DATA607_Project3_Wrangling

Josh Iden

2022-10-20

DATA WRANGLING

```
# load packages
library(tidyverse)
library(readxl)
library(stringr)
library(lubridate)
```

We stored the original multi-sheet .xlsx file from Kaggle in the project GitHub repo. Due to the structure of the file, we extracted the individual sheets locally, which we then wrote to CSVs and stored to the project repo for database ingestion:

```
# read file locally
file <- "/Users/joshiden/Documents/Classes/CUNY SPS/Fall 2022/DATA 607/Projects/Project 3/Data Science C
excel <- read_excel(file)

# store sheet names
sheets <- excel_sheets(file)

# read sheets into dataframes
ds_skills <- read_excel(file, sheet = sheets[1])
ds_software <- read_excel(file, sheet = sheets[2])
education <- read_excel(file, sheet=sheets[7])
```

Previewing the output dataframes,

```
head(ds_skills)
```

```
## # A tibble: 6 x 5
##   Keyword          LinkedIn Indeed SimplyHired Monster
##   <chr>               <dbl>  <dbl>       <dbl>   <dbl>
## 1 machine learning     5701   3439        2561    2340
## 2 analysis             5168   3500        2668    3306
## 3 statistics           4893   2992        2308    2399
## 4 computer science     4517   2739        2093    1900
## 5 communication        3404   2344        1791    2053
## 6 mathematics          2605   1961        1497    1815
```

```
head(ds_software)
```

```
## # A tibble: 6 x 12
##   Keyword LinkedIn Indeed SimplyHired Monster ‘LinkedIn %‘ ‘Indeed %‘
##   <chr>      <dbl>  <dbl>       <dbl>   <dbl>        <dbl>      <dbl>
## 1 Python      6347   3818        2888    2544        0.737      0.743
## 2 R           4553   3106        2393    2365        0.529      0.605
## 3 SQL         3879   2628        2056    1841        0.451      0.511
## 4 Spark       2169   1551        1167    1062        0.252      0.302
## 5 Hadoop      2142   1578        1164    1200        0.249      0.307
## 6 Java        1944   1377        1059    1002        0.226      0.268
## # ... with 5 more variables: ‘SimplyHired %‘ <dbl>, ‘Monster %‘ <dbl>,
## #   ‘Avg %‘ <dbl>, ‘GlassDoor Self Reported % 2017‘ <dbl>, Difference <dbl>
```

```
head(education)
```

```
## # A tibble: 6 x 6
##   Keyword   LinkedIn Monster Indeed SimplyHired AngelList
##   <chr>        <dbl>   <dbl>  <dbl>       <dbl>     <dbl>
## 1 MS            2917    1821   2532        1985       288
## 2 PhD           3242    1468   2221        1629       230
## 3 masters       1568     887   2704        2033       165
## 4 bachelors      677     578   3326        2631        97
## 5 MBA           1186     675    788         634        63
## 6 bootcamp        31      14    129          74        18
```

