Safa Alasady

Exploring Patterns in Data Science Salaries

Overview of the dataset

- "*Data Science Salaries 2023*" from Kaggle
- Has 3,755 rows and 11 columns
- Includes information on salaries, work years, etc.
- Goal: To understand and gather insights from the dataset, particularly for salaries
- Tools used to explore data: Python, Excel, and Tableau
    - **Python** for viewing data, data cleaning, and data analysis
    - **Excel** for further data cleaning and standardization
    - **Tableau** for data visualizations

The Importance of Data Cleaning

- Data stays accurate and consistent
- Removes **errors, duplicates, missing values, and inconsistencies**
- Allows the data analysis to be **valid**, **consistent**, and **reliable**
- Creates better decision making and reduces incorrect decisions based on data

Cleaning the Data (1 of 3)

Steps:

1. Viewed data prior to cleaning
    I. Initially 3,755 rows in total
2. Removed many duplicate rows
3. Checked for any missing values
4. Found and removed outliers using IQR method

Results:

- Found 1,171 duplicate rows and 95 outliers
- 2,489 rows remaining

```
# Counting how many duplicates there are before dropping duplicates
count_duplicates = data.duplicated().sum()
print("An initial check showed that there are", count_duplicates,
      "duplicate rows in the dataset.")
```

```
An initial check showed that there are 1171 duplicate rows in the dataset.
```

```
# Dropping duplicate rows in dataset.
data = data.drop_duplicates()
after_count_duplicates = data.duplicated().sum()
print("After removal, there are now", after_count_duplicates,
      "duplicate rows in the dataset.")
```

```
After removal, there are now 0 duplicate rows in the dataset.
```

```
# Check for missing values
print(data.isnull().sum())
```

```
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64
```

Cleaning the Data (2 of 3)

- **IQR (Interquartile Range):** Measures the spread of the middle 50% of the data
- Lower Bound: Q1 – 1.5 × IQR = –48,843.75
- Upper Bound: Q3 + 1.5 × IQR = 321,406.25
- IQR is a good method
  - Resistant to outliers
  - Focuses on 50% of data

```
# Mathematical calculations for the process of removing outliers.

# Calculating upper and quartile
Q1 = data['salary'].quantile(.25)
Q3 = data['salary'].quantile(.75)
IQR_salary = Q3 - Q1

# Calculating lower and upper bound
lowerbound_salary = Q1 - 1.5 * IQR_salary
upperbound_salary = Q3 + 1.5 * IQR_salary

print("SALARY column")
print("-------------")
print("IQR:", IQR_salary)
print("Lower and upper bounds:", lowerbound_salary, upperbound_salary)
```

```
SALARY column
-------------
IQR: 92562.5
Lower and upper bounds: -48843.75 321406.25
```

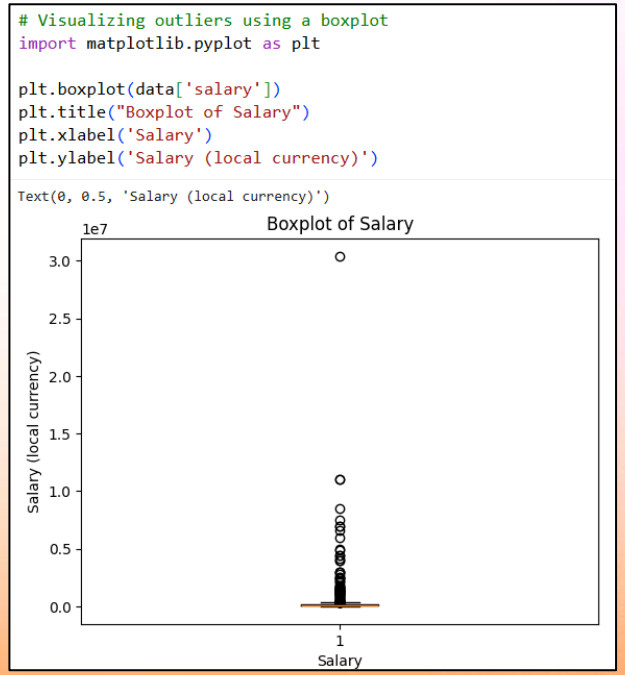
```
# Filtering out "salary" rows with outliers
data = data[(data['salary'] >= lowerbound_salary) & (data['salary'] <= upperbound_salary)]
```

