Data Analyst Job Posting Analysis

Moonyoung Hwamg

2023-10-13

Abstract

In this project, I analyze job postings from Google's search results for Data Analyst positions in the United States between November 4th, 2022, and July 31st, 2023. The motivation for this project was somewhat personal. As a job seeker in the field of data analytics and a learner in this domain, this project provides me with two significant benefits: 1) honing my R skills and 2) gaining insights into the current job market for data analytics. Anyone interested in the types of data analytic jobs available today will find valuable information in this research.

I ask the following 4 questions: 1) What is the ratio between remote and non-remote jobs? 2) What are the top 20 frequently mentioned skills? 3) Where are the top 20 online platforms with the most job openings advertised? 4) What are the minimum, maximum, and average hourly and yearly salaries advertised?

Load packages and data

```
library(readr)
library(tidyverse)

jobs<-read_csv("gsearch_jobs.csv")
```

```
jobs
```

```
## # A tibble: 24,446 × 27
##     ...1 index title   company_name location via   description extensions job_id
##    <dbl> <dbl> <chr>   <chr>        <chr>    <chr> <chr>       <chr>      <chr>
##  1     0     0 Data A… Robert Half  Oklahom… via … "Descripti… ['24 hour… eyJqb…
##  2     1     1 Data A… Apex Health… United … via … "Data Anal… ['21 hour… eyJqb…
##  3     2     2 Market… Ledger Benn… Anywhere via … "At Ledger… ['21 hour… eyJqb…
##  4     3     3 Boolea… IT Pros      Anywhere via … "Company D… ['14 hour… eyJqb…
##  5     4     4 Produc… The Toro Co… Perry, … via … "Who Are W… ['22 hour… eyJqb…
##  6     5     5 Associ… Talentify.io Anywhere via … "Talentify… ['18 hour… eyJqb…
##  7     6     6 Experi… Upwork       Anywhere via … "We are ac… ['10 hour… eyJqb…
##  8     7     7 Data A… WEBTPA       United … via … "Job Summa… ['11 hour… eyJqb…
##  9     8     8 Clinic… Medical Ass… Fayette… via … "Overview:… ['22 hour… eyJqb…
## 10     9     9 Data A… Centene Cor… Kansas … via … "You could… ['7 hours… eyJqb…
## # ℹ 24,436 more rows
## # ℹ 18 more variables: thumbnail <chr>, posted_at <chr>, schedule_type <chr>,
## #   work_from_home <lgl>, salary <chr>, search_term <chr>, date_time <dttm>,
## #   search_location <chr>, commute_time <lgl>, salary_pay <chr>,
## #   salary_rate <chr>, salary_avg <dbl>, salary_min <dbl>, salary_max <dbl>,
## #   salary_hourly <dbl>, salary_yearly <dbl>, salary_standardized <dbl>,
## #   description_tokens <chr>
```

1. What is the ratio between remote and non-remote jobs?

Location flexibility is of great importance to me as the mother of a toddler and as the spouse of someone who might need to relocate for his job in the future. Therefore, I am curious to know what percentage of job postings are for remote positions.

```
#Counting remote and non-remote jobs
count_work_from_home <- jobs %>%
  select(work_from_home) %>%
  count(work_from_home,sort=TRUE)
count_work_from_home
```

```
## # A tibble: 2 × 2
##   work_from_home     n
##   <lgl>          <int>
## 1 NA             13437
## 2 TRUE           11009
```

```
#Changing the variables of "work_from_home" column
count_work_from_home <- count_work_from_home %>%
  mutate(work_from_home=as.character(work_from_home)) %>%
  mutate(work_from_home=replace_na(work_from_home,"Unknown"))

count_work_from_home$work_from_home<-gsub("TRUE","Yes",count_work_from_home$work_from_home)

count_work_from_home
```

