Project_Final

Kate O'Rourke

2023-11-14

Load libraries.

```
#library(tidyverse)
library(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
```

```
library(tidyr)
library(ggplot2)
```

Read in the csv file.

```
salary_metadata <- read.csv("Data/ORourke.module05RProject.csv")
```

Data Wrangling

Check for any NA in the data set and filter out not full-time data and job titles that do not include Lead or Manager:

```
#Check for any NA in data set
sum(is.na(salary_metadata))
```

```
## [1] 0
```

```
#Filter in full-time and job titles including manager or lead
salary_ft_lead_or_manager <- salary_metadata %>%
  filter(employment_type=="FT") %>%
  filter(grepl('(?i)Manager', job_title) | grepl('(?i)Lead', job_title))
```

```r
#Create Data Set that only includes US companies.
salary_ft_lead_or_manager_us <- salary_ft_lead_or_manager %>%
  filter(company_location=='US')
```

Group data frames by country and by remote_ratio:

```r
#Group the data by company_location (country the company is in).
salary_group_by_country <- salary_ft_lead_or_manager %>%
  group_by(company_location)

#Group the data by remote_ratio.
salary_group_by_remote_ratio <- salary_ft_lead_or_manager_us %>%
  group_by(remote_ratio)

#Group data by experience_level
salary_group_by_experience_level <- salary_ft_lead_or_manager_us %>%
  group_by(experience_level)
```

Data Analysis

Summarize overall data, the data grouped by country, and the data grouped by experience level to include mean, median, IQR, Q1, Q3, minimum, and maximum.

```r
salary_ft_lead_or_manager_us %>%
    summarize(mean = mean(salary_in_usd), median = median(salary_in_usd),
              iqr = IQR(salary_in_usd), q1 = quantile(salary_in_usd, prob=.25, type = 1),
              q3 = quantile(salary_in_usd, prob=.75, type = 1),
              minimum_Value = min(salary_in_usd), maximum_Value = max(salary_in_usd))
```

```
##       mean median     iqr     q1     q3 minimum_Value maximum_Value
## 1 161637.7 152500 45429.5 120000 174000         54094        405000
```

Determine the proportions of leads or managers in the US by remote ratio.

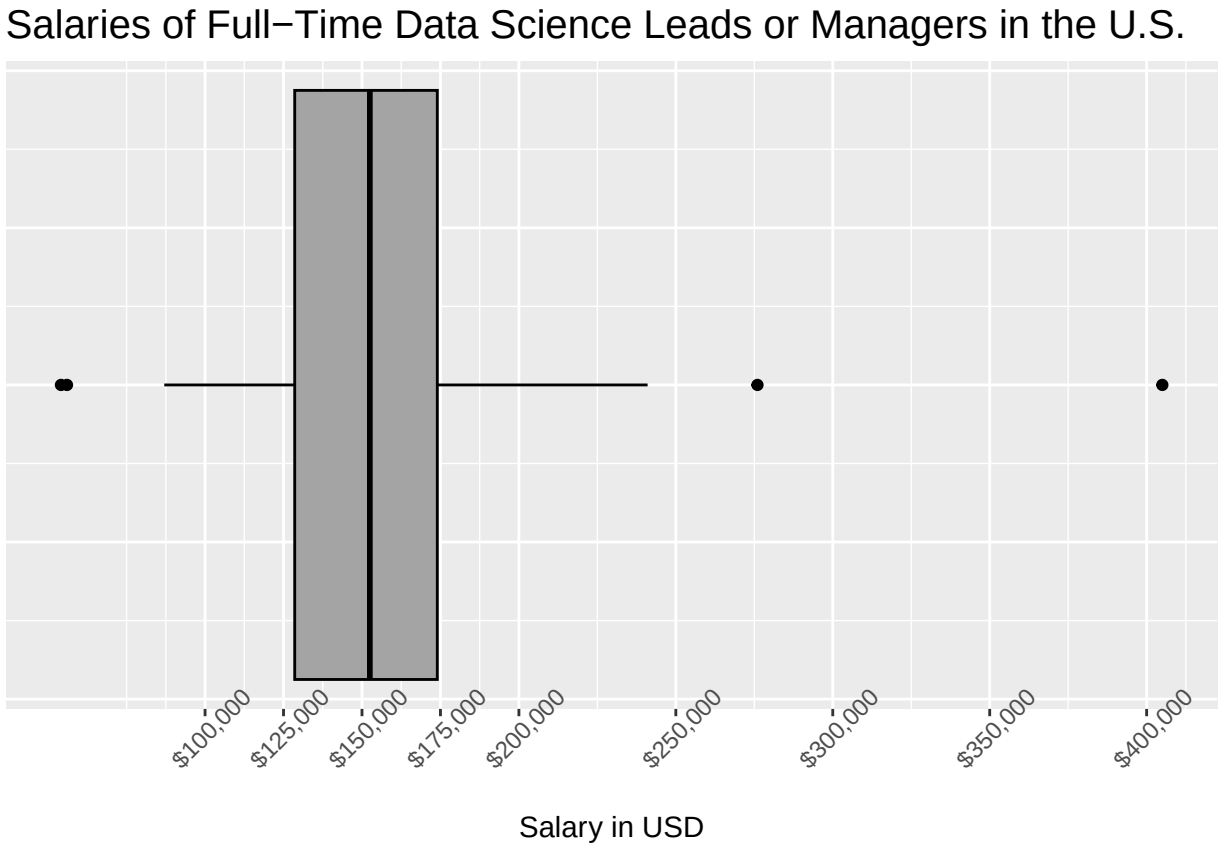
```r
#Calculate counts for each remote_ratio then divide by the total number of data
#points to get the percentage by remote ratio.
group_counts_by_remote_ratio <- table(salary_group_by_remote_ratio$remote_ratio)
total_count <- sum(group_counts_by_remote_ratio)
percentage_by_group <- (group_counts_by_remote_ratio / total_count) * 100
percentage_by_group
```

```
##
##         0        50       100
## 14.814815  7.407407 77.777778
```

