Data Scientist Salaries

Aksana Sutyrka

2023-08-26

Installing packages

```
library ("tidyverse")
library ("here")
library ("ggplot2")
library ("janitor")
library ("qrng")
library ("RColorBrewer")
```

Setting my favourite plot theme

```
theme_set (theme_classic())
```

Importing the dataset

```
ds_sal <- read.csv ("Latest_Data_Science_Salaries.csv")
```

Problem Statement/Questions to answer:

- What are top 10 company locations with the highest data scientist salaries: for entry level, for senior/executive?
- How does salary change from 2020 to 2023?
- DS salaries vs experience level
- DS salaries vs employee residence
- DS salaries vs company size
- What is the demand of small, medium and large companies in specialists of different level?

Exploring the dataset

```
glimpse (ds_sal)
```

```
## Rows: 3,300
## Columns: 11
## $ Job.Title          <chr> "Data Engineer", "Data Engineer", "Data Engineer", ~
## $ Employment.Type    <chr> "Full-Time", "Full-Time", "Full-Time", "Full-Time",~
## $ Experience.Level   <chr> "Senior", "Senior", "Senior", "Senior", "Senior", "~
## $ Expertise.Level    <chr> "Expert", "Expert", "Expert", "Expert", "Expert", "~
## $ Salary             <int> 210000, 165000, 185900, 129300, 140000, 126000, 170~
## $ Salary.Currency    <chr> "United States Dollar", "United States Dollar", "Un~
## $ Company.Location   <chr> "United States", "United States", "United States", ~
## $ Salary.in.USD      <int> 210000, 165000, 185900, 129300, 140000, 126000, 170~
```

```
## $ Employee.Residence <chr> "United States", "United States", "United States", ~
## $ Company.Size       <chr> "Medium", "Medium", "Medium", "Medium", "Medium", "~
## $ Year               <int> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202~
```

Cleaning the dataset

We don't need columns with salary not in USD, so we remove them.

Removing unnecessary columns ("Salary" and "Salary currency")

```
ds_sal <- ds_sal [,-c (5,6)]
```

Creating factors

```
ds_sal <- ds_sal%>%
   mutate_if(sapply(ds_sal, is.character), as.factor)
glimpse (ds_sal)
```

```
## Rows: 3,300
## Columns: 9
## $ Job.Title          <fct> Data Engineer, Data Engineer, Data Engineer, Data E~
## $ Employment.Type    <fct> Full-Time, Full-Time, Full-Time, Full-Time, Full-Ti~
## $ Experience.Level   <fct> Senior, Senior, Senior, Senior, Senior, Senior, Sen~
## $ Expertise.Level    <fct> Expert, Expert, Expert, Expert, Expert, Expert, Exp~
## $ Company.Location   <fct> "United States", "United States", "United States", ~
## $ Salary.in.USD      <int> 210000, 165000, 185900, 129300, 140000, 126000, 170~
## $ Employee.Residence <fct> "United States", "United States", "United States", ~
## $ Company.Size       <fct> Medium, Medium, Medium, Medium, Medium, Medium, Med~
## $ Year               <int> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202~
```

```
summary (ds_sal)
```

```
##                      Job.Title    Employment.Type  Experience.Level
##  Data Engineer            :702   Contract :  15   Entry    : 292
##  Data Scientist           :635   Freelance:  11   Executive: 146
##  Data Analyst             :459   Full-Time:3261   Mid      : 797
##  Machine Learning Engineer:300   Part-Time:  13   Senior   :2065
##  Analytics Engineer       :132
##  Research Scientist       :104
##  (Other)                  :968
##      Expertise.Level       Company.Location Salary.in.USD
##  Director    : 146   United States :2495    Min.   : 15000
##  Expert      :2065   United Kingdom: 251    1st Qu.: 90000
##  Intermediate: 797   Canada        : 104    Median :136000
##  Junior      : 292   Germany       :  65    Mean   :142096
##                      Spain         :  47    3rd Qu.:185000
##                      India         :  44    Max.   :450000
##                      (Other)       : 294
##       Employee.Residence Company.Size       Year
##  United States :2453     Large : 442   Min.   :2020
##  United Kingdom: 245     Medium:2707   1st Qu.:2022
##  Canada        : 101     Small : 151   Median :2023
##  Germany       :  58                   Mean   :2022
##  India         :  57                   3rd Qu.:2023
##  Spain         :  50                   Max.   :2023
```

```
##  (Other)       : 336
```

We see, that most of our employees have full-time job, have senior experience level, work for medium companies in the US.

Renaming columns (cleaning names)

```
ds_sal <- clean_names (ds_sal)
```

It would be more convenient to express annual salaries in terms of thousands of dollars.

