Problem Set 1

Zhengting (Johnathan) He

2021/9/2

```
# set working directory
require(tidyverse)

## Loading required package: tidyverse

## — Attaching packages ——————————————————————————— tidyverse
    1.3.1 —

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## — Conflicts ———————————————————————————————— tidyverse_conflicts
    () —
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

setwd(”D:/OneDrive␣−␣Johns␣Hopkins/Course/140.621.81␣−␣Statistical␣Methods␣in␣
    Public␣Health␣I/Problem␣set/jhsphbiostat621−assignment/Problem␣set␣1”)
```

Problem set 1: Displaying and Thinking About Public Health Data EDA

Problem 1. Air Pollution and Mortality in Baltimore

Section 1: Exploratory Data Analysis

```
# import data
data <- read_csv(”./data/baltps11.csv”)

## Rows: 40 Columns: 2

## — Column specification
    ————————————————————————————————————————
## Delimiter: ”,”
## dbl (2): group, deaths

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
    message.
```

```
data$group <- factor(data$group, levels = c("1", "2"))
```

a. Create stem and leaf displays by group:

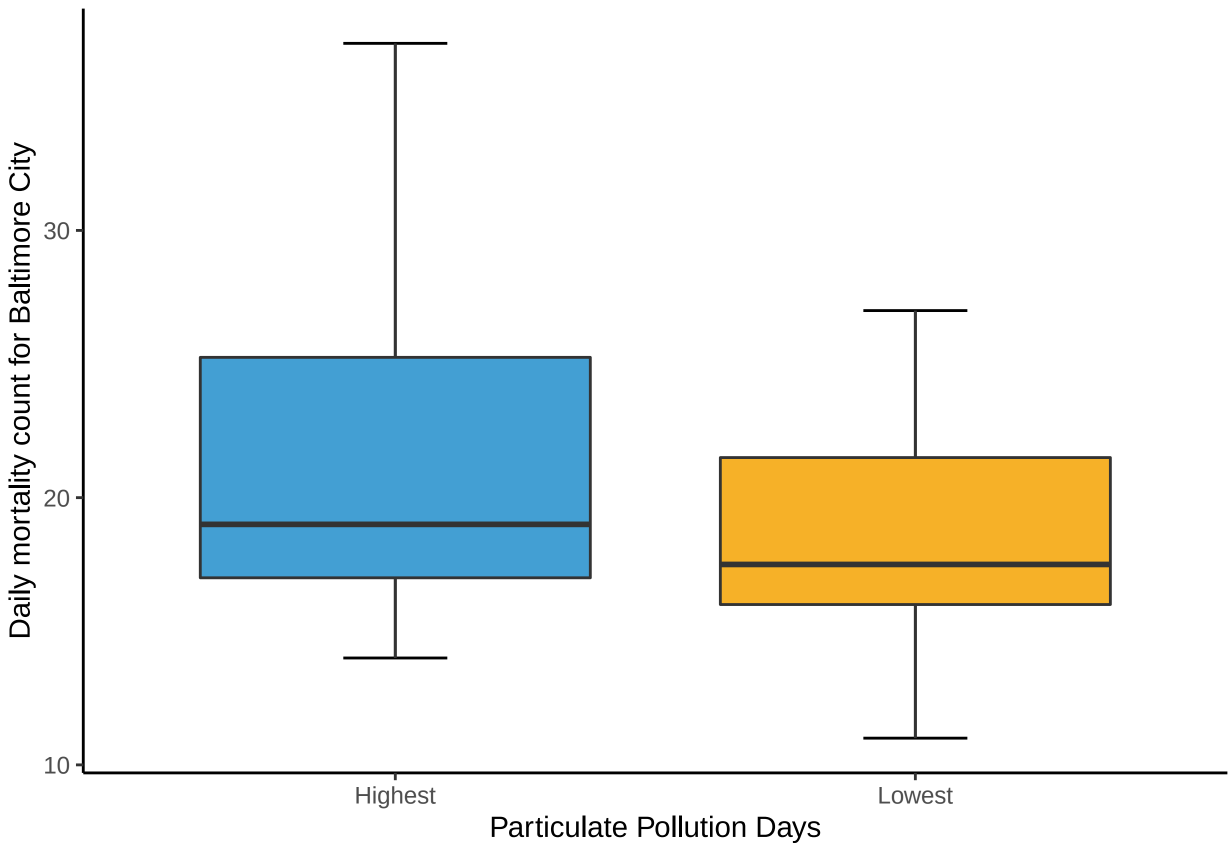
```
for (pollut in c("1", "2")) {
    print(ifelse(pollut == "Highest", "#␣Daily␣mortality␣count␣on␣Highest␣
        Particulate␣Pollution␣Days",
        "#␣Daily␣mortality␣count␣on␣Lowest␣Particulate␣Pollution␣Days"))
    stem(filter(data, group == pollut)$deaths)
}
```

```
## [1] "# Daily mortality count on Lowest Particulate Pollution Days"
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 44
##   1 | 567778899
##   2 | 334
##   2 | 5678
##   3 | 3
##   3 | 7
##
## [1] "# Daily mortality count on Lowest Particulate Pollution Days"
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 134
##   1 | 5666667899
##   2 | 0134
##   2 | 577
```

