Project 2

Justin Williams

2024-10-12

Packages, data, etc.

```
#import libraries and data
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggplot2)
hrdata_df <- read.csv("/Users/justinwilliams/Code/9050advresearch/Project 4/HRData.csv")
```

1.Report the full covariance and correlation matrix for the following variables: GenderID, PerfScoreID, Salary, Age, EngagementSurvey, EmpSatisfaction, and Absences.

```
#select the columns

selected_columns <- hrdata_df[, c("GenderID", "PerfScoreID", "Salary", "Age", "EngagementSurvey", "EmpSa
```

```r
#covariance matrix
cov_matrix <- cov(selected_columns)
print("Covariance Matrix:")
```

```
## [1] "Covariance Matrix:"
```

```r
print(cov_matrix)
```

```
##                       GenderID   PerfScoreID       Salary           Age
## GenderID           0.24644746   -0.01600456 7.005695e+02 -5.871715e-02
## PerfScoreID       -0.01600456    0.34465304 1.933267e+03  4.219337e-01
## Salary           700.56948449 1933.26707810 6.328564e+08  2.116953e+04
## Age               -0.05871715    0.42193372 2.116953e+04  7.910491e+01
## EngagementSurvey  -0.01422581    0.25270968 1.291017e+03  4.450113e-01
## EmpSatisfaction   -0.02013277    0.16204751 1.434585e+03 -5.131745e-01
## Absences          -0.01329738    0.16021160 1.212926e+04 -1.905516e+00
##                   EngagementSurvey EmpSatisfaction      Absences
## GenderID              -0.01422581     -0.02013277 -1.329738e-02
## PerfScoreID            0.25270968      0.16204751  1.602116e-01
## Salary              1291.01690323   1434.58479411  1.212926e+04
## Age                    0.44501129     -0.51317446 -1.905516e+00
## EngagementSurvey       0.62400129      0.13438710 -4.054839e-02
## EmpSatisfaction        0.13438710      0.82671922  4.002904e-01
## Absences              -0.04054839      0.40029043  3.425288e+01
```

```r
#correlation matrix
cor_matrix <- cor(selected_columns)
print("Correlation Matrix:")
```

```
## [1] "Correlation Matrix:"
```