Cleaning the Data (3 of 3)

```
# Visualizing outliers using a boxplot
import matplotlib.pyplot as plt

plt.boxplot(data['salary'])
plt.title("Boxplot of Salary")
plt.xlabel('Salary')
plt.ylabel('Salary (local currency)')
```

```
Text(0, 0.5, 'Salary (local currency)')
```

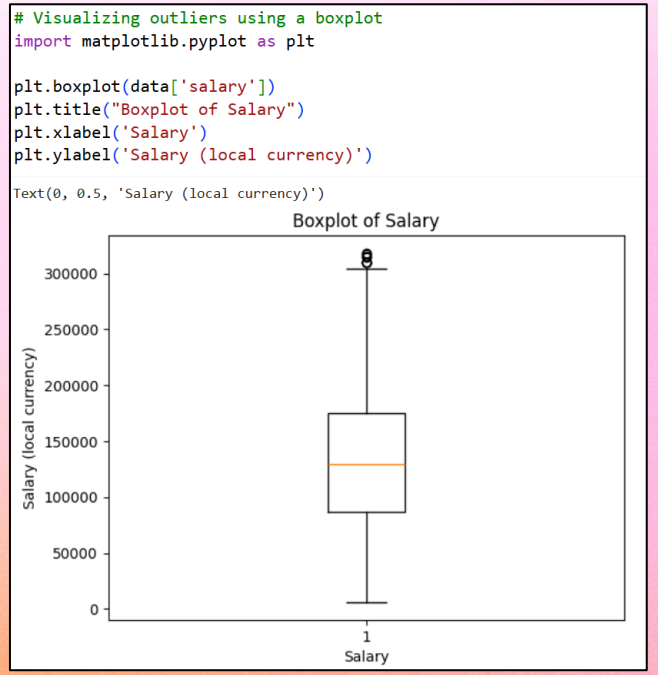


Before removing outliers

```
# Visualizing outliers using a boxplot
import matplotlib.pyplot as plt

plt.boxplot(data['salary'])
plt.title("Boxplot of Salary")
plt.xlabel('Salary')
plt.ylabel('Salary (local currency)')
```

```
Text(0, 0.5, 'Salary (local currency)')
```