b. Create a box and whiskers plot by group:

```
ggplot(aes(x = group, y = deaths), data = data) + stat_boxplot(geom = "
   errorbar",
    width = 0.2) + geom_boxplot(aes(fill = group)) + xlab("Particulate␣
        Pollution␣Days") +
    ylab("Daily␣mortality␣count␣for␣Baltimore␣City") + scale_x_discrete(labels
         = c("Highest",
    "Lowest")) + scale_fill_manual(values = c("#439fd3", "#f6b128"), guide =
        FALSE) +
    theme_classic()
```

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide.
   Please
## use `guide = "none"` instead.
```



c. Please check that your results for steps a. and b. give you the same results as you obtained by hand in Lab Exercise 1.

```
# Yes.
```

d. In a sentence or two, as if for a public health journal, summarize the evidence from these data relevant to whether mortality tends to be higher on high particulate pollution days, as compared to low particulate pollution days. Be quantitative, that is use numeric results, to describe your findings in the context of air pollution and mortality, the topic of greatest interest to your readers. Remember units!

```
# The median daily mortality count for Baltimore city is 1.5 person higher on
   high particulate pollution days, compared with low particulate pollution
   days. Since the daily mortality counts on both high particulate pollution
   days and low particulate pollution days are on a right-skewed distribution
   , it is reasonable to use median as a summary measurement to compare.
```

e. To assure reproducible research, save your log file. Please include the relevant parts of your edited log file with your results as part of your write-up.

Problem 2. Costs of Carotid Endarterectomy in Maryland

Section 1: Exploratory Data Analysis

i. We have drawn a random sample of 100 male and 100 female patients from the HSCRC carotid endarterectomy data set which is located in the Stata data file named CE621.dta on the course website. Refer to your Class Dataset Code-Book for the file format. Open the data

set. And, don't forget to open a log in order to save your work. Make stem and leaf plots of the male and female CE costs.

```
# import data
data2 <- read_csv("./data/ce621.csv")

## Rows: 200 Columns: 12

## -- Column specification
    ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (7): provnum, sex, race, volhosp, volsurg, dthstrk, smoker
## dbl (5): patid, ccscore, totchg, age, year

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
    message.

data2$provnum <- as.numeric(data2$provnum)
data2$sex <- factor(data2$sex, levels = c("Male", "Female"))
data2$race <- factor(data2$race, levels = c("Caucas", "AfroAm", "Other"))
data2$smoker <- factor(data2$smoker, levels = c("No", "Yes"))

for (sex in c("Male", "Female")) {
    print(paste("#␣CE␣costs␣for␣", sex, sep = ""))
    stem(filter(data2, sex == sex)$totchg)
}

