Analysing Data Scientist Salaries

Data Science Presentation

Mibella Villafana

PST107_22021

October 18

Purpose

- These findings will generate an analysis that will guide future data scientists into becoming aware of their career prospects in terms of salary, and may either inspire them or allow them to choose a different career path.
- They also provide geographical findings that can help dictate where getting a data science job may be suitable to their desired lifestyle.

Hypothesis

- We can predict that there will be a significant difference in salaries within the US and globally.
- We will use a t- test to find out if the findings are statistically significant
- Ho- There is no difference in salaries
- H1- There is a difference in salaries within the US versus globally.

Data Set ==ds_salaries

- The table imported is a representation of data scientist and related career salaries including machine learning engineers, data analyst, data engineer, research scientist, etc.
- We compare experience level with salary in USD amongst US populations and globally.
- We utilize scatterplots to visualize the data

Cleaning data: Experience Level

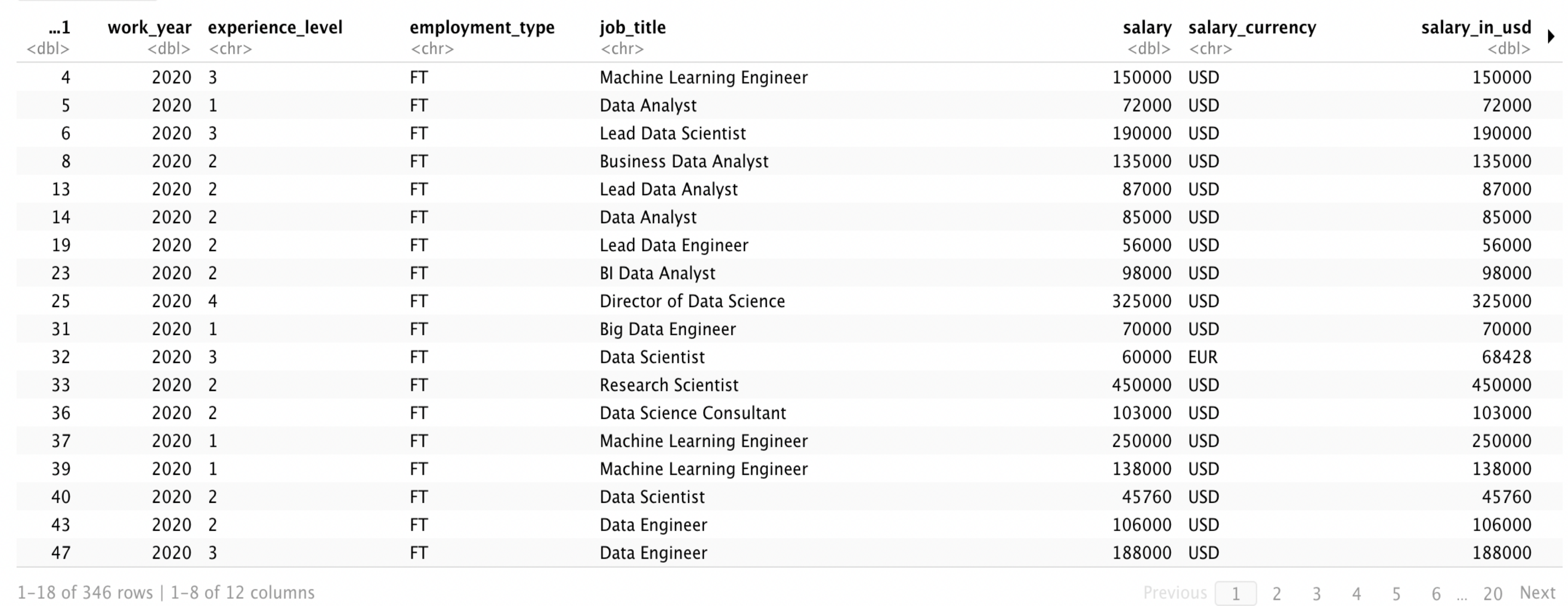
- Used a for loop to switch out character variables of EN,MI, SR etc. to numbers 1,2,3,4 to help with visualisation

```
for(i in 1:length(ds_salaries$experience_level)){
  if ((ds_salaries$experience_level[i]  == "EN")){
    ds_salaries$experience_level[i]<-1
  }
  else if((ds_salaries$experience_level[i]=='MI')){
    ds_salaries$experience_level[i]<-2
  }
  else if ((ds_salaries$experience_level[i]=='SE')){
    ds_salaries$experience_level[i]<-3
  }
  else{
    ds_salaries$experience_level[i]<-4
  }
}
```

