Salaries Projection

Melissa Nooney

2024-07-24

```
library(tinytex)
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(ggplot2)
library(stringr)
library(dplyr)
#loading all libraries I intend to use
```

```
#file.choose()
salary_data <- read.csv("C:\\Users\\mnoon\\OneDrive\\Desktop\\R and Python Programming\\R project\\r pro
#bringing data frame into my environment
```

General data wrangling and initial analysis

```
US_based_salaries <- filter(salary_data, company_location == "US")

International_Salaries <- filter(salary_data, company_location != "US")

# Since the company is, I'm assuming US based, I want to look at just US based
#salaries. SO I am filtering in just what I need, but I want to keep the
#international just for fun. Filtering in this way will make my future
#wrangling a little easier..hopefully.
#The hire can work offshore, but the company is interested in US rates.
```

```
summary(US_based_salaries$salary_in_usd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5679  100000  135000  144055  170000  600000
```

```
summary(International_Salaries$salary_in_usd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   39263   62689   67560   87932  260000
```

```
which(US_based_salaries$salary_in_usd == 5679)
```

```
## [1] 88
```

```
which(International_Salaries$salary_in_usd == 260000)
```

```
## [1] 2
```

```
# the which function came in handy because there were a few weird numbers
#that popped up, and helped in rectify my code, and/or told a deeper story of
#where and why for particular salary.
#want to get an overall picture here of average salaries from 2020-2022.
#As I get further into data, I think I will focus on just 2022, so that I can
#project inflation based on most current salaries.
mean(US_based_salaries$salary_in_usd, trim = .10)
```

```
## [1] 137711.4
```

```
mean(International_Salaries$salary_in_usd, trim = .10)
```

```
## [1] 63288.25
```

```
#trimmed to account for potential outliers, brings closer to median actually,
#which is more robust statistic.
```

```
US_company_size <- aggregate(US_based_salaries$salary_in_usd, list(US_based_salaries$company_size), summ
  arrange(factor(Group.1, levels = c('S', 'M', 'L')))
US_company_size
```

```
##   Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
## 1       S   5679.0   59000.0  90000.0 104570.5  120000.0 416000.0
## 2       M  12000.0  105615.0 135500.0 141446.8  167656.2 450000.0
## 3       L  20000.0  105250.0 150000.0 160967.2  197000.0 600000.0
```

```
US_experience <- aggregate(US_based_salaries$salary_in_usd, list(US_based_salaries$experience_level), su
  arrange(factor(Group.1, levels = c('EN', 'MI', 'SE', 'EX')))
US_experience
```

```
##   Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
## 1      EN  12000.0   70000.0  90000.0  93112.9  102500.0 250000.0
## 2      MI   5679.0   87750.0 111887.5 125780.2  150000.0 450000.0
## 3      SE  25000.0  115233.5 145500.0 151527.6  180000.0 412000.0
## 4      EX 110000.0  163406.2 220000.0 243742.2  268500.0 600000.0
```

```r
US_yearly <- aggregate(US_based_salaries$salary_in_usd, list(US_based_salaries$work_year), summary) %>%
  arrange(factor(Group.1, levels = c('2020', '2021', '2022')))

US_yearly
```

```
##   Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
## 1    2020  45760.0   88000.0 108000.0 143251.3  147087.5 450000.0
## 2    2021   5679.0   86250.0 125000.0 141991.0  172000.0 600000.0
## 3    2022  25000.0  110606.2 140000.0 145066.2  170000.0 405000.0
```

```r
#to show all years but with relation to company size and experience level.
#organizing data here, so I can get a broad overlook of data using specific
#variables
```

```r
size_exp_us <- US_based_salaries %>%
  group_by(company_size, experience_level) %>%
  summarize_at("salary_in_usd", list(mean = mean,
                                   median = median,
                                   max = max)) %>%
  arrange(factor(experience_level, levels = c('EN', 'MI', 'SE', 'EX'))) %>%
  arrange(factor(company_size, levels = c('S', 'M', 'L')))


size_exp_us
```

```
## # A tibble: 12 x 5
## # Groups:   company_size [3]
##    company_size experience_level    mean median    max
##    <chr>        <chr>              <dbl>  <dbl>  <int>
##  1 S            EN                84250   90000 138000
##  2 S            MI                69298.  58000 120000
##  3 S            SE               132333. 120000 256000
##  4 S            EX               416000  416000 416000
##  5 M            EN                79625   80000 125000
##  6 M            MI               130835. 120000 450000
##  7 M            SE               143844. 140000 266400
##  8 M            EX               192388. 187500 324000
##  9 L            EN               112591.  91000 250000
## 10 L            MI               133135. 112000 450000
## 11 L            SE               181686. 170000 412000
## 12 L            EX               312000  250000 600000
```

