Problem Set 2

Introduction to R | University of Oxford Sociology

Casey Breen

Problem Set 2

Complete the following questions in `R` within a `Quarto` document.

Exercise 1: Work with Real-World Data

For this exercise, download the CenSoc-Numident Demo file (as .CSV) and the accompanying codebook (as PDF) from Harvard Dataverse. The CenSoc-Numident is an individual-level data with information on individual-level mortality and sociodemographic characteristics.

1.1 Read in the dataset using `read_csv()` from the tidyverse package.

```
## install tidyverse package if it's not already installed
install.packages(tidyverse)

## library tidyverse, gives us the read_csv() function
library(tidyverse)

## read in data file
censoc <- fread_csv("/path/to/your/dataset.csv")
```

1.2 How many columns are in the dataset?

```
## check number of columns
ncol(censoc)
```

```
[1] 39
```

1.3 How many rows are in the dataset?

```
## check number of rows
nrow(censoc)
```

```
[1] 85865
```

1.4 List the column names. What are a few research questions that could be addressed using this dataset?

```
## print names
names(censoc)
```

```
 [1] "histid"                        "byear"
 [3] "bmonth"                        "dyear"
 [5] "dmonth"                        "death_age"
 [7] "race_first"                    "race_first_cyear"
 [9] "race_last"                     "bpl_string"
[11] "zip_residence"                 "socstate"
[13] "socstate_string"               "age_first_application"
[15] "link_abe_exact_conservative"   "weight"
[17] "weight_conservative"           "perwt"
[19] "age"                           "sex"
[21] "bpl"                           "mbpl"
[23] "fbpl"                          "educd"
[25] "empstatd"                      "hispan"
[27] "incnonwg"                      "incwage"
[29] "marst"                         "nativity"
[31] "occ"                           "occscore"
[33] "ownershp"                      "pernum"
[35] "race"                          "rent"
[37] "serial"                        "statefip"
[39] "urban"
```

***Answer**: This dataset has information both about sociodemographic characteristics and mortality from the `death_age` variable. This dataset could be used to study mortality disparities.*

Exercise 1: Data manipulation

2.1 Filter the `censoc` data frame to include only women (for `sex` variable, 1 = men, 2 = women). Use the `filter` command.

```r
## filter to women
censoc_women <- censoc %>%
  filter(sex == 2)
```

2.2 Filter the original `censoc` data.frame to only include people born between 1905 and 1920 using the `byear` variable.

```r
## filter to people born in 1905 to 1920
censoc_byear_filter <- censoc %>%
  filter(byear %in% 1905:1920)
```

2.3 Select the columns `histid`, `death_age`, `sex`, and `ownershp` from the original censoc data.frame

```r
## select columns histid, death_age, sex, and ownershp
censoc_select_vars <- censoc %>%
  select(histid, death_age, sex, ownershp)
```

2.4 Calculate the average age of death for women (hint: use `filter`)

```r
## calculate average age of death for women
censoc %>%
  filter(sex == 2) %>%
  summarize(mean(death_age))
```

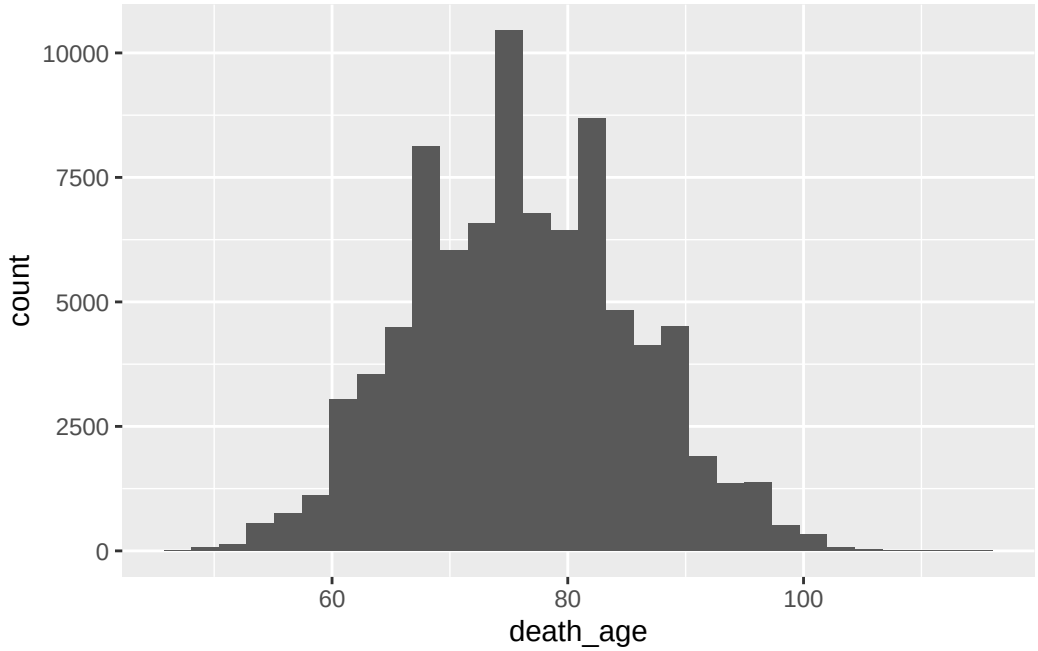
```
# A tibble: 1 x 1
  `mean(death_age)`
              <dbl>
1              78.2
```