Tidy technical and soft skills for category output tables. **Keyword** column contains skill names from first row until cell value **Total**.

```
# Find row index of Total
totalIdx <- which(ds_skills$Keyword == "Total")

skills <- ds_skills |>
          # Grab beginning rows until and excluding Total entry
          slice_head(n = totalIdx - 1) |>
          # select Keyword column
          select(Keyword) |>
          # drop NA
          filter(! is.na(Keyword)) |>
          # uppercase
          mutate(Keyword = str_to_upper(Keyword)) |>
          # add T_GENERAL and T_SOFT categories
          mutate(Category=ifelse(Keyword == "COMMUNICATION", "T_SOFT",
                   ifelse(Keyword == "PROJECT MANAGEMENT", "T_SOFT", "T_GENERAL")))
nrow(skills)
```

```
## [1] 15
```

```
tail(skills)
```

```
## # A tibble: 6 x 2
##   Keyword              Category
##   <chr>                <chr>
## 1 NLP COMPOSITE        T_GENERAL
## 2 SOFTWARE DEVELOPMENT T_GENERAL
## 3 NEURAL NETWORKS      T_GENERAL
## 4 DATA ENGINEERING     T_GENERAL
## 5 PROJECT MANAGEMENT   T_SOFT
## 6 SOFTWARE ENGINEERING T_GENERAL
```

```
totalIdx <- which(ds_software$Keyword == "Total")

software <- ds_software |>
           # Grab beginning rows until and excluding Total entry
           slice_head(n = totalIdx - 1) |>
           # select Keyword column
           select(Keyword) |>
           # drop NA
           filter(! is.na(Keyword)) |>
           # uppercase
           mutate(Keyword = str_to_upper(Keyword)) |>
           # add T_SOFTWARE category
           mutate(Category = "T_SOFTWARE")
nrow(software)
```

```
## [1] 37
```

```
tail(software)
```

```
## # A tibble: 6 x 2
##   Keyword   Category
##   <chr>     <chr>
## 1 GIT       T_SOFTWARE
## 2 MYSQL     T_SOFTWARE
## 3 MONGODB   T_SOFTWARE
## 4 CASSANDRA T_SOFTWARE
## 5 PYTORCH   T_SOFTWARE
## 6 CAFFE     T_SOFTWARE
```

Applying transformations to standardize and retain the desired data for data tables:

```
# keep only first 15 rows
# Keyword to upper
# pivot columns to column: source
ds_skills_transformed <- ds_skills |>
  head(15) |>
  mutate(Keyword = toupper(Keyword)) |>
  rename(KEYWORD = Keyword) |>
  pivot_longer(cols=("LinkedIn":"Monster"), names_to="SOURCE", values_to="COUNT") |>
  mutate(SOURCE = toupper(SOURCE), SURVEY_DATE=ymd("2018-06-15")) |>
  arrange(KEYWORD,SOURCE)

ds_skills_transformed
```

```
## # A tibble: 60 x 4
##    KEYWORD       SOURCE      COUNT SURVEY_DATE
##    <chr>         <chr>       <dbl> <date>
##  1 AI COMPOSITE  INDEED       1125 2018-06-15
##  2 AI COMPOSITE  LINKEDIN     1568 2018-06-15
##  3 AI COMPOSITE  MONSTER       687 2018-06-15
##  4 AI COMPOSITE  SIMPLYHIRED   811 2018-06-15
##  5 ANALYSIS      INDEED       3500 2018-06-15
##  6 ANALYSIS      LINKEDIN     5168 2018-06-15
##  7 ANALYSIS      MONSTER      3306 2018-06-15
##  8 ANALYSIS      SIMPLYHIRED  2668 2018-06-15
##  9 COMMUNICATION INDEED       2344 2018-06-15
## 10 COMMUNICATION LINKEDIN     3404 2018-06-15
## # ... with 50 more rows
```

```
# keep top 37 rows
# keyword to upper
# pivot columns to source
# source column to upper
# add date column
ds_software_transformed <- ds_software |>
  select(c("Keyword":"Monster")) |>
  head(37) |>
  mutate(Keyword = toupper(Keyword)) |>
  rename(KEYWORD = Keyword) |>
  pivot_longer(cols=("LinkedIn":"Monster"), names_to="SOURCE", values_to="COUNT") |>
  mutate(SOURCE = toupper(SOURCE), SURVEY_DATE=ymd("2018-06-15")) |>
  arrange(KEYWORD,SOURCE)

ds_software_transformed
```

```
## # A tibble: 148 x 4
##    KEYWORD SOURCE      COUNT SURVEY_DATE
##    <chr>   <chr>       <dbl> <date>
##  1 AWS     INDEED        791 2018-06-15
##  2 AWS     LINKEDIN      947 2018-06-15
##  3 AWS     MONSTER       467 2018-06-15
##  4 AWS     SIMPLYHIRED   607 2018-06-15
##  5 AZURE   INDEED        416 2018-06-15
##  6 AZURE   LINKEDIN      578 2018-06-15
##  7 AZURE   MONSTER       272 2018-06-15
##  8 AZURE   SIMPLYHIRED   285 2018-06-15
##  9 C       INDEED        492 2018-06-15
## 10 C       LINKEDIN      795 2018-06-15
## # ... with 138 more rows
```

```
# keyword to uppercase
# pivot columns to source
# source column to uppercase
# add date column
# drop AngelList column
# drop NA values
education_transformed <- education |>
```

```
  mutate(Keyword = toupper(Keyword)) |>
  rename(KEYWORD = Keyword) |>
  pivot_longer(cols=("LinkedIn":"SimplyHired"), names_to="SOURCE", values_to="COUNT") |>
  mutate(SOURCE = toupper(SOURCE), SURVEY_DATE=ymd("2018-06-15")) |>
  subset(select = -c(AngelList)) |>
  drop_na() |>
  arrange(KEYWORD,SOURCE)

education_transformed
```

```
## # A tibble: 28 x 4
##    KEYWORD   SOURCE      COUNT SURVEY_DATE
##    <chr>     <chr>       <dbl> <date>
##  1 BACHELORS INDEED       3326 2018-06-15
##  2 BACHELORS LINKEDIN      677 2018-06-15
##  3 BACHELORS MONSTER       578 2018-06-15
##  4 BACHELORS SIMPLYHIRED  2631 2018-06-15
##  5 BOOTCAMP  INDEED        129 2018-06-15
##  6 BOOTCAMP  LINKEDIN       31 2018-06-15
##  7 BOOTCAMP  MONSTER        14 2018-06-15
##  8 BOOTCAMP  SIMPLYHIRED    74 2018-06-15
##  9 KAGGLE    INDEED         49 2018-06-15
## 10 KAGGLE    LINKEDIN       74 2018-06-15
## # ... with 18 more rows
```