Plots

Plot boxplot of overall data:

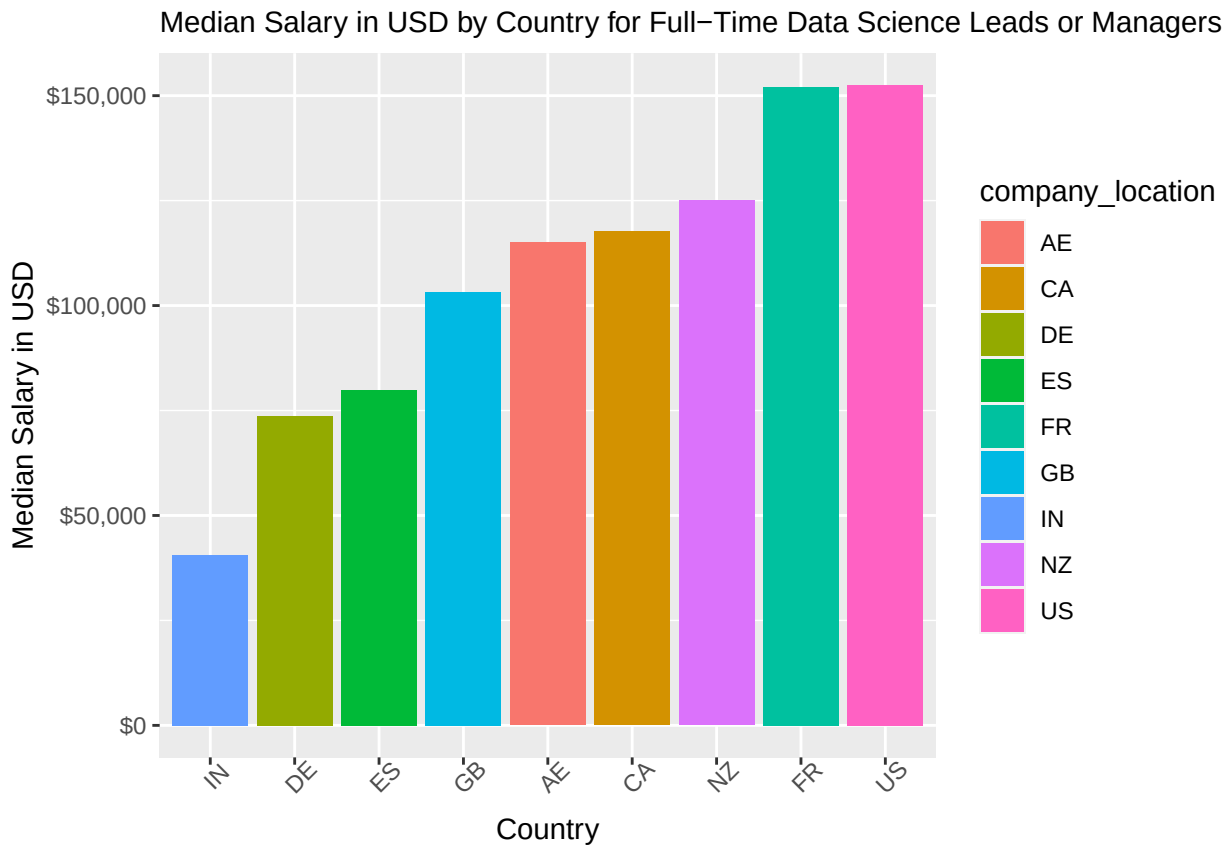
```r
#Plot box plot of full-time data science leads or managers in the U.S.
salary_ft_lead_or_manager_us %>%
  ggplot(aes(x=salary_in_usd)) +
          geom_boxplot(fill='#A4A4A4', color="black") +
          labs(x='Salary in USD') +
          labs(title='Salaries of Full-Time Data Science Leads or Managers in the U.S.') +
          scale_x_continuous(labels=scales::dollar_format(), breaks=c(100000, 125000, 150000, 175000,
                                                                      200000, 250000, 300000, 350000,
                                                                      400000)) +
          theme(plot.title = element_text(size = 15)) +
          theme(
           axis.title.y = element_blank(),
           axis.text.y = element_blank(),
           axis.ticks.y = element_blank()
          ) +
          theme(axis.text.x = element_text(angle = 45))
```



