Data 608

Story 4

Fredrick Jones

2024-03-17

Data source:

The salary data for this story was taken from kaggle which is available at the link:

https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023?resource=download

More information about the wages and employment can be found at U.S. Bureau of Labour Statistics: https://www.bls.gov/oes/current/oes_nat.htm

Load the required library

```
library(ggplot2)
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 lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- 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
```

Load the dataset

```
df <- read.csv("ds_salaries.csv")
head(df)
```

```
##   work_year experience_level employment_type                job_title salary
## 1      2023               SE              FT Principal Data Scientist  80000
## 2      2023               MI              CT              ML Engineer  30000
## 3      2023               MI              CT              ML Engineer  25500
## 4      2023               SE              FT           Data Scientist 175000
## 5      2023               SE              FT           Data Scientist 120000
## 6      2023               SE              FT        Applied Scientist 222200
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             EUR         85847                 ES          100
## 2             USD         30000                 US          100
## 3             USD         25500                 US          100
## 4             USD        175000                 CA          100
## 5             USD        120000                 CA          100
```

```
## 6             USD        222200                  US            0
##   company_location company_size
## 1               ES            L
## 2               US            S
## 3               US            S
## 4               CA            M
## 5               CA            M
## 6               US            L
```

Select the required variables

Considering salary in USD therefore selecting the salary work year, experience_level, job_title, salary_in_usd, company_location, and company_size

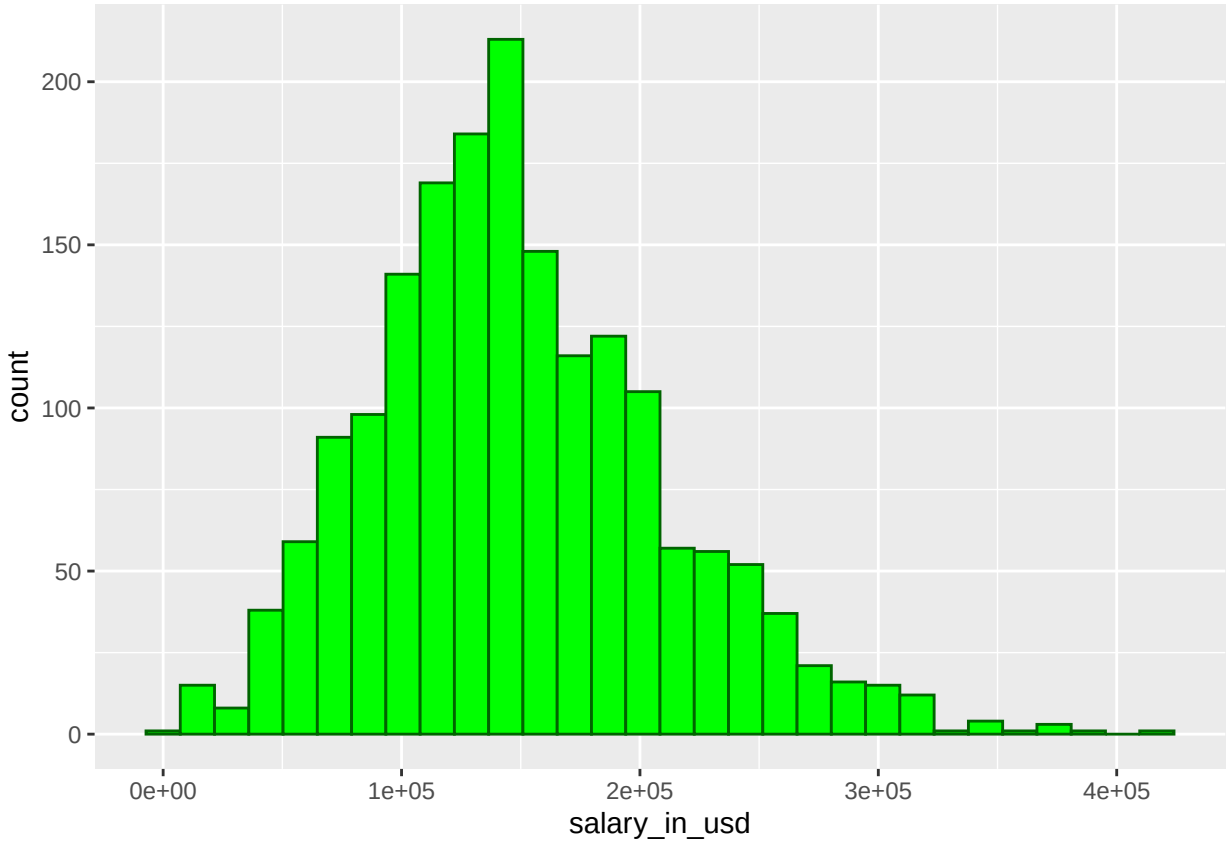
```
df <- df|>select(work_year, experience_level, employment_type, job_title, salary_in_usd, company_locati
head(df)
```

```
##   work_year experience_level employment_type                job_title
## 1      2023               SE              FT Principal Data Scientist
## 2      2023               MI              CT              ML Engineer
## 3      2023               MI              CT              ML Engineer
## 4      2023               SE              FT           Data Scientist
## 5      2023               SE              FT           Data Scientist
## 6      2023               SE              FT        Applied Scientist
##   salary_in_usd company_location company_size
## 1         85847               ES            L
## 2         30000               US            S
## 3         25500               US            S
## 4        175000               CA            M
## 5        120000               CA            M
## 6        222200               US            L
```

