Identifying salary ranges for certain jobs

true

2019-10-29

Program directory

This directory contains all programs necessary to run the cleaning, analysis, etc. They can be run separately, or in a single run.

Setup

Most parameters are set in the `config.R`:

```
source(file.path(rprojroot::find_rstudio_root_file(),"pathconfig.R"),echo=TRUE)
```

```
##
## > basepath <- rprojroot::find_rstudio_root_file()
##
## > acquired <- file.path(basepath, "data", "acquired")
##
## > interwrk <- file.path(basepath, "data", "interwrk")
##
## > generated <- file.path(basepath, "data", "generated")
##
## > outputs <- file.path(basepath, "analysis")
##
## > programs <- file.path(basepath, "programs")
##
## > for (dir in list(acquired, interwrk, generated, outputs)) {
## +     if (file.exists(dir)) {
## +     }
## +     else {
## +         dir.create(file.path(dir))
##  .... [TRUNCATED]
```

```
source(file.path(programs,"config.R"), echo=TRUE)
```

```
##
## > source(file.path(programs, "global-libraries.R"),
## +     echo = FALSE)
```

```
## Loading required package: 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
## Loading required package: devtools
```

```
## Loading required package: usethis
```

```
## Loading required package: rprojroot
```

```
## Loading required package: tictoc
```

```
##
## > source(file.path(programs, "libraries.R"), echo = FALSE)
```

```
## Loading required package: tidyr
```

```
## Loading required package: readxl
```

```
## Loading required package: fuzzyjoin
```

```
## Loading required package: knitr
```

```
##
## > onet.src.base <- "https://www.onetcenter.org/dl_files/database/"
##
## > onet.src.version <- "23_2"
##
## > onet.src.file <- paste("db", onet.src.version, "excel.zip",
## +     sep = "_")
##
## > oes.src.base <- "https://www.bls.gov/oes/special.requests/"
##
## > oes.src.version <- "18"
##
## > oes.src.file <- paste("oesm", oes.src.version, "nat.zip",
## +     sep = "")
##
## > soc_definitions_loc <- "https://www.bls.gov/soc/soc_structure_2010.xls"
```

Note that the path `interwrk` is transitory, and is only kept during processing. It will be empty in the replication archive.

Download and unpack the O*NET data

- Input data: O*NET
- Output data: path 'acquired'
- this part is only run on-demand (manually)

```
source(file.path(programs,"01_download_onet.R"),echo=TRUE)
```

Download SOC data

We can download definitions from https://www.bls.gov/soc/soc_structure_2010.xls. In particular, we might be interested in the major groupings. (not done yet)