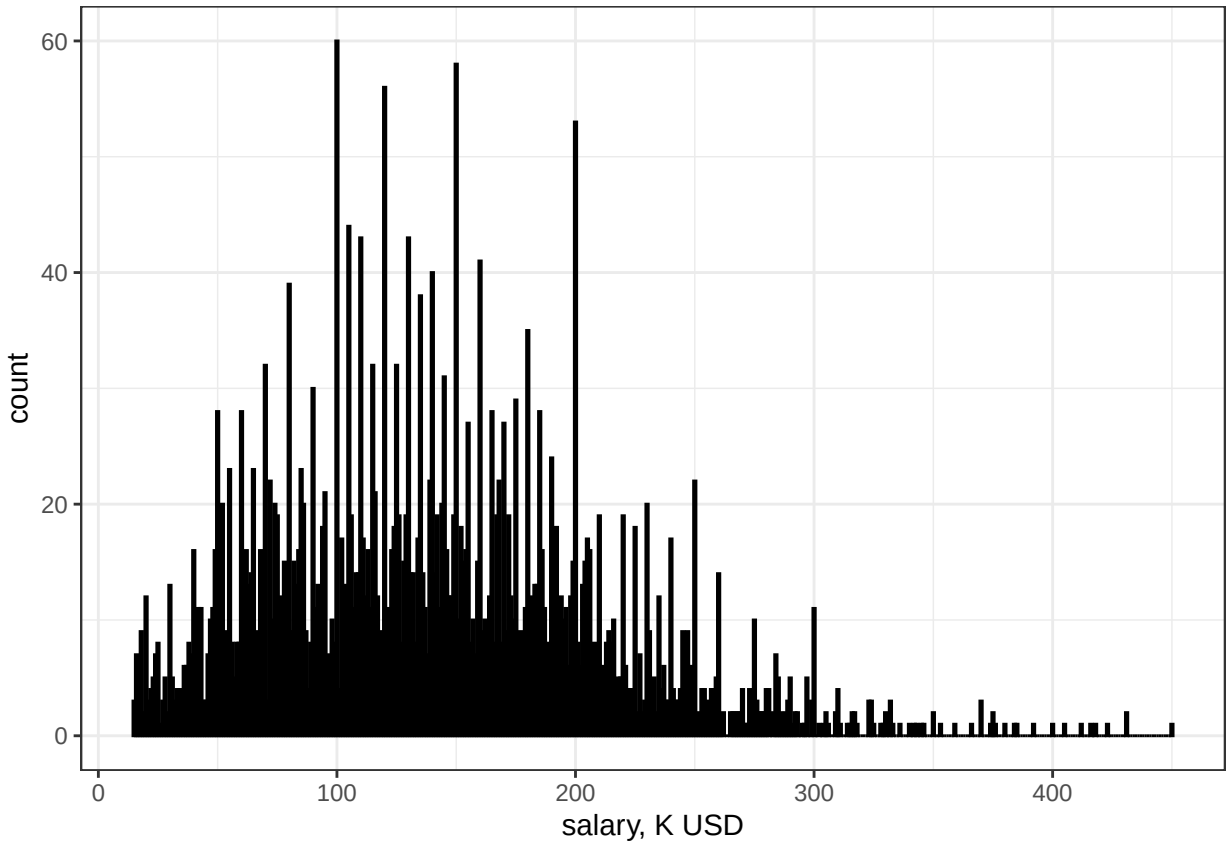
Converting salaries to K

```
ds_sal <- ds_sal %>%
  mutate (salary_k = round (salary_in_usd/1000,1))
```

Analysing the dataset. Answering the questions.

How salaries are distributed?

```
ds_sal%>%
    ggplot( aes(x = salary_k)) +
  geom_histogram(binwidth = 1,color = "#000000", fill = "#0099F8", alpha = 0.6) +
theme_bw()+ labs (x="salary, K USD")
```



We have a lot of outliers, so we need to use median for our dataset.

DS salaries vs experience level.

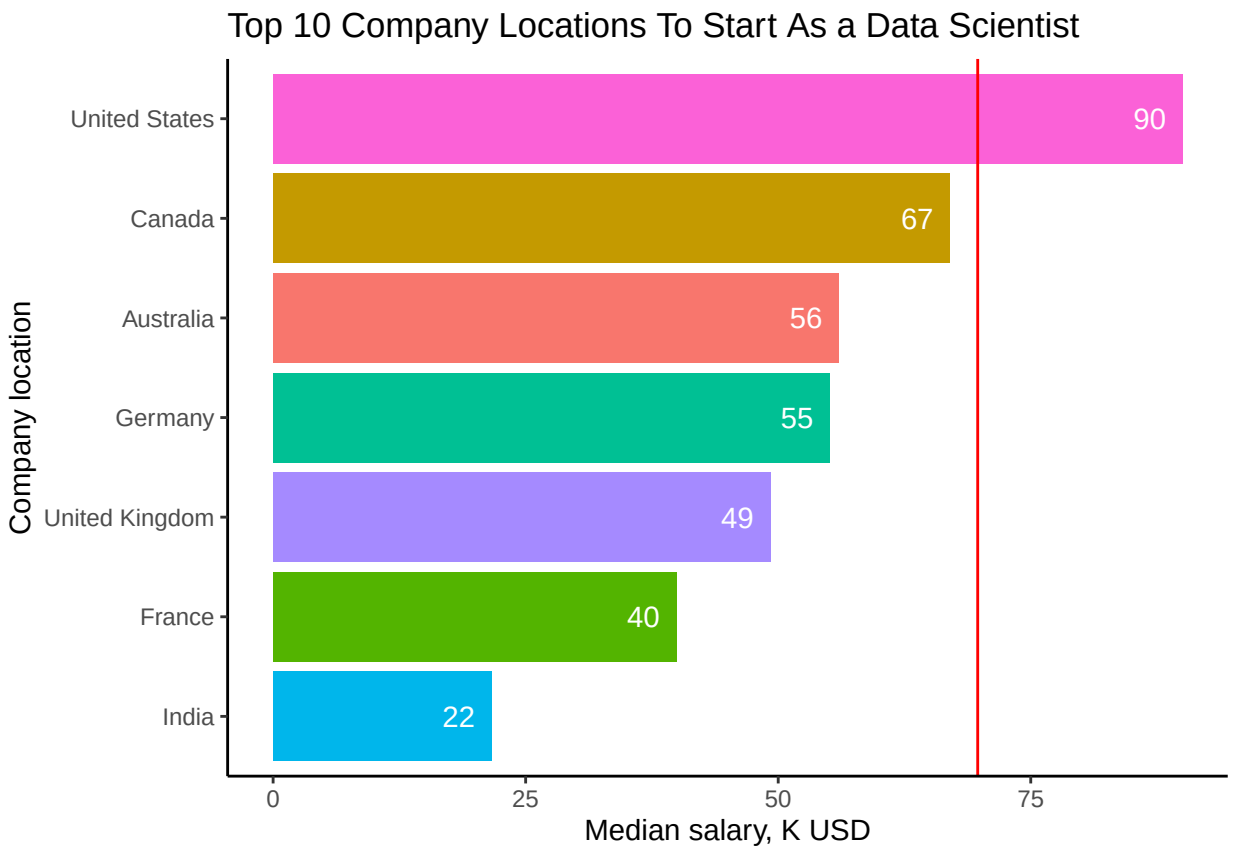
```
ds_sal %>%
  group_by (experience_level) %>%
    summarise (median_sal = median (salary_k)) %>%
    ggplot (aes (x = reorder(experience_level,+median_sal), y=median_sal, fill = experience_level)) +
    geom_hline(yintercept = median (ds_sal$salary_k), color = 'red')+
  labs (title = "Data Scientist Salaries vs Experience Level",x = "experience level", y="median salary,
  theme(legend.position = "none")+
  geom_text (aes(label = round (median_sal,0)), vjust = 1.5, color = "white")
```

