Employee Attrition Codes

Satwik

2022-10-07

Code

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Reading the dataset and chekcing dimensions
df <- read.csv('data/WA_Fn-UseC_-HR-Employee-Attrition.csv')
dim(df)
```

```
## [1] 1470   35
```

```
df <- subset(df, select = -c(EmployeeCount, Over18, Gender, StandardHours, Department, MaritalStatus, Em
dim(df)
```

```
## [1] 1470   13
```

```
df %>% summarise_all(n_distinct)
```

```
##   Age Attrition Education EnvironmentSatisfaction JobInvolvement JobLevel
## 1  43         2         5                        4              4        5
##   JobSatisfaction MonthlyIncome NumCompaniesWorked OverTime PercentSalaryHike
## 1               4          1349                 10        2                15
##   RelationshipSatisfaction WorkLifeBalance
## 1                        4               4
```

```
sample_n(df, 5)
```

```
##   Age Attrition Education EnvironmentSatisfaction JobInvolvement JobLevel
## 1  26       Yes         3                       1             3        1
## 2  47       Yes         4                       1             3        3
## 3  39        No         1                       3             3        3
## 4  34        No         4                       2             4        2
## 5  34        No         2                       1             3        2
##   JobSatisfaction MonthlyIncome NumCompaniesWorked OverTime PercentSalaryHike
## 1               3          2377                  1       No                20
## 2               2         11849                  1      Yes                12
## 3               3         13464                  7       No                21
## 4               4          9725                  0       No                11
## 5               2          4325                  1       No                15
##   RelationshipSatisfaction WorkLifeBalance
## 1                        3               2
## 2                        4               2
## 3                        3               3
## 4                        4               2
## 5                        3               3
```

```
df <- mutate(df, gen = cut(Age,
                    c(-Inf, 23, 38, 54, Inf),
                    c('gen_z', 'millennial', 'gen_x', 'boomers')))
df <- df %>%
  mutate( OverTime = ifelse(OverTime=="Yes", 1, 0 ))
df %>% group_by(gen) %>%
  summarise(avg_Education = median(Education),
            avg_EnvironmentSatisfaction = median(EnvironmentSatisfaction),
            avg_JobInvolvement = median(JobInvolvement))
```

```
## # A tibble: 4 x 4
##   gen        avg_Education avg_EnvironmentSatisfaction avg_JobInvolvement
##   <fct>              <int>                       <int>              <int>
## 1 gen_z                  2                           3                  3
## 2 millennial             3                           3                  3
## 3 gen_x                  3                           3                  3
## 4 boomers                3                           3                  3
```

```
df %>% group_by(gen) %>%
  summarise(
            avg_JobLevel = median(JobLevel),
            avg_JobSatisfaction = median(JobSatisfaction),
            count_OverTime = sum(OverTime))
```

```
## # A tibble: 4 x 4
##   gen        avg_JobLevel avg_JobSatisfaction count_OverTime
##   <fct>             <int>               <int>          <dbl>
## 1 gen_z                 1                   3             23
## 2 millennial            2                   3            225
## 3 gen_x                 2                   3            141
## 4 boomers               3                   3             27
```

```
df %>% group_by(gen) %>%
  summarise(
            avg_MonthlyIncome = median(MonthlyIncome),
```

```
            avg_NumCompaniesWorked = median(NumCompaniesWorked),
            avg_PercentSalaryHike = median(PercentSalaryHike))
```

```
## # A tibble: 4 x 4
##   gen        avg_MonthlyIncome avg_NumCompaniesWorked avg_PercentSalaryHike
##   <fct>                  <int>                  <int>                  <int>
## 1 gen_z                   2500                      1                     15
## 2 millennial              4377                      1                     14
## 3 gen_x                   6811                      3                     14
## 4 boomers                10312                      4                     14
```

