MATH215 - Assignment 5

Pierre Visconti

Problem 1:

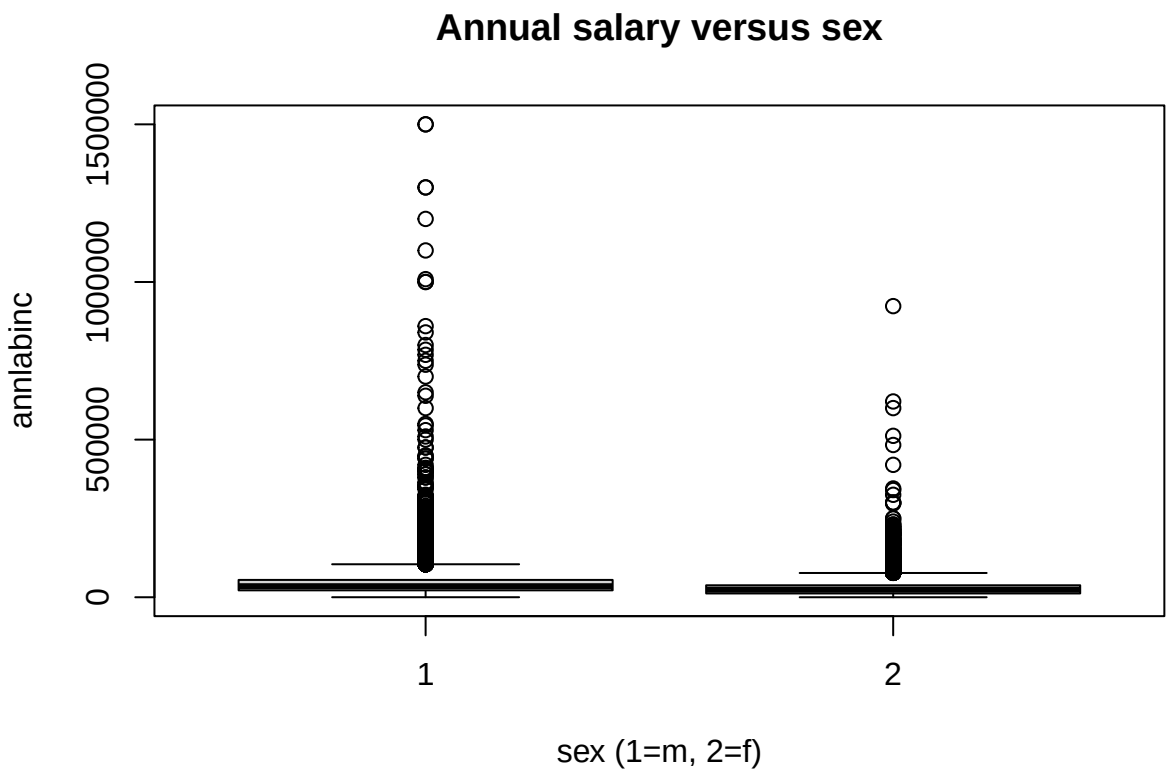
```
pg = read.csv("PanelStudyIncomeDynamics.csv")
max.salary = max(pg$annlabinc)
max.salary
```

```
## [1] 1500000
```

```
# pg[which(pg$annlabinc==max.salary),] # commented out because of size
```

The largest salary is 1,500,000 yearly. There are two individuals with this salary. One of them is a male, 53 years old, with 35 years of experience, from the northeast, with occupation code 001. The other is male, 35 years old, with 6 years of experience, from the south, with occupation code 301. Note: I found conflicting information online with the occupation codes so I decided not to include what they could mean.

```
boxplot(annlabinc~sex, data=pg, main="Annual salary versus sex", xlab="sex (1=m, 2=f)")
```

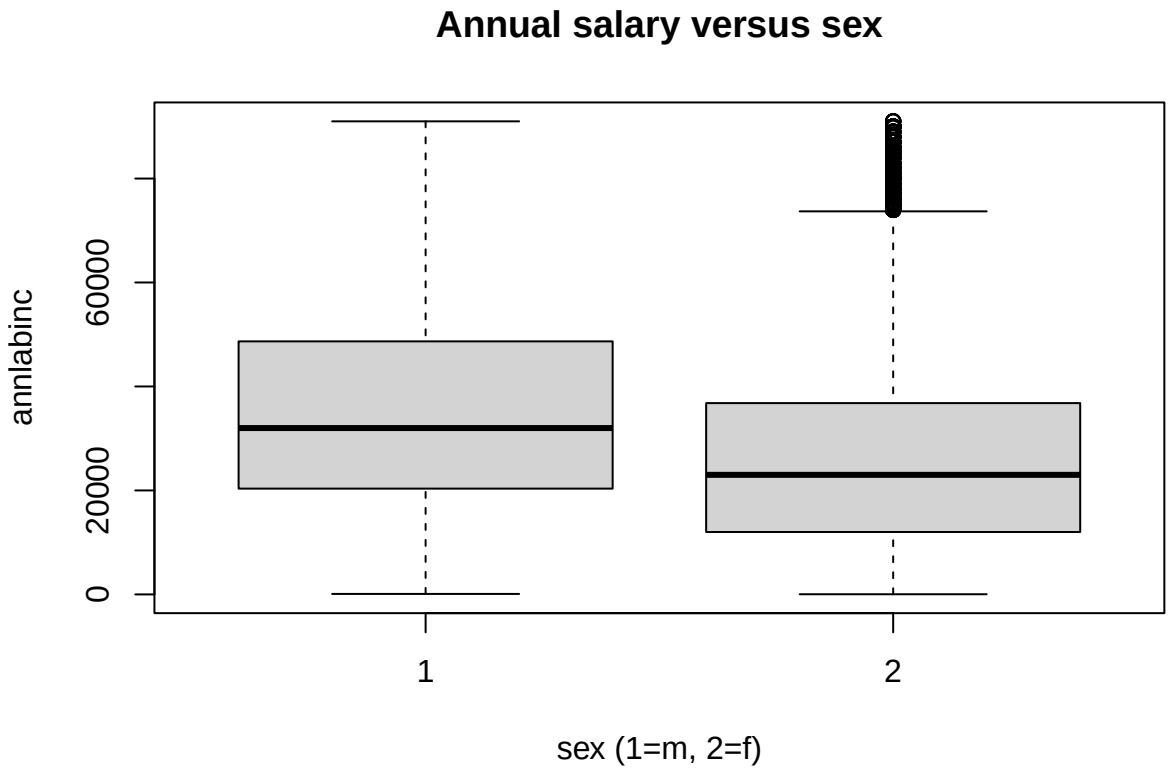



```
quantile(pg$annlabinc)
```

```
##      0%     25%     50%     75%    100%
##      30   16000   29000   46000 1500000
```

```
IQR=46000-16000
lower=16000 - 1.5*IQR
upper = 46000 + 1.5*IQR
pg.no=pg[which(pg$annlabinc <= upper),] # no need for lower bound
# since it is negative and the min salary is 30.
```

```
boxplot(annlabinc~sex, data=pg.no, main="Annual salary versus sex", xlab="sex (1=m, 2=f)")
```



The mean salary for males appears to be slightly higher than for females. In fact the entire 1st to 3rd quartile range is shifted up higher for the males in comparison with the females. What is note worthy is that in this boxplot comparison the males have no outliers while the females do which means that there are salaries on the upper end for males that are considered within the range and not outliers while the same salary for females is considered as an outlier.

Two tailed test for mean with two populations:

```
t.test(annlabinc~sex, data=pg.no,alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 43.298, df = 30280, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  9040.535 9897.851
## sample estimates:
## mean in group 1 mean in group 2
##        35946.87        26477.68
```

Our p-value is extremely small which means there is strong statistical evidence to reject the null hypothesis that the mean annual salaries between males and females are the same.