Mapping job titles to SOC

We merge the alternate titles (as defined by O*Net) and the BLS data (merged in via occupational code). We keep all observations in our normative list of NLM-state related occupations.

```
##
## > source(file.path(rprojroot::find_rstudio_root_file(),
## +     "pathconfig.R"), echo = FALSE)
##
## > source(file.path(programs, "config.R"), echo = FALSE)
##
## > Alternate_Titles <- read_excel(file.path(acquired,
## +     "Alternate Titles.xlsx"))
##
## > Occupation_Data <- read_excel(file.path(acquired,
## +     "Occupation Data.xlsx")) %>% select(-Title)
##
## > BLS.data <- read_excel(file.path(acquired, paste0("national_M20",
## +     oes.src.version, "_dl.xlsx")))
##
## > job_titles.main <- read_excel(file.path(generated,
## +     "job_titles.xlsx"), sheet = "Main", col_names = TRUE)
##
## > job_titles.plus <- read_excel(file.path(generated,
## +     "job_titles.xlsx"), sheet = "Inclusions", col_names = TRUE) %>%
## +     rename(`Job Title`  .... [TRUNCATED]
##
## > job_titles <- bind_rows(job_titles.main, job_titles.plus %>%
## +     select("Alternate Title") %>% rename(`Job Title` = "Alternate Title")) %>%
## +    .... [TRUNCATED]
##
## > job_titles.minus <- read_excel(file.path(generated,
## +     "job_titles.xlsx"), sheet = "Exclusions", col_names = TRUE) %>%
## +     rename(`Job Title` .... [TRUNCATED]
##
## > soc_job_titles <- Alternate_Titles %>% select("O*NET-SOC Code",
## +     Title) %>% distinct()
##
## > soc_job_alttitles <- Alternate_Titles %>% select("O*NET-SOC Code",
## +     "Alternate Title") %>% distinct()
##
## > primary <- stringdist_inner_join(y = soc_job_titles,
## +     x = job_titles %>% select("Job Title"), by = c(`Job Title` = "Title"),
## +     method = " ..." ... [TRUNCATED]
##
## > secondary.raw <- stringdist_inner_join(y = soc_job_alttitles,
## +     x = job_titles %>% select("Job Title"), by = c(`Job Title` = "Alternate Title") .... [TRUNCATE
##
## > secondary <- secondary.raw
##
## > nlm.titles.raw <- bind_rows(primary, secondary) %>%
## +     left_join(Occupation_Data, by = "O*NET-SOC Code") %>% separate("O*NET-SOC Code",
## +     s .... [TRUNCATED]
##
## > nlm.titles.raw2 <- nlm.titles.raw %>% left_join(job_titles.plus %>%
## +     select(-`O*NET-SOC Code`, -SOC), by = c(`Job Title` = "Alternate Title")) .... [TRUNCATED]
##
## > nlm.titles <- anti_join(nlm.titles.raw2, job_titles.minus %>%
## +     select(SOC), by = "SOC")
```

```
##
## > saveRDS(nlm.titles, file = file.path(outputs, "nlm.titles.RDS"))
##
## > write.csv(nlm.titles, file = file.path(outputs, "nlm.titles.csv"))
```

Note: we have added a few "alternate titles" where we believe that the O*Net descriptions do not capture the right name, and we have explicitly removed a few SOC codes because we believe they do not apply in this context. These are listed in the Appendix.

Results

The following table lists the annual salaries by job title (median, and the 25% and 75% percentile). Blank salaries ("NA") indicate that no occupation code could be found on O*Net based on the normative description. We only print one line per normative job title - these might map to the same occupation code (SOC).

```
nlm.titles <- readRDS(file.path(outputs,"nlm.titles.RDS"))
nlm.extract <- nlm.titles %>%
  distinct(`Job Title`,SOC,.keep_all = TRUE) %>%
  select("Job Title","Title","SOC", "Alternate Title","A_PCT25","A_MEDIAN","A_PCT75")
kable(nlm.extract)
```

Job Title	Title	SC
Researcher	Industrial Ecologists	19-
Researcher	Anthropologists	19-
Researcher	Historians	19-
Researcher	Biofuels/Biodiesel Technology and Product Development Managers	11-
Researcher	Mathematicians	15-
Researcher	Chemical Engineers	17-
Researcher	Nanosystems Engineers	17-
Researcher	Manufacturing Engineering Technologists	17-
Researcher	Biologists	19-
Researcher	Biochemists and Biophysicists	19-
Researcher	Bioinformatics Scientists	19-
Researcher	Medical Scientists, Except Epidemiologists	19-
Researcher	Chemists	19-
Researcher	Hydrologists	19-
Researcher	Remote Sensing Scientists and Technologists	19-
Researcher	Geographers	19-
Data Librarian	Librarians	25-
Data Librarian	Library Science Teachers, Postsecondary	25-
Data Librarian	Archivists	25-
Metadata Librarian	Librarians	25-
Metadata Librarian	Library Science Teachers, Postsecondary	25-
Metadata Librarian	Archivists	25-
Records Management Specialist	Librarians	25-
Records Management Specialist	Library Science Teachers, Postsecondary	25-
Records Management Specialist	Archivists	25-
Curator	Curators	25-
Curator	Archivists	25-
Curator	Archeologists	19-
Research Domain Curator	Biofuels/Biodiesel Technology and Product Development Managers	11-
Research Domain Curator	Mathematicians	15-
Research Domain Curator	Chemical Engineers	17-
Research Domain Curator	Nanosystems Engineers	17-

