Final Report

due November 16, 2021 by 11:59 PM

Team TBD: Maggie Lundberg, Riya Mohan, and Izzy Kjaerulff

11/16/2021

Reading Data and Data Clean Up:

```
spending <- read.csv("../data/spending_data_unzip/IHME_DEX_ED_SPENDING_2006_2016_DATA_Y2021M09D23.CSV")
```

Emergency spending

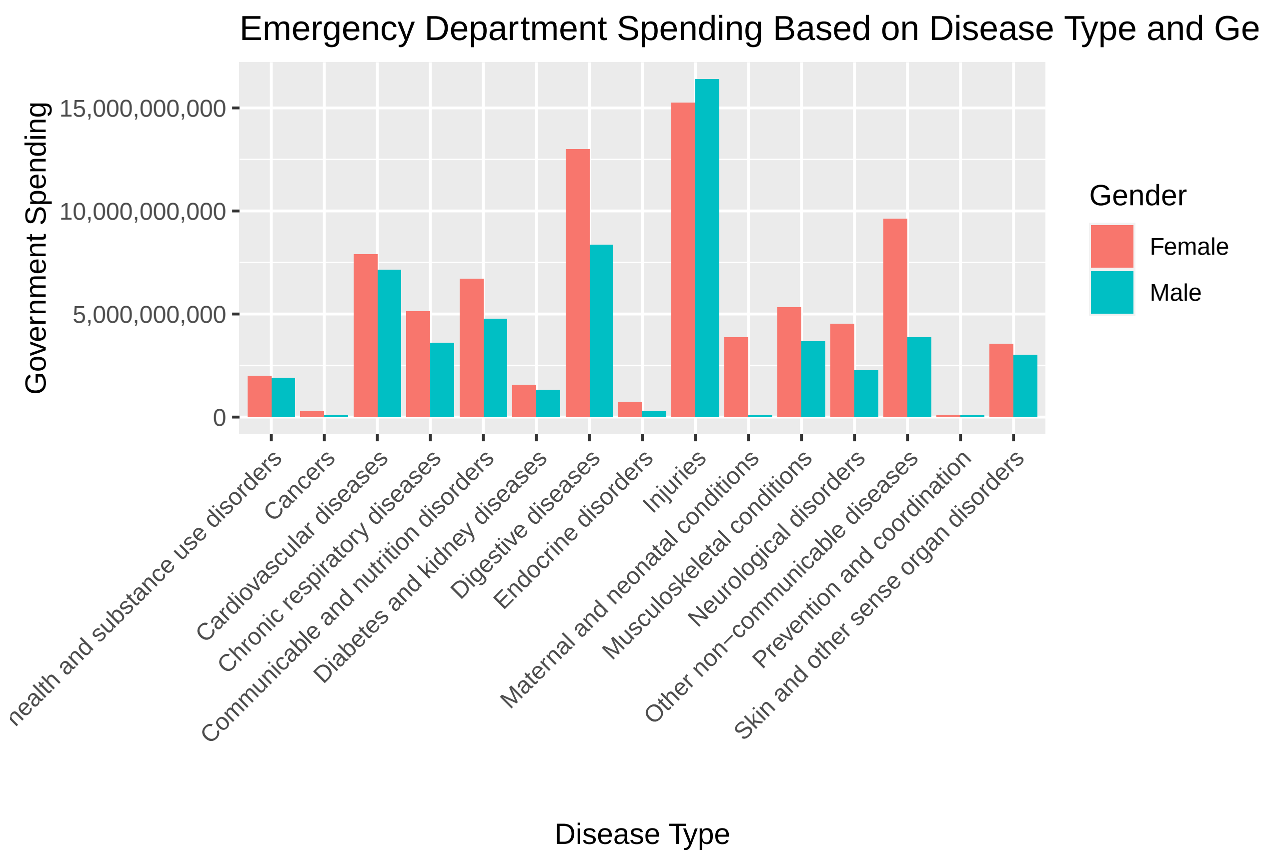
Gender

Does the emergency department spend a different amount of money on males and females? This is looking at all spending, not taking into account type of insurance.

