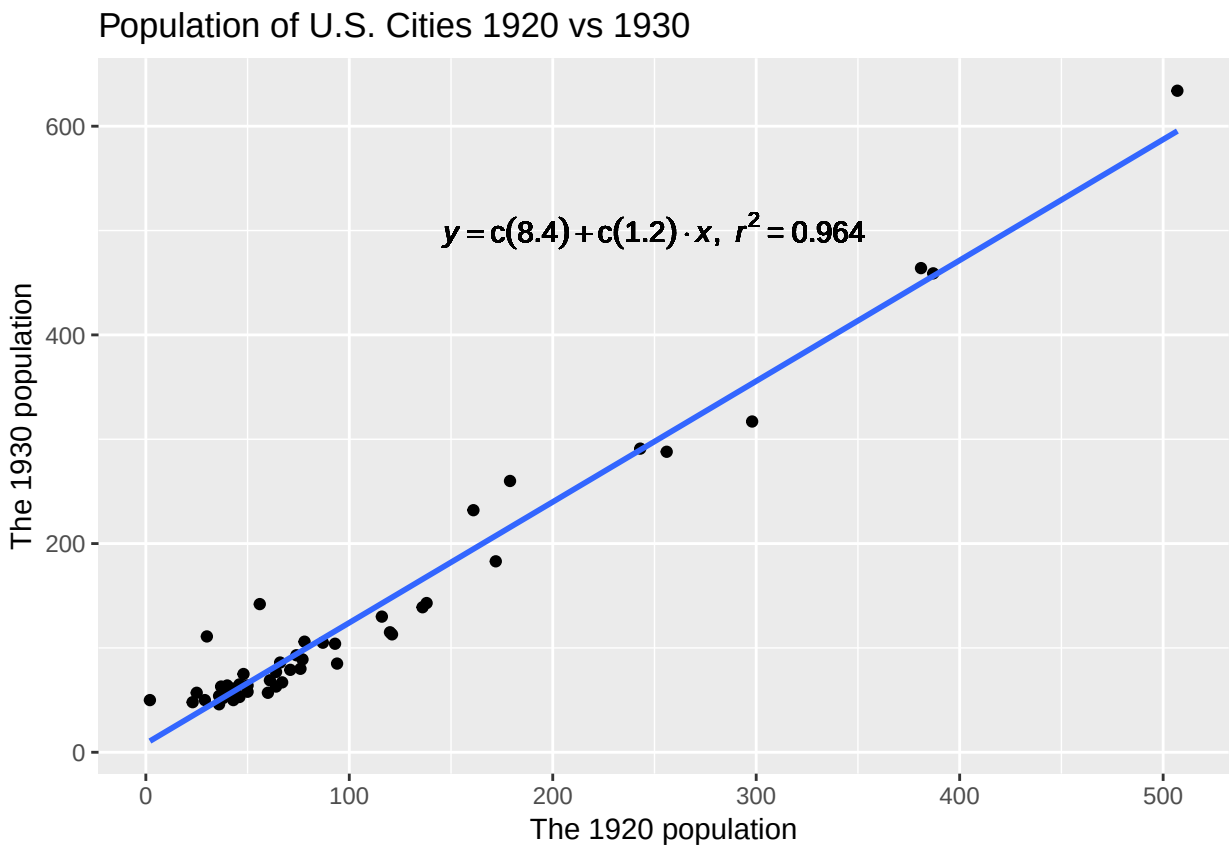
Dataset3 bigcity

```
data(bigcity)
city_tb <- as_tibble(bigcity)

p<-
  ggplot(data = city_tb,mapping = aes(x = u, y= x)) +
  geom_point()+
  geom_smooth(method="lm",se=FALSE)+
  xlab("The 1920 population")+
  ylab("The 1930 population")+
  labs(title ="Population of U.S. Cities 1920 vs 1930")

lm_eqn <- function(city_tb){
    m <- lm(x ~ u, city_tb);
    eq <- substitute(italic(y) == a + b %.% italic(x)*","~~italic(r)^2~"="~r2,
         list(a = format(coef(m)[1], digits = 2),
              b = format(coef(m)[2], digits = 2),
             r2 = format(summary(m)$r.squared, digits = 3)))
    as.character(as.expression(eq));
```

```
}
p1 <- p + geom_text(x = 250, y = 500, label = lm_eqn(city_tb), parse = TRUE);p1
```




```
city2 <- mutate(city_tb,
                growth = x-u,
                growth_rate = (x-u)/u)
arrange(city2,desc(growth_rate))
```

```
## # A tibble: 49 x 4
##        u     x growth growth_rate
##    <dbl> <dbl>  <dbl>       <dbl>
##  1     2    50     48      24
##  2    30   111     81       2.7
##  3    56   142     86       1.54
##  4    25    57     32       1.28
##  5    23    48     25       1.09
##  6    29    50     21       0.724
##  7    37    63     26       0.703
##  8    40    64     24       0.6
##  9    48    75     27       0.562
## 10    40    60     20       0.5
## # ... with 39 more rows
```

Results

This figure was derived from the "bigcity" data set in R,which is derived from "Cochran, W.G. (1977) Sampling Techniques. Third edition. John Wiley". It shows the population of 49 U.S.cities in 1920 and 1930. The 49 cities are randomly chosen from 196 largest cities in 1920. In this data set, u represents the 1920 population, and x represents the 1930 population. The fitted slop 1.2 implies that, in our sample of 49 cities, the overall population growth rate from 1920 to 1930 is approximately 1.2-1=0.2. R^2 =0.964 indicate a very good fit of the model, showing that the population growth can be well approximated by a linear function shown in the figure. We can use this linear function to predict 1930 population of a city given its 1920 population.