Job Title	Title	SC
Research Domain Curator	Manufacturing Engineering Technologists	17-
Research Domain Curator	Biologists	19-
Research Domain Curator	Biochemists and Biophysicists	19-
Research Domain Curator	Bioinformatics Scientists	19-
Research Domain Curator	Medical Scientists, Except Epidemiologists	19-
Research Domain Curator	Chemists	19-
Research Domain Curator	Climate Change Analysts	19-
Research Domain Curator	Hydrologists	19-
Research Domain Curator	Remote Sensing Scientists and Technologists	19-
Research Domain Curator	Anthropologists	19-
Research Domain Curator	Geographers	19-
Research Domain Project Manager	Biofuels/Biodiesel Technology and Product Development Managers	11-
Research Domain Project Manager	Mathematicians	15-
Research Domain Project Manager	Chemical Engineers	17-
Research Domain Project Manager	Nanosystems Engineers	17-
Research Domain Project Manager	Manufacturing Engineering Technologists	17-
Research Domain Project Manager	Biologists	19-
Research Domain Project Manager	Biochemists and Biophysicists	19-
Research Domain Project Manager	Bioinformatics Scientists	19-
Research Domain Project Manager	Medical Scientists, Except Epidemiologists	19-
Research Domain Project Manager	Chemists	19-
Research Domain Project Manager	Climate Change Analysts	19-
Research Domain Project Manager	Hydrologists	19-
Research Domain Project Manager	Remote Sensing Scientists and Technologists	19-
Research Domain Project Manager	Anthropologists	19-
Research Domain Project Manager	Geographers	19-
Informatician	Computer Systems Analysts	15-
Informatician	Information Technology Project Managers	15-
Data Wrangler	Information Technology Project Managers	15-
Education Specialist	Health Educators	21-
Education Specialist	Special Education Teachers, Secondary School	25-
Education Specialist	Instructional Coordinators	25-
Communication Specialist	Public Relations Specialists	27-
Software Engineer	Computer and Information Research Scientists	15-
Software Engineer	Software Developers, Applications	15-
Software Engineer	Software Developers, Systems Software	15-
IT Security Specialist	Security Management Specialists	13-
IT Systems Engineer	Computer and Information Systems Managers	11-
IT Systems Engineer	Information Technology Project Managers	15-
IT Project Manager	Computer and Information Systems Managers	11-
IT Project Manager	Information Technology Project Managers	15-
Project Manager	Construction Managers	11-
Project Manager	Architectural and Engineering Managers	11-
Project Manager	Managers, All Other	11-
Project Manager	Information Technology Project Managers	15-
Project Manager	Environmental Engineers	17-
Project Manager	Wind Energy Engineers	17-
Project Manager	Environmental Restoration Planners	19-
Project Manager	Social Science Research Assistants	19-
Project Manager	Remote Sensing Technicians	19-
Project Manager	Technical Directors/Managers	27-
Project Manager	Intelligence Analysts	33-