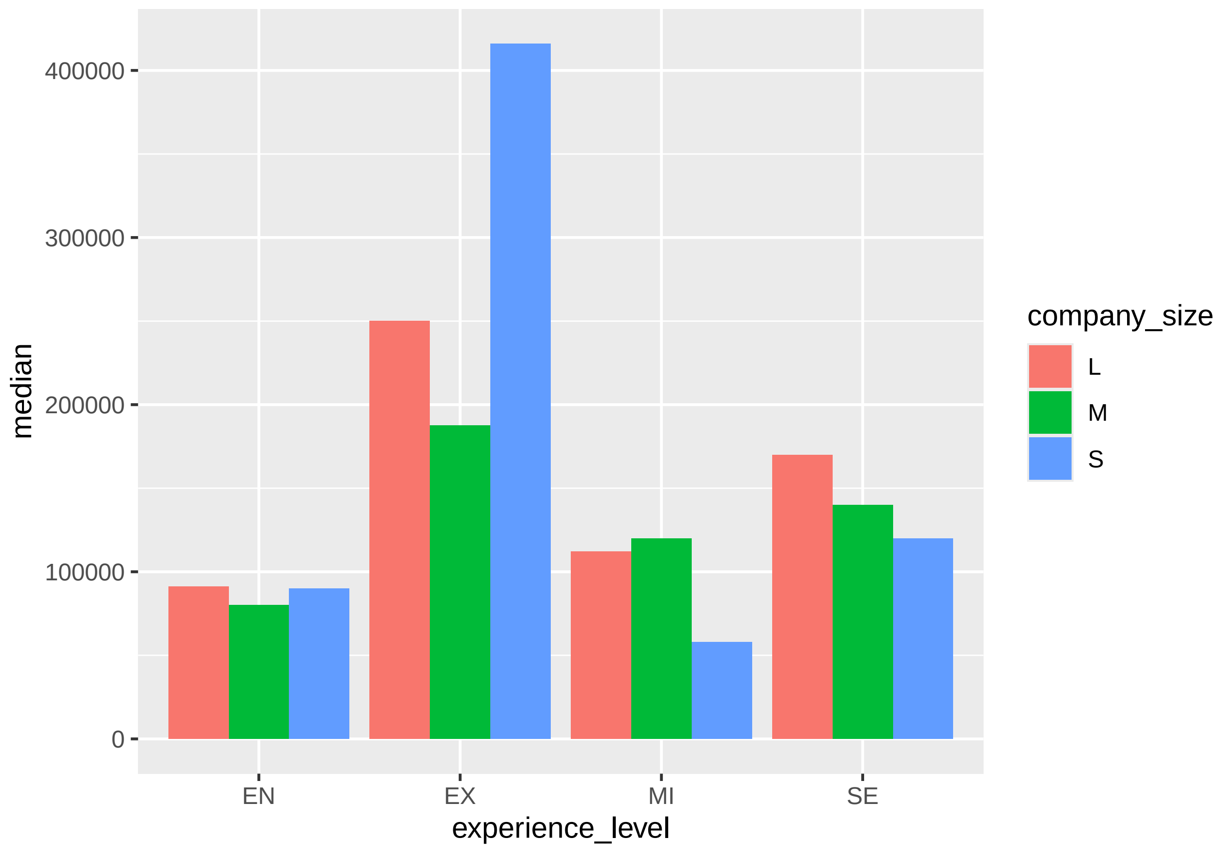
```r
#organizing data in order S-L , and Entry-Exec. I guess I don't really need to
#do this, but I personally prefer to look at the data in this order.

all_factors <-US_based_salaries %>%
  group_by(company_size, experience_level, work_year) %>%
  summarize_at("salary_in_usd", list(mean = mean,
                                   median = median,
                                   max = max)) %>%
  arrange(factor(experience_level, levels = c('EN', 'MI', 'SE', 'EX'))) %>%
  arrange(factor(company_size, levels = c('S', 'M', 'L'))) %>%
```