Exercise 3 - Data visualization

3.1 Make a histogram of the variable `death_age`. When are most people dying?

```r
ggplot(data = censoc) +
  geom_histogram(aes(x = death_age))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

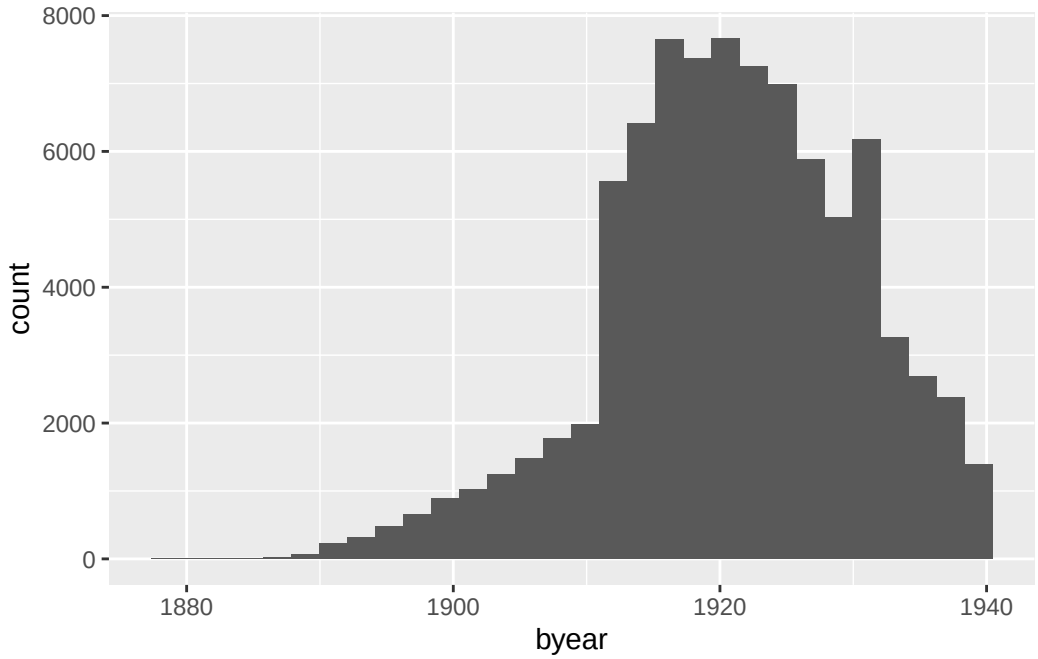


***Answer**: Most people are dying approximately between the ages of 60 and 90.*

3.2 Make a histogram of the variable `byear`. When are most people born?

```
ggplot(data = censoc) +
  geom_histogram(aes(x = byear))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



***Answer**: Most people are born between 1910 and 1930.*

3.3 Recode the variable `sex` from numeric (1, 2) to take values "men" and "women"

```
## recode sex
censoc <- censoc %>%
  mutate(sex_recode = case_when(
    sex == 1 ~ "men",
    sex == 2 ~ "women"
  ))

## look at first few rows to check our recode worked
censoc %>%
  select(sex, sex_recode) %>%
  head()
```

```
# A tibble: 6 x 2
    sex sex_recode
  <dbl> <chr>
1     1 men
2     1 men
3     1 men
```

```
4     2 women
5     1 men
6     2 women
```

3.4 Calculate the mean of of death for both men and women using `group_by()` and `summarize()`. Use the `death_age` variable. Do men or women live longer in this sample?

```
## calculate mean age of death
censoc %>%
  group_by(sex) %>%
  summarize(mean(death_age))
```

```
# A tibble: 2 x 2
    sex `mean(death_age)`
  <dbl>             <dbl>
1     1              73.9
2     2              78.2
```