## [1] "# CE costs for Male"
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 29
##    2 | 5667788990002223344445666677778899999
##    4 | 011222333333444444444555556777789990000011122222233344445566667778999
##    6 | 0011222333444555666677789999001222223459
##    8 | 35577889123555779
##   10 | 12524567889
##   12 | 2588908
##   14 | 57
##   16 | 177
##   18 | 12244489
##   20 |
##   22 |
##   24 | 0
##   26 |
##   28 |
##   30 | 6
##   32 | 2
##   34 |
##   36 |
##   38 |
##   40 | 1
##
```

```
## [1] "# CE costs for Female"
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 29
##    2 | 5667788990002223344445666677778899999
##    4 | 011222333333444444444455555677778999000001112222223334444556666778999
##    6 | 0011222333444555666677789999001222223459
##    8 | 35577889123555779
##   10 | 12524567889
##   12 | 2588908
##   14 | 57
##   16 | 177
##   18 | 12244489
##   20 |
##   22 |
##   24 | 0
##   26 |
##   28 |
##   30 | 6
##   32 | 2
##   34 |
##   36 |
##   38 |
##   40 | 1
```

ii. Make the necessary calculations using Stata to complete the summary table below regarding CE costs by sex.

```
# summary statistics
data2.summary <- data2 %>%
     group_by(sex) %>%
     summarise(Mean = mean(totchg, na.rm = TRUE), Median = median(totchg, na.rm
          = TRUE),
          IQR = quantile(totchg, probs = c(0.75), na.rm = TRUE) - quantile(
               totchg,
                probs = c(0.25), na.rm = TRUE), SD = sd(totchg, na.rm = TRUE))

# transform dataset to meet with the required format
data2.summary <- data2.summary %>%
     gather(Statistic, Value, c(Mean:SD)) %>%
     spread(sex, Value, c("Male", "Female")) %>%
     mutate(Statistic = factor(Statistic, levels = c("Mean", "Median", "IQR", "
        SD"))) %>%
     arrange(Statistic) %>%
     mutate(Male = round(as.numeric(Male), 2), Female = round(as.numeric(Female
        ),
         2))

## Warning: attributes are not identical across measure variables;
## they will be dropped

## Warning in if (!is.na(fill)) {: the condition has length > 1 and only the
    first
## element will be used
```

```
print(data2.summary)

## # A tibble: 4 x 3
##   Statistic  Male Female
##   <fct>     <dbl>  <dbl>
## 1 Mean      6484.  8099.
## 2 Median    5598.  5644.
## 3 IQR       2862.  5187
## 4 SD        3278.  6679.
```

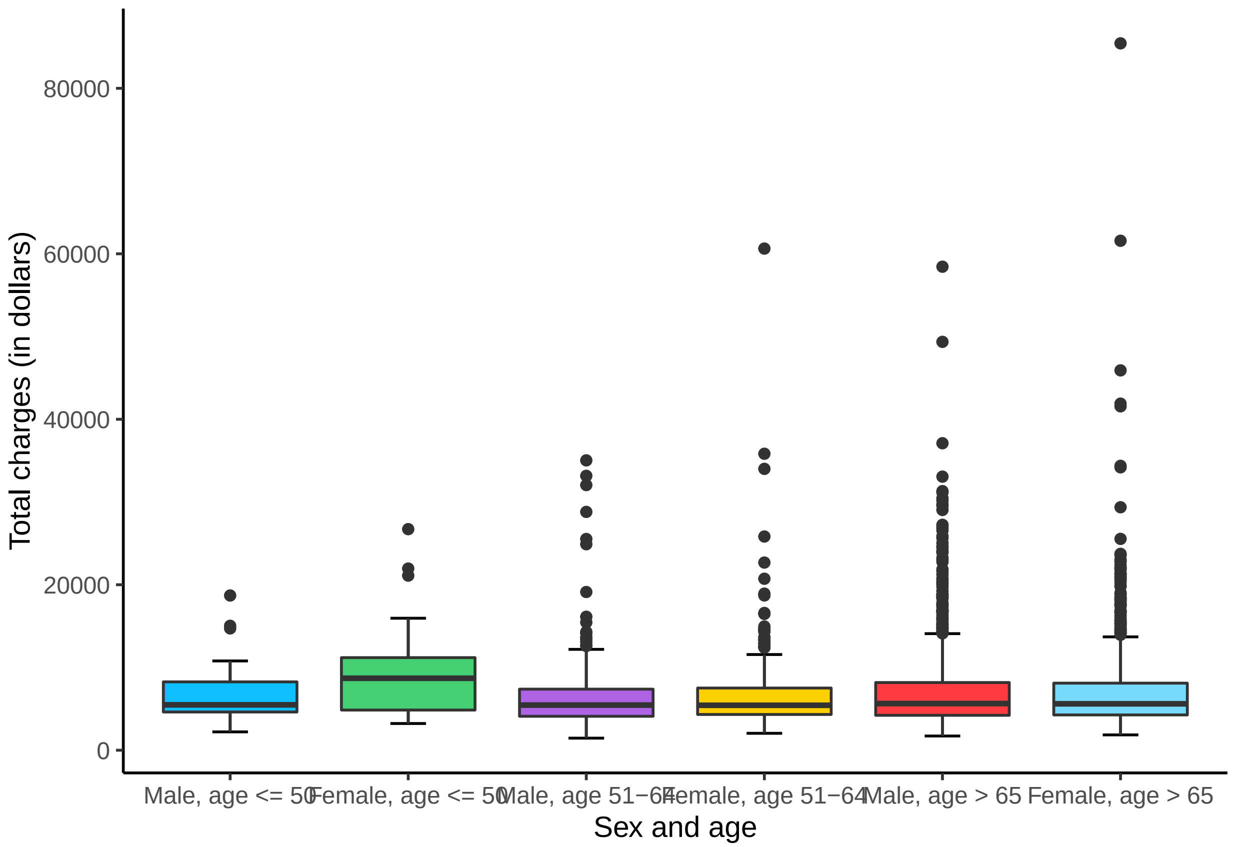
iii. In a brief paragraph, as if written for a health services research journal, compare the distribution of costs for men and women. Be quantitative by referring to the average cost, variability, and the shape of the distribution.