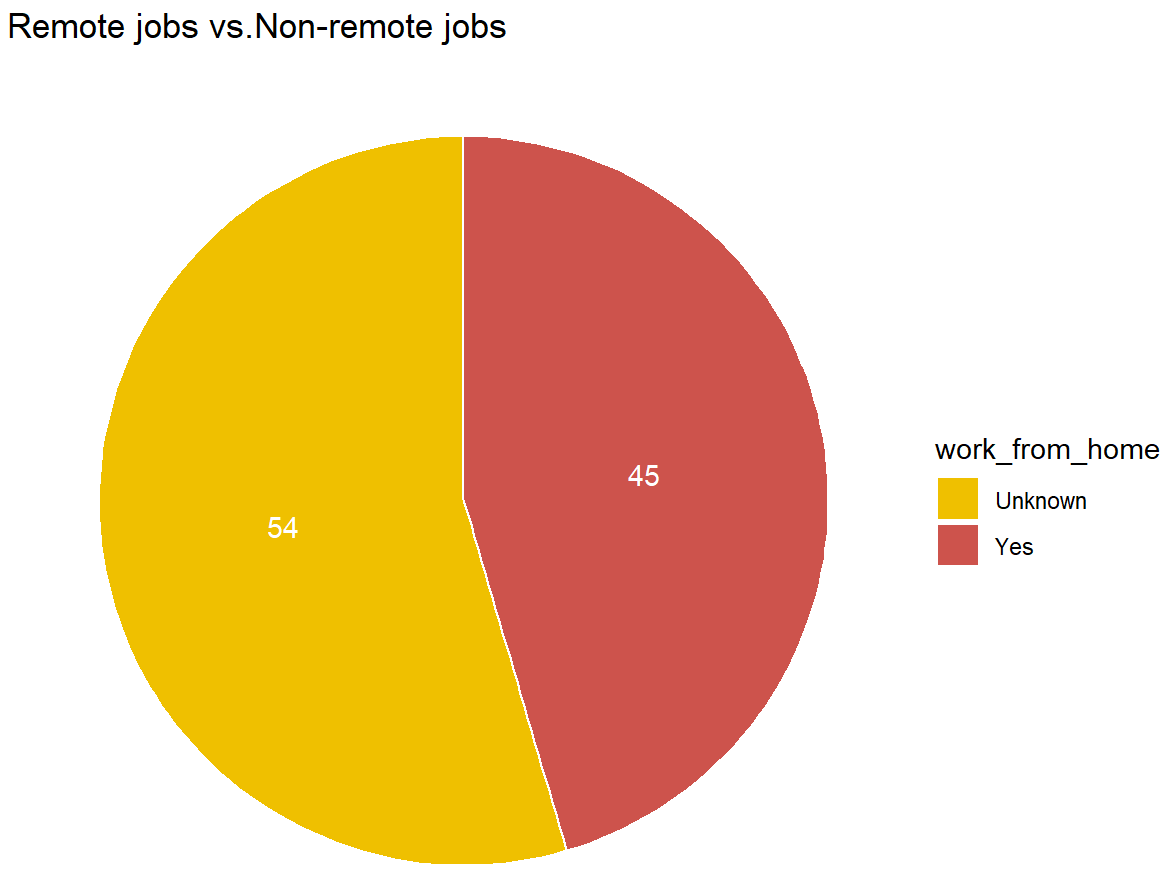
```
## # A tibble: 2 × 2
##   work_from_home     n
##   <chr>          <int>
## 1 Unknown        13437
## 2 Yes            11009
```

```
#Visualizing the number of remote vs. non_remote jobs
count_work_from_home<-count_work_from_home %>%
  arrange(desc(n))%>%
  mutate(prop=n/sum(count_work_from_home$n)*100) %>%
  mutate(ypos = cumsum(prop)- 0.5*prop)

count_work_from_home$prop<-as.integer(count_work_from_home$prop)

mycols <-c("#EFC000FF", "#CD534CFF")
plotting_work_from_home<-ggplot(
  count_work_from_home,
  aes(x = "", y = prop, fill = work_from_home)) +
  geom_col(width = 1, color = "white") +
  geom_text(
    aes(label = prop),
    color = "white",
    position = position_stack(vjust = .5)) +
  scale_fill_manual(values = mycols) +
  coord_polar("y", start = 0) +
  theme_void() +
  labs(title = "Remote jobs vs.Non-remote jobs")


plotting_work_from_home
```



The result was surprising. Nearly half of the job postings were for remote positions, which exceeded my expectations. Great!