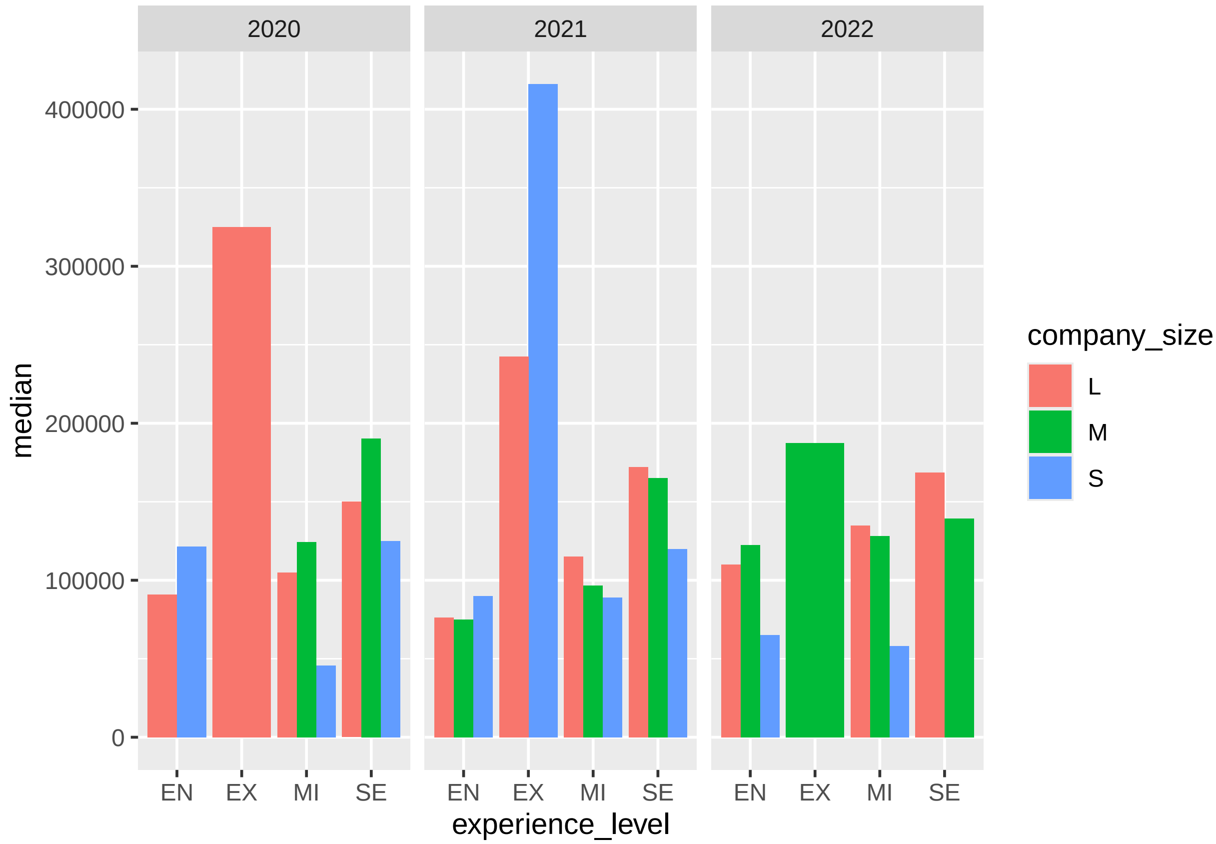
```
  arrange(factor(work_year, levels = c('2020', '2021', '2022')))

#this way of organizing let me some interesting information. This could help
#in further analysis.Some years only large or medium companies had executive
#level, not sure if that will play out somewhere further, but interesting
#to keep in mind.
```

```
ex_size_plot <- ggplot(size_exp_us,
       mapping = aes( x= experience_level, y = median, fill = company_size)
       ) + geom_bar(stat = 'identity', position = 'dodge')
ex_size_plot + scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```




```
facet_ex_size_plot <- ggplot(all_factors,
       mapping = aes( x= experience_level, y = median, fill = company_size)
       ) + geom_bar(stat = 'identity', position = 'dodge') +
        facet_wrap(~work_year)
facet_ex_size_plot + scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