```
df %>% group_by(gen) %>%
  summarise(
            avg_RelationshipSatisfaction = median(RelationshipSatisfaction),
            avg_WorkLifeBalance = median(WorkLifeBalance),
            )
```

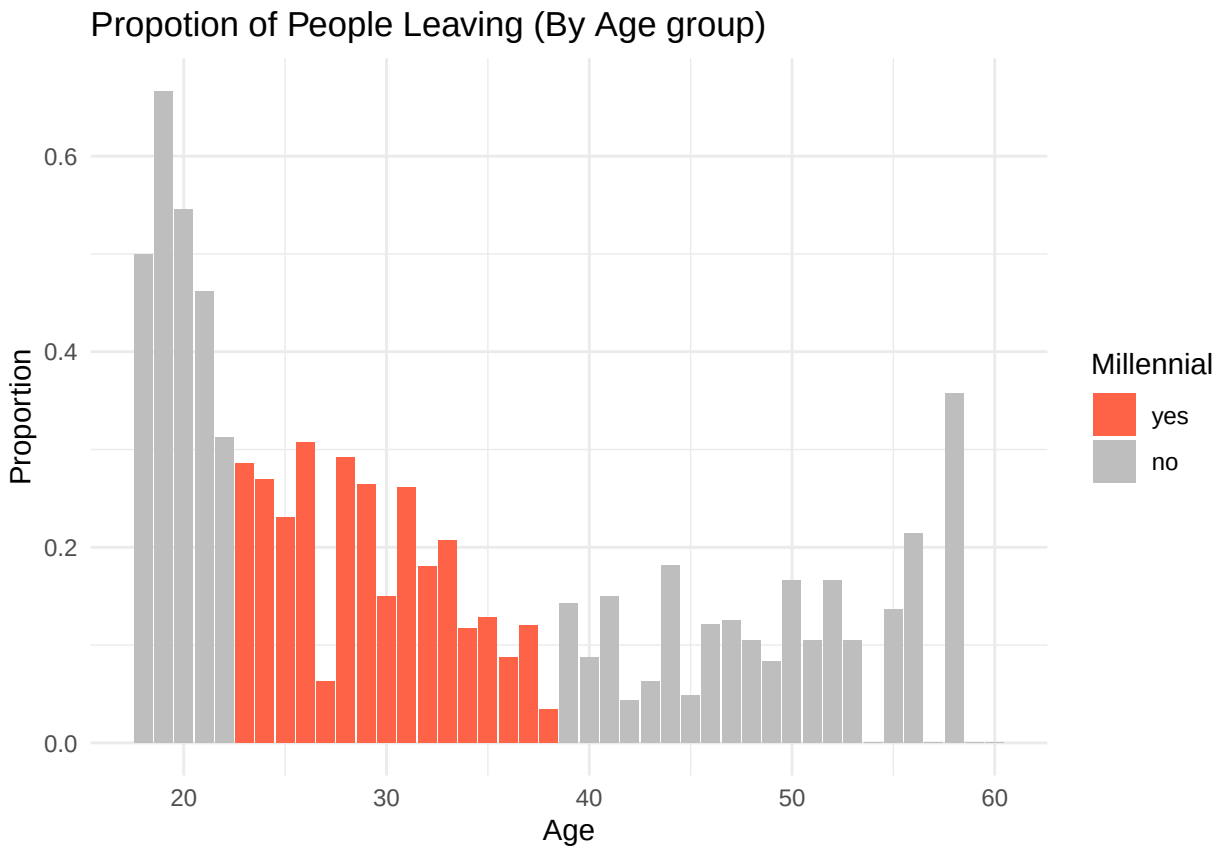
```
## # A tibble: 4 x 3
##   gen        avg_RelationshipSatisfaction avg_WorkLifeBalance
##   <fct>                             <int>               <int>
## 1 gen_z                                 3                   3
## 2 millennial                            3                   3
## 3 gen_x                                 3                   3
## 4 boomers                               3                   3
```

```
# Since our outcome variable is attrition, and the predictor of interest is age, let's focus on these t

df$Attrition <- factor(df$Attrition)

count_attr <- df %>%
    group_by(Age) %>%
    summarize(count_n = n(),
              attr_count = sum(Attrition=="Yes"))

count_attr$prop <- count_attr$attr_count/count_attr$count_n
count_attr[order(-count_attr$prop), ]
```

```
## # A tibble: 43 x 4
##      Age count_n attr_count  prop
##    <int>   <int>      <int> <dbl>
##  1    19       9          6 0.667
##  2    20      11          6 0.545
##  3    18       8          4 0.5
##  4    21      13          6 0.462
##  5    58      14          5 0.357
##  6    22      16          5 0.312
##  7    26      39         12 0.308
##  8    28      48         14 0.292
##  9    23      14          4 0.286
## 10    24      26          7 0.269
## # ... with 33 more rows
```

```
count_attr <- count_attr %>%
  mutate( Millennial = ifelse((Age >=23 & Age <= 38), "yes", "no" ))

ggplot(count_attr, aes(x = Age, y = prop, fill = Millennial)) +
```

```
  geom_col() +
  theme_minimal() +
  ggtitle("Propotion of People Leaving (By Age group)") +
  xlab("Age") +
  ylab("Proportion") +
  scale_fill_manual(values = c("yes"="tomato", "no"="gray"))
```




```
colnames(df)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "Education"                "EnvironmentSatisfaction"
##  [5] "JobInvolvement"           "JobLevel"
##  [7] "JobSatisfaction"          "MonthlyIncome"
##  [9] "NumCompaniesWorked"       "OverTime"
## [11] "PercentSalaryHike"        "RelationshipSatisfaction"
## [13] "WorkLifeBalance"          "gen"
```