2. What are the top 20 frequently mentioned skills?

If you were to conduct a Google search, you would come across various skills associated with data analytics. However, time and energy are limited resources for everyone. It would be immensely helpful to identify the skills most frequently mentioned in job advertisements. I am currently learning R, but is it truly valuable for securing a job? I will investigate the top 20 frequently mentioned skills.

```
#Selecting job_description column and cleaning data
Job_descriptions<-jobs %>%
  select(description_tokens)

Job_descriptions$description_tokens<-gsub("[[:punct:]]","",as.character(Job_descriptions$description_tokens))

Job_descriptions$description_tokens<-strsplit(Job_descriptions$description_tokens,split=" ")

unnested_skills<-Job_descriptions %>%
  unnest(description_tokens)
unnested_skills
```

```
## # A tibble: 79,696 × 1
##    description_tokens
##    <chr>
##  1 go
##  2 azure
##  3 excel
##  4 powerbi
##  5 sql
##  6 excel
##  7 tableau
##  8 sql
##  9 assembly
## 10 excel
## # ℹ 79,686 more rows
```

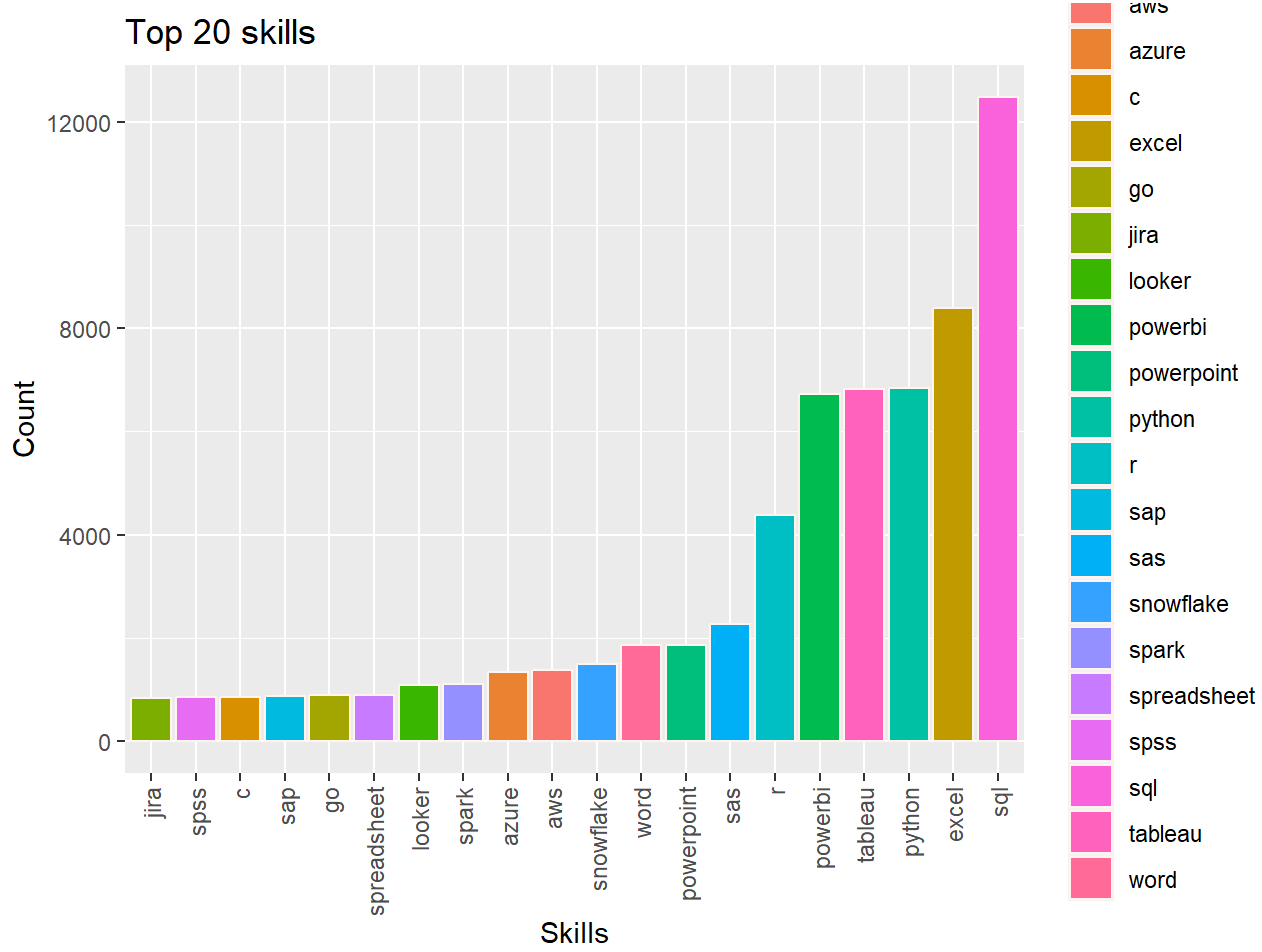
```
#Counting skills and leaving top 20 results
count_skills<-unnested_skills %>%
  count(description_tokens,sort=TRUE)
count_skills
```

```