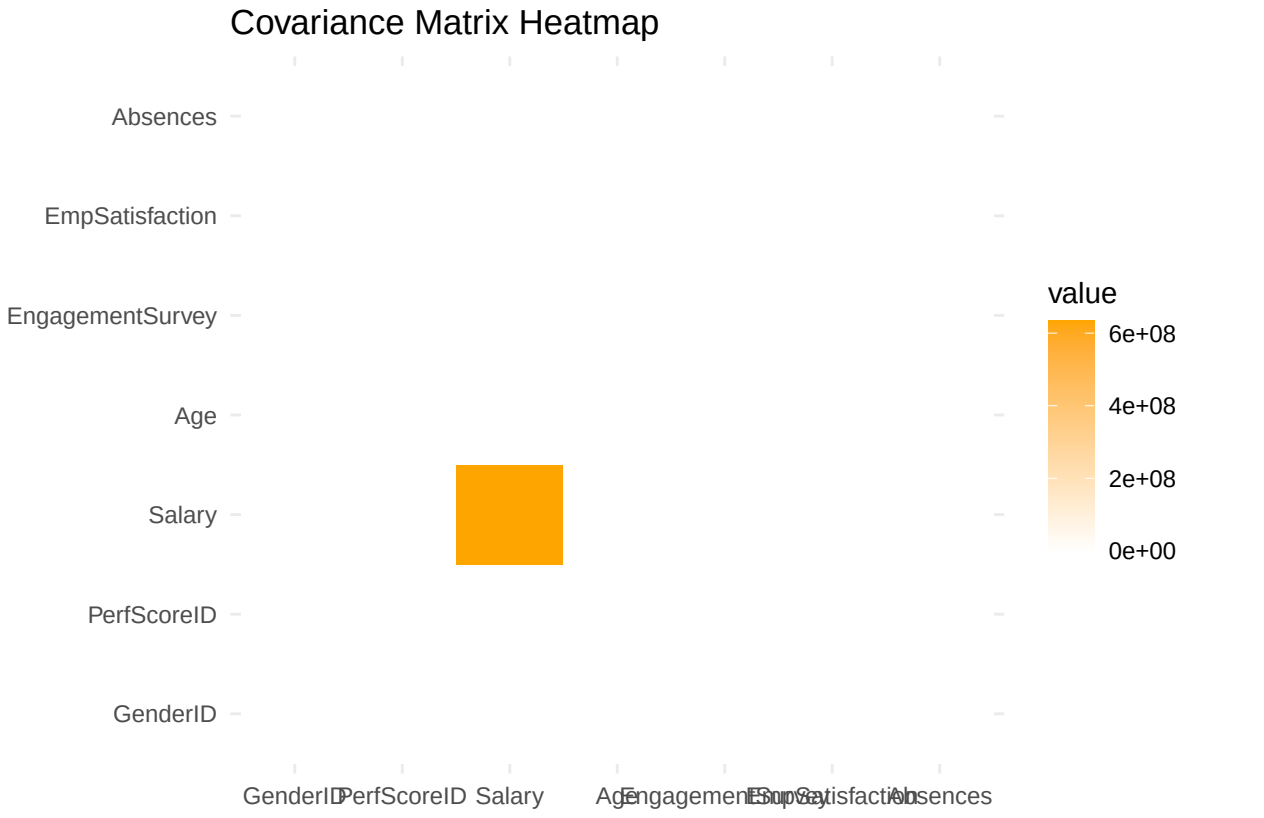
```r
print(cor_matrix)
```

```
##                        GenderID PerfScoreID     Salary         Age
## GenderID           1.000000000 -0.05491495 0.05609659 -0.01329845
## PerfScoreID       -0.054914952  1.00000000 0.13090258  0.08080746
## Salary             0.056096589  0.13090258 1.00000000  0.09461433
## Age               -0.013298450  0.08080746 0.09461433  1.00000000
## EngagementSurvey -0.036276216  0.54492668 0.06496607  0.06333978
## EmpSatisfaction  -0.044602814  0.30357938 0.06271835 -0.06345770
## Absences         -0.004576729  0.04662881 0.08238216 -0.03660685
##                   EngagementSurvey EmpSatisfaction     Absences
## GenderID              -0.036276216     -0.04460281 -0.004576729
## PerfScoreID            0.544926678      0.30357938  0.046628812
## Salary                 0.064966071      0.06271835  0.082382156
## Age                    0.063339782     -0.06345770 -0.036606847
## EngagementSurvey       1.000000000      0.18710518 -0.008770661
## EmpSatisfaction        0.187105180      1.00000000  0.075222480
## Absences              -0.008770661      0.07522248  1.000000000
```