Distribution of salaries in 2023

```
df2023<- df|> filter(work_year==2023)
ggplot(df2023, aes(x=salary_in_usd))+
  geom_histogram(col='darkgreen', fill='green')
```

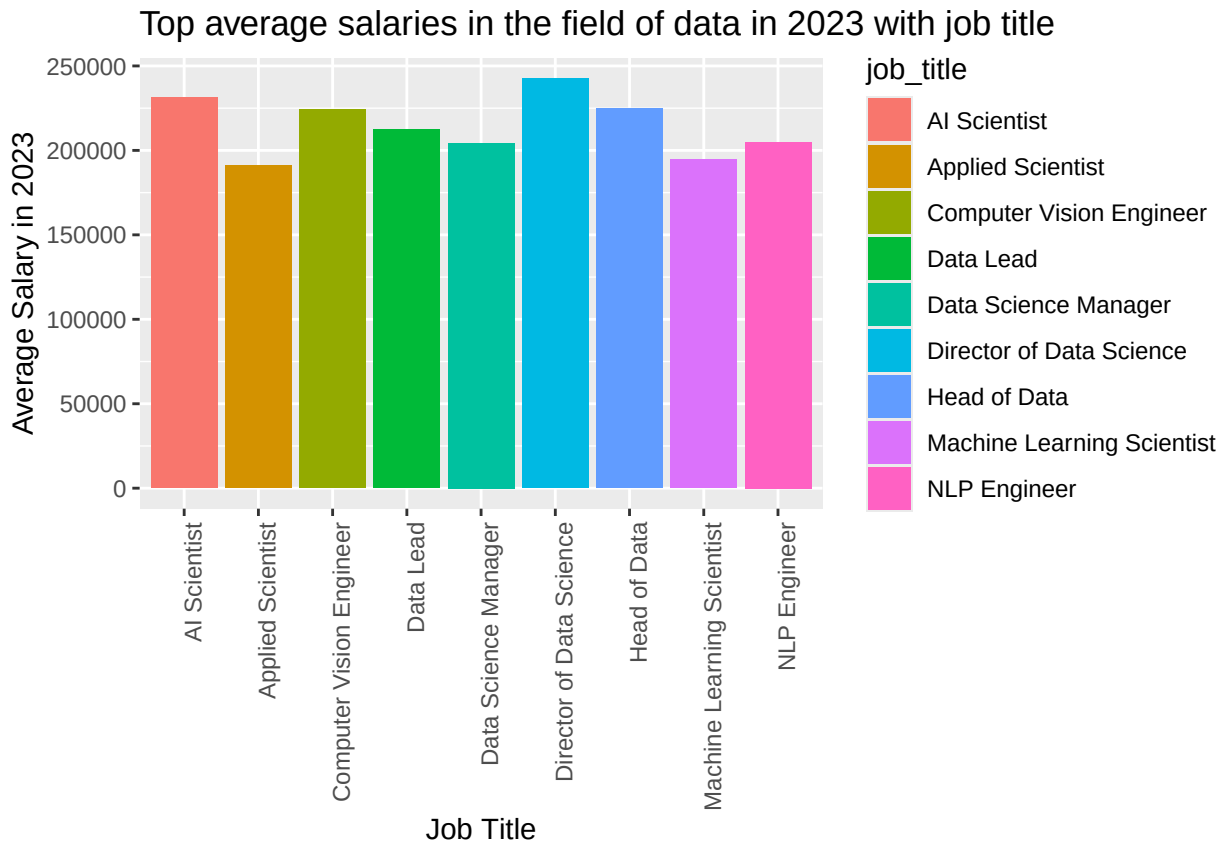
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Mean salary of all the occupations related to data is around 1.5E05 and its range is from 0 to 4E05 usd per year.