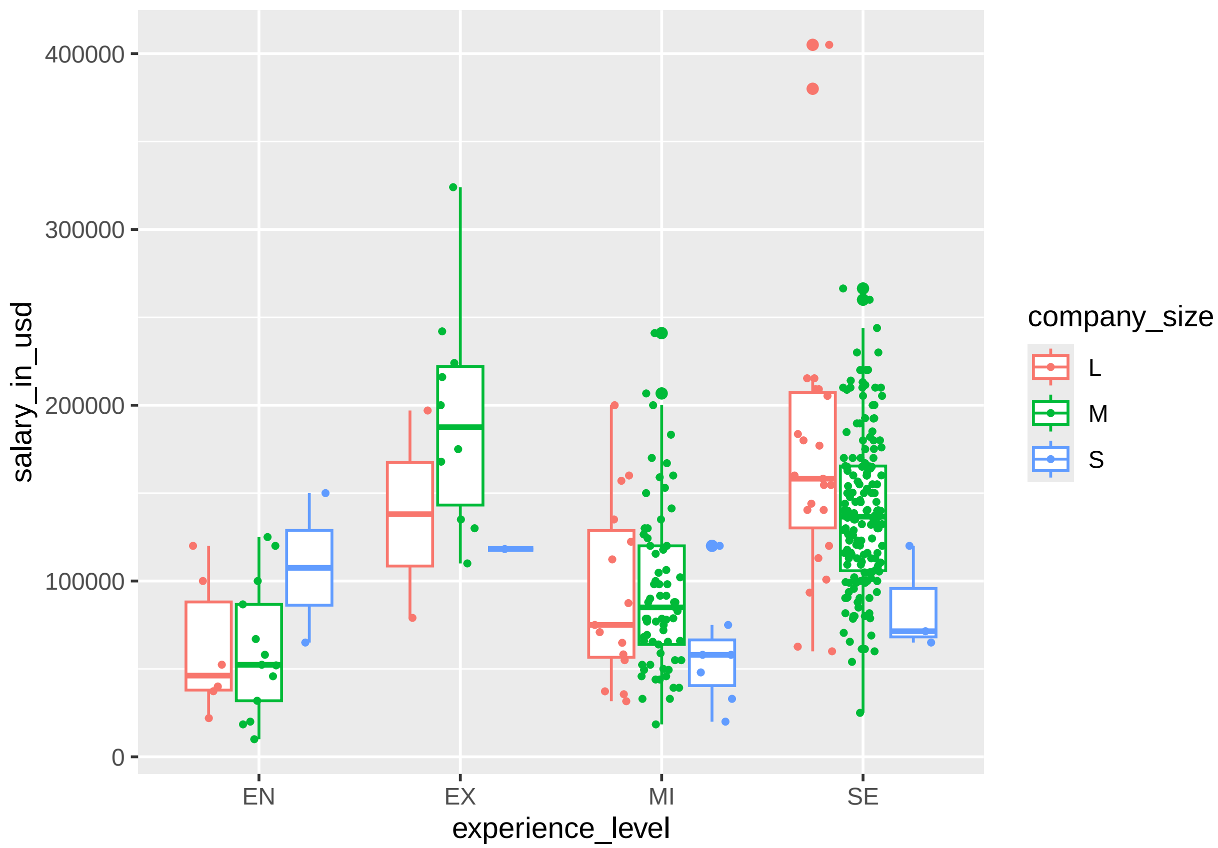

```
#general overview plots to visualize median salaries across all factors
```

More Focused Analysis

```
#This area is to breakdown 2022 rates so that I can later make my projections
#based on inflation and cost of living

US_based_2022 <- filter(salary_data, work_year == "2022")

 US_based_2022 %>%
  mutate(experience_level  = as.factor(experience_level)) %>%
  ggplot(mapping = aes(x= experience_level, y = salary_in_usd, colour = company_size)
         ) +
  geom_boxplot(outlier.shape = 19 , varwidth = F) +
  geom_jitter( position = position_jitterdodge(0.2), size = 0.75) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```




```
#using this plot to visualize the majority of salaries in a given area.
 #Mid and Senior at Medium companies, with entry level following, and executive
 #level seems to be "rare".