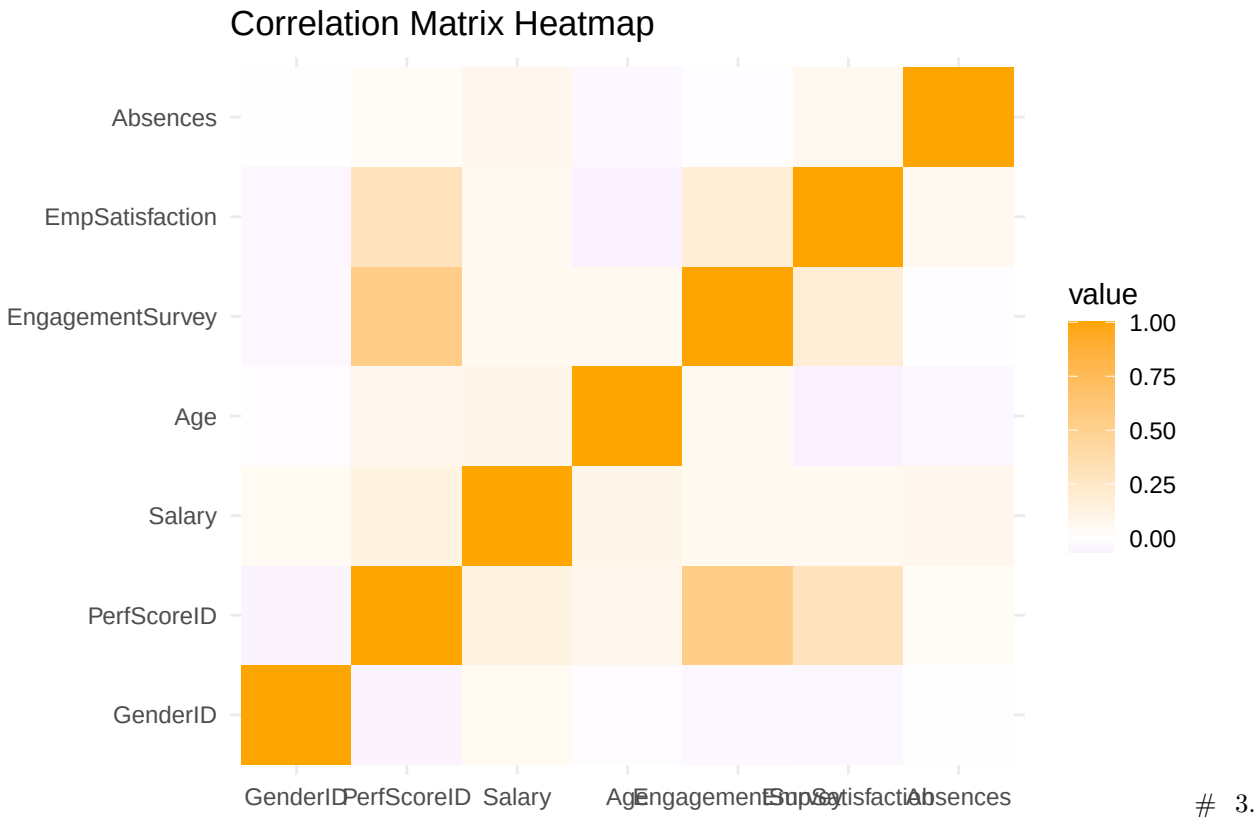
2. Visualize the correlations you have computed. There are a few different ways to do this (including scatterplots, heat maps, etc.). Use the visualization method that you think best conveys the relations.

```
#converting matrices to long format for ggplot
cov_melt <- melt(cov_matrix)
cor_melt <- melt(cor_matrix)

#heatmap for covariance matrix
ggplot(data = cov_melt, aes(Var1, Var2, fill = value)) +
    geom_tile() +
    scale_fill_gradient2(low = "purple", high = "orange", mid = "white", midpoint = 0) +
    theme_minimal() +
    labs(title = "Covariance Matrix Heatmap", x = "", y = "")
```




```
# heatmap for correlation matrix
ggplot(data = cor_melt, aes(Var1, Var2, fill = value)) +
    geom_tile() +
    scale_fill_gradient2(low = "purple", high = "orange", mid = "white", midpoint = 0) +
    theme_minimal() +
    labs(title = "Correlation Matrix Heatmap", x = "", y = "")
```



3. Provide a summary of what you see in these results. Explain the nature of the bivariate relations (or lack thereof) in these data. Write this summary like you might see reported in a journal article.

Create a subset of the data in which you only include employees who are from the Production Department. Also, create a subset of everyone NOT in Production.

4. Run correlations separately for each of these two subsamples. Report the results for each and provide an interpretation of what you see as you did in Questions 2 and 3. Are the correlations the same across the subsamples and/or from the entire sample? I'm not asking for a test of statistical significance, but your own interpretation based on the values you compute.