Plot a bar graph of salaries in USD for full-time data science leads or managers by country.

```r
#Plot a bar graph of median salaries in USD for full-time data science leads or managers by country.
salary_group_by_country %>%
  summarize(median_salary = median(salary_in_usd)) %>%
  ggplot(aes(x = reorder(company_location, median_salary), y=median_salary,
             fill = company_location)) +
          geom_bar(stat='identity') +
          labs(x='Country', y='Median Salary in USD') +
```

```
          labs(title=c('Median Salary in USD by Country for Full-Time Data Science Leads or Managers'))
          theme(axis.text.x = element_text(angle = 45)) +
          scale_y_continuous(labels=scales::dollar_format()) +
          theme(plot.title = element_text(size = 11))
```



Plot a bar graph of median salaries by experience level.

```
salary_group_by_remote_ratio %>%
  ggplot(aes(x=factor(remote_ratio), y=salary_in_usd, fill=factor(remote_ratio)), group=1) +
          geom_boxplot() +
          labs(x='Remote Ratio', y='Salary in USD') +
          labs(title='Salary by Ratio of Remote Work Done by Full-Time Data Science Leads or Managers')
          theme(legend.position='none') +
          theme(plot.title = element_text(size = 10)) +
          scale_y_continuous(labels=scales::dollar_format())
```

Plot a box plot of salaries in USD for full-time data science leads or managers by remote_ratio and experience level.

```
#Plot a bar graph of salaries by remote ratio for full-time data science leads or managers by country.
salary_group_by_experience_level %>%
  ggplot(aes(x=experience_level, y=salary_in_usd, fill=experience_level), group=1) +
           geom_boxplot() +
           labs(x='Experience Level', y='Salary in USD') +
           labs(title='Salary by Experience Level of Full-Time Data Science Leads or Managers') +
           theme(legend.position='none') +
           theme(plot.title = element_text(size = 12)) +
           scale_y_continuous(labels=scales::dollar_format())
```


Salary by Experience Level of Full−Time Data Science Leads or Managers
$400,000
$300,000
$200,000
$100,000
Salary in USD
MI
SE
Experience Level