Combining ds_skills_transformed and ds_software_transformed -> skills_in_demand

```
#skills in demand
skills_in_demand <- rbind(ds_skills_transformed,ds_software_transformed)
skills_in_demand
```

```
## # A tibble: 208 x 4
##    KEYWORD       SOURCE      COUNT SURVEY_DATE
##    <chr>         <chr>       <dbl> <date>
##  1 AI COMPOSITE  INDEED       1125 2018-06-15
##  2 AI COMPOSITE  LINKEDIN     1568 2018-06-15
##  3 AI COMPOSITE  MONSTER       687 2018-06-15
##  4 AI COMPOSITE  SIMPLYHIRED   811 2018-06-15
##  5 ANALYSIS      INDEED       3500 2018-06-15
##  6 ANALYSIS      LINKEDIN     5168 2018-06-15
##  7 ANALYSIS      MONSTER      3306 2018-06-15
##  8 ANALYSIS      SIMPLYHIRED  2668 2018-06-15
##  9 COMMUNICATION INDEED       2344 2018-06-15
## 10 COMMUNICATION LINKEDIN     3404 2018-06-15
## # ... with 198 more rows
```

The files were then written to CSV and added to the project repository for database ingestion.

```
# write to software_skills.csv
# outputDir <- "/Users/jawaidhakim/Downloads/"
# outputFile <- str_c(inputDir, "software_skills.csv")
# write.csv(software, outputFile)
# write to general_skills.csv
# outputFile <- str_c(inputDir, "general_skills.csv")
# write.csv(skills, outputFile)

# write to skills_in_demand.csv
# sid <- "/Users/joshiden/Documents/Classes/CUNY SPS/Fall 2022/DATA 607/Projects/skills_in_demand.csv"
# write.csv(skills_in_demand, sid)

# write to education_in_demand.csv
# eid <- "/Users/Melissa/OneDrive/Documents/CUNY/Fall 2022/Data 607/Project 3/education_in_demand.csv"
# write.csv(education_transformed, eid)
```