Here is a boxplot showing the distribution of Emergency Department Government spending based on disease type and gender. ADD interpretation

```
spending_malefemale <- spending %>%
  filter(sex %in% c("Female", "Male"))
```

```
ggplot(data = spending_malefemale, aes(x = agg_cause, y = mean_all, fill = sex)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme(axis.text.x = element_text(angle = 45,hjust = 1)) +
  scale_y_continuous(labels = scales::comma) +
  labs(
    x = "Disease Type",
    y = "Government Spending",
    title = "Emergency Department Spending Based on Disease Type and Gender",
    fill = "Gender"
       )
```




```
t.test(spending_malefemale$mean_all~spending_malefemale$sex) %>%
  print()
```

```
##
##  Welch Two Sample t-test
##
## data:  spending_malefemale$mean_all by spending_malefemale$sex
## t = 4.0269, df = 6416.2, p-value = 5.717e-05
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##   56638431 164092573
## sample estimates:
## mean in group Female   mean in group Male
##            413610916            303245414
```

Linear regression model for gender and government spending model

```
spending_malefemale_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(mean_all~sex, data = spending_malefemale)
```

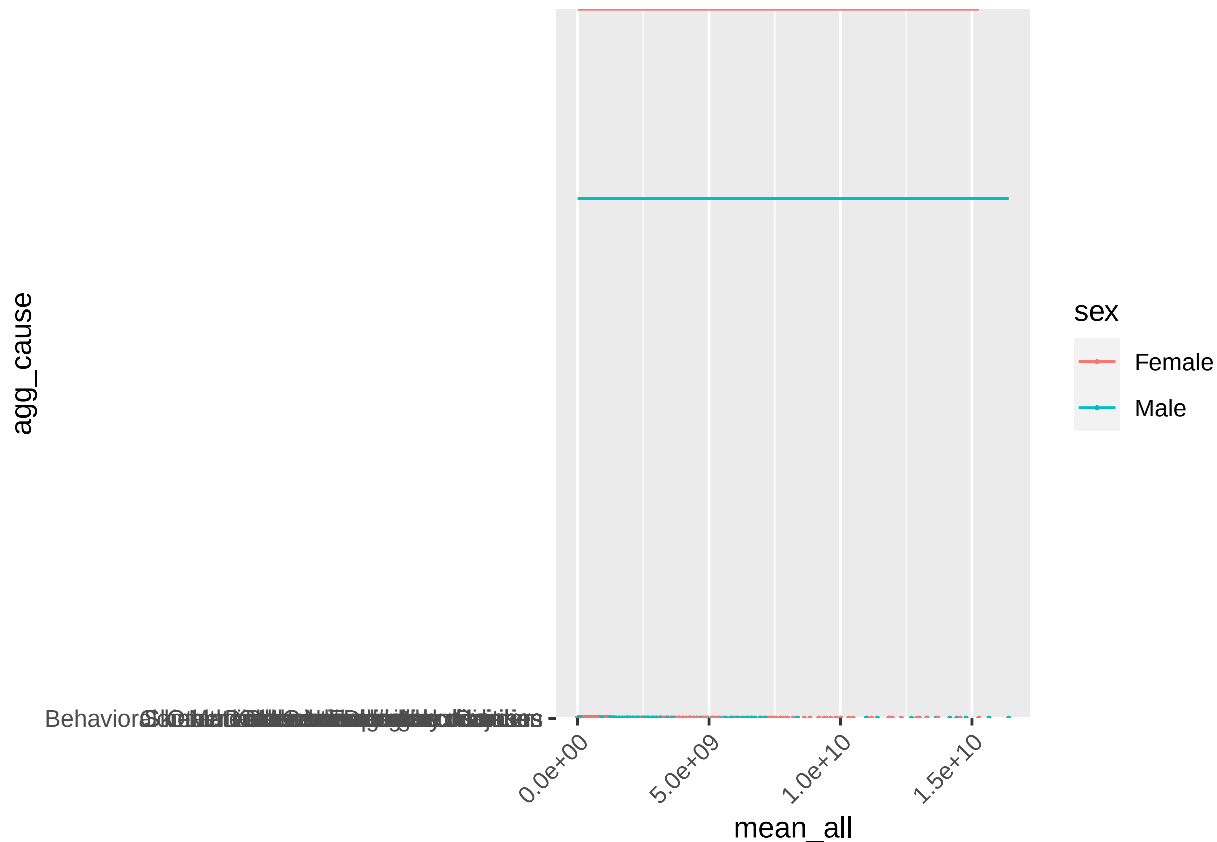
Correlation for males vs females given disease type:

```
spending_malefemale_disease_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(mean_all~sex*agg_cause, data = spending_malefemale) %>%
  tidy()
```