## # A tibble: 124 × 2
##    description_tokens     n
##    <chr>              <int>
##  1 sql                12478
##  2 excel               8396
##  3 python              6846
##  4 tableau             6822
##  5 powerbi             6721
##  6 r                   4390
##  7 sas                 2264
##  8 powerpoint          1871
##  9 word                1858
## 10 snowflake           1490
## # ℹ 114 more rows
```

```
top_20_skills<-head(count_skills,n=20)
top_20_skills
```

```
## # A tibble: 20 × 2
##    description_tokens     n
##    <chr>              <int>
##  1 sql                12478
##  2 excel               8396
##  3 python              6846
##  4 tableau             6822
##  5 powerbi             6721
##  6 r                   4390
##  7 sas                 2264
##  8 powerpoint          1871
##  9 word                1858
## 10 snowflake           1490
## 11 aws                 1371
## 12 azure               1339
## 13 spark               1104
## 14 looker              1084
## 15 spreadsheet          897
## 16 go                   886
## 17 sap                  870
## 18 c                    863
## 19 spss                 860
## 20 jira                 839
```

```
#Plotting
ggplot(top_20_skills,aes(x=fct_reorder(description_tokens,n),y=n,fill=description_tokens))+geom_bar(stat="identity",position
=position_dodge(),colour="seashell")+theme(axis.text.x=element_text(angle=90,vjust=0.5,hjust=1))+xlab("Skills")+ylab("Coun
t")+ggtitle("Top 20 skills")
```