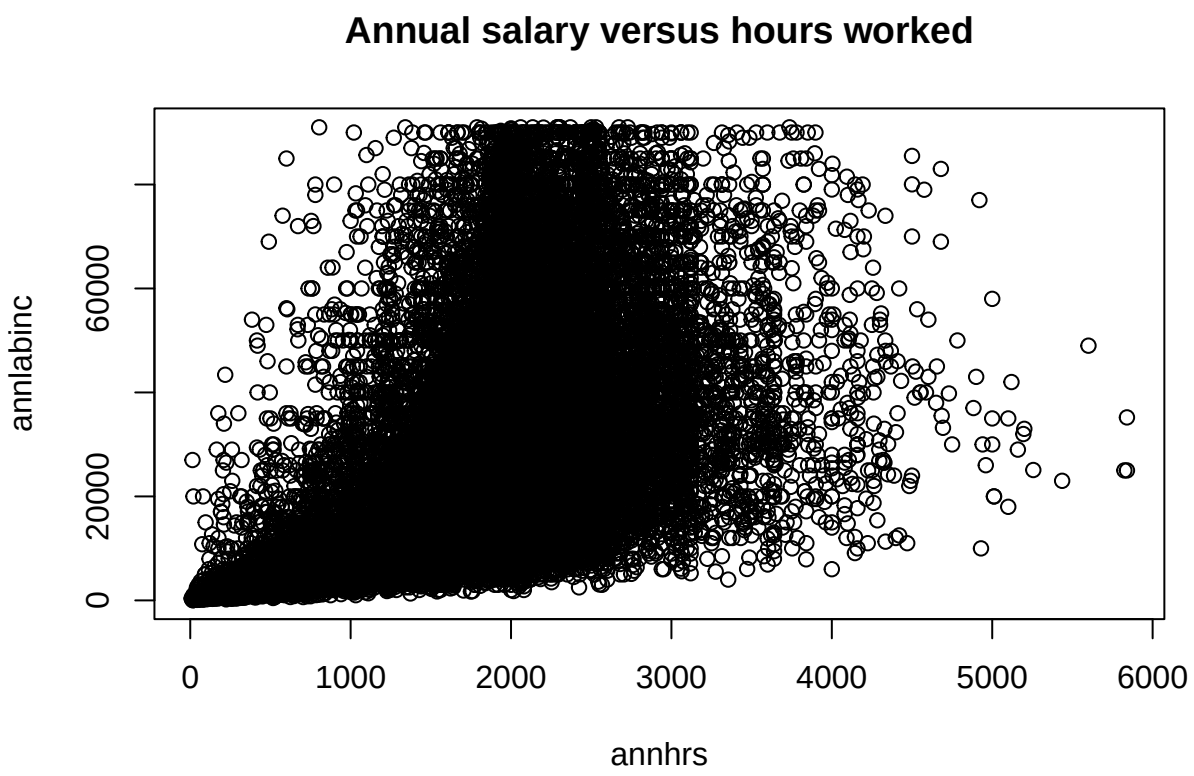
Problem 2

```
t.test(annhrs~sex, data=pg.no, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  annhrs by sex
## t = 55.868, df = 31423, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  359.4003 385.5352
## sample estimates:
## mean in group 1 mean in group 2
##        2164.755        1792.287
```

The p-value is very small which means that there is strong statistical evidence to reject the null hypothesis that the mean annual working hours between males and females are the same. This does not effect the conclusion from the previous test, yet it could explain why the mean salaries may not be the same.

```
plot(annlabinc~annhrs, data=pg.no, main="Annual salary versus hours worked")
```



It's pretty hard to say whether a relationship exists from looking at this scatter plot. There does seem to be a relationship for the first few thousand hours but then the data is all over the place and the scale of

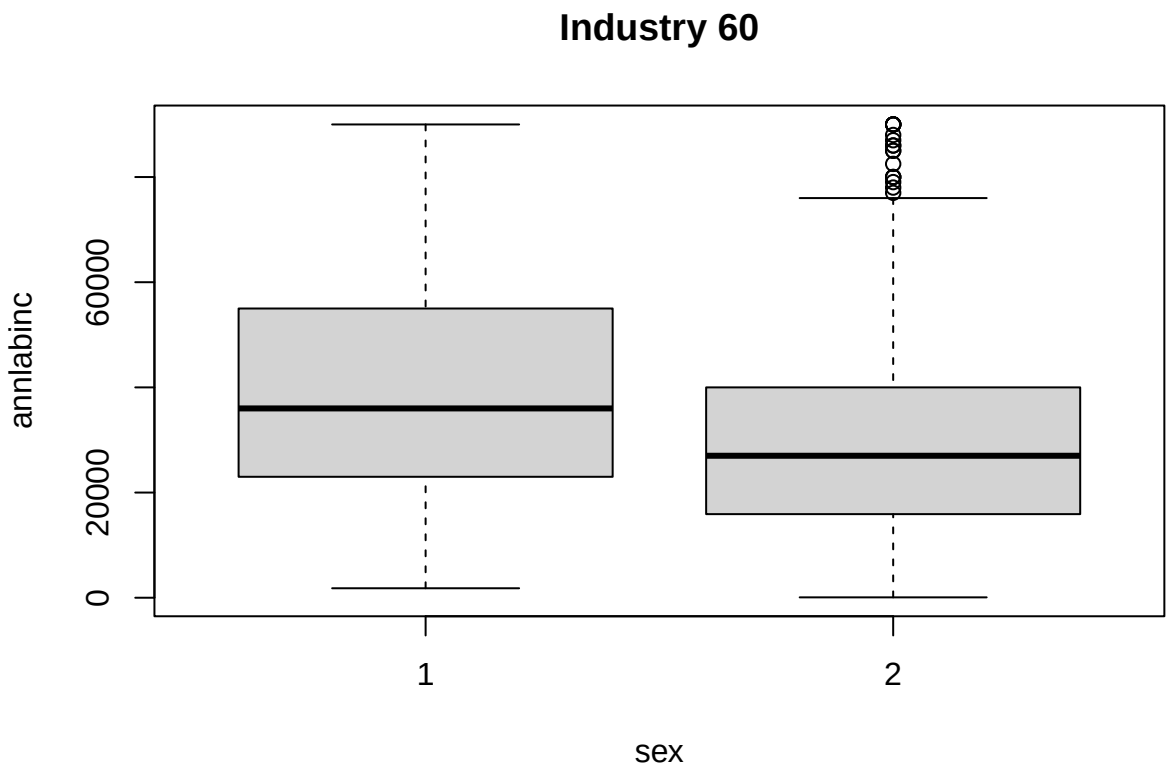
the plot makes it hard to clearly see. If we look at the lowest income levels though, we can see them slowly increase with labor hours. This contributes to our understanding of the gender pay gap by guiding us in a direction for further testing.