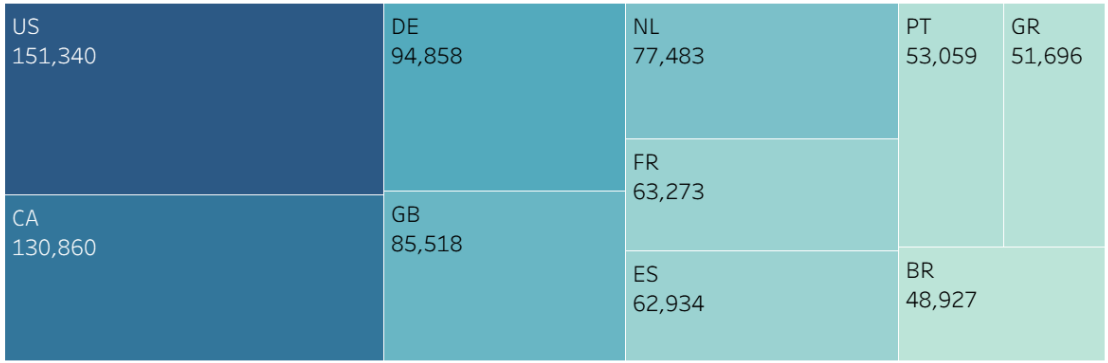
After removing outliers

Questions

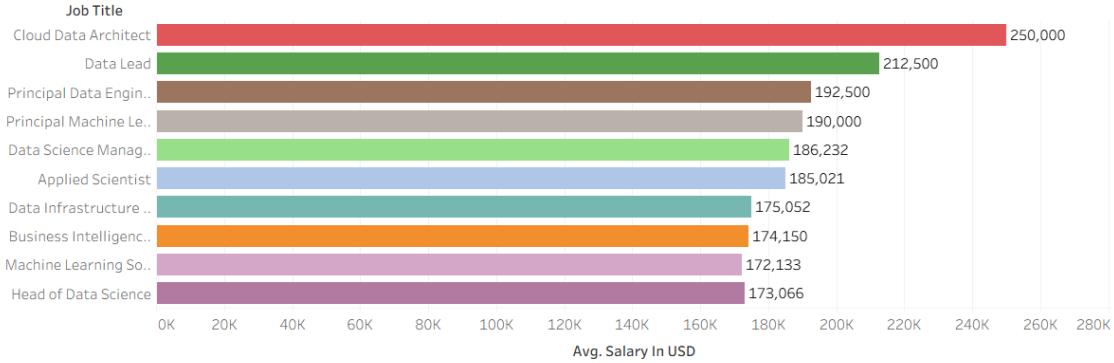
- How do salaries for the same job title and experience level vary across different countries?
  - Specific job titles: Data Engineer, Data Scientist, Data Analyst
- Did global events like COVID-19 affect salaries by experience level?
- What job titles saw the fastest salary growth over time?

Overview of Dataset

Top 10 average salaries across employee residences



Top 10 Roles with the Highest Average Salaries



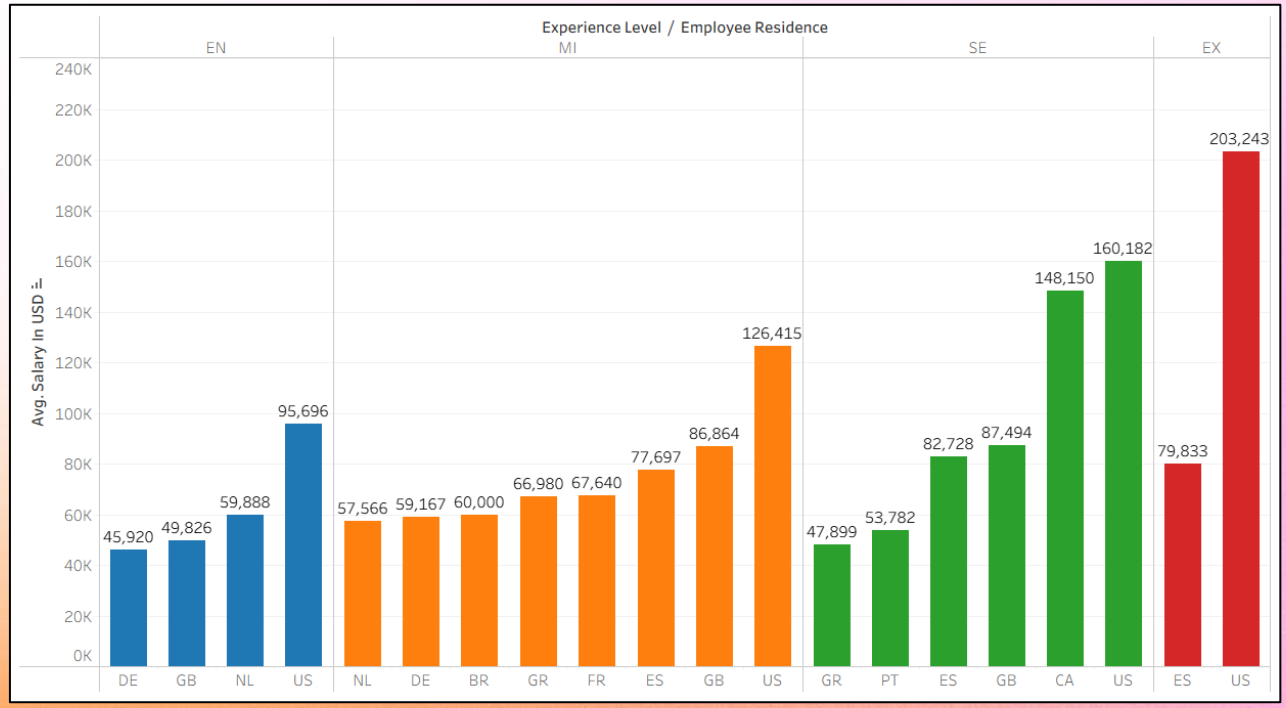
Average and Total Salaries for the Top 30 Job Titles