What are top 10 company locations to start as a data scientist?

We will filter only those company locations, where we have more than 5 values for entry level.

```
ds_sal_top_entry <- ds_sal %>%
  filter (experience_level=="Entry") %>%
  group_by (company_location) %>%
  filter (n()>5) %>%
  summarize (median_salary = median (salary_k)) %>%
  arrange (-median_salary) %>%
  top_n (10, median_salary)

ds_sal_entry <- ds_sal %>%
  filter (experience_level == "Entry")

ds_sal_top_entry %>%
```

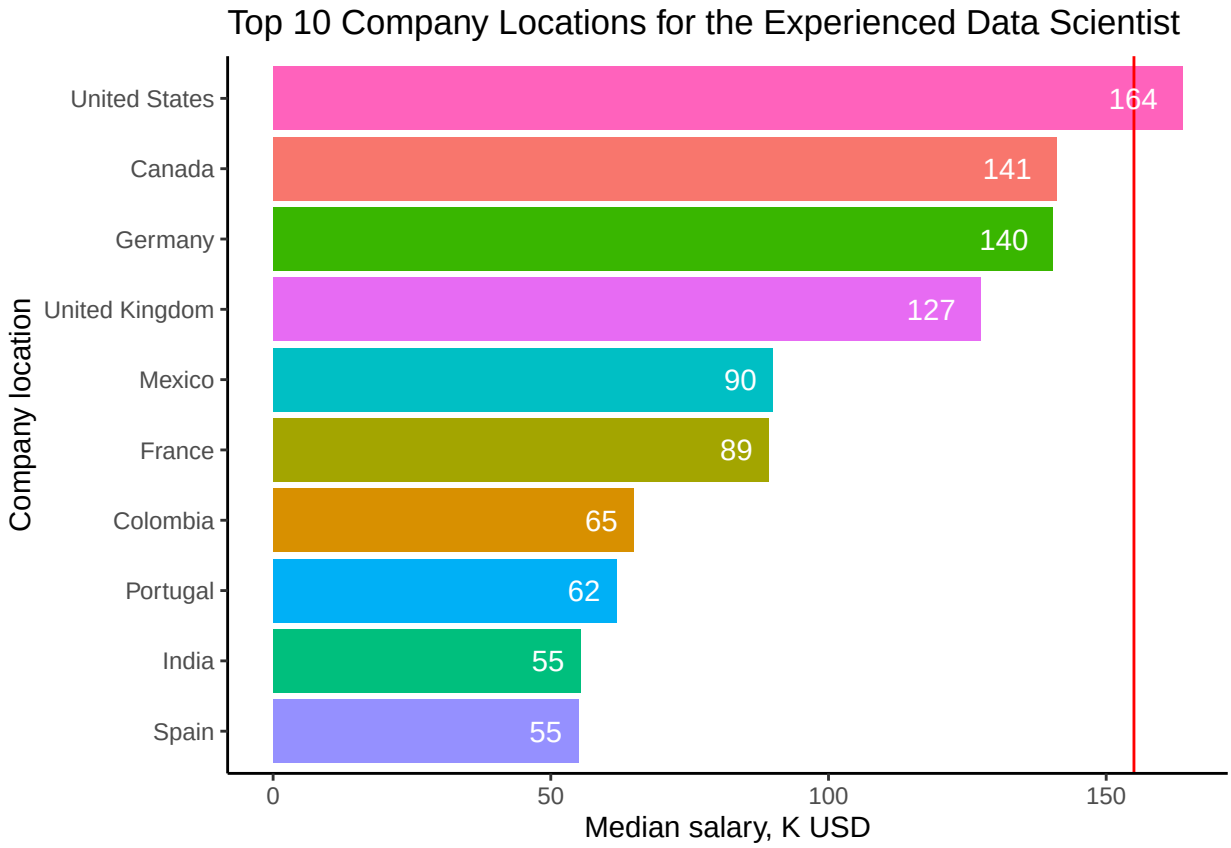
```
ggplot (aes (x=reorder(company_location, +median_salary), y= median_salary, fill = company_location)) +
  geom_col ()+
  geom_hline(yintercept = median (ds_sal_entry$salary_k), color = 'red')+
  coord_flip()+
  labs (title = "Top 10 Company Locations To Start As a Data Scientist",
  y= "Median salary, K USD",
  x= "Company location")+
  theme(legend.position = "none")+
  geom_text (aes(label = round (median_salary,0)), hjust = 1.5, color = "white")
```



What are 10 top company locations to work as a senior/executive specialist?