Problem 3

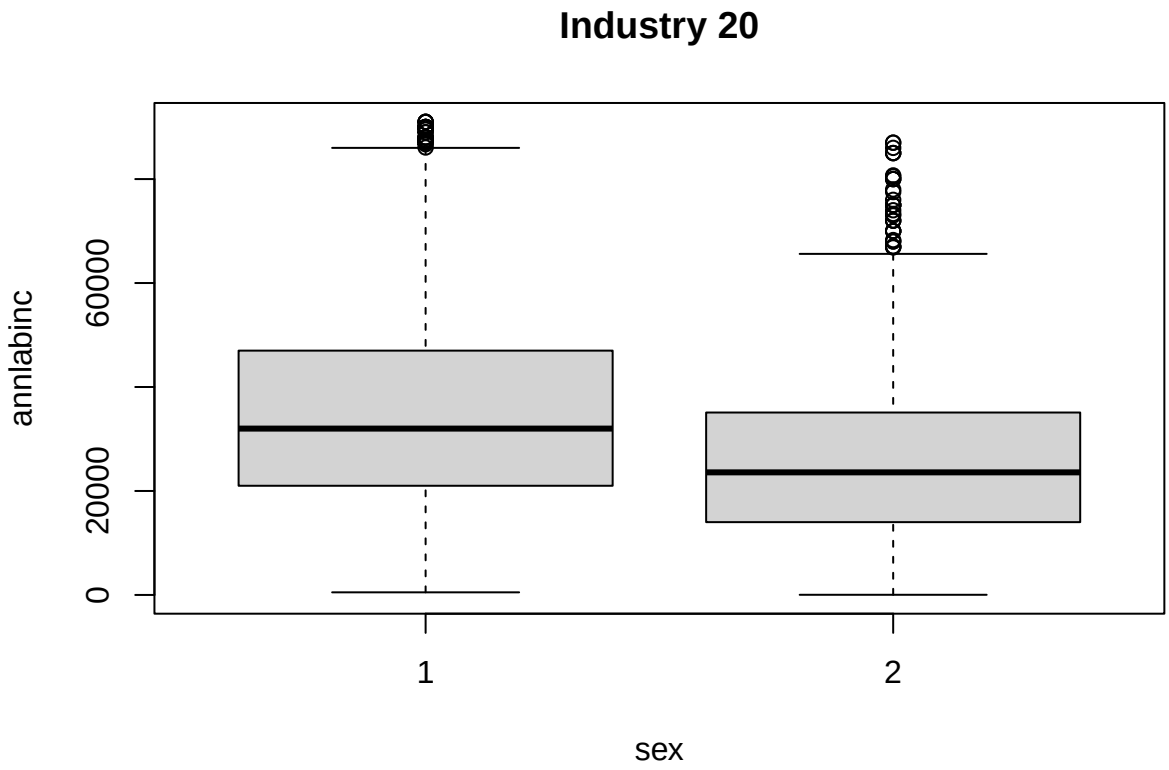
```
ind = unique(pg.no$ind2)
ind
```

```
##  [1] 60 20 62 52 30 70 50 72 12 46 68 42 80 10 74 44
```

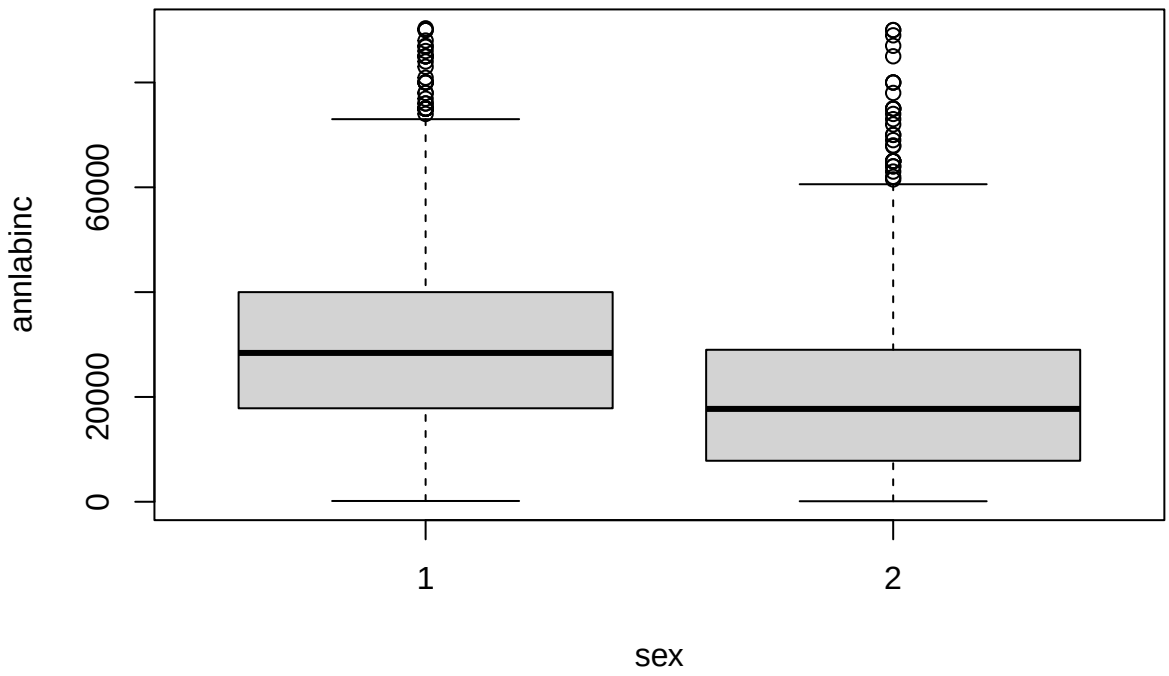
```
p.val.vec = c(1:(length(ind)))
index=1
for (i in ind) {
  #print(i)
  pg.temp = pg.no[which(pg.no$ind2 == i),]
  boxplot(annlabinc~sex, data=pg.temp, main=paste("Industry", i, sep=" "))
  print(t.test(annlabinc~sex, data=pg.temp,alternative="two.sided"))
  p.val.vec[index]=t.test(annlabinc~sex,data=pg.temp,alternative="two.sided")$p.value
  index = index + 1
}
```



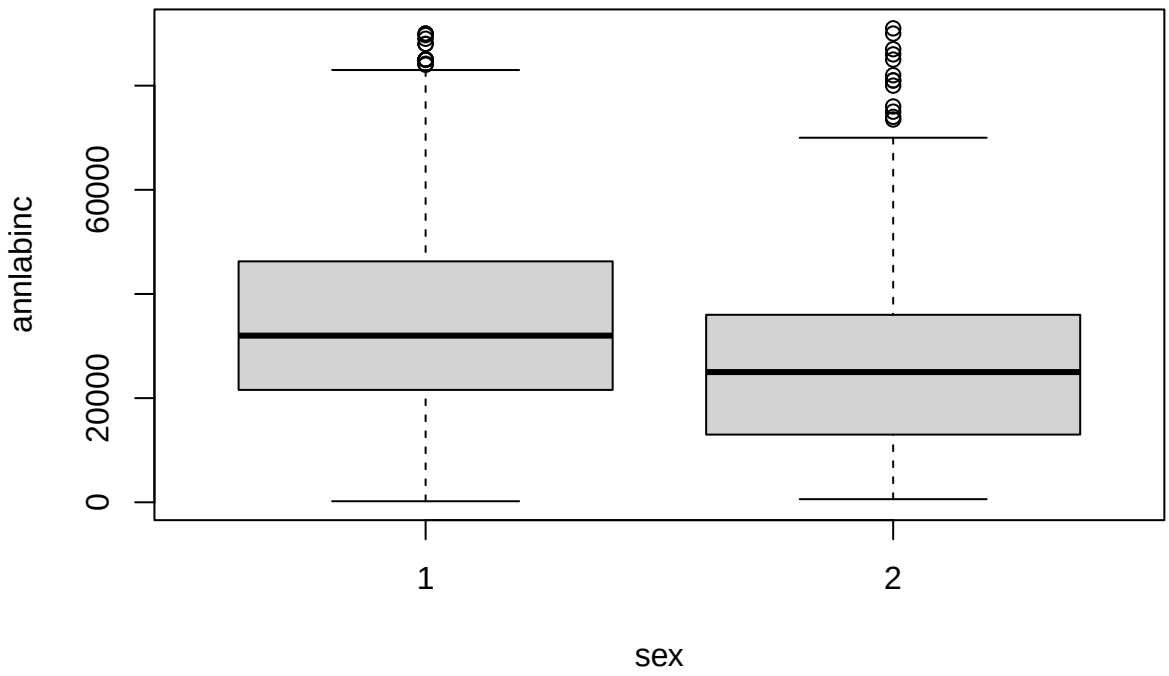
```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 10.44, df = 937.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   8701.338 12730.102
## sample estimates:
## mean in group 1 mean in group 2
##        40103.64        29387.92
```