Job Title	Avg. Salary In USD	Salary In USD
Data Engineer	140,881	83,119,638
Data Scientist	136,296	69,783,663
Data Analyst	105,225	40,827,346
Machine Learning Engineer	149,544	34,843,722
Analytics Engineer	150,152	13,663,809
Data Architect	161,523	10,014,413
Data Science Manager	186,232	9,311,610
Research Scientist	142,203	8,816,614
Applied Scientist	185,021	5,550,620
Research Engineer	165,909	5,475,010
Machine Learning Scientist	163,220	4,243,722
Data Manager	114,982	2,644,587
Data Analytics Manager	140,630	2,531,340
Computer Vision Engineer	139,059	2,224,943
Data Science Consultant	91,958	1,931,122
BI Data Analyst	82,104	1,806,293
Machine Learning Software Engineer	172,133	1,549,200
Machine Learning Infrastructure Engineer	143,012	1,573,130
Head of Data	167,675	1,509,075
AI Developer	136,666	1,503,327
Director of Data Science	163,150	1,468,348
Data Specialist	124,583	1,495,000
BI Developer	130,727	1,438,000
Head of Data Science	173,066	1,384,530
AI Scientist	94,289	1,320,047
Data Science Lead	156,334	1,250,675
Principal Data Scientist	167,053	1,169,369
Data Infrastructure Engineer	175,052	1,050,310
ETL Developer	135,572	949,003
NLP Engineer	132,785	929,497

How do salaries for the same job title and experience level vary across different countries?

Data Engineer

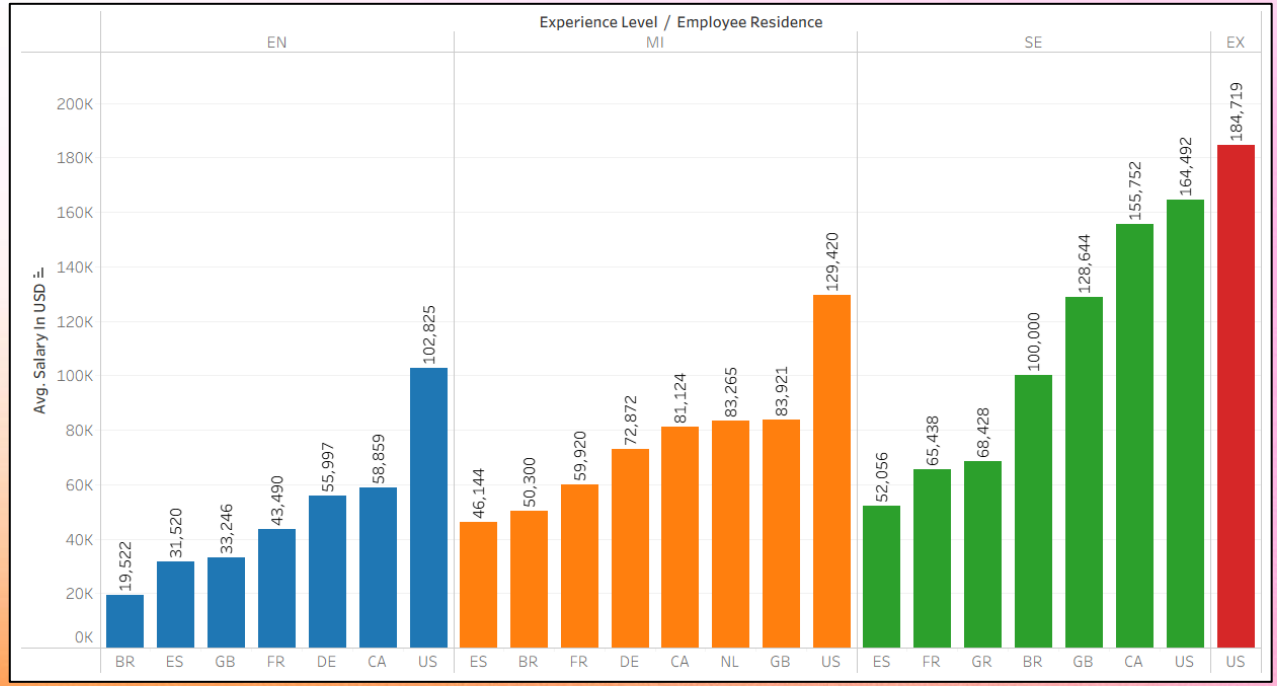
- **Entry-level**: U.S. salaries are nearly double those in Europe
- **Mid-level**: U.S. salary is up to 120% higher than other countries
- **Senior-level**: Gaps expand, with the U.S. paying over 200% more than Greece and Portugal
- **Executive-level**: U.S. earn about 2.5x Spain's salaries



How do salaries for the same job title and experience level vary across different countries?