FT employees

- created a subset to remove all employees that aren't full time to process less biased salaries

- ds_salaries<-subset(ds_salaries,employment_type=="FT")



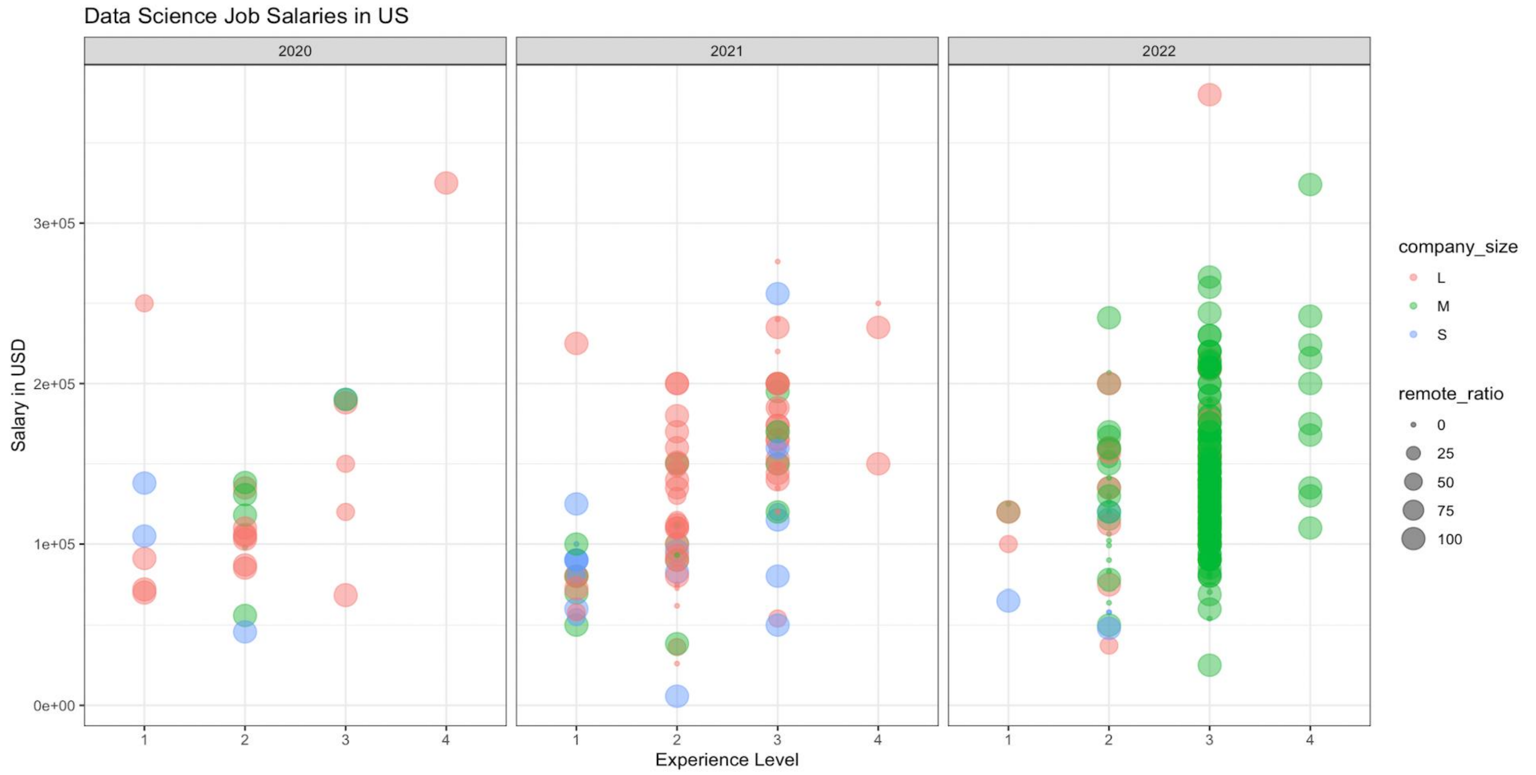
...1	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd
<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<chr>	<dbl>
4	2020	3	FT	Machine Learning Engineer	150000	USD	150000
5	2020	1	FT	Data Analyst	72000	USD	72000
6	2020	3	FT	Lead Data Scientist	190000	USD	190000
8	2020	2	FT	Business Data Analyst	135000	USD	135000
13	2020	2	FT	Lead Data Analyst	87000	USD	87000
14	2020	2	FT	Data Analyst	85000	USD	85000
19	2020	2	FT	Lead Data Engineer	56000	USD	56000
23	2020	2	FT	BI Data Analyst	98000	USD	98000
25	2020	4	FT	Director of Data Science	325000	USD	325000
31	2020	1	FT	Big Data Engineer	70000	USD	70000
32	2020	3	FT	Data Scientist	60000	EUR	68428
33	2020	2	FT	Research Scientist	450000	USD	450000
36	2020	2	FT	Data Science Consultant	103000	USD	103000
37	2020	1	FT	Machine Learning Engineer	250000	USD	250000
39	2020	1	FT	Machine Learning Engineer	138000	USD	138000
40	2020	2	FT	Data Scientist	45760	USD	45760
43	2020	2	FT	Data Engineer	106000	USD	106000
47	2020	3	FT	Data Engineer	188000	USD	188000