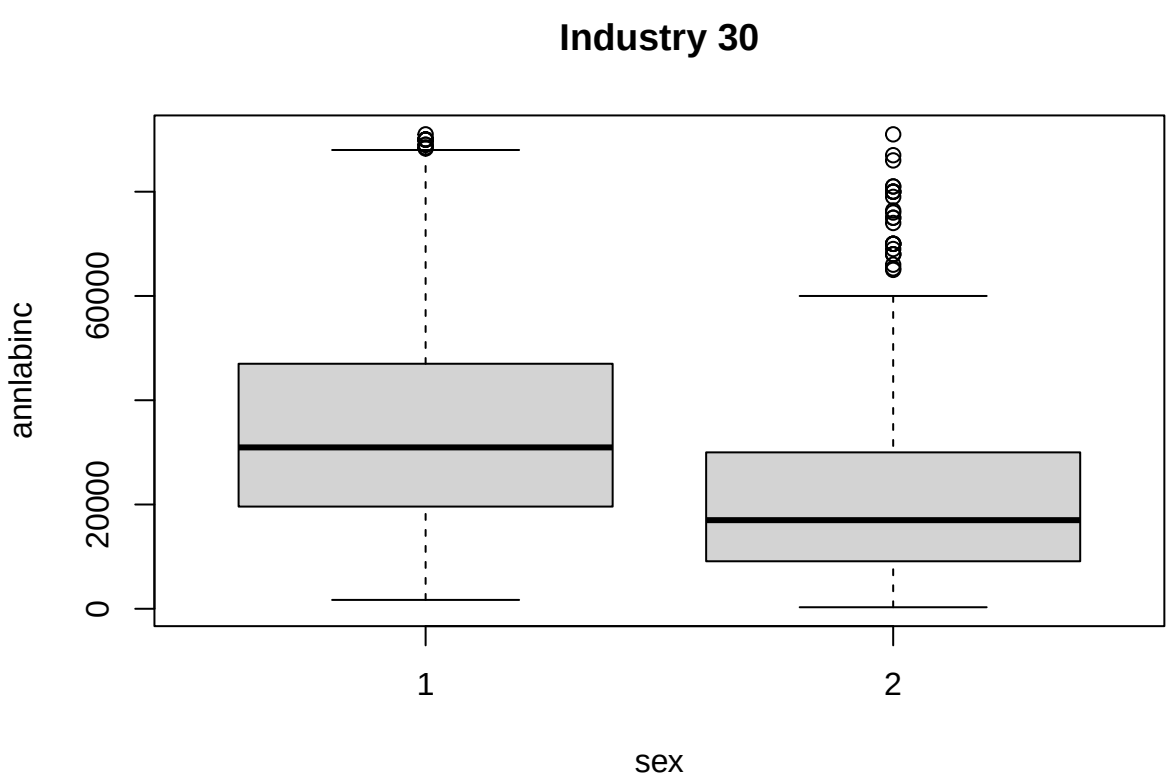

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 13.424, df = 2030.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   7920.409 10630.574
## sample estimates:
## mean in group 1 mean in group 2
##        35943.21        26667.72
```



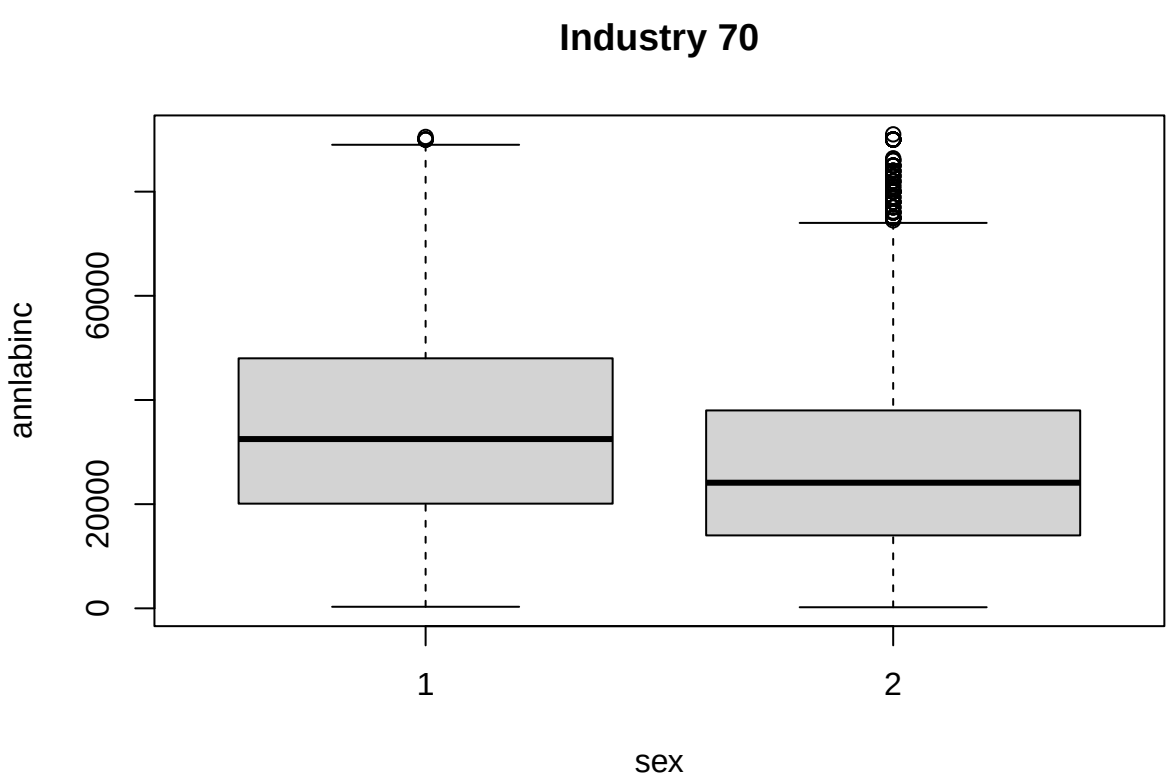

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 13.497, df = 1584.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   9118.504 12219.557
## sample estimates:
## mean in group 1 mean in group 2
##        31301.83        20632.80
```