Sql turns out to be the most frequently mentioned skill. Python, tableau, and powerbi are within top 5 skills, which is expected. What is surprising to me is that excel is the second most mentioned skill. Since I thought excel is somewhat outdated tool to data analytic, I did not expect this result. R was ranked as 6th skill. Now, I know what other skills are worth learning.

3. Where are the top 20 online platforms with the most job openings advertised?

Now, where can I find possible job opportunities? What website does companies use the most to advertise their jobs?

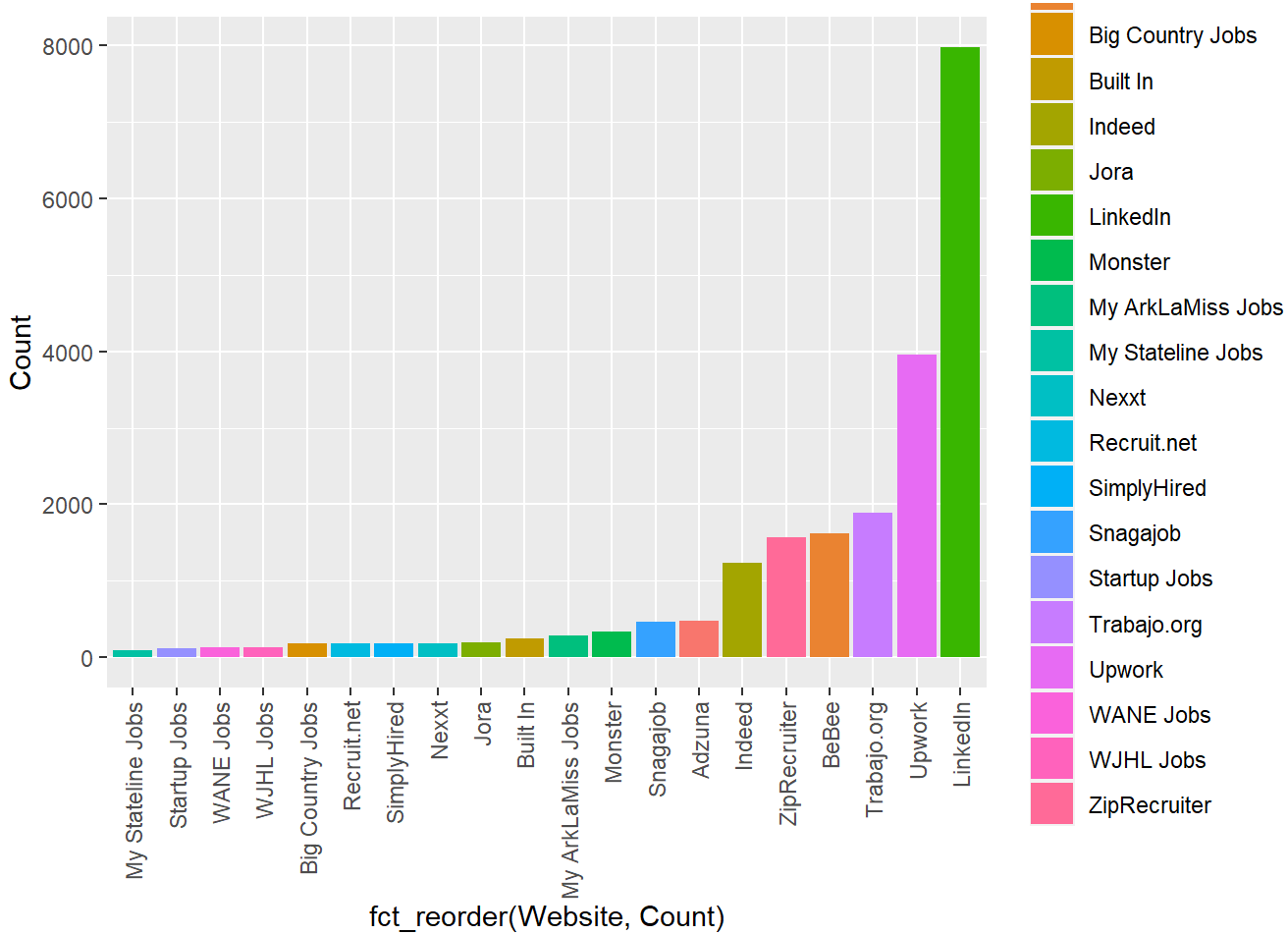
```
#Selecting columns that are relevant to my interest and data cleaning
jobs_interest<-jobs %>%
  select(company_name,location,via,schedule_type,work_from_home,salary)

jobs_interest$via<-gsub("via ","",jobs_interest$via)

#Counting websites and Cleaning data
count_via<-jobs_interest %>%
  group_by(via) %>%
  count(via,sort=TRUE) %>%
  head(n=20)

count_via<-count_via %>%
  rename("Website"="via","Count"="n")

#Plotting the result
ggplot(count_via,aes(x=fct_reorder(Website,Count),y=Count,fill=Website))+ geom_bar(stat="identity")+ theme(axis.text.x=eleme
nt_text(angle=90,vjust=0.5,hjust=1))
```



Linkedin and Upwork are top 2

platforms. Linkedin is a expected result. However, Upwork was surprising to me. Since I know that Upwork is the platform for freelancers, I could figure that the market for temporary contractor jobs in this field is somewhat big.

4. What are salaries advertised?

What are the salaries advertised? The “salary” column in this dataset includes both hourly and yearly based salaries, So we need to saparate them. Then, I will see what each case’s minimum, maximum, and average salary is.