Top salaries in 2023

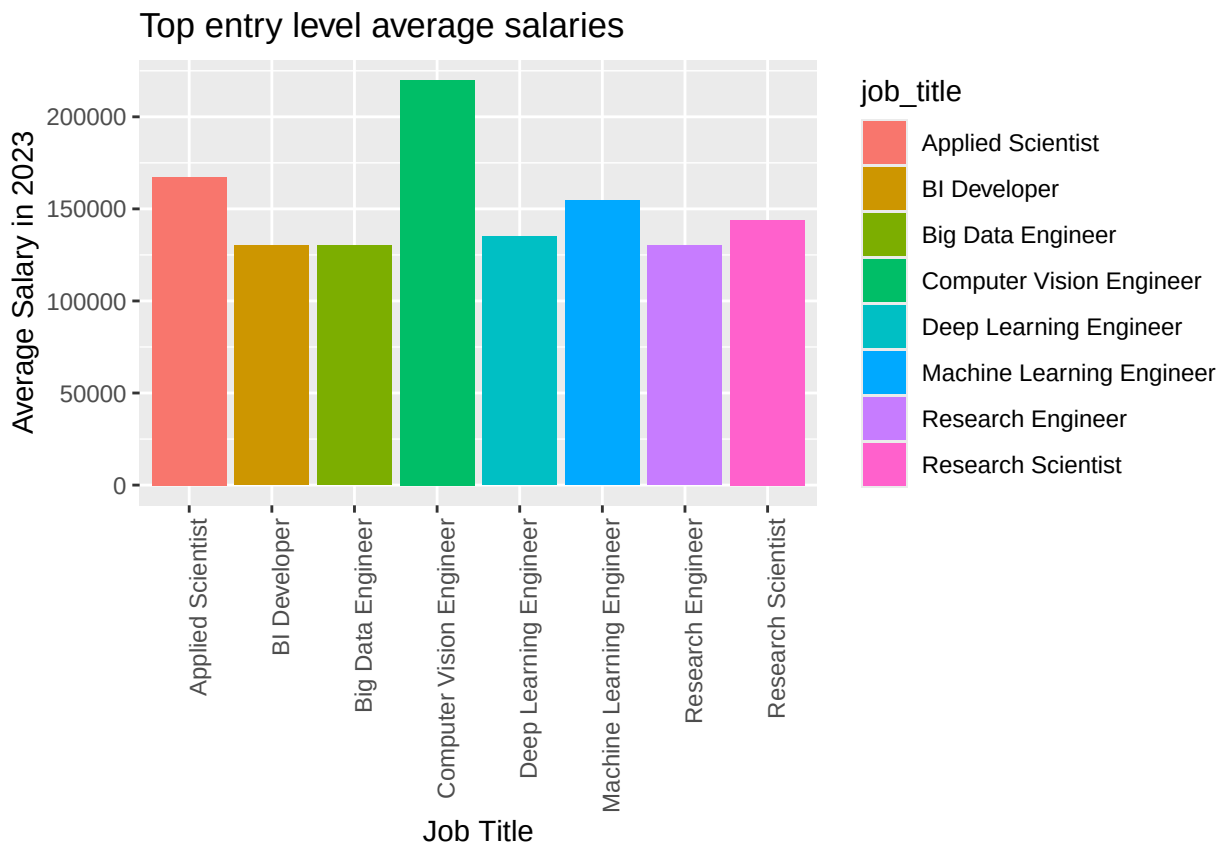
```
df2023|>
  group_by(job_title)|>
  summarize(mean_salary = mean(salary_in_usd))|>
  filter(mean_salary> 1.5*mean(mean_salary))|>
  ggplot(aes(x=job_title, y=mean_salary, fill= job_title))+
    geom_col()+ labs(x="Job Title",
                     y="Average Salary in 2023",
                     title = "Top average salaries in the field of data in 2023 with job title")+
    theme(axis.text.x = element_text(angle=90, hjust=1))
```



It can be seen that the highest average salary was for the position of director of data science and top five fields of highest paying designations are AI Scientist, computer vision engineer, head of data and director of data science.