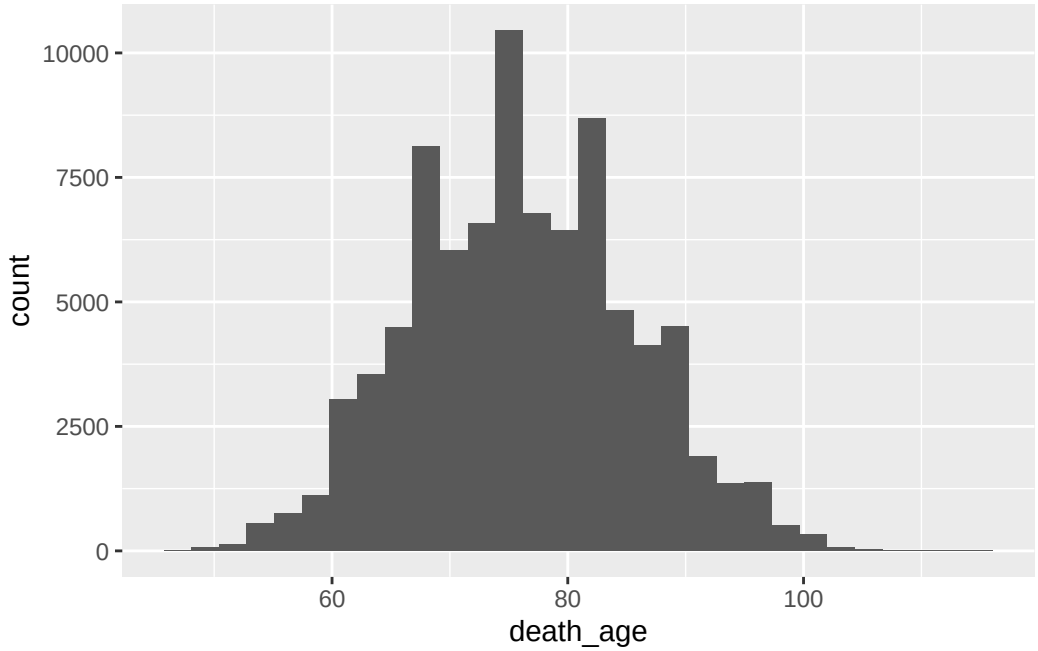
Answer: *The women in our sample live approximately 4.3 years longer than men.*

3.5 Make a histogram of the variable `death_age` for both men and women. Use the `filter()` command.

```
## filter to men
censoc_men <- censoc %>%
  filter(sex_recode == "men")

## histogram for men
ggplot(censoc) +
  geom_histogram(aes(x = death_age))
```

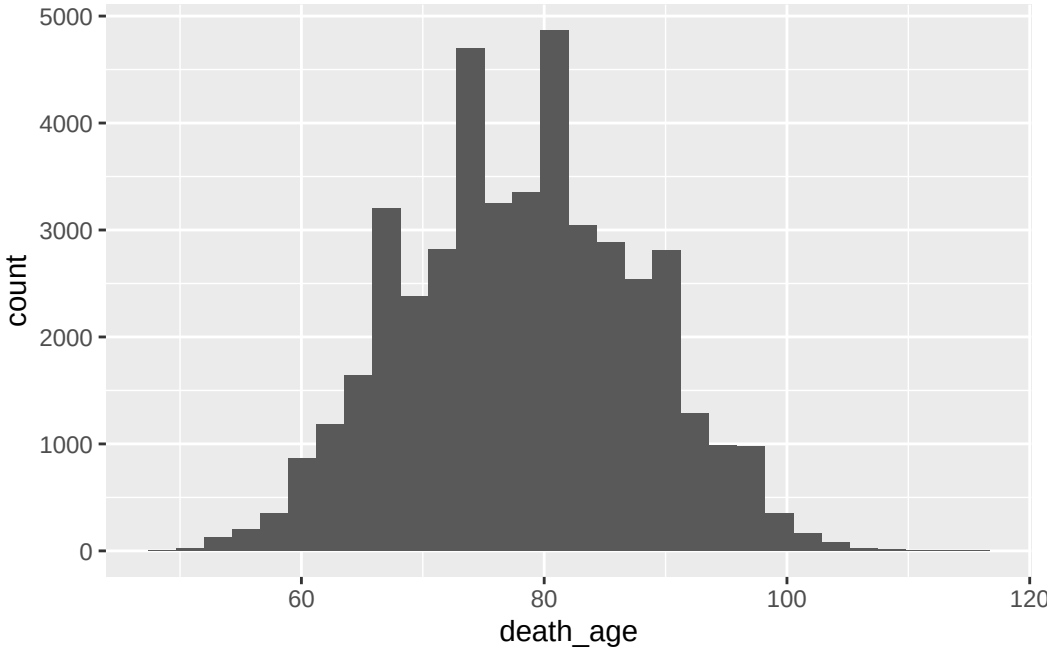
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## histogram for women
censoc_women <- censoc %>%
  filter(sex_recode == "women")