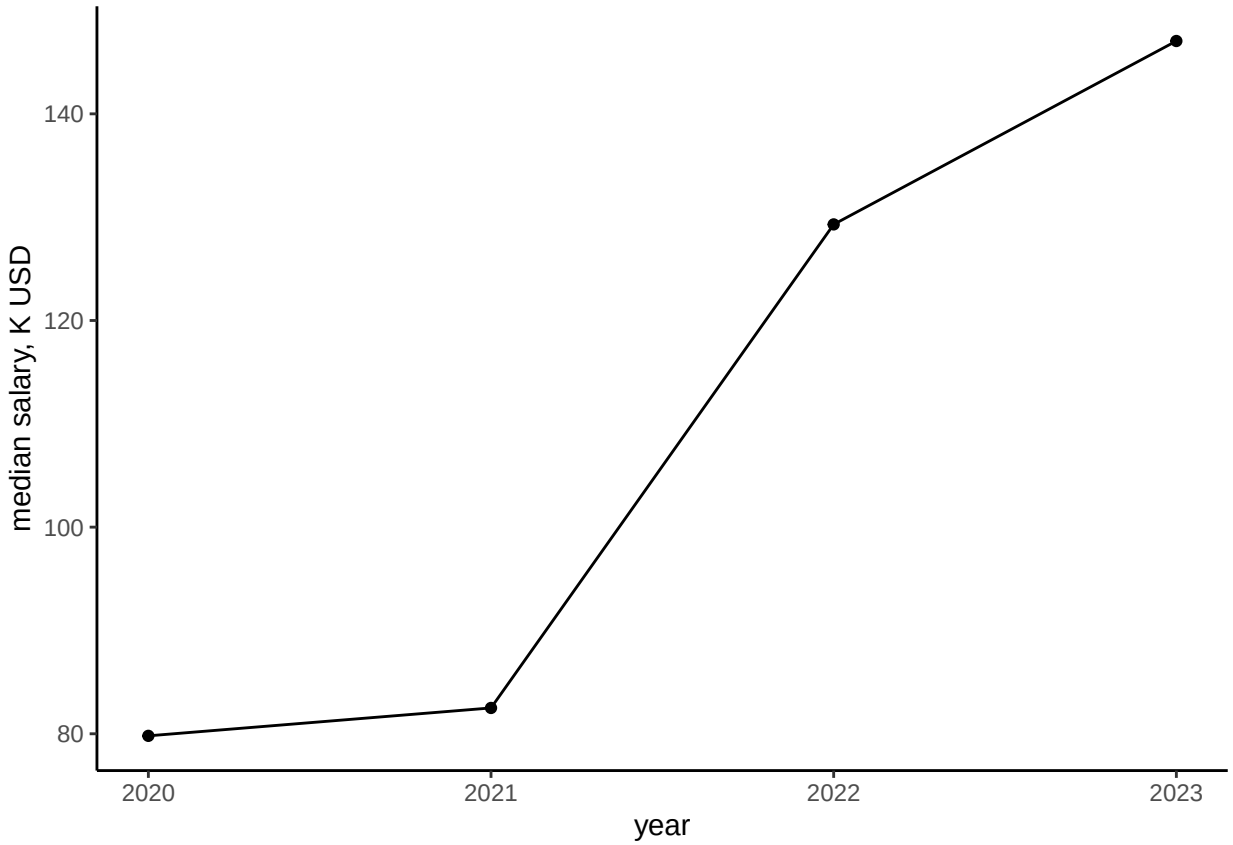
```
ds_sal_top_senior <- ds_sal %>%
  filter (experience_level=="Senior"| experience_level == "Executive") %>%
  group_by (company_location) %>%
  filter (n()>5) %>%
  summarize (median_salary = median (salary_k)) %>%
  arrange (-median_salary) %>%
  top_n (10, median_salary)

ds_sal_senior <- ds_sal %>%
  filter (experience_level == "Senior"| experience_level== "Executive")

ds_sal_top_senior %>%
ggplot (aes (x=reorder(company_location,+median_salary), y= median_salary, fill = company_location))+
geom_col ()+
```

```
  geom_hline(yintercept = median (ds_sal_senior$salary_k), color = 'red')+
    coord_flip()+
          labs (title = "Top 10 Company Locations for the Experienced Data Scientist",
y= "Median salary, K USD",
x= "Company location")+
  theme(legend.position = "none")+
  geom_text (aes(label = round (median_salary,0)), hjust = 1.5, color = "white")
```