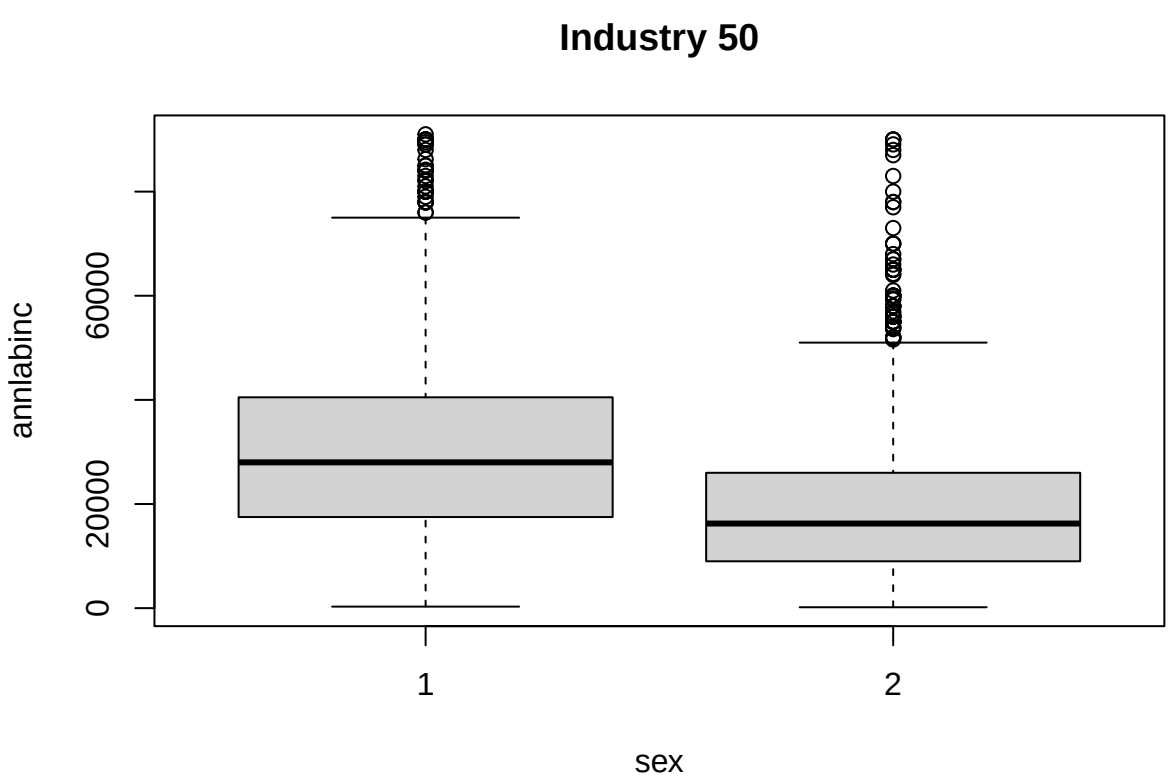

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 5.9631, df = 530.97, p-value = 4.519e-09
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   5319.399 10545.961
## sample estimates:
## mean in group 1 mean in group 2
##        35584.85        27652.17
```



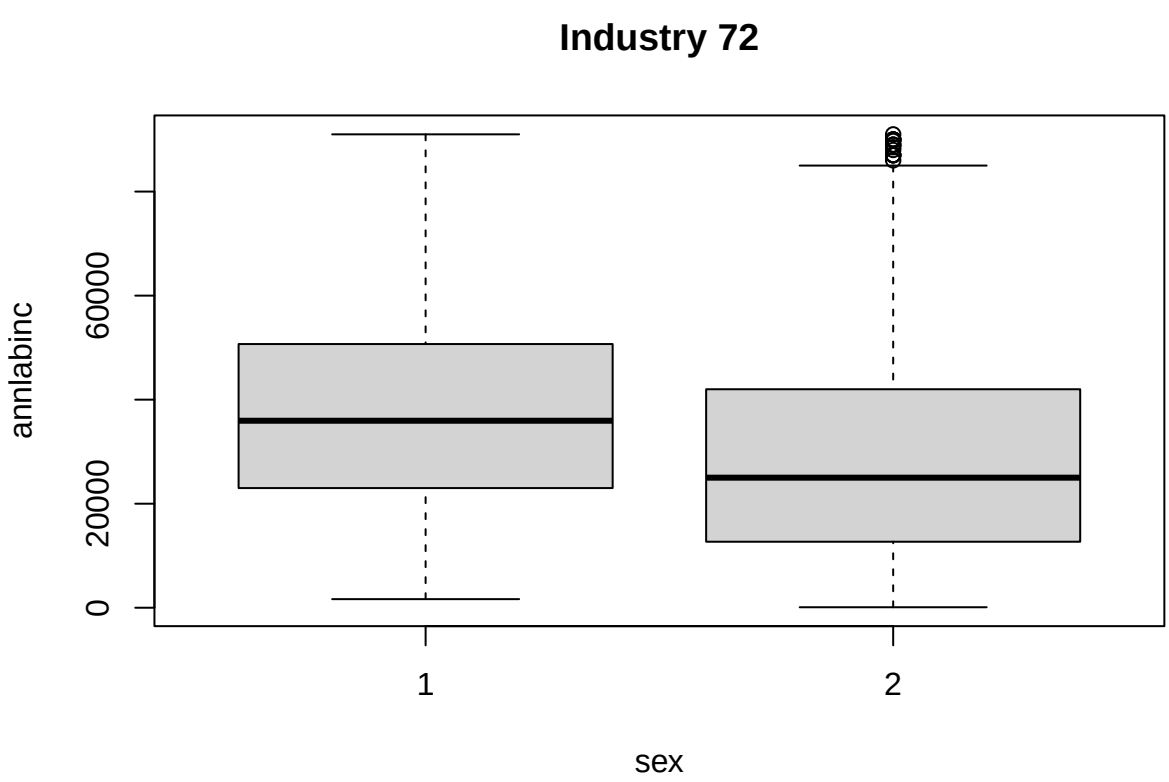

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 17.286, df = 2135, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  11923.81 14975.46
## sample estimates:
## mean in group 1 mean in group 2
##        34817.15        21367.51
```



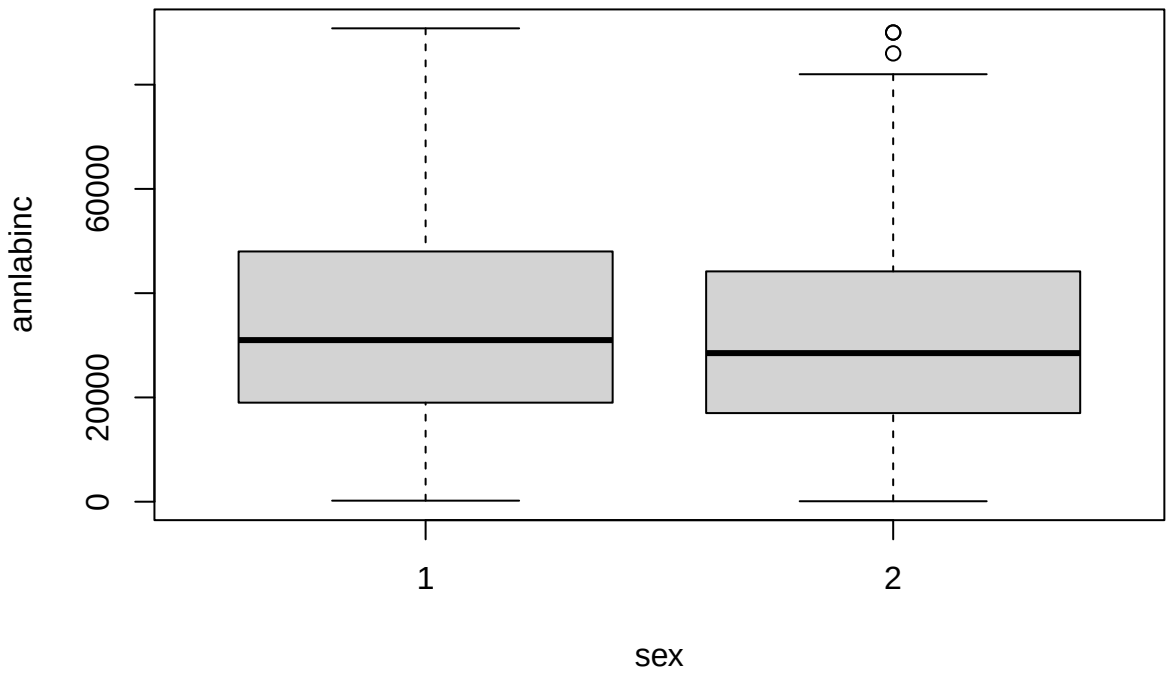

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 9.4459, df = 856.84, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  6379.031 9725.301
## sample estimates:
## mean in group 1 mean in group 2
##        35914.32        27862.15
```



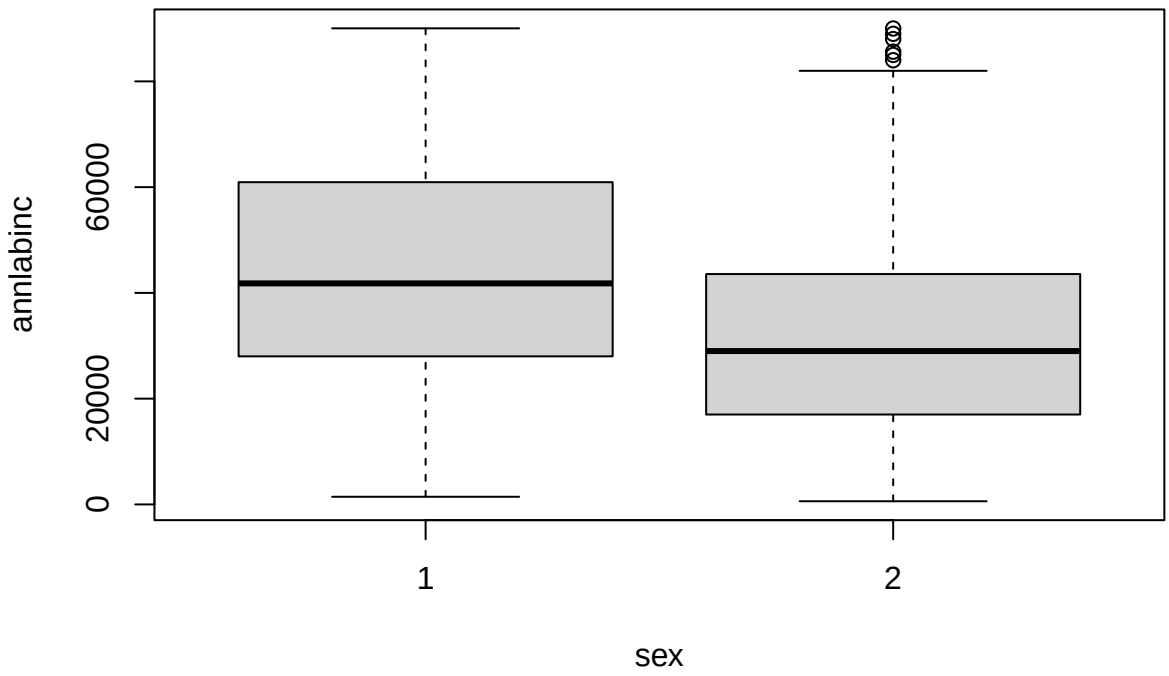

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 17.464, df = 2428.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  10221.06 12806.79
## sample estimates:
## mean in group 1 mean in group 2
##        31041.99        19528.06
```