Hourly salary summary

```
#Filtering hourly salary data

hourly<-jobs_interest %>%
  filter(salary != "NA") %>%
  filter(grepl(' an hour', salary))

#cleaning the data
hourly$salary<-gsub(" an hour","",hourly$salary)
hourly<-hourly %>%
  separate(salary,into=c("min_salary","max_salary"),sep="-",fill="right",convert=TRUE)

hourly
```

```
## # A tibble: 2,817 × 7
##    company_name           location via   schedule_type work_from_home min_salary
##    <chr>                  <chr>    <chr> <chr>         <lgl>                <dbl>
##  1 Upwork                 Anywhere Upwo… Contractor    TRUE                    37
##  2 Upwork                 Anywhere Upwo… Contractor    TRUE                    18
##  3 Upwork                 Anywhere Upwo… Contractor    TRUE                    18
##  4 Upwork                 Anywhere Upwo… Contractor    TRUE                    18
##  5 Upwork                 Anywhere Upwo… Contractor    TRUE                    45
##  6 Upwork                 Anywhere Upwo… Contractor    TRUE                    18
##  7 Insight Global         Anywhere Link… Full-time     TRUE                    40
##  8 Apex Systems           Anywhere Link… Contractor    TRUE                    30
##  9 Upwork                 Anywhere Upwo… Contractor    TRUE                    20
## 10 Global Network Techno… Eastbor… Adzu… Full-time     NA                      32
## # ℹ 2,807 more rows
## # ℹ 1 more variable: max_salary <dbl>
```

```
#Average hourly salary
hourly %>%
  summarise(average_hourly_salary=mean(c(min_salary,max_salary),na.rm=TRUE))
```

```
## # A tibble: 1 × 1
##   average_hourly_salary
##                   <dbl>
## 1                  44.5
```

```
#Maximum hourly salary
max(hourly$max_salary,na.rm=TRUE)
```

```
## [1] 500
```

```
#Minimum
min(hourly$min_salary,na.rm=TRUE)
```

```
## [1] 8
```

The maximum hourly salary is 500 dollars and the minimum hourly salary is 8. 500 dollars seems to be extremley high for an hourly salary. 8 dollar is too low as well. What kind of jobs do suggest these extreme hourly salaries?

```
hourly %>%
  filter(max_salary==500)
```

```
## # A tibble: 1 × 7
##   company_name location via   schedule_type work_from_home min_salary max_salary
##   <chr>        <chr>    <chr> <chr>         <lgl>               <dbl>      <dbl>
## 1 Upwork       Anywhere Upwo… Contractor    TRUE                  100        500
```

```
hourly %>%
  filter(min_salary==8)
```

```
## # A tibble: 30 × 7
##    company_name location via    schedule_type work_from_home min_salary
##    <chr>        <chr>    <chr>  <chr>         <lgl>                <dbl>
##  1 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  2 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  3 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  4 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  5 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  6 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  7 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  8 Upwork       Anywhere Upwork Contractor    TRUE                     8
##  9 Upwork       Anywhere Upwork Contractor    TRUE                     8
## 10 Upwork       Anywhere Upwork Contractor    TRUE                     8
## # ℹ 20 more rows
## # ℹ 1 more variable: max_salary <dbl>
```

Both cases are contractor jobs posted on Upwork. It makes sense that project-based temporary jobs are more diverse in salaries than regular full-time jobs. I was wondering what kind of project pays 500 dollars per hour. Unfortunately, the description column is so long that R console does not display the full description.

Meanwhile, the average hourly salary is 44.5 dollars, which is higher than other kinds of jobs. I was wondering if extreme cases in Upwork influences the average hourly salary. So, I calculated the average salary again, excluding Upwork posted jobs.

```
hourly %>%
  filter(via!="via Upwork") %>%
  summarise(average_hourly_salary=mean(c(min_salary,max_salary),na.rm=TRUE))
```

```
## # A tibble: 1 × 1
##   average_hourly_salary
##                   <dbl>
## 1                  44.5
```

The average salary, excluding Upwork postings is 50.8, which higher than 44.5. This is a surprising result to me. I thought extremely high salaries lifts up the overall average salary. However, the opposite was the case.