!! ok obviously this isnt right

```
augment_spendinggenderfit <- augment(spending_malefemale_fit$fit)
```

```
ggplot(data = spending_malefemale, mapping = aes(x = mean_all, y = agg_cause, color = sex)) +
  geom_point(size = 0.25) +
  geom_line(data = augment_spendinggenderfit, mapping = aes(x = mean_all, y = .fitted))+
  theme(axis.text.x = element_text(angle = 45,hjust = 1))
```



!! Could we do anova for disease type??

Age

We wonder whether there is a correlation between government healthcare expenditures in the emergency department and age. The age variable is categorical, split into 19 groups that generally include 5 years each, apart from the first (<1 year) and last (85 plus) groups.

To address this question, we began by rearranging and subsetting our given data set. We collapsed group observations divided by years into one and created two distinct age groups divided at the 45-year mark (below 45 and 45 plus) instead of the original 19. Doing so made it possible to do an F test on the newly defined age groups, which necessitates that there are 2 levels in the grouping factor.

!! trying to figure out how to turn all those groups into 2

```
age_year_combined <- spending_malefemale %>%
  group_by(age_group_name) %>%
  summarize_at(vars(mean_all),funs(mean(.,na.rm=TRUE))) %>%
  print()
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
## # A tibble: 20 x 2
##    age_group_name    mean_all
##    <chr>                <dbl>
##  1 <1 year          37373144.
##  2 1 to 4          110266529.
##  3 10 to 14        110034260.
##  4 15 to 19        182987055.
##  5 20 to 24        268905287.
##  6 25 to 29        279296088.
##  7 30 to 34        254696363.
##  8 35 to 39        240303298.
##  9 40 to 44        245704951.
## 10 45 to 49        253869656.
## 11 5 to 9           84242670.
## 12 50 to 54        248987505.
## 13 55 to 59        224429718.
## 14 60 to 64        195831906.
## 15 65 to 69        188573358.
## 16 70 to 74        167344507.
## 17 75 to 79        153410249.
## 18 80 to 84        148316306.
## 19 85 plus         180252733.
## 20 All Ages       3572334040.
```

```
under45 <- age_year_combined %>%
  filter(age_group_name %in% c("<1 year", "1 to 4", "5 to 9", "10 to 14", "15 to 19", "20 to 24", "25 to
  summarize_at(vars(mean_all),funs(mean(.,na.rm=TRUE))) %>%
  print()
```

```
## # A tibble: 1 x 1
##     mean_all
##        <dbl>
## 1 181380965.
```

```
above45 <- age_year_combined %>%
  filter(age_group_name %in% c("45 to 49", "50 to 54", "55 to 59", "60 to 64", "65 to 69", "70 to 74",
  summarize_at(vars(mean_all),funs(mean(.,na.rm=TRUE))) %>%
  print()
```

```
## # A tibble: 1 x 1
##     mean_all
##        <dbl>
```

```
## 1 195668438.
```

!! this won't work until there are only 2 levels for the grouping factor

```
var.test(mean_all ~ age_group_name, age_year_combined,
         alternative = "two.sided")
```