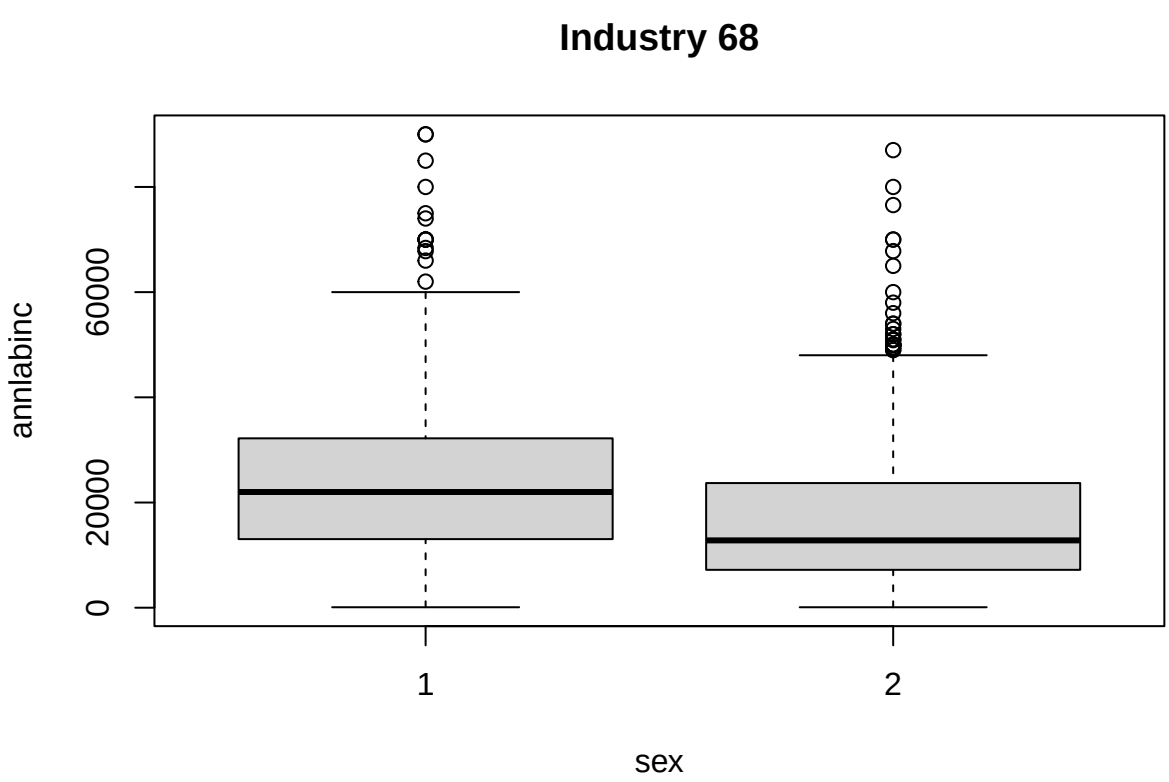

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 12.436, df = 1701.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   7655.745 10522.865
## sample estimates:
## mean in group 1 mean in group 2
##        37967.94        28878.63
```



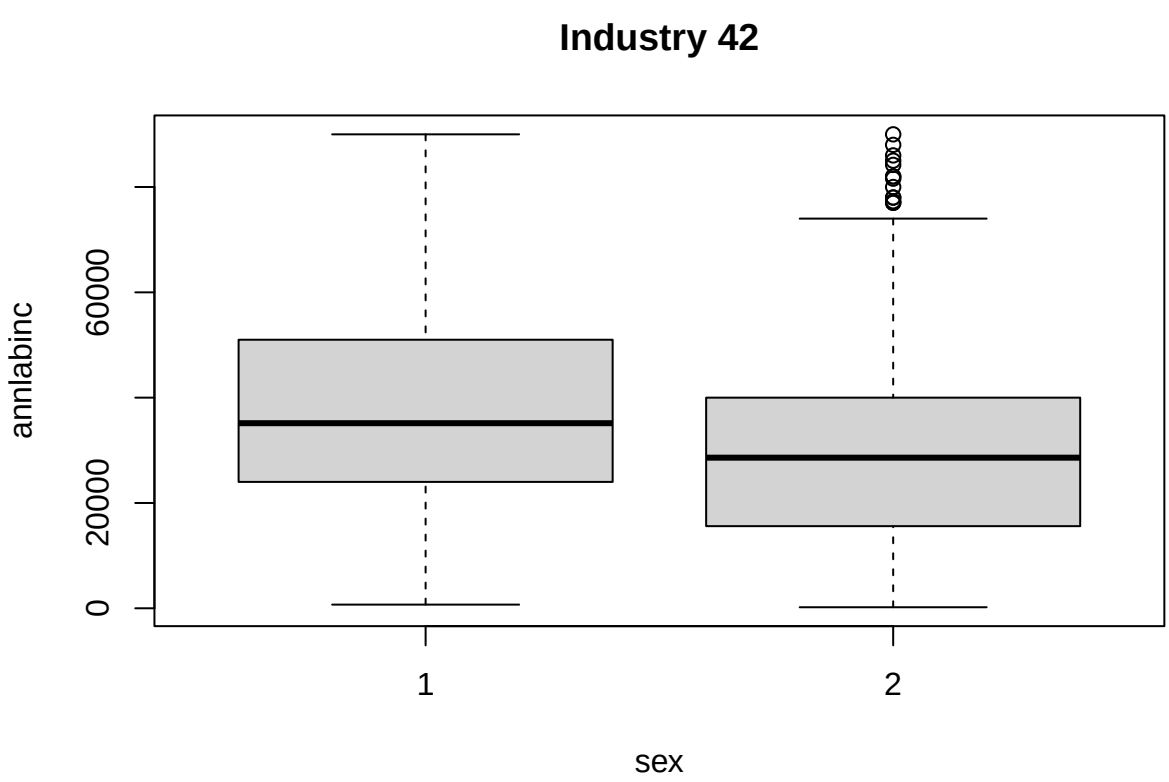

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 2.5008, df = 281.71, p-value = 0.01296
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   746.4325 6266.5518
## sample estimates:
## mean in group 1 mean in group 2
##        34580.22        31073.73
```



