

# practice\_\_exercise

*Ekta Chaudhary*

*28/06/2020*

```
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()
```

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

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

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

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

```
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_visit)
  mutate(
    age_cat = case_when(
      age < 30 ~ 1,
      age <= 45 ~ 2,
      age > 45 ~ 3
    )
  )
```

```
practice_data %>%
  select(
    employee_id, age_cat, quarter
  ) %>%
  group_by(
    quarter, employee_id
  ) %>%
  summarise(
    total = sum(employee_id)
  )
```

```
## # A tibble: 19,103 x 3
## # Groups:   quarter [12]
##   quarter employee_id total
##   <dbl>      <dbl> <dbl>
## 1     1         1     1
## 2     1         2     2
## 3     1         9     9
## 4     1        10    10
## 5     1        14    14
## 6     1        16    16
## 7     1        18    18
## 8     1        27    27
## 9     1        28    28
## 10    1        29    29
## # ... with 19,093 more rows
```

```
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
```

```
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>
```

```
supply(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
```

```
supply(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
```

```
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
```

```
sqldf("SELECT employee_id, COUNT(employee_id) AS missing FROM practice_data WHERE sex IS NULL GROUP BY employee_id")
```

```
##   employee_id missing
## 1         1994      10
## 2         1995       9
## 3         1996      12
## 4         1997      11
## 5         1998      12
## 6         1999       7
## 7         2000      10
```

```
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

```

```

sqldf("SELECT employee_id, COUNT(employee_id) FROM practice_data WHERE race IS NULL
      GROUP BY employee_id")

```

```

##   employee_id COUNT(employee_id)
## 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
## 11         60                 9
## 12         76                11
## 13         79                 6
## 14         89                 8
## 15        104                 4
## 16        105                 6
## 17        119                 9
## 18        132                12
## 19        169                12
## 20        170                 4
## 21        173                12
## 22        188                11
## 23        192                12
## 24        197                 8
## 25        210                12
## 26        236                12
## 27        257                 9
## 28        276                 8
## 29        277                 8
## 30        283                12

```

## 31	308	10
## 32	313	8
## 33	318	10
## 34	320	6
## 35	324	12
## 36	325	8
## 37	327	6
## 38	338	8
## 39	346	11
## 40	358	11
## 41	369	11
## 42	375	5
## 43	378	12
## 44	379	12
## 45	386	12
## 46	401	12
## 47	416	8
## 48	422	11
## 49	426	12
## 50	430	12
## 51	432	12
## 52	434	10
## 53	436	9
## 54	445	9
## 55	449	12
## 56	454	12
## 57	455	6
## 58	460	12
## 59	476	12
## 60	477	12
## 61	480	12
## 62	485	12
## 63	499	11
## 64	505	9
## 65	509	9
## 66	517	8
## 67	530	12
## 68	543	10
## 69	557	12
## 70	583	12
## 71	586	12
## 72	593	12
## 73	597	8
## 74	616	1
## 75	622	7
## 76	628	7
## 77	650	7
## 78	664	12
## 79	665	12
## 80	671	12
## 81	689	11
## 82	709	8
## 83	713	12
## 84	716	6

## 85	722	12
## 86	728	12
## 87	732	12
## 88	734	12
## 89	736	12
## 90	737	5
## 91	774	12
## 92	793	10
## 93	820	12
## 94	824	8
## 95	828	8
## 96	829	1
## 97	832	12
## 98	848	6
## 99	851	12
## 100	865	10
## 101	873	10
## 102	875	9
## 103	878	12
## 104	900	8
## 105	906	12
## 106	914	12
## 107	918	7
## 108	941	12
## 109	977	8
## 110	990	10
## 111	992	12
## 112	995	12
## 113	1001	12
## 114	1012	12
## 115	1027	12
## 116	1036	8
## 117	1046	12
## 118	1049	7
## 119	1064	8
## 120	1070	1
## 121	1080	12
## 122	1082	9
## 123	1094	10
## 124	1098	10
## 125	1109	12
## 126	1120	12
## 127	1139	4
## 128	1146	11
## 129	1172	11
## 130	1177	12
## 131	1188	7
## 132	1218	7
## 133	1231	9
## 134	1233	9
## 135	1237	9
## 136	1247	8
## 137	1248	8
## 138	1255	12

## 139	1268	11
## 140	1281	5
## 141	1308	12
## 142	1316	6
## 143	1317	11
## 144	1318	9
## 145	1337	6
## 146	1353	9
## 147	1364	12
## 148	1373	8
## 149	1390	11
## 150	1394	2
## 151	1397	4
## 152	1432	12
## 153	1434	9
## 154	1438	9
## 155	1439	12
## 156	1453	11
## 157	1466	11
## 158	1470	6
## 159	1476	12
## 160	1482	9
## 161	1491	12
## 162	1505	6
## 163	1512	9
## 164	1543	6
## 165	1548	9
## 166	1564	10
## 167	1580	1
## 168	1584	11
## 169	1587	12
## 170	1591	12
## 171	1597	9
## 172	1607	11
## 173	1613	12
## 174	1624	9
## 175	1628	10
## 176	1638	12
## 177	1654	7
## 178	1660	11
## 179	1662	10
## 180	1676	7
## 181	1685	12
## 182	1711	11
## 183	1712	10
## 184	1723	8
## 185	1731	2
## 186	1738	10
## 187	1740	12
## 188	1745	9
## 189	1757	12
## 190	1764	9
## 191	1786	10
## 192	1792	5

```
## 193      1795      12
## 194      1797       7
## 195      1817       9
## 196      1822       9
## 197      1841       8
## 198      1851       7
## 199      1854       8
## 200      1855      10
## 201      1863       8
## 202      1864      12
## 203      1872      12
## 204      1887      12
## 205      1890      11
## 206      1900      12
## 207      1906       9
## 208      1909      12
## 209      1912      10
## 210      1924       9
## 211      1926      11
## 212      1931      12
## 213      1942      12
## 214      1944      12
## 215      1948      12
## 216      1949      12
## 217      1961      11
## 218      1966       9
## 219      1997      11
## 220      1999       7
```

```
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
```

