# DATA SCIENCE SALARY ANALYSIS

Prepared by Brendan OConnell

April 14, 2024

# REQUIREMENTS REVIEW

## Deliverable

Recommend a competitive salary range to acquire a top-talent Data Scientist.

## Key Points of Interest

- Full-time Data Scientist
- U.S. employees vs. Offshore employees
- Examine company size
- Investigate salary trends over time
- Data Science team

# REQUIREMENTS REVIEW

## Full-time Data Scientist

- Someone to drive data science within the organization
- Potentially lead a team in the future
- "Top talent" → extensive experience

## U.S. employees vs. Offshore employees

- Highlight the differences in salary data

## Examine company size

- Currently small, rapidly expanding to medium size
- Observe salaries across various company sizes

## Investigate salary trends over time

- "Salaries are going up due to the great recession"
- Does the data support this statement?

## Data Science team

- Explore data points for future hiring needs

# RESEARCH QUESTIONS

BASED ON REQUIREMENTS REVIEW

Can the job titles be categorized to provide more useful analysis?

Which employment types are most relevant?

Are all job titles in the data a good fit based on the requirements?

Is there any external data from reliable sources that can be incorporated?

How do salaries across experience levels match up?

How does company size impact the analysis?

Is the distribution of salary data relatively normal, skewed, or other shape?

Does the data support the theory that salaries have increased over recent years?

What are the salary trends by job title? By experience level? By country?

Which statistics best illustrate salary differences for U.S. vs. offshore employees?

Are there extractable insights to help develop a future data science team?

# CAN THE JOB TITLES BE CATEGORIZED TO PROVIDE MORE USEFUL ANALYSIS?

The job titles in the data can be grouped together based on their related fields of interest.

These *job fields* help narrow the focus on jobs most relevant to the requirements, which in turn will yield more relevant insights.

The following slide breaks down some high-level descriptions of each job field that will be added to the dataset.

# JOB FIELDS

## Data Science

Asks questions and studies data to extract insights for solutions.

Combines scientific thinking with mathematics, computer science, and data analytics.

## Artificial Intelligence / Machine Learninging (A.I. / M.L.)

A.I. refers to technology that simulates human intelligence.

M.L. is a branch of A.I. using data models & algorithms to help computers learn.

## Data Engineering / Architecture

Data architecture is the design and implementation of data storage & access.

Data engineering builds/maintains the pipelines that allow data to be available for analysis.

## Data Analysis

Analyze data to identify patterns/relationships and build predictive models.

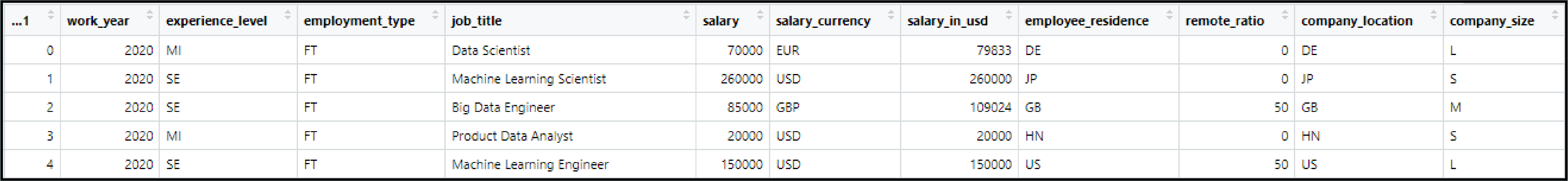
Create analytics reports with data visualizations to communicate findings.

## Data Consulting

Assist organizations in the development of data-driven solutions.