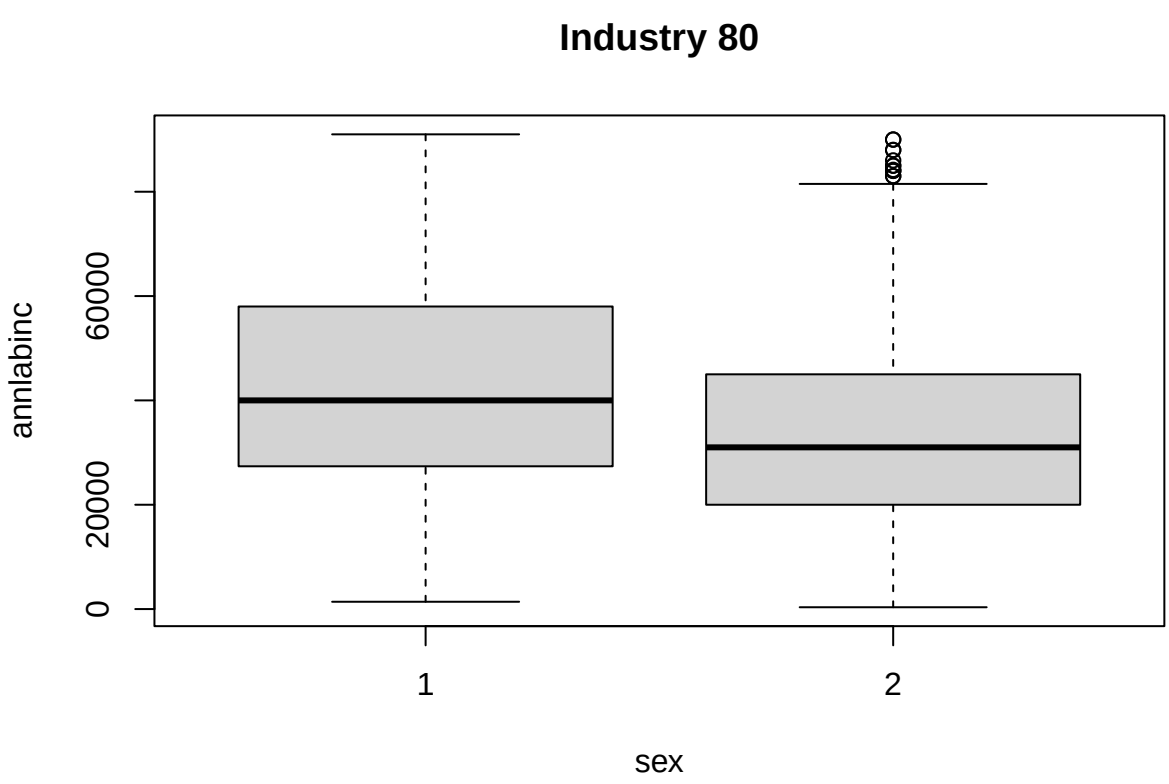

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 9.0959, df = 769.58, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  10261.20 15909.25
## sample estimates:
## mean in group 1 mean in group 2
##        44603.83        31518.60
```



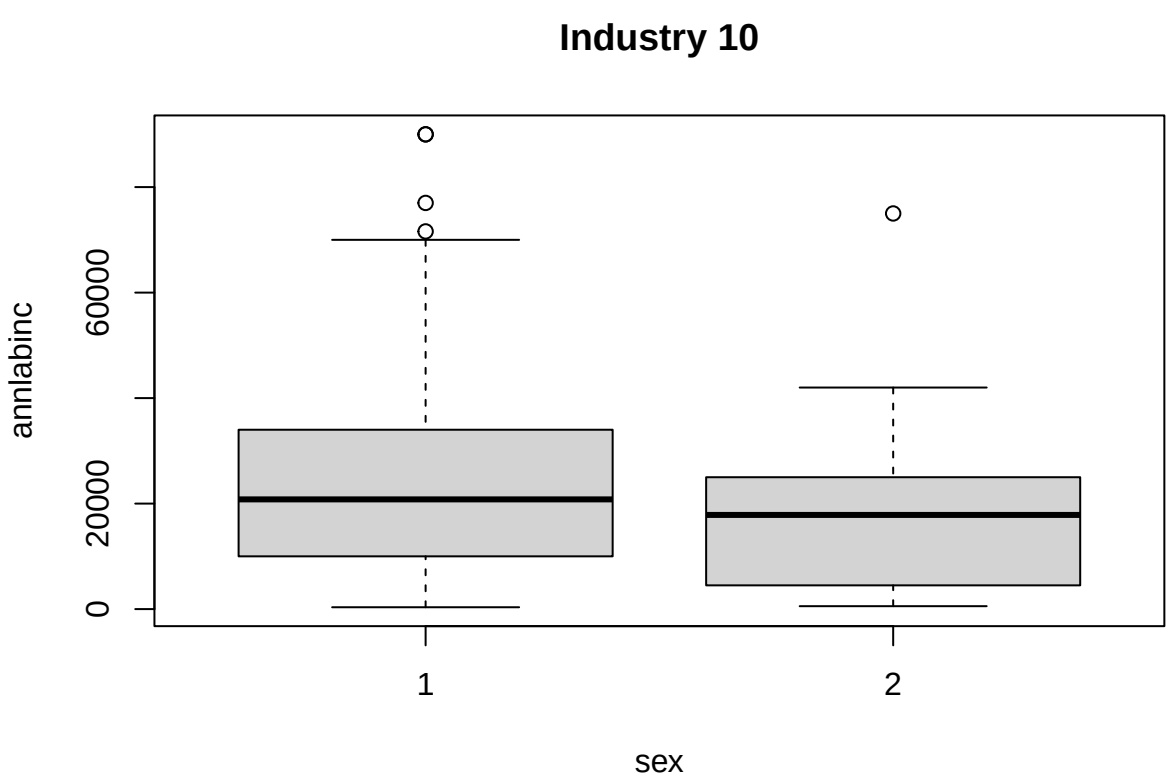

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 10.123, df = 748.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   6896.188 10214.333
## sample estimates:
## mean in group 1 mean in group 2
##        25047.62        16492.36
```