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

```
## # A tibble: 12 x 2
##   quarter avg_score
##   <dbl>    <dbl>
## 1       1      3.43
## 2       2      3.48
## 3       3      3.48
## 4       4      3.49
## 5       5      3.49
## 6       6      3.54
## 7       7      3.59
## 8       8      3.58
## 9       9      3.62
## 10      10      3.59
## 11      11      3.66
## 12      12      3.86
```

```
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.
```

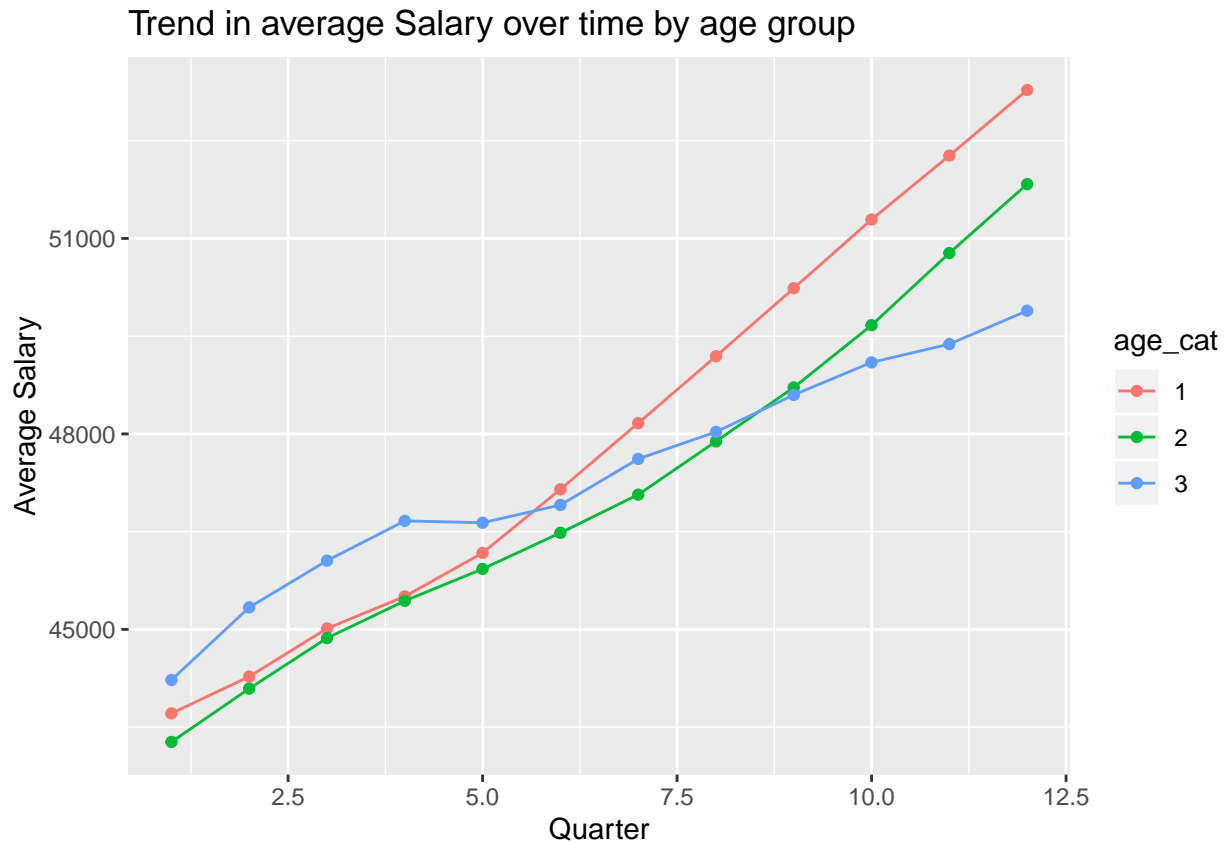
```
practice_data %>%
  select(
    health_score, quarter, age_cat
  ) %>%
  group_by(
    quarter, age_cat
  ) %>%
  summarise(
    avg_score = mean(health_score)
  )
```

```
## # 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
```

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

```
salary = salary %>%
  mutate(
    age_cat = factor(age_cat))
```

```
p <- ggplot(salary, 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
```



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

```
health_sc = health_sc %>%
  mutate(
    age_cat = factor(age_cat))
```

```
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 Mean Health Score over time by age group")
p1
```