Job Title	Title	SC
Senior Staff	NA	NA
Policy Specialist	NA	NA
Administrative Staff	First-Line Supervisors of Office and Administrative Support Workers	43-
Administrative Staff	Executive Secretaries and Executive Administrative Assistants	43-
Administrative Staff	Secretaries and Administrative Assistants, Except Legal, Medical, and Executive	43-
Administrative Staff	Business Operations Specialists, All Other	13-
Administrative Staff	Billing and Posting Clerks	43-
Administrative Staff	New Accounts Clerks	43-
Administrative Staff	Medical Secretaries	43-
Facilities Manager	General and Operations Managers	11-
Facilities Manager	Administrative Services Managers	11-
Facilities Manager	Property, Real Estate, and Community Association Managers	11-
Facilities Manager	First-Line Supervisors of Housekeeping and Janitorial Workers	37-
Facilities Manager	First-Line Supervisors of Office and Administrative Support Workers	43-
Facilities Manager	First-Line Supervisors of Mechanics, Installers, and Repairers	49-
Facilities Manager	Maintenance and Repair Workers, General	49-
Data Scientist	Computer and Information Research Scientists	15-

```r
saveRDS(nlm.extract,file = file.path(outputs,"nlm.extract.RDS"))
write.csv(nlm.extract,file=file.path(outputs,"nlm.extract.csv"))
```

We collapse the raw data into the minimum “PCT25” number, the median “MEDIAN” number, and the maximum “PCT75” number to get a range:

```r
nlm.collapsed <- nlm.extract  %>% group_by(`Job Title`) %>%
  summarise(PCT25 = min(as.numeric(A_PCT25),na.rm = TRUE),
            MEDIAN = median(as.numeric(A_MEDIAN),na.rm = TRUE),
            PCT75=max(as.numeric(A_PCT75),na.rm = TRUE)
            )
```

```
## Warning in min(as.numeric(A_PCT25), na.rm = TRUE): no non-missing arguments
## to min; returning Inf

## Warning in min(as.numeric(A_PCT25), na.rm = TRUE): no non-missing arguments
## to min; returning Inf

## Warning in max(as.numeric(A_PCT75), na.rm = TRUE): no non-missing arguments
## to max; returning -Inf

## Warning in max(as.numeric(A_PCT75), na.rm = TRUE): no non-missing arguments
## to max; returning -Inf
```

```r
kable(nlm.collapsed)
```

Job Title	PCT25	MEDIAN	PCT75
Administrative Staff	28930	37800	94890
Communication Specialist	44490	60000	81550
Curator	38090	53780	80230
Data Librarian	38090	59050	90550
Data Scientist	91650	118370	149470
Data Wrangler	66410	90270	117070
Education Specialist	39800	60600	82860
Facilities Manager	29560	58340	157120

Job Title	PCT25	MEDIAN	PCT75
Informatician	66410	89505	117070
IT Project Manager	66410	116400	180190
IT Security Specialist	52200	70530	94890
IT Systems Engineer	66410	116400	180190
Metadata Librarian	38090	59050	90550
Policy Specialist	Inf	NA	-Inf
Project Manager	35450	87620	173180
Records Management Specialist	38090	59050	90550
Research Domain Curator	47500	80300	173180
Research Domain Project Manager	47500	80300	173180
Researcher	40670	79945	173180
Senior Staff	Inf	NA	-Inf
Software Engineer	79340	110000	149470

```r
saveRDS(nlm.collapsed,file = file.path(outputs,"nlm.collapsed.RDS"))
write.csv(nlm.collapsed,file=file.path(outputs,"nlm.collapsed.csv"))
```

We crosscheck that the categories are right, by looking at the median median in each category:

```r
nlm.categories <- nlm.titles %>%
  distinct(`Job Title`,SOC,.keep_all = TRUE) %>%
  group_by(`Relative Salary`) %>%
  mutate(Missing=is.na(A_MEDIAN)) %>%
  summarise(PCT25 = min(as.numeric(A_PCT25),na.rm = TRUE),
            MEDIAN = median(as.numeric(A_MEDIAN),na.rm = TRUE),
            PCT75=max(as.numeric(A_PCT75),na.rm = TRUE),
            N=n(),Missing=sum(Missing)) %>%
  arrange(MEDIAN)
kable(nlm.categories)
```

Relative Salary	PCT25	MEDIAN	PCT75	N	Missing
L	28930	37800	94890	7	0
M	29560	61505	173180	34	0
H	40670	80300	180190	50	1
VH	52200	103620	180190	10	1

```r
saveRDS(nlm.categories,file = file.path(outputs,"nlm.categories.RDS"))
write.csv(nlm.extract,file=file.path(outputs,"nlm.categories.csv"))
```