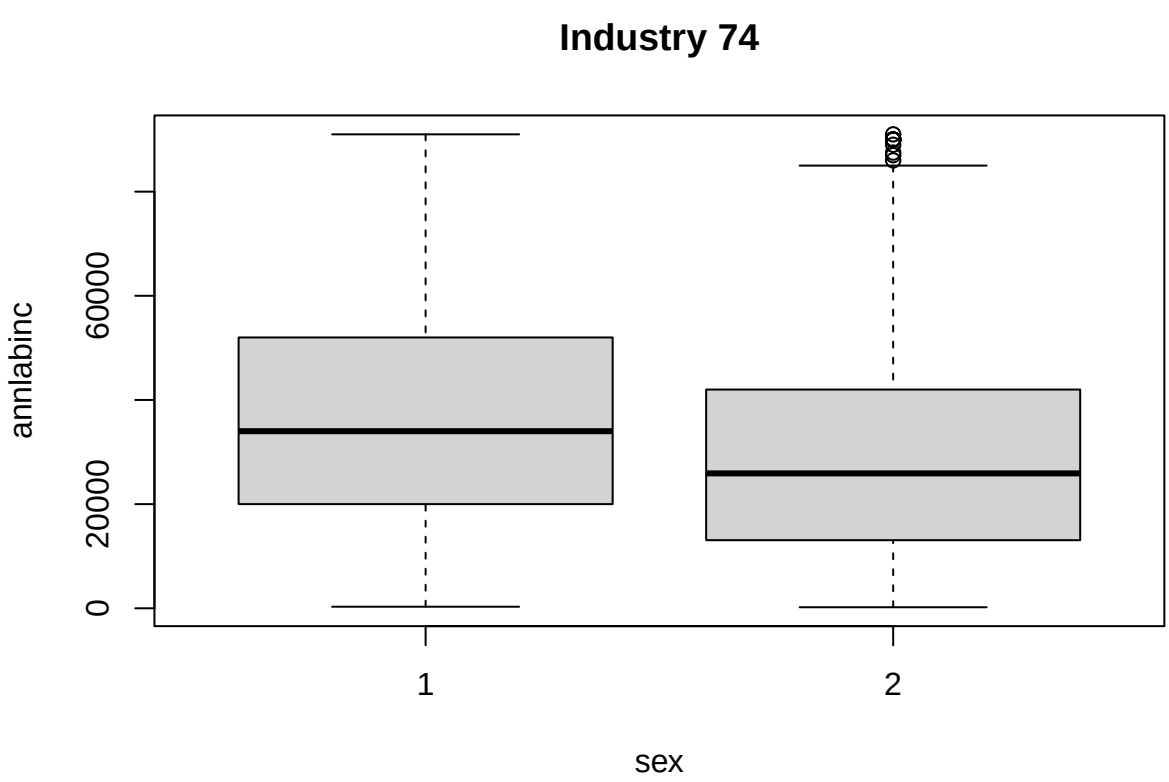

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 8.2166, df = 961.67, p-value = 6.717e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   6261.022 10190.182
## sample estimates:
## mean in group 1 mean in group 2
##        38485.89        30260.29
```



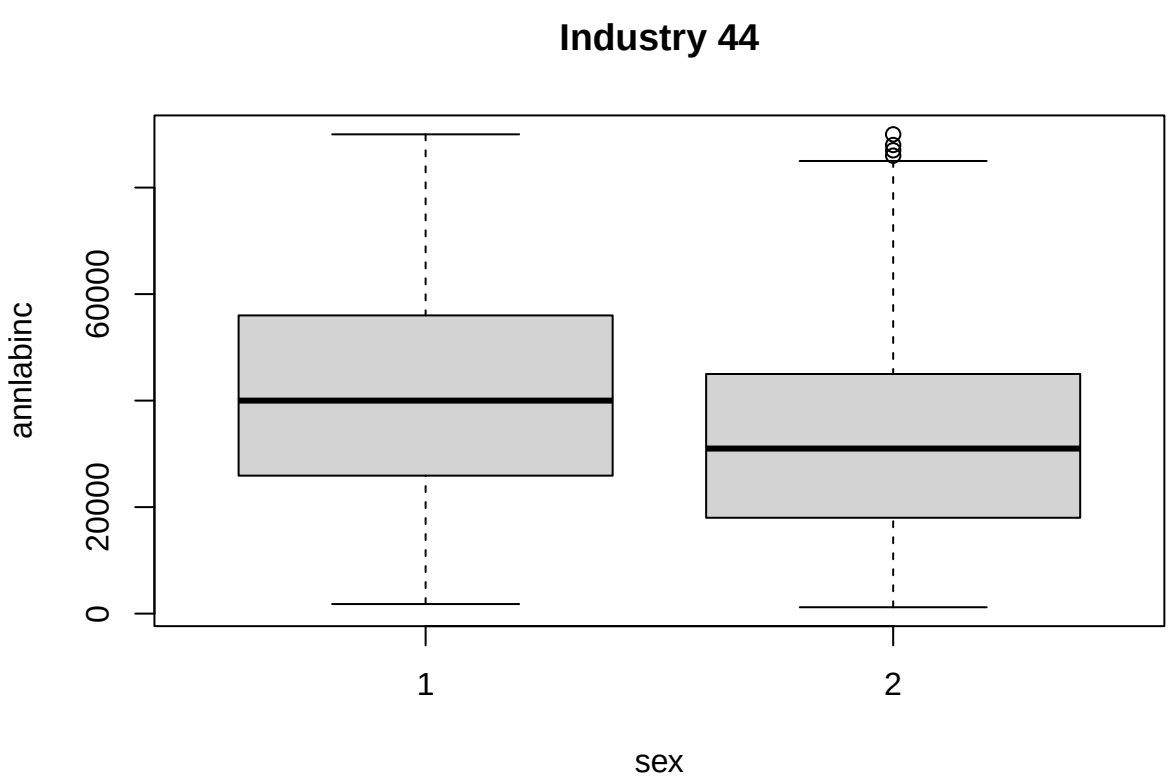

```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 11.391, df = 2146.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   7904.008 11191.365
## sample estimates:
## mean in group 1 mean in group 2
##        43149.65        33601.96
```




```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 2.2378, df = 52.035, p-value = 0.02954
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##    608.3045 11167.6855
## sample estimates:
## mean in group 1 mean in group 2
##        23901.44        18013.45
```




```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 8.4441, df = 1943.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  6103.239 9795.871
## sample estimates:
## mean in group 1 mean in group 2
##        37421.46        29471.91
```




```
##
##  Welch Two Sample t-test
##
## data:  annlabinc by sex
## t = 3.7258, df = 226.05, p-value = 0.000246
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   3902.168 12663.225
## sample estimates:
## mean in group 1 mean in group 2
##        42630.25        34347.55
```

```
print(p.val.vec)
```

```
##  [1] 3.264456e-24 1.996066e-39 2.243360e-39 4.518603e-09 9.151920e-63
##  [6] 3.252080e-20 1.989271e-64 4.803484e-34 1.296078e-02 7.870821e-19
## [11] 1.158033e-22 6.716949e-16 3.133424e-29 2.953987e-02 5.904517e-17
## [16] 2.459562e-04
```

This analysis tells us that for every industry present we can reject the null hypothesis that the mean annual salaries between both sex are the same. This also tells us that in this case we are not seeing Simpson's Paradox unlike with the UC Berkley dataset.

Problem 4

```
library(vcd)
```

```
## Loading required package: grid
```

```
austin = read.csv("Austin_Profiling.csv")
mosaic(PERSON_SEARCHED~apd.race, data=austin)
```


```
my.table <- table(austin$apd.race,austin$PERSON_SEARCHED)
```

```
my.results = chisq.test(my.table)
```

```
## Warning in chisq.test(my.table): Chi-squared approximation may be incorrect
```

```
my.results
```

```
##
##  Pearson's Chi-squared test
##
## data:  my.table
## X-squared = 464.38, df = 7, p-value < 2.2e-16
```

The p-value is very small so we can reject the null hypothesis that proportion of searched to not searched is the same for every each race. I got a warning that states that the "Chi-squared approximation may be incorrect."

```
my.results$expected
```

```
##
##           NO = 2      YES = 1
##   A  1116.406438   57.5935622
##   B  6490.182230  334.8177700
##   H 11979.022059  617.9779412
##   I    23.773561    1.2264387
##   M   406.052427   20.9475733
##   P     2.852827    0.1471726
##   U   167.365871    8.6341286
##   W 21587.344587 1113.6554134
```

The conditions appear to be violated for races I and P.

```
austin.new = austin[which(austin$apd.race=='A' | austin$apd.race=='B' | austin$apd.race=='H' | austin$ap
my.table.new <- table(austin.new$apd.race,austin.new$PERSON_SEARCHED)
my.results.new = chisq.test(my.table.new)
my.results.new
```

```
##
##  Pearson's Chi-squared test
##
## data:  my.table.new
## X-squared = 456.03, df = 3, p-value < 2.2e-16
```

```
my.results.new$expected
```

```
##
##        NO = 2    YES = 1
##   A  1116.082   57.91773
##   B  6488.298  336.70231
##   H 11975.544  621.45627
##   W 21581.076 1119.92369
```

The warning is now gone. The assumption of sufficiently large data is satisfied as can be seen by checking the new expected values table. All that we can conclude is that the proportion of search to not-searched is not the same for every race. In other words, at least one race has a proportion that is different than the others. We arrive to this conclusion since the p-value is very small and we reject the null hypothesis that all the proportions of searched to not-searched are the same between all races.

```
pairwise.prop.test(my.table.new)
```

```
##
##  Pairwise comparisons using Pairwise comparison of proportions
##
## data:  my.table.new
```

```
##
##   A       B       H
## B 2.9e-14 -       -
## H 9.5e-10 5.4e-06 -
## W 0.039   < 2e-16 < 2e-16
##
## P value adjustment method: holm
```

The p-values are small enough for every cell which indicates that there is a statistically significant difference among every race in the proportion of searched to not searched.

```
non.search = my.table.new[,"NO = 2"]
search = my.table.new[,"YES = 1"]
index = 1
for (i in non.search) {
  print(search[index] / i)
  index = index +1
}
```

```
##          A
## 0.01998262
##          B
## 0.0928743
##          H
## 0.07126456
##          W
## 0.0315823
```

The differences do not seem significant to me for several reasons. The differences are in the hundredth decimal place which is pretty insignificant. Second, the populations are not massive. If there were billions of people getting searched then I think it could be argued that the differences are significant as there would be a lot more people who identify as B or H getting searched than W or A (relative to their population size). But I do not think the population of people getting searched are large enough for it to be considered significant.