practice_exercise

Ekta Chaudhary

28/06/2020

```r
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2
```

```
## v ggplot2 3.2.1      v purrr   0.3.2
## v tibble  2.1.3      v dplyr   0.8.3
## v tidyr   1.0.0      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0
```

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

```r
library(readxl)
library(dplyr)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```r
practice_data = read_excel("./data/Practice_exercise.xlsx", sheet = "Data") %>%
  janitor::clean_names() %>%
  select(observation_number,quarter,employee_id, sex = sex_male_1, race, age, hospital_visit = hospital_
  mutate(
    age_cat = case_when(
      age < 30 ~ 1,
      age <= 45 ~ 2,
      age > 45 ~ 3
    )
  )
```

#Checking for missing data

```r
sapply(practice_data, function(x) sum(is.na(x)))
```

```
## observation_number             quarter          employee_id
##                  0                   0                    0
##                sex                race                  age
##                 71                2123                    0
##     hospital_visit              salary         health_score
##                  0                   0                    0
##            age_cat
##                  0
```

#Checking for missing data

```
practice_data %>%
  select(everything()) %>%  # replace to your needs
  summarise_all(funs(sum(is.na(.))))
```

```
## Warning: funs() is soft deprecated as of 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 per session.
```

```
## # A tibble: 1 x 10
##   observation_num~ quarter employee_id   sex  race   age hospital_visit
##              <int>   <int>       <int> <int> <int> <int>          <int>
## 1                0       0           0    71  2123     0              0
## # ... with 3 more variables: salary <int>, health_score <int>,
## #   age_cat <int>
```

#finding the minimum and maximum values of each variable

```
sapply(practice_data, function(x) min(x))
```

```
## observation_number            quarter        employee_id
##       1.000000e+00       1.000000e+00       1.000000e+00
##                sex               race                age
##                 NA                 NA       7.000000e+00
##     hospital_visit             salary       health_score
##       0.000000e+00       2.835070e+04       6.265991e-01
##            age_cat
##       1.000000e+00
```

```
sapply(practice_data, function(x) max(x))
```

```
## observation_number            quarter        employee_id
##           19103.00              12.00            2000.00
##                sex               race                age
##                 NA                 NA             172.00
##     hospital_visit             salary       health_score
##               1.00           68826.34              10.00
##            age_cat
##               3.00
```

#checking the number of employees with health score outside the range of data

```r
practice_data %>%
  count(
   health_sc_6 = ifelse(health_score > 6, 1, 0)
  )
```

```
## # A tibble: 2 x 2
##   health_sc_6     n
##         <dbl> <int>
## 1           0 17865
## 2           1  1238
```

```r
practice_data %>%
  select(
    employee_id, sex
  ) %>%
  filter(
  is.na(sex)
  ) %>%
  group_by(
    employee_id
  ) %>%
  summarise(
    missing = sum(is.na(sex))
  )
```

```
## # A tibble: 7 x 2
##   employee_id missing
##         <dbl>   <int>
## 1        1994      10
## 2        1995       9
## 3        1996      12
## 4        1997      11
## 5        1998      12
## 6        1999       7
## 7        2000      10
```

```r
practice_data %>%
  select(
    employee_id, race
  ) %>%
  filter(
     is.na(race)
  ) %>%
  group_by(
    employee_id
  ) %>%
  summarise(
    miss = sum(is.na(race))
  )
```

```
## # A tibble: 220 x 2
##   employee_id  miss
```

```
##             <dbl> <int>
##   1             8    10
##   2            10    12
##   3            13     9
##   4            22     9
##   5            36    12
##   6            38    12
##   7            48    10
##   8            49     7
##   9            51     8
## 10            55     9
## # ... with 210 more rows
```

#Calculating the number of employees in each age group for each quarter

```r
emp_data = practice_data %>%
  mutate(
    quarter = factor(
    quarter),
    age_cat = factor(age_cat)
  )
```

```r
emp_data = emp_data %>%
  select(
    employee_id, quarter, age_cat
  ) %>%
  group_by(
    quarter, age_cat
  ) %>%
  tally()
```

```r
e <- ggplot(emp_data, aes(x = quarter, y = n, group = age_cat)) +
  geom_line(aes(color = age_cat)) +
  geom_point(aes(color = age_cat)) + labs(x = "Quarter", y = "Number of employees", title = "Number of
e
```

