Project 1

PJ Grant

2024-12-01

#Import Libraries  
library(tidyverse)

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

library(readxl)  
library(writexl)  
library(scales)

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

#File path for excel outputs to create charts in PPT  
file\_path = '~/MM/DSE5002/Week\_5/Project 1/proj\_1\_summary\_stats.xlsx'

## Load and prepare data for use

#Load Data  
raw\_data = read.csv('~/MM/DSE5002/Week\_5/Project 1/r project data.csv')  
  
#Update name and begin data cleaning  
data = raw\_data  
   
#dropping unnecessary columns  
drop\_cols = c('X', 'salary', 'salary\_currency')  
  
data = data |>   
 select(-all\_of(drop\_cols))  
  
#Rename salary column  
data = data |>   
 rename(  
 salary = salary\_in\_usd  
 )  
  
#Select only FT roles  
data = data |>   
 filter(  
 employment\_type == "FT"  
 )

#EDA

#Determine how many submissions we have for each role  
summarystats = data |>   
 group\_by(job\_title) |>   
 summarise(  
 count = n(),  
 .groups = 'drop'  
 )

# We will analyze the following roles: Data Scientist, Data Engineer, Data Analyst, and Machine Learning Engineer. These account for 70% of full-time submissions, with the rest being variations of these roles.

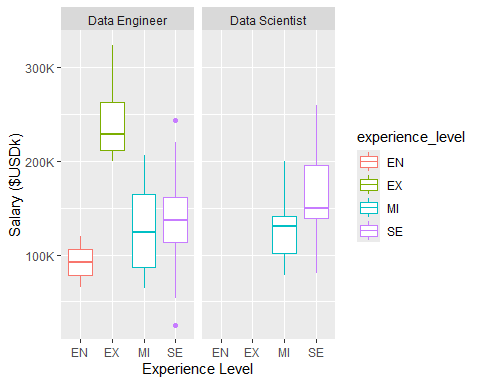
#Isolating the 4 roles above for the rest of the analysis  
kept\_roles = c('Data Scientist', 'Data Engineer', 'Data Analyst', 'Machine Learning Engineer')  
ds\_data = data |>   
 filter(  
 job\_title %in% kept\_roles  
 )  
  
#Distinguish which submissions are US based vs Offshore  
ds\_data = ds\_data |>   
 mutate(  
 off\_onshore = ifelse(employee\_residence == 'US','On','Off')  
 )

# Aggregations for Slide 1

# Determine the average onshore vs offshore salary for each role in 2020, 2022 and the growth rate  
avg\_salary = ds\_data |>   
 filter(  
 work\_year != '2021'  
 ) |>   
 group\_by(job\_title, off\_onshore, work\_year) |>   
 summarize(  
 avg\_salary = round(mean(salary),0),  
 .groups = 'drop'  
 ) |>   
 pivot\_wider(  
 names\_from = work\_year,  
 values\_from = avg\_salary,  
 names\_prefix = 'year\_'  
 ) |>   
 mutate(  
 cagr = round(((year\_2022 / year\_2020)^(1/2)) - 1,2)  
 )  
  
#write data to excel to create chart in ppt  
write\_xlsx(avg\_salary, path = '~/MM/DSE5002/Week\_5/Project 1/slide1.xlsx')

# Aggregations for Slide 2

#Filter for only 2022 Data for DS and DE roles in the US for small and medium sized companies  
slide2 = ds\_data |>   
 filter(  
 work\_year == '2022',  
 job\_title %in% c('Data Scientist', 'Data Engineer'),  
 employee\_residence == 'US',  
 company\_size %in% c('S', 'M')  
 )  
  
# Calculate the mean and median salaries for each of these roles, these will serve as the salary ranges we expect to pay these two roles + 20% for overhead. The recommendation on the page will be to hire one SE Data scientist and a MI Data Engineer  
salary\_ranges = slide2 |>   
 group\_by(job\_title, experience\_level) |>   
 summarize(  
 median\_salary = median(salary),  
 avg\_salary = mean(salary),  
 .groups = 'drop'  
 )  
  
#Create a histogram, faceted by job type showing the salaries for each of these roles  
ggplot(slide2, aes(x = experience\_level, y = salary, colour = experience\_level)) +  
 geom\_boxplot() +  
 scale\_y\_continuous(labels = label\_number(scale = 1e-3, suffix = "K")) +  
 labs(y = "Salary ($USDk)", x = 'Experience Level') +  
 facet\_wrap(~job\_title)



#Work from home % by job title and experience level  
wfh = slide2 |>   
 group\_by(job\_title, experience\_level, remote\_ratio) |>   
 summarize(  
 count = n(),  
 .groups = 'drop'  
 )  
  
#write data to excel to create chart in ppt  
write\_xlsx(salary\_ranges, path = '~/MM/DSE5002/Week\_5/Project 1/slide2.xlsx')

# Aggregations for Slide 3

#Looking to fill out the rest of the team with a mid level analyst and data engineer who lives offshore  
slide3 = ds\_data |>   
 filter(  
 job\_title %in% c('Data Analyst', 'Data Engineer', 'Machine Learning Engineer'),  
 company\_size %in% c('S', 'M'),  
 experience\_level == "MI",  
 work\_year != '2020',  
 employee\_residence != 'US'  
 ) |>   
 group\_by(job\_title, employee\_residence) |>   
 summarise(  
 avg\_salary = mean(salary),  
 median\_salary = median(salary),  
 .groups = 'drop'  
 )  
  
#write data to excel to create chart in ppt  
write\_xlsx(slide3, path = '~/MM/DSE5002/Week\_5/Project 1/slide3.xlsx')