Best Paying Branch to enter in 2023

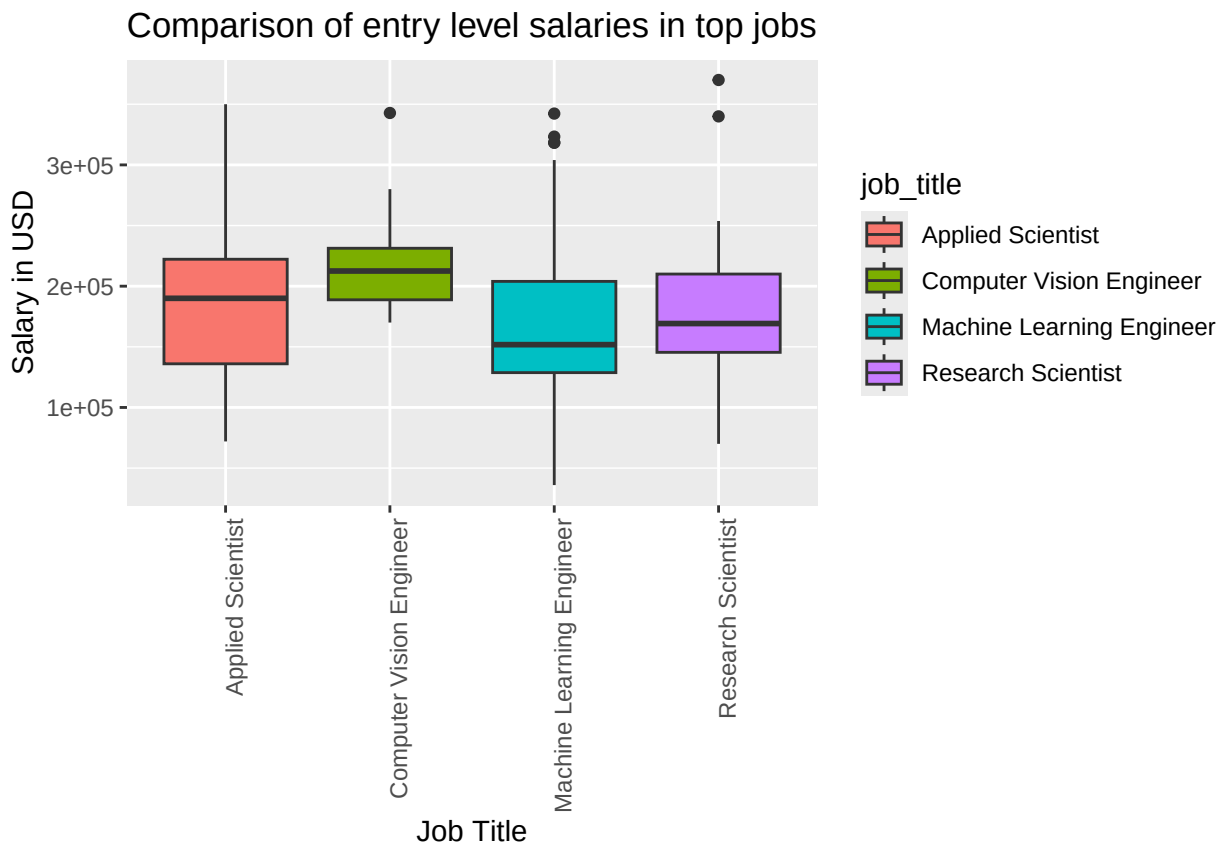
```
df2023|> filter(experience_level=="EN")|>
  group_by(job_title)|>
  summarize(mean_salary = mean(salary_in_usd))|>
  filter(mean_salary> 1.3*mean(mean_salary))|>
  ggplot(aes(x=job_title, y=mean_salary, fill= job_title))+
    geom_col()+ labs(x="Job Title",
                     y="Average Salary in 2023",
                     title = "Top entry level average salaries")+
    theme(axis.text.x = element_text(angle=90, hjust=1))
```



It can be seen that the top high paying entry level jobs in the field of data are Computer Vision Engineer and applied scientists.

Let's look at the range of salaries for these des'ignations.