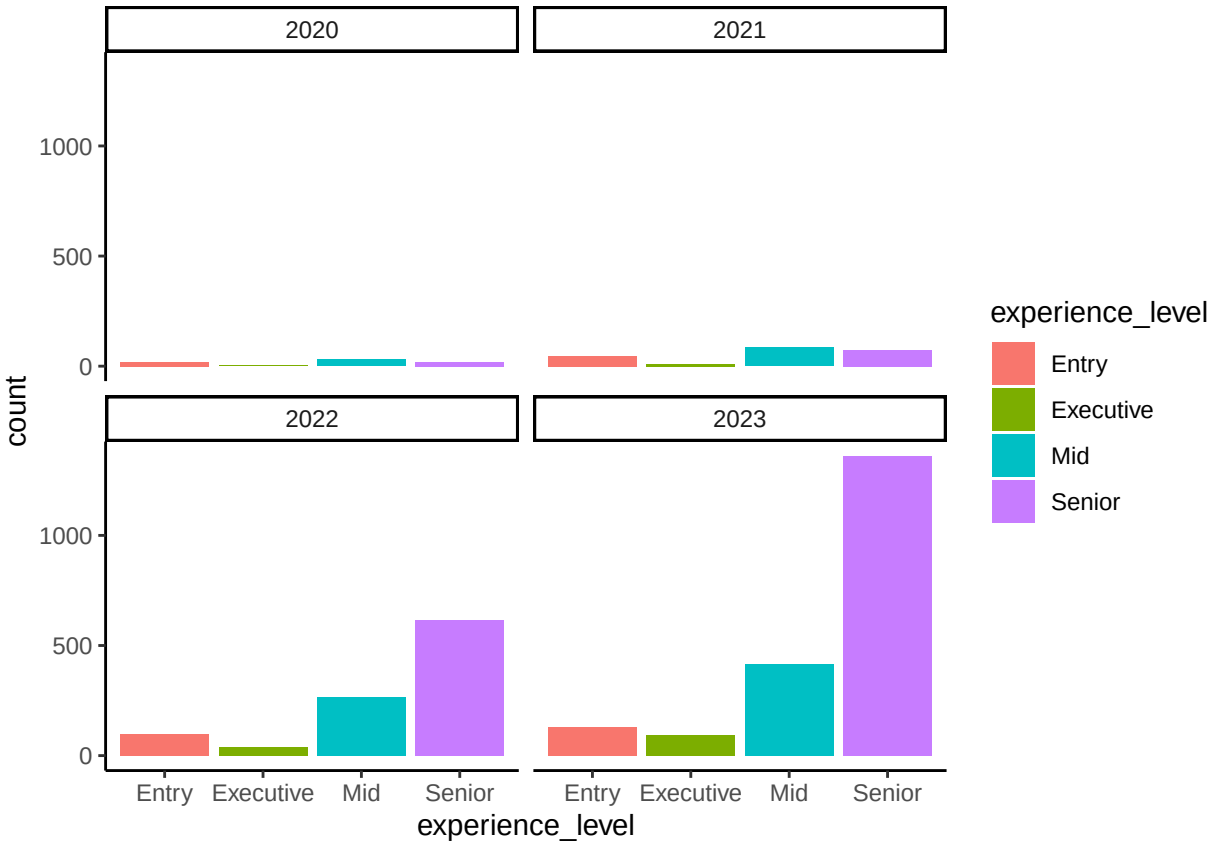
What are the salary trends?

```
ds_sal %>%
  group_by(year) %>%
  summarize (median_salary = median (salary_k)) %>%
  ggplot (aes (x= year, y = median_salary)) + geom_point ()+
  geom_line()+ labs (y="median salary, K USD")
```



There's no significant difference between years 2020 and 2021 and there's an increase in the years 2022-2023. We'll check if this increase was due to higher proportion of senior data scientists in the year 2023.

```
ds_sal %>%
      ggplot (aes (x = experience_level, fill = experience_level)) + geom_bar()+
      facet_wrap (~year)
```



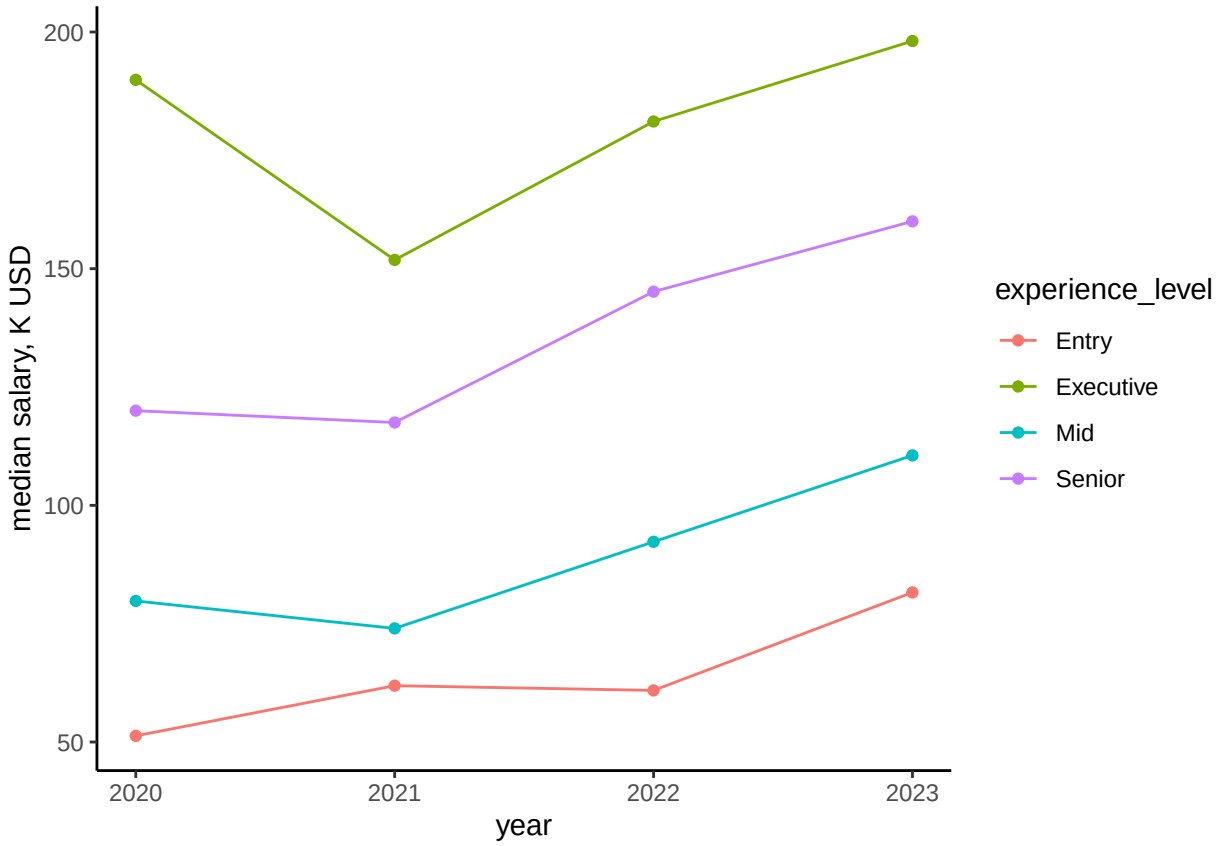

```
   labs (title = "Data Scientists Experience Level Proportion from 2020 to 2023",x = "experience level"
```

```
## $x
## [1] "experience level"
##
## $y
## [1] "count"
##
## $title
## [1] "Data Scientists Experience Level Proportion from 2020 to 2023"
##
## attr(,"class")
## [1] "labels"
```