Data Scientist

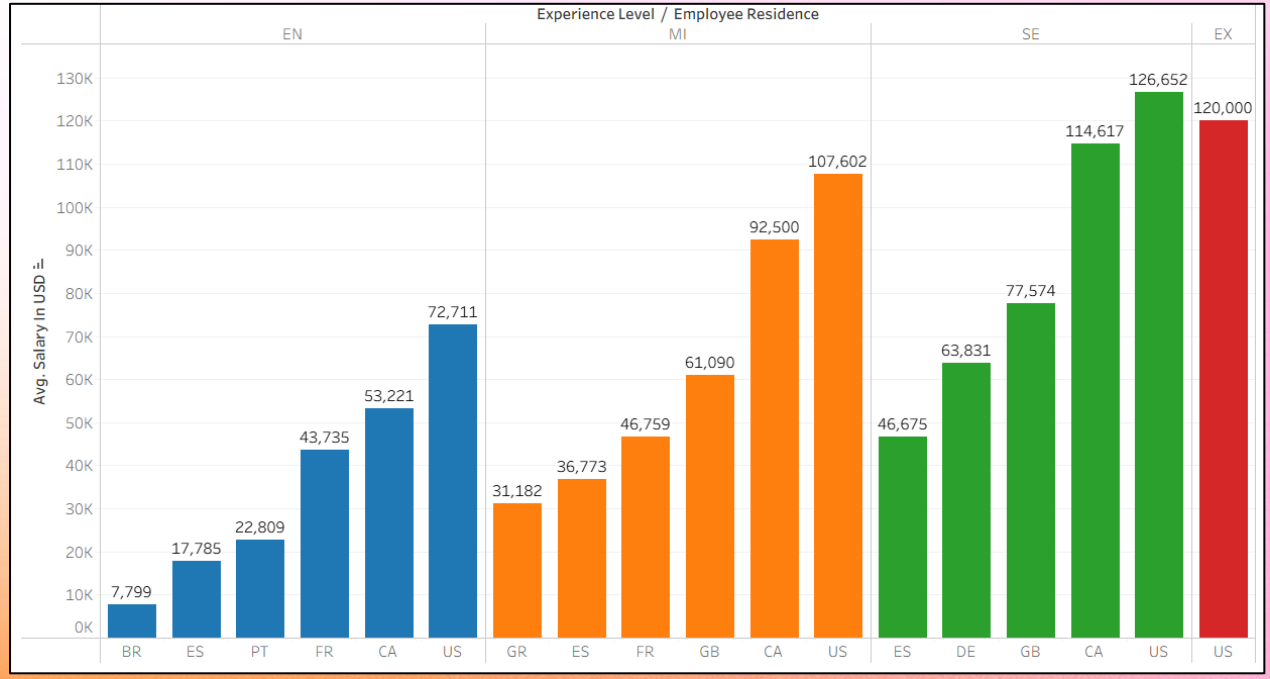
- **Entry-level**: U.S. pays up to 5x more than Brazil and almost double Europe
- **Mid-level**: U.S. salaries ($129K) are 50 to 100% higher
- **Senior-level**: U.S. and Canada surpass $150K, while others stay under $100K
- **Executive-level**: U.S. leads at $185K, much higher than other countries



How do salaries for the same job title and experience level vary across different countries?