1–18 of 346 rows | 1–8 of 12 columns

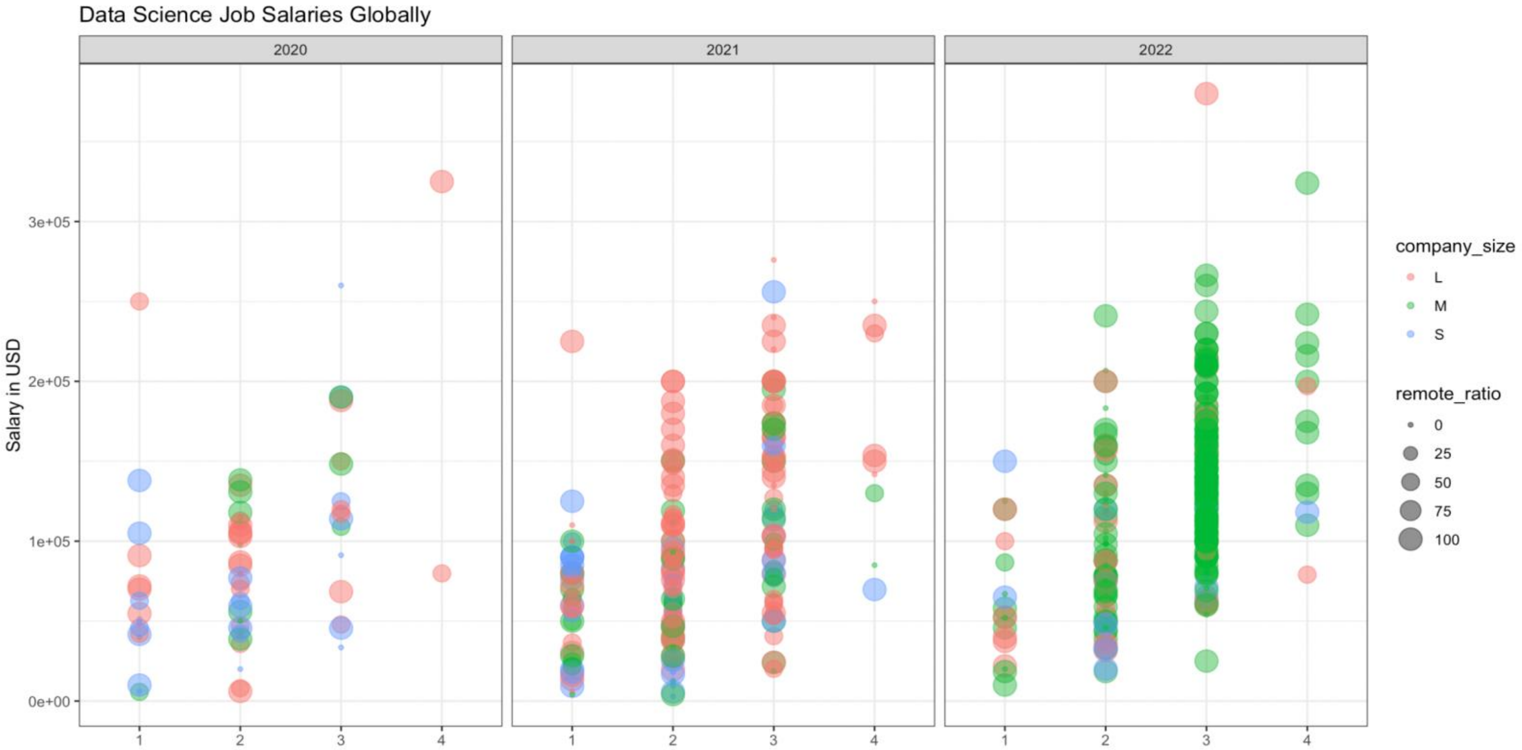
Plotting

- The algorithms used for plotting were fun as they produced an immediate visual result
- We used #ggplot
- ds_salaries %>% #pipe operator to include code below
- filter(salary_in_usd<4e+05) %>% # to remove outliers
- ggplot(aes(experience_level,salary_in_usd))+ #aesthetic to include x and y values
- geom_point(aes(colour=company_size, # to visualize the company size in colour and add values to plot
-         size=remote_ratio),
-       alpha=0.5)+ # to make the dots transparent
- facet_wrap(~work_year, nrow = 1)+ # to divide all 3 graphs by year
- labs(x="Experience Level",
-   y="Salary in USD",
-   title= "Data Science Job Salaries")+ # to label the graph
- theme_bw() #themes change the visuals

Mean salary in US= $144,638



Mean salary globally (including US)= $113,468



Sample T test

- Two smple t- test preferred to compare and return significance between mean salary values within US and globally
- However:
- we have produced a one sample t- test using the code
- ds_salaries %>%
- filter(company_location=="US") %>%
- select(salary_in_usd) %>%
- t.test(mu=144638)

One sample t-test

- data: .
- t = -0.15706, df = 354, p-value = 0.8753
- alternative hypothesis: true mean is not equal to 144638
- 95 percent confidence interval:
- 136758.3 151352.2
- sample estimates:
- mean of x
- 144055.3

Add-ons

- It would be beneficial to calculate and plot the mean salaries of the global population without the US
- We also produced a new table that included only Data Scientist salaries using the subset function
- ds_salaries<-subset(ds_salaries,job_title=="Data Scientist")
- This could lead us to compare salaries of all data and machine learning related careers in original table with solely Data Scientists within US and globally.

Data Scientists

- Because there are various career types in the data table we have filtered for data scientist jobs and compared the means of both
- For data scientists: global mean == $112,298
- For all related careers in the data sample: mean == $113,468
- There seems to be no significant difference

Source validity

- Data sourced from Kaggle which is one of the largest most reputable data sources acquired by Google.

Conclusion

- Mean US salary : $144,638
- Mean Gobal salary: $113,468
- Mean Data Scientist Global salary: $112,298
- We have plotted data based on salary and experience level
- "Bureau of Labor Statistics (BLS), the median wage for workers in the United States in the second quarter of 2022 was about $1,041 per week or **$54,132 per year** (assuming 52 weeks of work per year)."
- Overall, we can say that this career produces a great salary, there is a big difference between US and global salaries, however we could have collected more data that compares salaries in other careers within other geographic locations to give more reference to what even is a good salary.
- Any Questions ?