Number of employees in different age group for each quarter
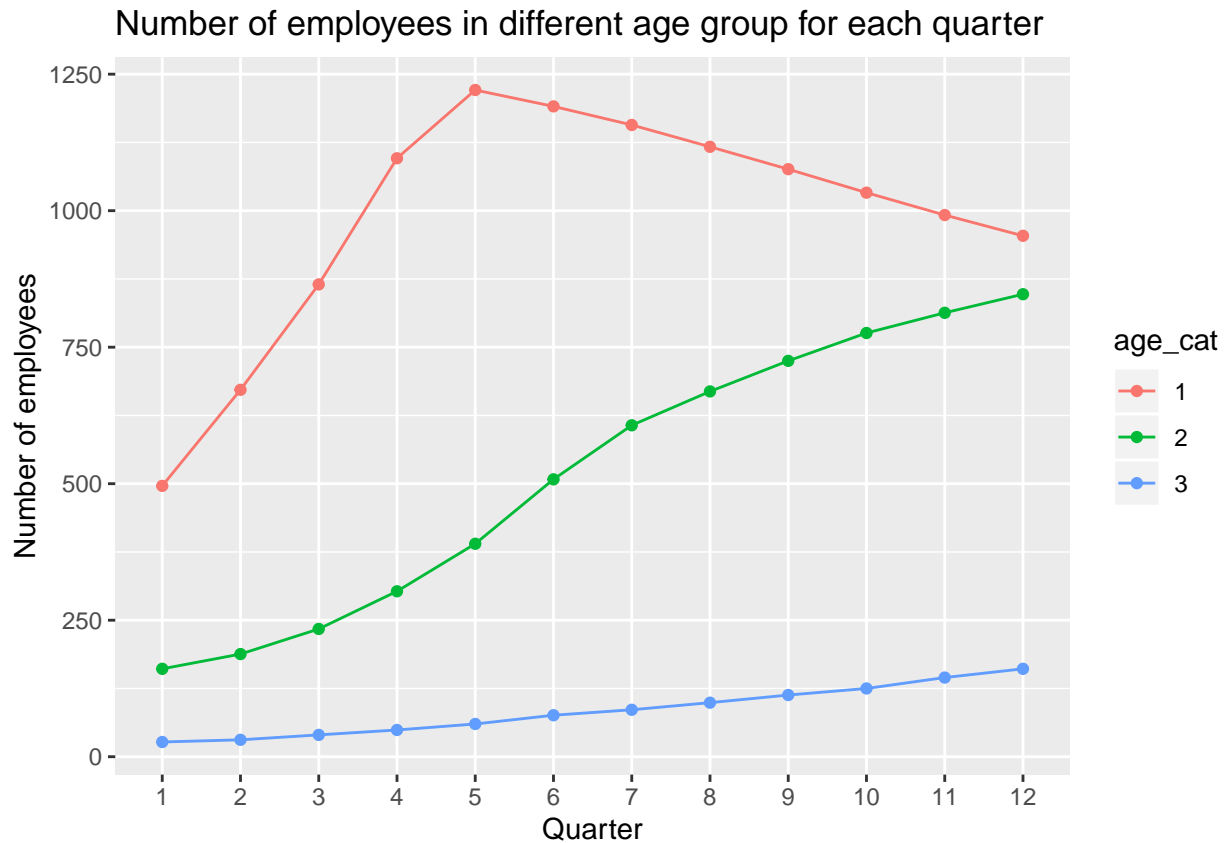
#checking the trend in average salary over time

```
practice_data %>%
  select(
    salary, quarter
  ) %>%
  group_by(
    quarter
  ) %>%
  summarise(
    avg_salary = mean(salary)
  )
```

```
## # A tibble: 12 x 2
##    quarter avg_salary
##      <dbl>      <dbl>
##  1       1     43628.
##  2       2     44274.
##  3       3     45021.
##  4       4     45531.
##  5       5     46133.
##  6       6     46948.
##  7       7     47780.
##  8       8     48667.
##  9       9     49562.
## 10      10     50498.
## 11      11     51433.
```

```
## 12      12     52376.
```

```
salary_data = practice_data %>%
  mutate(
    age_cat = factor(age_cat),
quarter = factor(quarter))
```