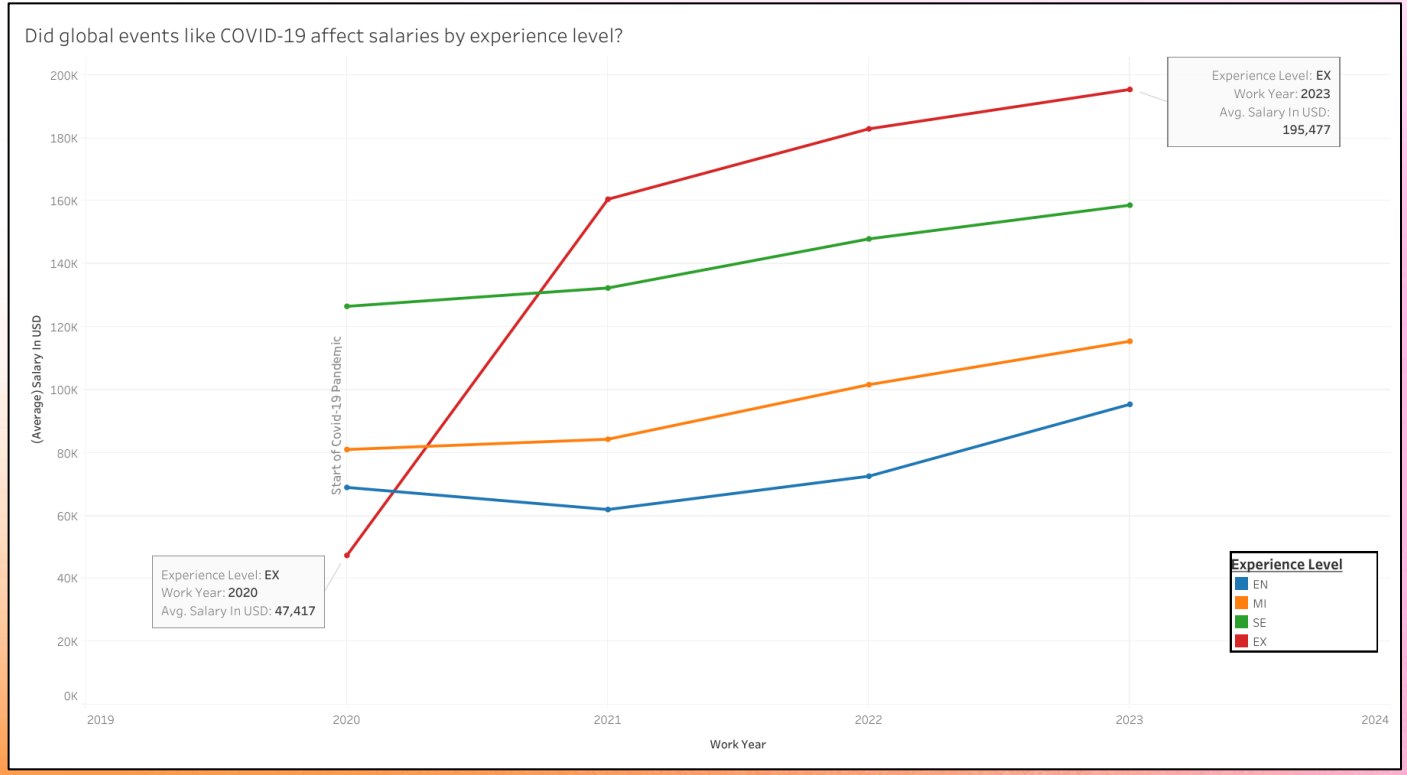
Data Analyst

- **Entry-level**: U.S. employees make 3x Europe and 9x Brazil
- **Mid-level**: $108K in the US compared to $30–60K in Europe
- **Senior-level**: While some stay around $65K, the US and Canada surpass $114K
- **Executive-level**: The U.S. leads at $120k due to executive-level data analysts only in the U.S