ggplot(censoc_women) +
  geom_histogram(aes(x = death_age))
```

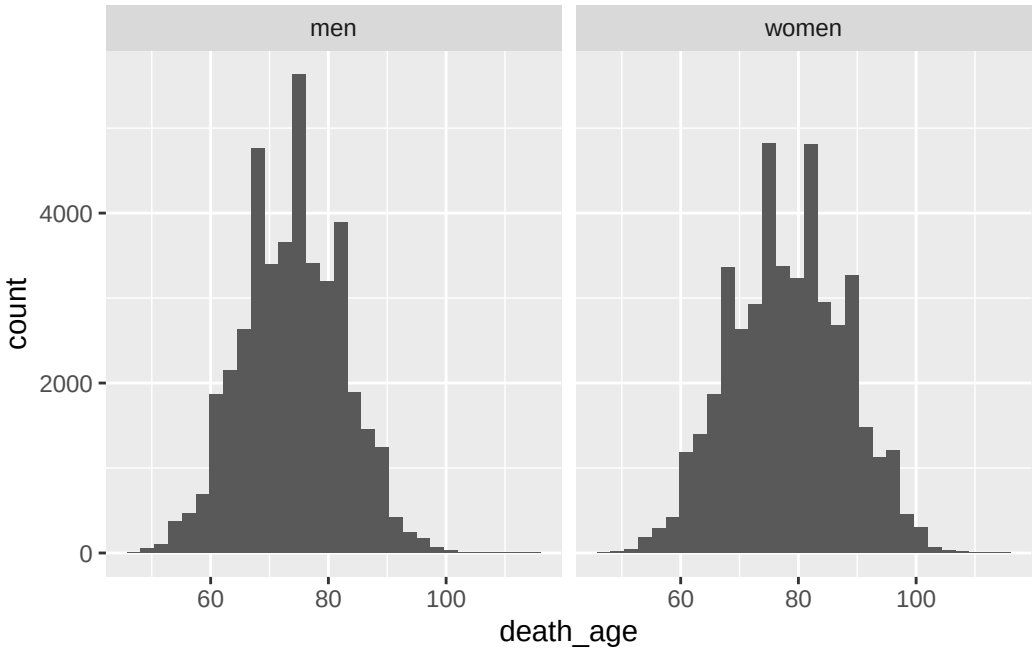
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.6 Try adding the following line to the histogram you made in question 1: `+ facet_wrap(~sex_recode)`

```
## create histogram by
ggplot(data = censoc, aes(x = death_age)) +
  geom_histogram() +
  facet_wrap(~sex_recode)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Exercise 4 - Mortality advantage of homeowners

Do homeowners in the United States live longer than renters in the United States?

4.1 Using the `censoc` data.frame, create a new data.frame `censoc_homeownership` that filters out any "not available values" for the `ownershp` variable (values of ownershp = 0). Use the `filter` command.

```
censoc_homeownership <- censoc %>%
  filter(ownershp != 0)
```

4.2 In the `censoc_homeownership` data.frame, create a new variable `homeowner` using the `mutate` command and the `case_when` command. Assign this new variable `homeowner` a value of "own" if `ownershp == 1` and a value of "rent" if `ownershp == 2`. Note: we can check the values for this variable here: https://usa.ipums.org/usa-action/variables/OWNERSHP#codes_section

```
## create new homeowner variable
censoc_homeownership <- censoc_homeownership %>%
  mutate(homeowner = case_when(
    ownershp == 1 ~ "own",
    ownershp == 2 ~ "rent"
```

```
  ))
```

4.3 Make a histogram on the age of death for "homeowner" and "renter" groups using `ggplot`. Use the `facet_wrap(~homeowner)` — and make sure you're using your `censoc_homeownership` data.frame

```
ggplot(data = censoc_homeownership) +
  geom_histogram(aes(x = death_age)) +
  facet_wrap(~homeowner)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

4.4 Calculate the average age of death for "homeowner" and "renter" groups. Which group lives longer, on average? Use the `group_by` and `summarize` commands. What are some possible explanations for homeowners living longer than renters in the US?

```
censoc_homeownership %>%
  group_by(homeowner) %>%
  summarize(mean(death_age))
```

```
# A tibble: 2 x 2
  homeowner `mean(death_age)`
  <chr>                 <dbl>
1 own                    76.5
2 rent                   75.8
```

Answer: *There are lots of reasons why the homeowners in our sample might be living longer than renters. First off, it's important that we are just looking at the unadjusted difference in life expectancy between homeowners and renters. So it's possible the difference in life expectancy could have nothing to do with homeownership per se, and all driven by unmeasured confounders.*

However, there are potentially a few different pathways in which owning a home could increase longevity:

1. *Stability: Ownership usually implies a more stable and less stressful living condition.*
2. *Neighborhood Factors: Amenities, safety, and community structures can be better in areas with more homeowners.*
3. *Wealth accumulation: homeownership is a key vessel for saving and wealth accumulation in the U.S. s*
4. *And others!*