From the table we observe that the growth rate ranges from -0.096 to 24.However a growth rate of 24 seems too high, and we need to double check if this record is correct.

As a next step, it would be interesting to look at the population variation 10 years later, e.g. in 1940,1950,1960, etc. We can observe if the growth rate is constant during each 10 year period or if the population saturates at some point.

Dataset4 Beaver

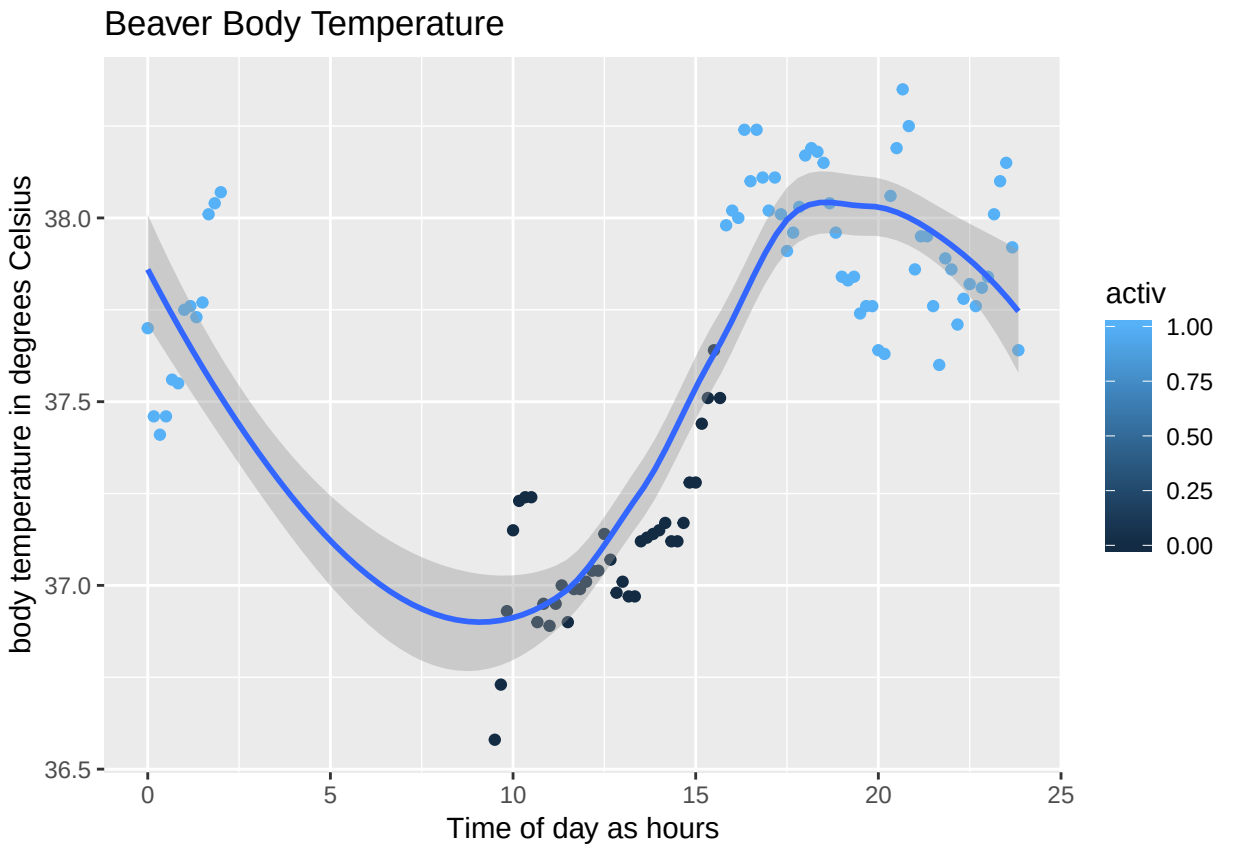
```
beaver_tb <- as_tibble(beaver)
beaver_tb
```

```
## # A tibble: 100 x 4
##      day  time  temp activ
##    <dbl> <dbl> <dbl> <dbl>
##  1   307   930  36.6     0
##  2   307   940  36.7     0
##  3   307   950  36.9     0
##  4   307  1000  37.2     0
##  5   307  1010  37.2     0
##  6   307  1020  37.2     0
##  7   307  1030  37.2     0
##  8   307  1040  36.9     0
##  9   307  1050  37.0     0
## 10   307  1100  36.9     0
## # ... with 90 more rows
```

```
beaver2 <- beaver_tb %>%
  mutate(
    record_hour = time %/% 100,
    record_min = time %% 100,
    time = record_hour + record_min / 60
  )

beaver2 %>%
  ggplot(aes(x=time,y=temp,colour = activ)) +
  geom_point()+
  geom_smooth()+
  xlab("Time of day as hours")+
  ylab("body temperature in degrees Celsius")+
  labs(title ="Beaver Body Temperature")
```

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



##Results

This figure was derived from the "beaver" data set in R, which is a part of a long study "Reynolds, P.S. (1994) Time-series analyses of beaver body temperatures."

The data reads the body temperature in degree Celsius in beavers every 10 minutes in day 307 and early 308 of the study. Activ 1 represents intensive activity when the beaver is outside of the retreat and 0 represents no such high-intensity activity. The figure shows that high intensity activities occur around and after 15:50 until early next day at end of the record 308, when beavers' body temperatures tend to be high. The beavers stay inside of the retreat during 9:30-15:50, and during this time body temperature tends to be low. The pattern indicates some assiciation between activity intensity, day of time and beaver's body temperature. It would be interesting to look at the full data from day 1 unitl the end day of study. We can observe whether this pattern persists through a long period of time. We can also further investigate what factor causes the bearver's body temperature to change.