The hypothesis was proven to be true: we can see higher proportion of senior data scientists in the year 2023. Let's examine if there is any trends over time for specialists at a specific level.

```
ds_sal %>%
  group_by(year, experience_level) %>%
  summarize (median_salary = median (salary_k)) %>%
  ggplot (aes (x= year, y = median_salary, color=experience_level)) + geom_point ()+
  geom_line()+ labs (y="median salary, K USD")
```

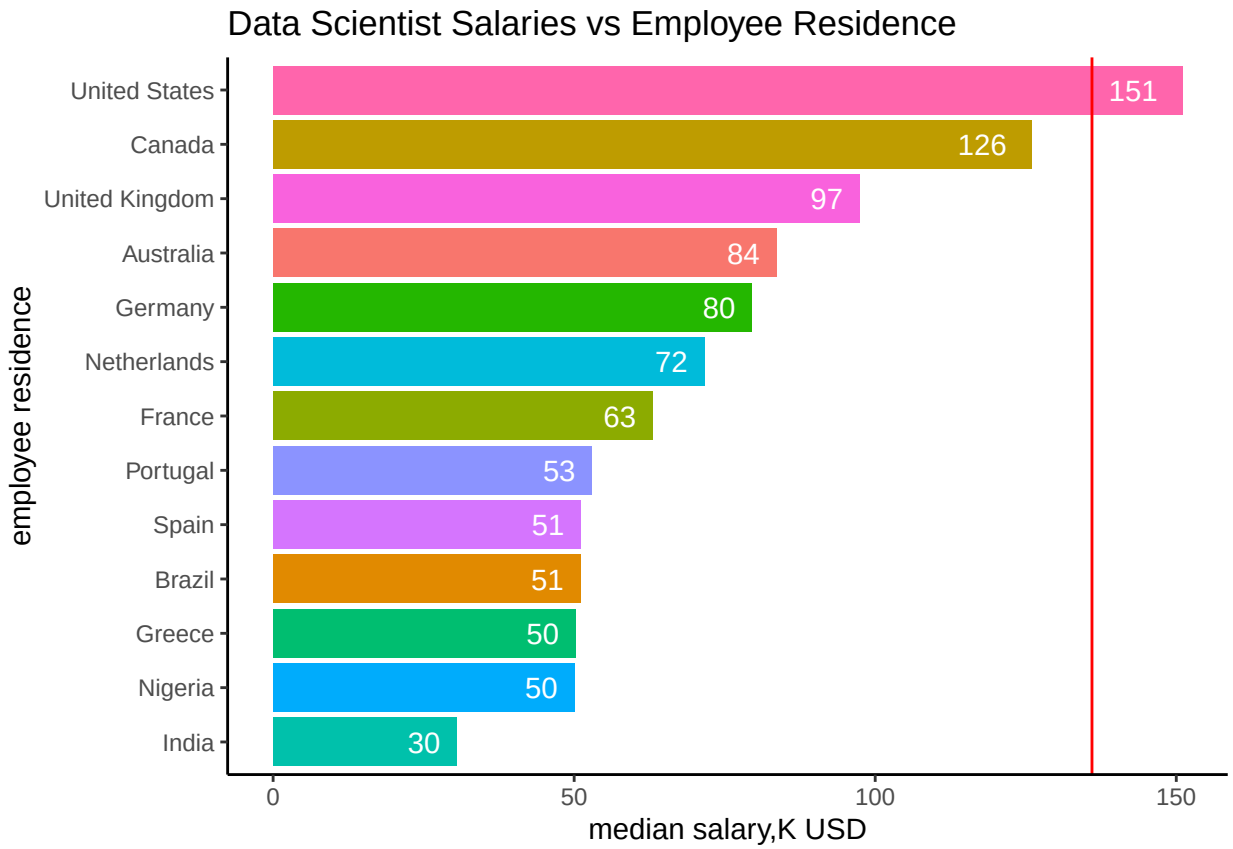
```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```



That's it! The significant salary increase in the data science industry is observed between 2022 and 2023.

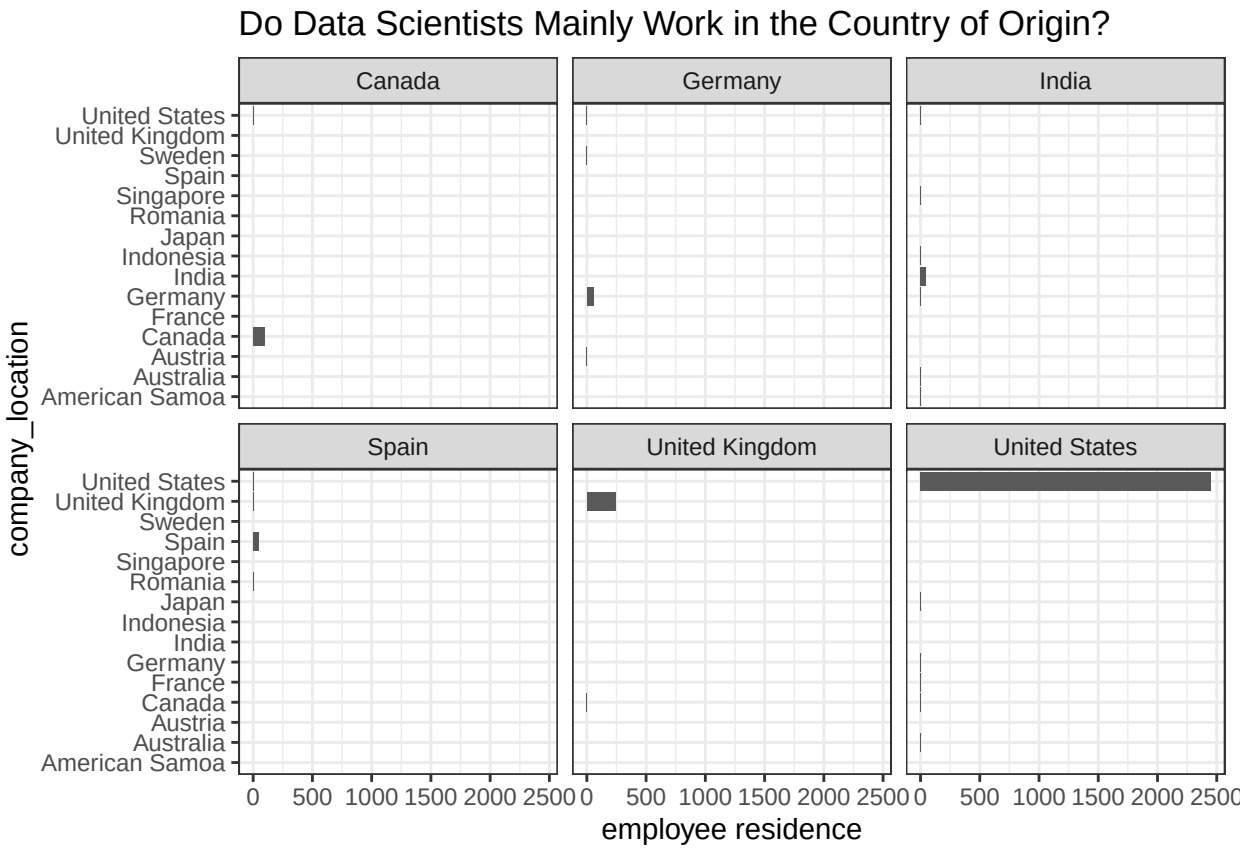
DS salaries vs employee residence