#checking the trend in average salary over time by age group

```
salary_data = salary_data %>%
  select(
    salary, quarter, age_cat
  ) %>%
  group_by(
    quarter, age_cat
  ) %>%
  summarise(
    avg_salary = mean(salary)
  )
```

```
p <- ggplot(salary_data, aes(x = quarter, y = avg_salary, group = age_cat)) +
  geom_line(aes(color = age_cat)) +
  geom_point(aes(color = age_cat)) + labs(x = "Quarter", y = "Average Salary", title = "Trend in average
p
```


Trend in average Salary over time by age group
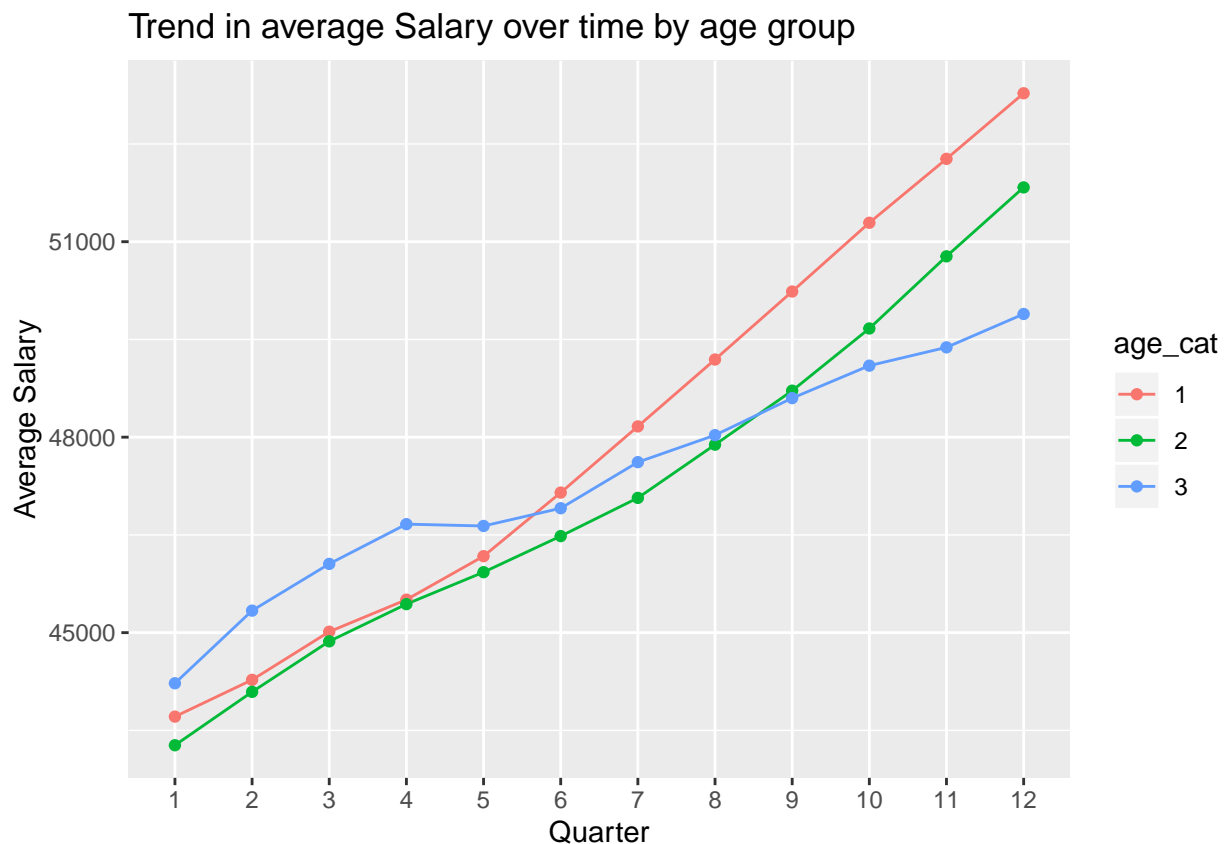
#checking the trend in mean health score over time

6

```
hc = practice_data %>%
  select(
    health_score, quarter
  ) %>%
  group_by(
    quarter
  ) %>%
  summarise(
    avg_score = mean(health_score)
  )
```

```
px <- ggplot(hc, aes(x = factor(quarter), y = avg_score)) +
  geom_bar(stat = "identity", position = "dodge") + labs(x = "Quarter", y = "Mean Health Score", title =
px
```