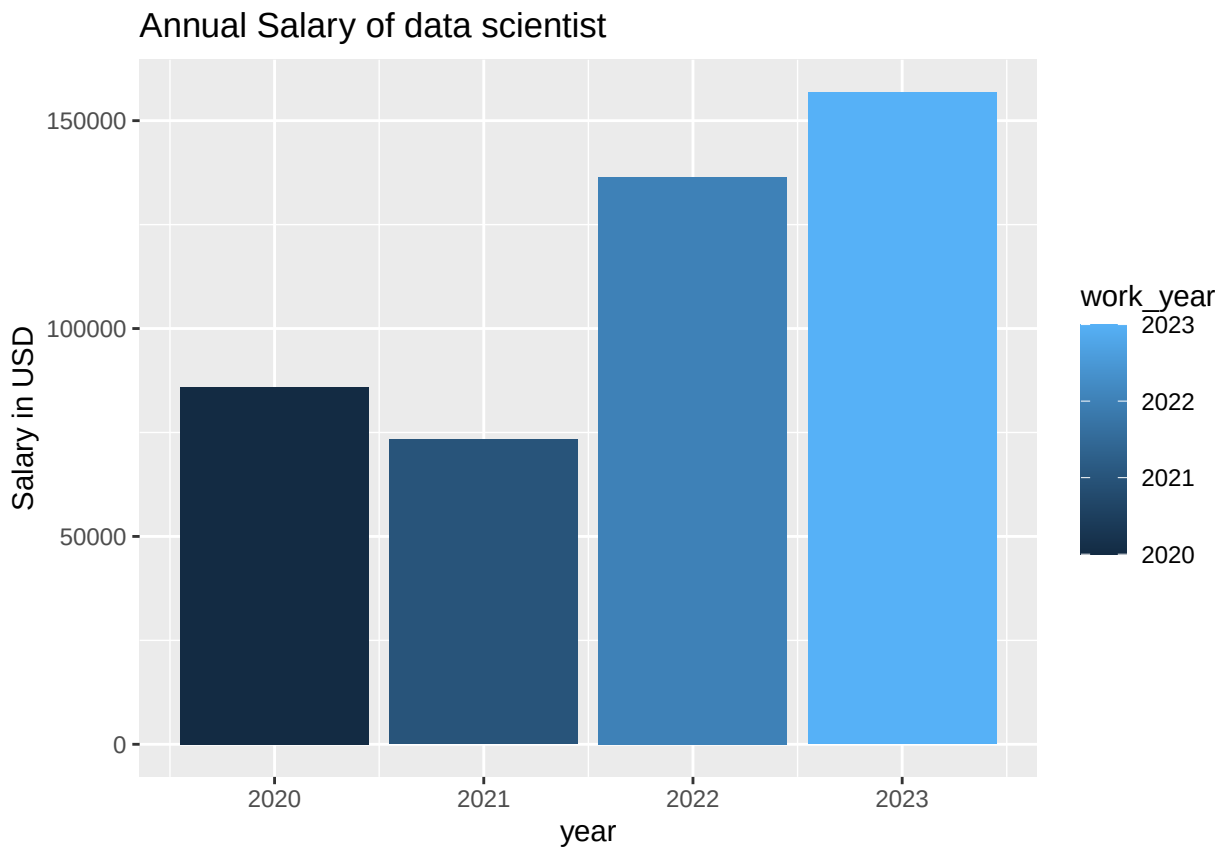
```
top_jobs <- c("Applied Scientist", "Computer Vision Engineer", "Machine Learning Engineer", "Research S
df2023|> filter (job_title %in% top_jobs )|>
  ggplot(aes(x=job_title, y=salary_in_usd, fill= job_title))+
    geom_boxplot()+ labs(
      x= "Job Title",
      y = "Salary in USD",
      title = "Comparison of entry level salaries in top jobs"
    )+theme(axis.text.x = element_text(angle=90, hjust=1))
```



It can be seen that the computer vision engineers are those who got highest paying entry level jobs.

Mean salary range of data scientist according to the year

```
df|> filter (job_title=="Data Scientist")|>
     group_by(work_year)|>
     summarize( mean_salary = mean(salary_in_usd))|>
     ggplot(aes(x= work_year, y=mean_salary, fill=work_year))+
        geom_col()+labs(
          x= "year",
          y="Salary in USD",
          title = "Annual Salary of data scientist"
        )
```


Annual Salary of data scientist
Salary in USD
150000
100000
50000
0
2020
2021
2022
2023
year
work_year
2023
2022
2021
2020