```
# The average cost of women is 1615.25 dollars higher than men, and the
  standard deviation of cost of female is 3401.34 dollars higher than men,
  indicating women have a higher variability. The shapes of the distribution
   of costs of both men and women are right skewed. Since the distribution
  of costs of both men and women are skewed, it may be more appropriate to
  use median and interquartile range to summary descriptive statistics. The
  median cost of female is slightly higher than male (46 dollars), while the
   interquartile range of women is 2325.5 dollars higher than men,
  indicating a large variability of women compared with men.
```

iv. Now use the entire CE Stata data file named CE621entire.dta. Choose only one year and stratify the population of costs into six sex-by-age strata where age is categorized: 50 years; 51-64; 65+. Create and add value labels to the new categorized age variable. Display side-by-side six box plots (50-male, female; 51-64-male, female; 65+-male, female). Order the box plots to facilitate the sex comparison within age-group.

```
# import dataset
data3 <- read_csv("./data/ce621entire.csv")

## Rows: 9918 Columns: 8

## ── Column specification
  ────────────────────────────────────────────────────────

## Delimiter: ","
## chr (4): provnum, sex, race, smoker
## dbl (4): patid, totchg, age, year

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
  message.

data3$provnum <- as.numeric(data3$provnum)
data3$sex <- factor(data3$sex, levels = c("Male", "Female"))
data3$race <- factor(data3$race, levels = c("Caucas", "AfroAm", "Other"))
data3$smoker <- factor(data3$smoker, levels = c("No", "Yes"))

# > max(data3$age)
# [1] 92

# Categorize age and sex into composite factor variable
```

```
data3 <- data3 %>%
                filter(year == 1995) %>% # choose year 1995
                    mutate(age_cat = cut(age, breaks = c(0, 50, 64, 92))) %>%
                        mutate(agesex = case_when(age_cat == "(0,50]" & sex == "
                            Male" ~ "1",
                                                          age_cat == "(0,50]" & sex == "
                                                              Female" ~ "2",
                                                          age_cat == "(50,64]" & sex == "
                                                              Male" ~ "3",
                                                          age_cat == "(50,64]" & sex == "
                                                              Female" ~ "4",
                                                          age_cat == "(64,92]" & sex == "
                                                              Male" ~ "5",
                                                          age_cat == "(64,92]" & sex == "
                                                              Female" ~ "6")) %>%
                            mutate(agesex = factor(agesex, levels = c("1", "2", "
                                3", "4", "5", "6"),
                                                                 labels = c("M␣<=␣50",
                                                                     "F␣<=␣50", "M␣
                                                                     51-64",
                                                                           "F␣51-64",
                                                                               "M␣>=␣
                                                                               65", "F
                                                                               ␣>=␣65"
                                                                               )))