Trend in Mean Health Score over time
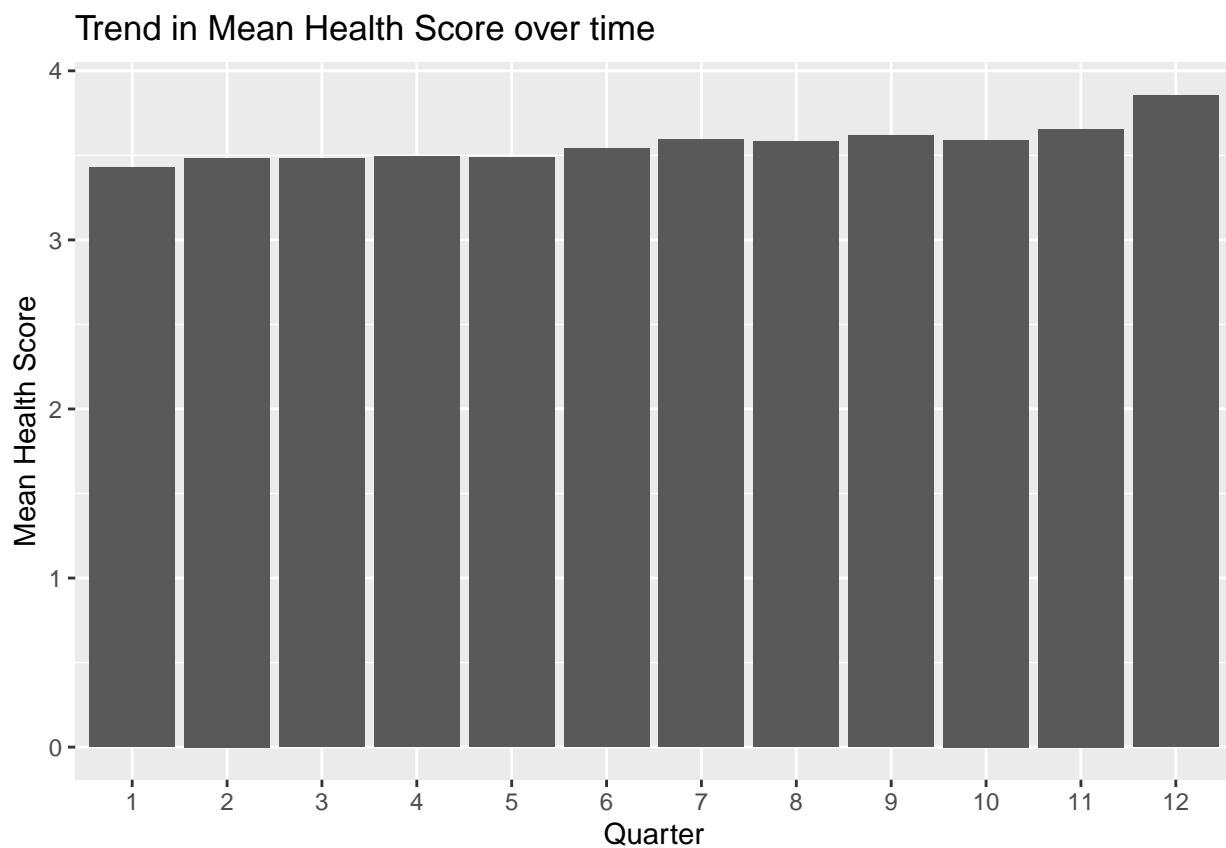
#Mean health score over time by age group
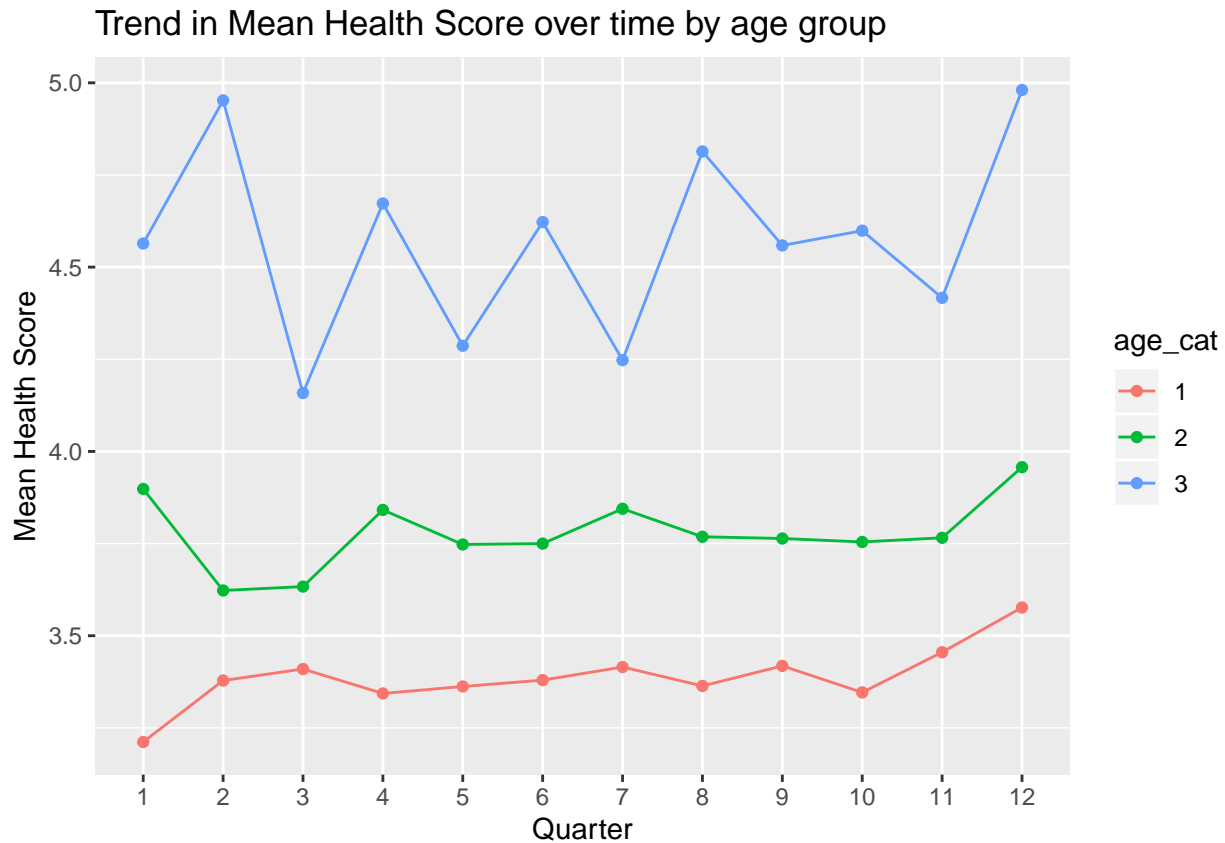
```
practice_data %>%
  select(
    health_score, quarter, age_cat
  ) %>%
  group_by(
    quarter, age_cat
  ) %>%
  summarise(
    avg_score = mean(health_score)
  )
```

```
## # A tibble: 36 x 3
## # Groups:   quarter [12]
##    quarter age_cat avg_score
##      <dbl>   <dbl>     <dbl>
##  1       1       1      3.21
##  2       1       2      3.90
##  3       1       3      4.56
##  4       2       1      3.38
##  5       2       2      3.62
##  6       2       3      4.95
##  7       3       1      3.41
##  8       3       2      3.63
##  9       3       3      4.16
## 10       4       1      3.34
## # ... with 26 more rows
```

```r
health_sc = practice_data %>%
  mutate(
    age_cat = factor(age_cat),
quarter = factor(quarter))
```

```r
health_sc = health_sc %>%
  select(
    health_score, quarter, age_cat
  ) %>%
  group_by(
    quarter, age_cat
  ) %>%
  summarise(
    avg_score = mean(health_score)
  )
```

```r
p1 <- ggplot(health_sc, aes(x = quarter, y = avg_score, group = age_cat)) +
  geom_line(aes(color = age_cat)) +
  geom_point(aes(color = age_cat)) + labs(x = "Quarter", y = "Mean Health Score", title = "Trend in Mea
p1
```