summary_2022 <- aggregate(US_based_2022$salary_in_usd, list(US_based_2022$experience_level, US_based_20
  arrange(factor(Group.2, levels = c('S', 'M', 'L'))) %>%
  arrange(factor(Group.1, levels = c('EN', 'MI', 'SE', 'EX')))
summary_2022 #wanted to see summary based on company size and experience
```

```
##    Group.1 Group.2    x.Min. x.1st Qu.  x.Median    x.Mean x.3rd Qu.    x.Max.
## 1       EN       S  65000.00  86250.00 107500.00 107500.00 128750.00 150000.00
## 2       EN       M  10000.00  31875.00  52351.00  60554.85  86703.00 125000.00
## 3       EN       L  21983.00  37975.00  46198.00  61946.50  88099.00 120000.00
## 4       MI       S  20000.00  40487.00  58000.00  58853.43  66500.00 120000.00
## 5       MI       M  18442.00  63900.00  85000.00  93973.78 120000.00 241000.00
## 6       MI       L  31615.00  56606.00  75000.00  93499.00 128673.00 200000.00
## 7       SE       S  65000.00  68222.00  71444.00  85481.33  95722.00 120000.00
## 8       SE       M  25000.00 105830.00 136600.00 139935.05 165400.00 266400.00
## 9       SE       L  60000.00 130200.00 158200.00 173120.78 207200.00 405000.00
## 10      EX       S 118187.00 118187.00 118187.00 118187.00 118187.00 118187.00
## 11      EX       M 110000.00 143218.75 187500.00 192387.50 222000.00 324000.00
## 12      EX       L  79039.00 108524.00 138009.00 138009.00 167494.00 196979.00
```

```
#levels, to give a better representation of wage ranges, this df is listing the
#wage ranges visualized on the box plot "boxplot_overall"
```

```r
medians_2022 <- US_based_2022 %>%
  group_by(company_size, experience_level) %>%
  summarize_at("salary_in_usd", list(mean = mean,
                                     median = median,
                                     max = max)) %>%
  arrange(factor(company_size, levels = c('S', 'M', 'L'))) %>%
  arrange(factor(experience_level, levels = c('EN', 'MI', 'SE', 'EX')))

size_2022 <- aggregate(US_based_2022$salary_in_usd, list(US_based_2022$company_size), summary)
size_2022
```

```
##   Group.1    x.Min. x.1st Qu.  x.Median    x.Mean x.3rd Qu.    x.Max.
## 1       L  21983.00  66364.75 121173.00 131129.57 172750.00 405000.00
## 2       M  10000.00  88966.00 123000.00 125731.40 160040.00 324000.00
## 3       S  20000.00  58000.00  65000.00  77046.54 118187.00 150000.00
```

```r
summary_2022 <- aggregate(US_based_2022$salary_in_usd, list(US_based_2022$experience_level, US_based_20
  arrange(factor(Group.2, levels = c('S', 'M', 'L'))) %>%
  arrange(factor(Group.1, levels = c('EN', 'MI', 'SE', 'EX')))
summary_2022 #wanted to see summary based on company size and experience levels,
```

```
##    Group.1 Group.2    x.Min. x.1st Qu.  x.Median    x.Mean x.3rd Qu.    x.Max.
## 1       EN       S  65000.00  86250.00 107500.00 107500.00 128750.00 150000.00
## 2       EN       M  10000.00  31875.00  52351.00  60554.85  86703.00 125000.00
## 3       EN       L  21983.00  37975.00  46198.00  61946.50  88099.00 120000.00
## 4       MI       S  20000.00  40487.00  58000.00  58853.43  66500.00 120000.00
## 5       MI       M  18442.00  63900.00  85000.00  93973.78 120000.00 241000.00
## 6       MI       L  31615.00  56606.00  75000.00  93499.00 128673.00 200000.00
## 7       SE       S  65000.00  68222.00  71444.00  85481.33  95722.00 120000.00
## 8       SE       M  25000.00 105830.00 136600.00 139935.05 165400.00 266400.00
## 9       SE       L  60000.00 130200.00 158200.00 173120.78 207200.00 405000.00
## 10      EX       S 118187.00 118187.00 118187.00 118187.00 118187.00 118187.00
## 11      EX       M 110000.00 143218.75 187500.00 192387.50 222000.00 324000.00
## 12      EX       L  79039.00 108524.00 138009.00 138009.00 167494.00 196979.00
```