Appendix

The full table (corresponds to Table x in report)

Job Title	Definition
Researcher	An individual who in the course of conducting research generates potentially shareable
Data Librarian	An individual who is trained in the technical aspects of data management
Metadata Librarian	An individual who is trained in the technical aspects of data standards
Records Management Specialist	An individual, often an archivist, who is trained in managing data throughout the dat

Job Title	Definition
Curator	An individual, often an archivist, who is trained in methods to describe and add value
Research Domain Curator	An individual who is trained in methods to describe and add value to data, and who i
Research Domain Project Manager	An individual who has the responsibility to plan, execute, and oversee a project, and v
Informatician	An individual who is trained in biology, medicine, or other health-related field and wh
Data Wrangler	An individual who is trained in methods for transforming data from one format into a
Education Specialist	An individual who is trained in the design and implementation of training materials re
Communication Specialist	An individual who is trained in effective methods for publicizing and disseminating inf
Software Engineer	An individual who is trained in the design, implementation, testing, and evaluation of
IT Security Specialist	An individual who is trained in methods to protect information technology systems ag
IT Systems Engineer	An individual who is trained in the implementation, monitoring, and maintenance of i
IT Project Manager	An individual who has the responsibility to plan, execute, and oversee a project, and v
Project Manager	An individual who has the responsibility to plan, execute, and oversee a project
Senior Staff	An individual who has a supervisory and decision-making role within an organization
Policy Specialist	An individual who is trained in relevant ethical, legal and regulatory requirements
Administrative Staff	An individual who provides a variety of support functions for a project or program
Facilities Manager	An individual who oversees and handles matters relating to the physical environment
Data Scientist	NA

The list of force-included occupations

Job Title	O*NET-SOC Code	SOC	Alternate Title
Scientist	19-1041.00	19-1041	Epidemiologist
Data Librarian	25-4021.00	25-4021	Librarian
Metadata Librarian	25-4021.00	25-4021	Librarian
Records Management Specialist	25-4021.00	25-4021	Librarian
Research Domain Curator	NA.00	NA	Scientist
Research Domain Project Manager	NA.00	NA	Scientist
Curator	NA.00	NA	Archivist
IT Project Manager	11-3021.00	11-3021	Computer and Information Systems Managers
Researcher	NA.00	NA	Scientist
IT Security Specialist	NA.00	NA	Security Management Specialists
IT Security Specialist	NA.00	NA	IT Security Analyst
Data Wrangler	NA.00	NA	IT Specialist
Informatician	NA.00	NA	IT Specialist
Informatician	15-1121.00	15-1121	Computer Systems Analysts
IT Systems Engineer	NA.00	NA	IT Specialist
IT Systems Engineer	11-3021.00	11-3021	Computer and Information Systems Managers
Administrative Staff	NA.00	NA	Office and Administrative Support Workers
Administrative Staff	NA.00	NA	Administrative Assistant
Administrative Staff	NA.00	NA	First-Line Supervisors of Office and Administrative
Administrative Staff	NA.00	NA	Executive Secretaries and Executive Administrative
Administrative Staff	43-6014.00.00	43-6014.00	Secretaries and Administrative Assistants, Except

The list of excluded occupations

Job Title	SOC
Nuclear Engineers	17-2161
Astronomers	19-2011

Job Title	SOC
Physicists	19-2012
Architectural Drafters	17-3011
File Clerks	43-4071
First-Line Supervisors of Construction Trades and Extraction Workers	47-1011
Travel Guides	39-7012
Park Naturalists	19-1031
Adapted Physical Education Specialists	25-2059
Tour Guides and Escorts	39-7011
Speech-Language Pathologists	29-1127
Switchboard Operators, Including Answering Service	43-2011