```
ds_sal %>%
  group_by (employee_residence) %>%
  filter (n()>10) %>%
  summarise (median_sal = median (salary_k)) %>%
    ggplot (aes (x = reorder (employee_residence,+median_sal), y=median_sal, fill = employee_residence)
  coord_flip()+
  geom_hline(yintercept = median (ds_sal$salary_k), color = 'red')+
  labs (title = "Data Scientist Salaries vs Employee Residence",x = "employee residence", y="median sala
  theme(legend.position = "none")+
  geom_text (aes(label = round (median_sal,0)), hjust = 1.5, color = "white")
```



Why is it so? The hypothesis is that employees mainly work at the country of origin. So we'll check now.

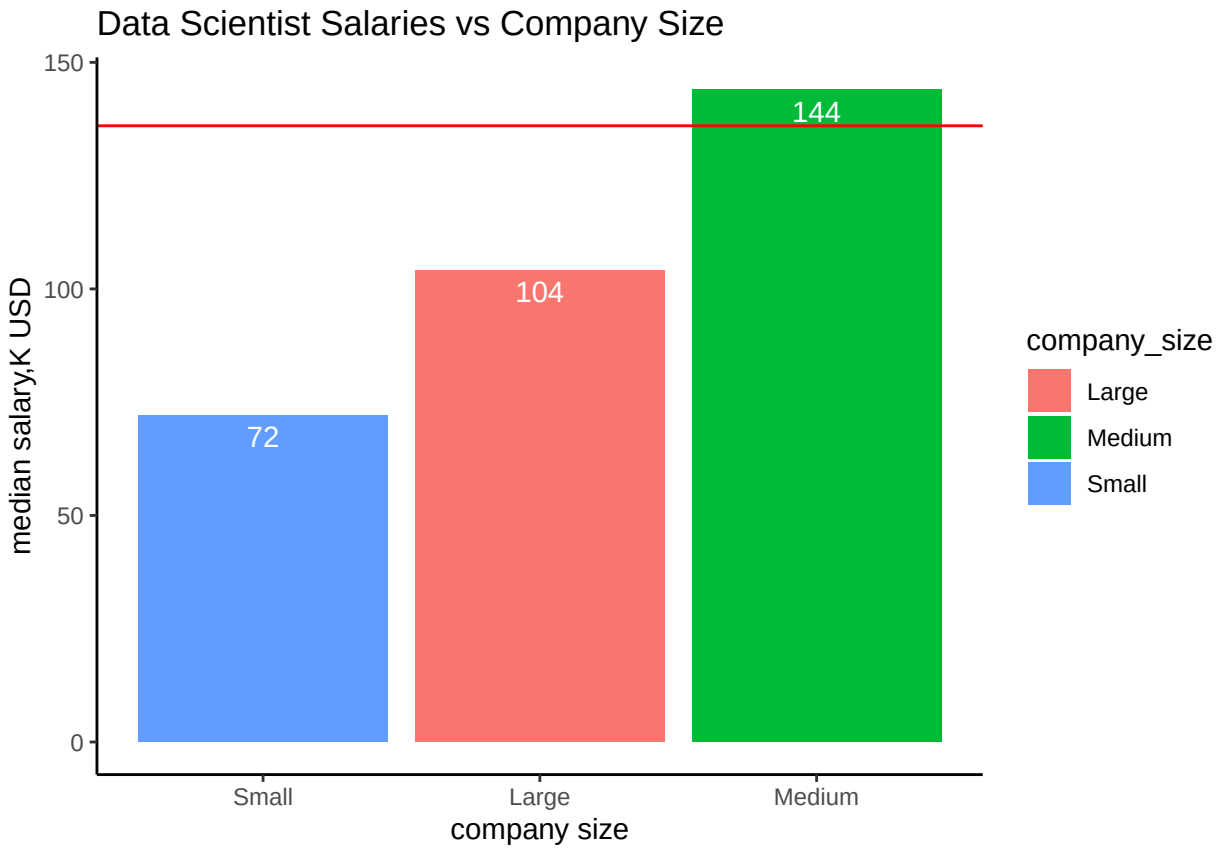
```
ds_sal %>%
filter (employee_residence=="United States"|employee_residence=="Canada"|employee_residence=="United Ki
      ggplot (aes (y= company_location)) + geom_bar()+
      labs (title = "Do Data Scientists Mainly Work in the Country of Origin?",x = "employee residence"
   facet_wrap(~employee_residence)+
  theme_bw()
```



We can see that people from this dataset mainly work in the country of origin, and their salary level correlates with the median salary in that country.

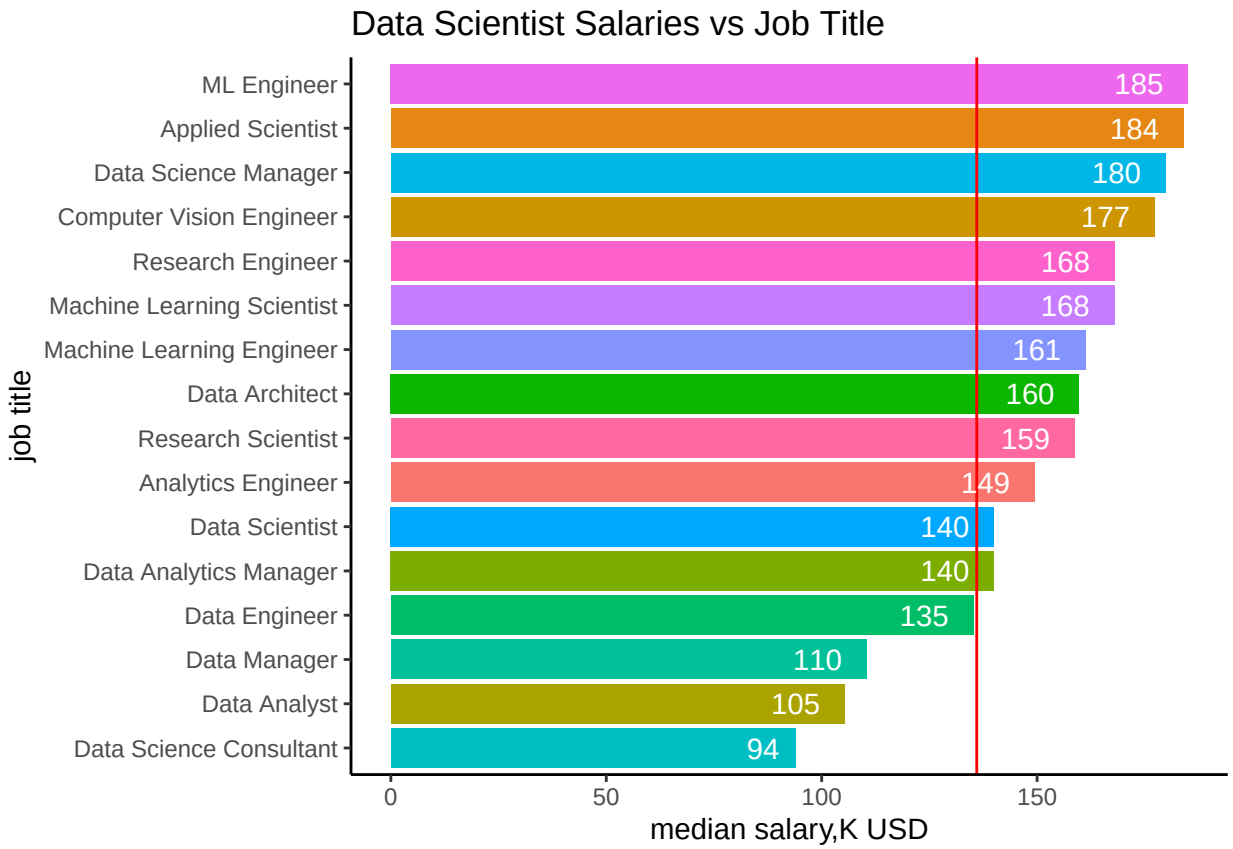
DS salaries vs company size.