```r
#to give a better representation of wage ranges
```

#Inflation and COLA Adjustments

```r
select_inflation <- select(summary_2022, -(3))
trial <-summary_2022[, sapply(summary_2022, is.numeric)] <- summary_2022[, sapply(summary_2022, is.nume
select_inflation_2 <-bind_cols(select_inflation, trial)
select_inflation_2[3:8] = lapply(select_inflation_2[3:8], "*", 1.152)
 #so here we have multiplied all columns by the adjusted rate of 15.2%

#size_inflation <- select(size_2022, -(2))
#trial_2 <-size_2022[, sapply(size_2022, is.numeric)] <- size_2022[, sapply(size_2022, is.numeric)]
#size_inflation_2 <-bind_cols(size_inflation, trial_2)
#size_inflation_2[2:7] = lapply(size_inflation_2[2:7], "*", 1.152)
```

#Building Team for growth. Who is the team at medium companies?

```r
#Since our CEO wants to grow from small to medium, what roles are those
#companies comprised off? What will be competitive salaries for a team of
#data scientists?
length(unique(US_based_2022$job_title))#33 unique job titles
```

```
## [1] 33
```

```r
table(US_based_2022$job_title) # I want to know the most popular job titles,
```

```
##
##                              AI Scientist
##                                         2
##                        Analytics Engineer
##                                         4
##                    Applied Data Scientist
##                                         3
##        Applied Machine Learning Scientist
##                                         2
##                     Business Data Analyst
##                                         2
##                  Computer Vision Engineer
##                                         2
##         Computer Vision Software Engineer
##                                         1
##                              Data Analyst
##                                        73
##                   Data Analytics Engineer
##                                         1
##                       Data Analytics Lead
##                                         1
##                    Data Analytics Manager
##                                         4
##                            Data Architect
##                                         8
##                             Data Engineer
##                                        89
##                     Data Science Engineer
##                                         1
##                      Data Science Manager
##                                         5
##                            Data Scientist
##                                        77
##                  Director of Data Science
##                                         1
##                             ETL Developer
##                                         2
##                    Financial Data Analyst
##                                         1
##                              Head of Data
##                                         2
##                      Head of Data Science
##                                         2
```

```
##                 Head of Machine Learning
##                                        1
##                       Lead Data Engineer
##                                        1
##           Lead Machine Learning Engineer
##                                        1
##               Machine Learning Developer
##                                        2
##                Machine Learning Engineer
##                                       18
## Machine Learning Infrastructure Engineer
##                                        1
##               Machine Learning Scientist
##                                        3
##                              ML Engineer
##                                        1
##                             NLP Engineer
##                                        1
##                   Principal Data Analyst
##                                        1
##                 Principal Data Scientist
##                                        1
##                       Research Scientist
##                                        4
```

```r
#so I can build a team and know proper wages
#Top 5 are Data Engineer, Data Scientist, Data Analyst, Machine Learning
#Engineer, Data Architect

US_based_2022$job_title=factor(US_based_2022$job_title)

medians_by_title <- by(US_based_2022$salary_in_usd,US_based_2022$job_title,median)
medians_by_title # I know my top 5 so I can reference this list to find my
```

```
## US_based_2022$job_title: AI Scientist
## [1] 160000
## ------------------------------------------------------------
## US_based_2022$job_title: Analytics Engineer
## [1] 179850
## ------------------------------------------------------------
## US_based_2022$job_title: Applied Data Scientist
## [1] 177000
## ------------------------------------------------------------
## US_based_2022$job_title: Applied Machine Learning Scientist
## [1] 53437.5
## ------------------------------------------------------------
## US_based_2022$job_title: Business Data Analyst
## [1] 44677
## ------------------------------------------------------------
## US_based_2022$job_title: Computer Vision Engineer
## [1] 67500
## ------------------------------------------------------------
## US_based_2022$job_title: Computer Vision Software Engineer
## [1] 150000
```