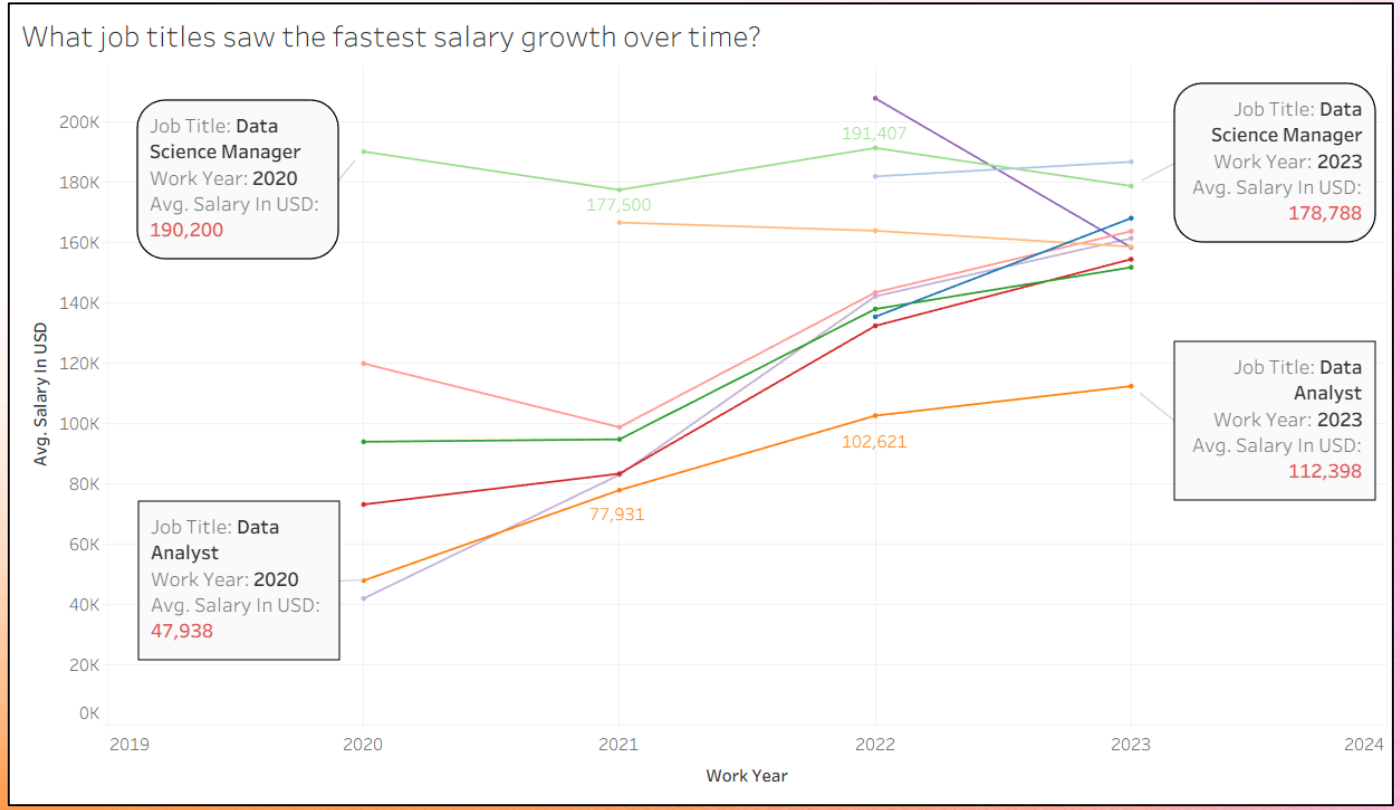
Did global events like COVID-19 affect salaries by experience level?

- **Entry-level**: Flat through 2020, steady rise after 2021
- **Mid-level**: Slight increase, steady increase after 2021
- **Senior-level**: Steady growth overall
- **Executive-level**: Jump from $47K in 2020 to $195K in 2023
- **Trend**: COVID-19 helped increased salaries for higher experience levels



What job titles saw the fastest salary growth over time?

- **Data Science Managers**: $190K in 2020 to $178K in 2023
- **Data Analysts**: Fastest salary increase, from $48K in 2020 to $112K in 2023
- **Other job titles**: Show steady but smaller increases compared to Data Analysts
- **Insight**: Demand is increasing entry and mid-level analyst salaries the most



Summary of Findings

- Salaries do grow with experience,
  - U.S. leads at all experience levels
- COVID-19 did help increase growth for salaries, especially for executive experience levels
- Data Analysts saw the fastest salary growth

Challenges, Insights, and Takeaways

- Learned a lot about the variations of salaries across different job titles
- Gained many insights from using Python, Excel, and Tableau
- Expanded my existing knowledge of Python and data analysis
- Had some challenges when using Tableau
- **Reached my goal** of gathering insights

Thank You!