Yearly salary summary

```
#Filtering yearly salary data
yearly<-jobs_interest %>%
  filter(salary != "NA") %>%
  filter(grepl(' a year', salary))

#cleaning the data
yearly$salary<-gsub(" a year","",yearly$salary)
yearly<-yearly %>%
  separate(salary,into=c("min_salary","max_salary"),sep="-",fill="right")

yearly
```

```
## # A tibble: 1,786 × 7
##    company_name           location via   schedule_type work_from_home min_salary
##    <chr>                  <chr>    <chr> <chr>         <lgl>          <chr>
##  1 Charles River Laborat… United … Inde… Full-time     NA             65K
##  2 Charles River Laborat… United … Inde… Full-time     NA             65K
##  3 Charles River Laborat… United … Inde… Full-time     NA             65K
##  4 Progressive            Anywhere Inde… Full-time     TRUE           56.3K
##  5 PSCU                   United … FOX … Full-time     NA             61.7K
##  6 Redaptive, Inc.        United … Ai-J… Full-time     NA             45,360
##  7 Glocomms               Anywhere Link… Full-time     TRUE           160K
##  8 Bayforce               Anywhere Link… Full-time     TRUE           45K
##  9 Sezzle                 Anywhere Inde… Full-time     TRUE           75K
## 10 Bosch Group            United … Ai-J… Full-time     NA             92,350
## # ℹ 1,776 more rows
## # ℹ 1 more variable: max_salary <chr>
```

```
quickfun <- function(x){
    yy <- readr::parse_number(x)
    ifelse(stringr::str_detect(x, "K"), yy*1e3, yy)
}
yearly<-yearly %>%
  mutate(across(c(min_salary, max_salary), ~quickfun(.x)))

yearly
```

```
## # A tibble: 1,786 × 7
##    company_name           location via   schedule_type work_from_home min_salary
##    <chr>                  <chr>    <chr> <chr>         <lgl>                <dbl>
##  1 Charles River Laborat… United … Inde… Full-time     NA                   65000
##  2 Charles River Laborat… United … Inde… Full-time     NA                   65000
##  3 Charles River Laborat… United … Inde… Full-time     NA                   65000
##  4 Progressive            Anywhere Inde… Full-time     TRUE                 56300
##  5 PSCU                   United … FOX … Full-time     NA                   61700
##  6 Redaptive, Inc.        United … Ai-J… Full-time     NA                   45360
##  7 Glocomms               Anywhere Link… Full-time     TRUE                160000
##  8 Bayforce               Anywhere Link… Full-time     TRUE                 45000
##  9 Sezzle                 Anywhere Inde… Full-time     TRUE                 75000
## 10 Bosch Group            United … Ai-J… Full-time     NA                   92350
## # ℹ 1,776 more rows
## # ℹ 1 more variable: max_salary <dbl>
```

```
#Average yearly salary
yearly %>%
  summarise(average_yearly_salary=mean(c(min_salary,max_salary),na.rm=TRUE))
```

```
## # A tibble: 1 × 1
##   average_yearly_salary
##                   <dbl>
## 1               101664.
```

```
#Maximum yearly salary
max(yearly$max_salary,na.rm=TRUE)
```

```
## [1] 283000
```

```
#Minimum yearly salary
min(yearly$min_salary,na.rm=TRUE)
```

```
## [1] 27519.63
```

When it comes to yearly salaries, the average salary is 101,664 dollars. The maximum is 283,000 dollars. The minumum is 275,19 dollars. These numbers are about what I expected. Since Upwork projects do not suggest “yearly” salaries, I believe these numbers show regular jobs’ data.

Reference

Data source(from Kaggle): https://www.kaggle.com/datasets/lukebarousse/data-analyst-job-postings-google-search (https://www.kaggle.com/datasets/lukebarousse/data-analyst-job-postings-google-search) Orinigal data souce: https://storage.googleapis.com/gsearch_share/gsearch_jobs.csv (https://storage.googleapis.com/gsearch_share/gsearch_jobs.csv)

On the original Kaggle posting, the author wrote that the data collection started on November 4th, 2022, and adds ~100 new job postings to this data set daily. I downloaded this document on July 31th 2023.