```
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Analyst
## [1] 105000
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Analytics Engineer
## [1] 20000
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Analytics Lead
## [1] 405000
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Analytics Manager
## [1] 127140
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Architect
## [1] 192482
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Engineer
## [1] 120000
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Science Engineer
## [1] 60000
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Science Manager
## [1] 159000
## ---------------------------------------------------------------
## US_based_2022$job_title: Data Scientist
## [1] 140000
## ---------------------------------------------------------------
## US_based_2022$job_title: Director of Data Science
## [1] 196979
## ---------------------------------------------------------------
## US_based_2022$job_title: ETL Developer
## [1] 54957
## ---------------------------------------------------------------
## US_based_2022$job_title: Financial Data Analyst
## [1] 1e+05
## ---------------------------------------------------------------
## US_based_2022$job_title: Head of Data
## [1] 116487
## ---------------------------------------------------------------
## US_based_2022$job_title: Head of Data Science
## [1] 195937.5
## ---------------------------------------------------------------
## US_based_2022$job_title: Head of Machine Learning
## [1] 79039
## ---------------------------------------------------------------
## US_based_2022$job_title: Lead Data Engineer
## [1] 118187
## ---------------------------------------------------------------
## US_based_2022$job_title: Lead Machine Learning Engineer
## [1] 87932
## ---------------------------------------------------------------
## US_based_2022$job_title: Machine Learning Developer
## [1] 78791
```

```
## ------------------------------------------------------------
## US_based_2022$job_title: Machine Learning Engineer
## [1] 120000
## ------------------------------------------------------------
## US_based_2022$job_title: Machine Learning Infrastructure Engineer
## [1] 58255
## ------------------------------------------------------------
## US_based_2022$job_title: Machine Learning Scientist
## [1] 153000
## ------------------------------------------------------------
## US_based_2022$job_title: ML Engineer
## [1] 21983
## ------------------------------------------------------------
## US_based_2022$job_title: NLP Engineer
## [1] 37236
## ------------------------------------------------------------
## US_based_2022$job_title: Principal Data Analyst
## [1] 75000
## ------------------------------------------------------------
## US_based_2022$job_title: Principal Data Scientist
## [1] 162674
## ------------------------------------------------------------
## US_based_2022$job_title: Research Scientist
## [1] 106713.5
```

```r
#median values, then I can use my dataframe "sum-by_title_2022' to reference
#the wages by company size.

sum_by_title_2022 <- US_based_2022 %>%
  group_by(company_size, job_title) %>%
  summarize_at("salary_in_usd", list(mean = mean,
                                     median = median,
                                     max = max,
                                     sd = sd
                                     )) %>%
  arrange(factor(company_size, levels = c('S', 'M', 'L')))

#accounting for inflation and cost of living code below

job_summary <- aggregate(US_based_2022$salary_in_usd, list(US_based_2022$experience_level, US_based_2022
  arrange(factor(Group.2, levels = c('S', 'M', 'L'))) %>%
  arrange(factor(Group.1, levels = c('EN', 'MI', 'SE', 'EX'))) %>%
  arrange(factor(Group.3, levels = c('Data Engineer', 'Data Scientist', 'Data Analyst', 'Machine Learnin


select_job_inflation <- select(job_summary, -(4))
test <-job_summary[, sapply(job_summary, is.numeric)] <- job_summary[, sapply(job_summary, is.numeric)]
job_inflation_2 <-bind_cols(select_job_inflation, test)
job_inflation_2[4:9] = lapply(job_inflation_2[4:9], "*", 1.152)
#so we have created summaries by job title, and then applied the inflation
#parameter for the values.
#The summary columns were not recognized as numeric, so i had to separate
#and combine so that I could apply the 1.152 rate
```

```
#some of these data frames I created I ended up not using, which I assume is
#part of the process. Not sure I needed to create so many, as it did cause
#some confusion, and a few times I re-used a dataframe or variable name and
#really threw off my previous data, causing me to have to re-do some dataframes.
```

#In Summary

```
#In summery, my goal was to take all the data and evaluate from a company
#size perspective, an experience level perspective, and job title perspective.
#Ultimately, I wanted to answer the basic question of what a competitive salary
#is with inflation based on current company needs, but I wanted to give a back story on what those numb
#In terms of growth, I wanted to address what that data would look like going
#from a small to medium company, as well as, what roles are most prevalent
#in those sized companies.
#I primarily used median, as it is more accurate in mitigating outliers.
#I did also provide quartile ranges.
```