```
ds_sal %>%
  group_by (company_size) %>%
    summarise (median_sal = median (salary_k)) %>%
    ggplot (aes (reorder(x = company_size,+median_sal), y=median_sal, fill = company_size)) + geom_col()
  geom_hline(yintercept = median (ds_sal$salary_k), color = 'red')+
      labs (title = "Data Scientist Salaries vs Company Size",x = "company size", y="median salary,K USD
  theme(legend.position = "none")+
  geom_text (aes(label = round (median_sal,0)), vjust = 1.5, color = "white")+
  theme_classic()
```



DS Salaries vs job_title

```
ds_sal %>%
  group_by (job_title) %>%
  filter (n()>20) %>%
  summarise (median_sal = median (salary_k)) %>%
    ggplot (aes (x = reorder (job_title, +median_sal), y=median_sal, fill = job_title)) + geom_col()+
    geom_hline(yintercept = median (ds_sal$salary_k), color = 'red')+coord_flip()+
  labs (title = "Data Scientist Salaries vs Job Title",x = "job title", y="median salary,K USD")+
  theme(legend.position = "none")+
  geom_text (aes(label = round (median_sal,0)), hjust = 1.5, color = "white")
```



Number of workers in companies by size and worker experience

```
ds_sal %>%
  group_by(experience_level, company_size)%>%
  count()%>%
  ggplot(mapping = aes(y = experience_level, x = n, fill = experience_level))+
  geom_bar(stat = 'Identity')+
  facet_wrap(~company_size)+
  theme_bw()+ labs(y = "experience level", x = "number of positions",
                   title = "Number of positions by experience level and company size") +
  theme  (axis.text.x = element_blank())+
  coord_flip()+
     geom_text (aes(label = n), hjust = 0.5, vjust =0.0,color = "black")
```

We can see that medium companies have a much higher proportion of high level data scientists, that's why we can observe higher median salary in such type of companies.