ggplot(aes(x = agesex, y = totchg), data = data3) + stat_boxplot(geom = "
    errorbar",
     width = 0.2) + geom_boxplot(aes(fill = agesex)) + xlab("Sex␣and␣age") +
         ylab("Total␣charges␣(in␣dollars)") +
     scale_x_discrete(labels = c("Male,␣age␣<=␣50", "Female,␣age␣<=␣50", "Male,
         ␣age␣51-64",
          "Female,␣age␣51-64", "Male,␣age␣>␣65", "Female,␣age␣>␣65")) + scale_
             fill_manual(values = c("#0ebeff",
     "#47cf73", "#ae63e4", "#fcd000", "#ff3c41", "#76daff"), guide = FALSE) +
         theme_classic()

## Warning: It is deprecated to specify `guide = FALSE` to remove a guide.
    Please
## use `guide = "none"` instead.
```



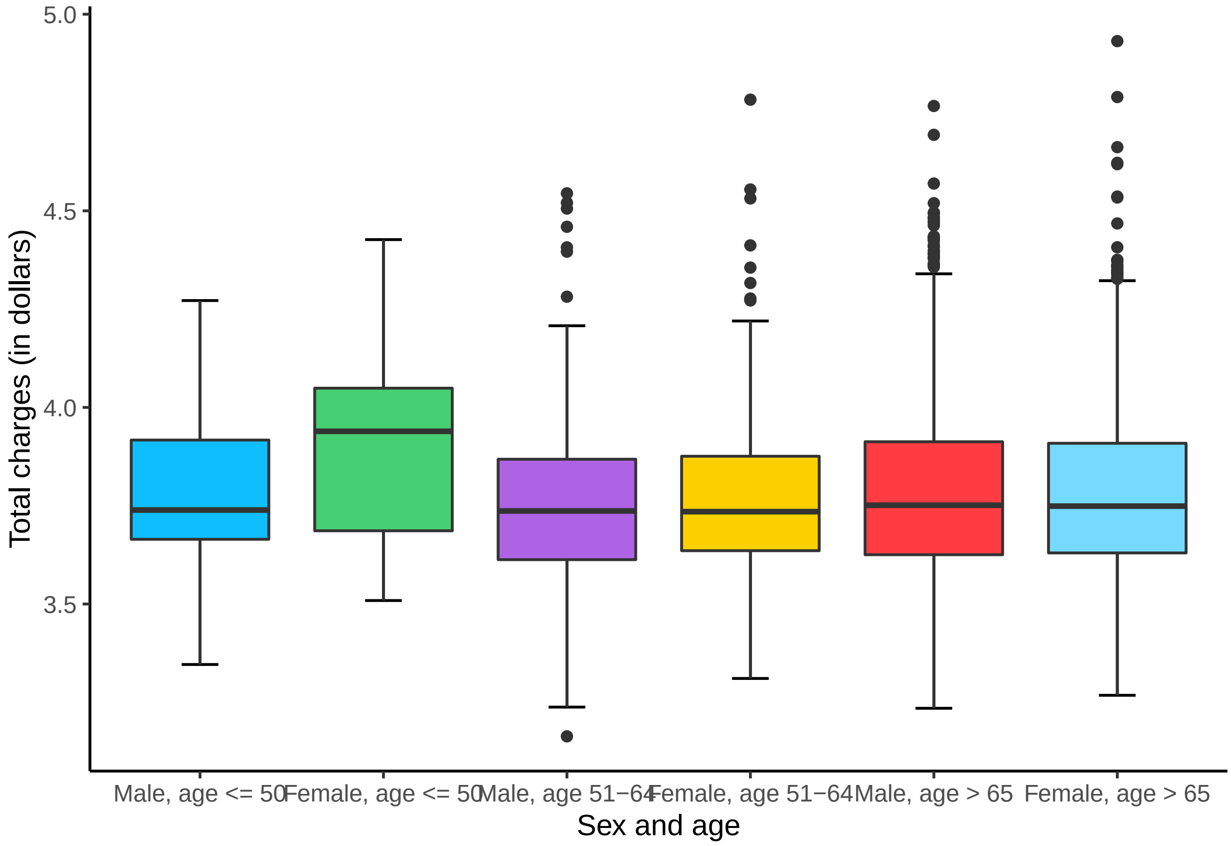
v. Write a sentence or two that describe how the CE cost distribution differs for men and women within each of the three age strata. From an equivalent but alternate perspective, describe how the distribution varies across age separately for men and women.