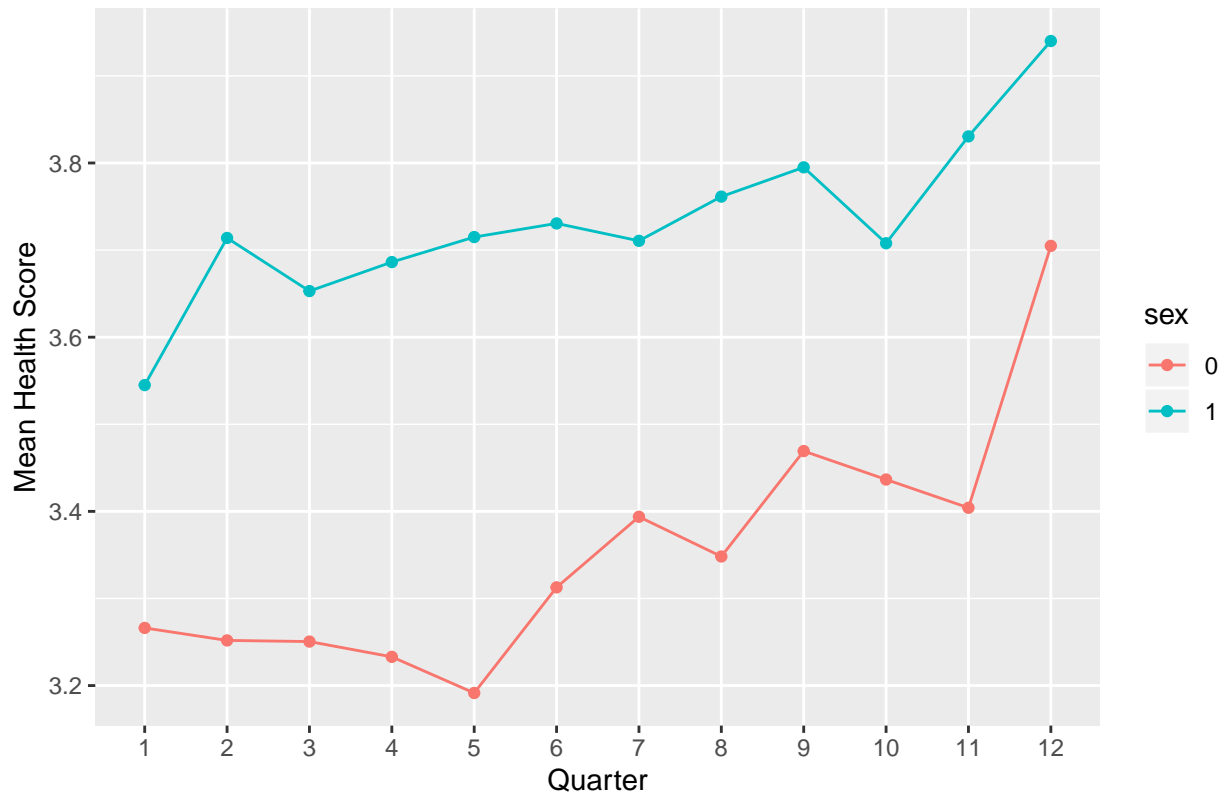## Trend in Mean Health Score over time by age group