```
# data3 %>% group_by(interaction(age_cat, sex)) %>% summarise(mean =
# mean(totchg), se = sd(totchg), median = median(totchg), iqr =
# quantile(totchg, c(0.75)) - quantile(totchg, c(0.25)))

## A tibble: 6 x 5 `interaction(age_cat, sex)` mean se median iqr <fct> <dbl>
# <dbl> <dbl> <dbl> 1 (0,50].Male 6758. 3500. 5484 3640. 2 (50,64].Male
    6461.
# 4276. 5454 3282. 3 (64,92].Male 7152. 5053. 5638 3959 4 (0,50].Female
# 9524. 6317. 8695 6337 5 (50,64].Female 6953. 5490. 5431 3192 6
# (64,92].Female 7203. 5638. 5609 3842.

# In the year 1995, in all age and sex groups, the distributions of total
    charges in dollars are right skewed.
# As for the distribution differs for men and women within each of the three
    age strata, among age <= 50 years, both median and interquartile range of
    men are lower than women, indicating a lower charges among men and lower
    variability. While among age group 51−64 and 65+, both median and
    interquartile range of men are slightly higher than women, indicating a
    higher charges and variability among men.
# As for the distribution various across age separately for men and women,
    among men, age group 51−64 has the lowest median and iterquartile range,
    follow by age group <= 50 and 65+, indicating the total charges with
    variability among age groups follow a sequence of 51−64, <= 50, 65+, from
```

```
small to large. Among women, age group 51−64 has the lowest median and
iterquartile range, follow by age group 65+ and <= 50, indicating the
total charges with variability among age groups follow a sequence of
51−64, 65+, <= 50, from small to large.
```

vi. Now repeat Step iv. using a logarithmic (base 10) scale for dollars. Specifically generate a new variable with log(base 10) dollars. How does this display compare with the previous one using the original scale? What additional patterns are made apparent?

```
data3 <− data3 %>%
      mutate(log10chg = log10(totchg))

ggplot(aes(x = agesex, y = log10chg), data = data3) + stat_boxplot(geom = "
     errorbar",
      width = 0.2) + geom_boxplot(aes(fill = agesex)) + xlab("Sex␣and␣age") +
           ylab("Total␣charges␣(in␣dollars)") +
      scale_x_discrete(labels = c("Male,␣age␣<=␣50", "Female,␣age␣<=␣50", "Male,
          ␣age␣51−64",
            "Female,␣age␣51−64", "Male,␣age␣>␣65", "Female,␣age␣>␣65")) + scale_
                 fill_manual(values = c("#0ebeff",
      "#47cf73", "#ae63e4", "#fcd000", "#ff3c41", "#76daff"), guide = FALSE) +
           theme_classic()

## Warning: It is deprecated to specify `guide = FALSE` to remove a guide.
     Please
## use `guide = "none"` instead.
```



```
# The distribution of total dollars change to an approximately normal
    distribution in each age and sex group. The differences in median and
    variability in each age and sex group are made apparent compared with
    privous boxplot.
```

vii. To assure reproducible research, save your log file. Please include the relevant parts of your edited log file with your results as part of your write-up.