#Mean health score over time by sex

```
health_sex = practice_data %>%
  mutate(
    sex = factor(sex),
quarter = factor(quarter))
```

```
health_sex = health_sex %>%
  drop_na() %>%
  select(
    health_score, quarter, sex
  ) %>%
  group_by(
    quarter, sex
  ) %>%
  summarise(
    avg_score = mean(health_score)
  )
```

```
pq <- ggplot(health_sex, aes(x = quarter, y = avg_score, group = sex)) +
  geom_line(aes(color = sex)) +
  geom_point(aes(color = sex)) + labs(x = "Quarter", y = "Mean Health Score", title = "Trend in Mean Hea
pq
```

Trend in Mean Health Score over time by Sex

#Mean health score over time by race

```r
health_race = practice_data %>%
  mutate(
    race = factor(race),
quarter = factor(quarter))
```

```r
health_race = health_race %>%
  drop_na() %>%
  select(
    health_score, quarter, race
  ) %>%
  group_by(
    quarter, race
  ) %>%
  summarise(
    avg_score = mean(health_score)
  )
```

```r
pe <- ggplot(health_race, aes(x = quarter, y = avg_score, group = race)) +
  geom_line(aes(color = race)) +
  geom_point(aes(color = race)) + labs(x = "Quarter", y = "Mean Health Score", title = "Trend in Mean He
pe
```

Trend in Mean Health Score over time by Race