# DATA PREPARATION

## Original Data:



| ...1 | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | company_size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2020 | MI | FT | Data Scientist | 70000 | EUR | 79833 | DE | 0 | DE | L |
| 1 | 2020 | SE | FT | Machine Learning Scientist | 260000 | USD | 260000 | JP | 0 | JP | S |
| 2 | 2020 | SE | FT | Big Data Engineer | 85000 | GBP | 109024 | GB | 50 | GB | M |
| 3 | 2020 | MI | FT | Product Data Analyst | 20000 | USD | 20000 | HN | 0 | HN | S |
| 4 | 2020 | SE | FT | Machine Learning Engineer | 150000 | USD | 150000 | US | 50 | US | L |

Add column for "Job Field" of each entry based on previous slide

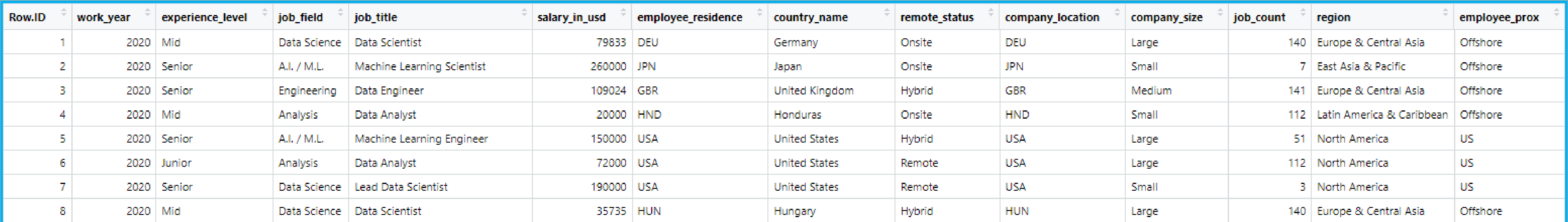
Fix ID column name and reset index to start at 1

Select only "Full-Time" job types, based on requirements review *(remove "job types" column after filtering to only FT)*

Drop "salary" and "salary currency" *(only need USD salary)*

More descriptive value names for "experience level", "remote ratio", and "company size"

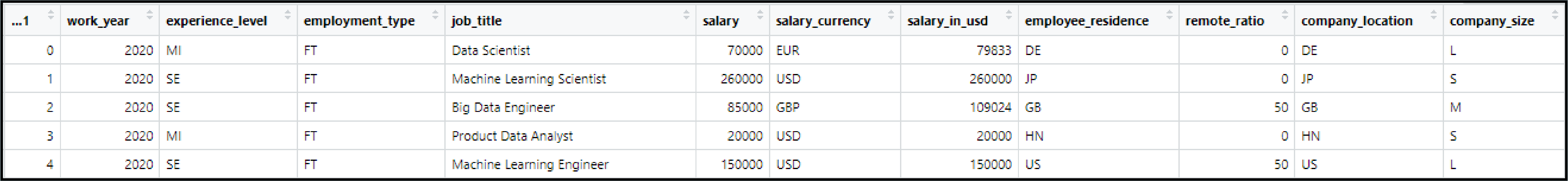
## Cleaned Data:



| Row.ID | work_year | experience_level | job_field | job_title | salary_in_usd | employee_residence | country_name | remote_status | company_location | company_size | job_count | region | employee_prox |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2020 | Mid | Data Science | Data Scientist | 79833 | DEU | Germany | Onsite | DEU | Large | 140 | Europe & Central Asia | Offshore |
| 2 | 2020 | Senior | A.I. / M.L. | Machine Learning Scientist | 260000 | JPN | Japan | Onsite | JPN | Small | 7 | East Asia & Pacific | Offshore |
| 3 | 2020 | Senior | Engineering | Data Engineer | 109024 | GBR | United Kingdom | Hybrid | GBR | Medium | 141 | Europe & Central Asia | Offshore |
| 4 | 2020 | Mid | Analysis | Data Analyst | 20000 | HND | Honduras | Onsite | HND | Small | 112 | Latin America & Caribbean | Offshore |
| 5 | 2020 | Senior | A.I. / M.L. | Machine Learning Engineer | 150000 | USA | United States | Hybrid | USA | Large | 51 | North America | US |
| 6 | 2020 | Junior | Analysis | Data Analyst | 72000 | USA | United States | Remote | USA | Large | 112 | North America | US |
| 7 | 2020 | Senior | Data Science | Lead Data Scientist | 190000 | USA | United States | Remote | USA | Small | 3 | North America | US |
| 8 | 2020 | Mid | Data Science | Data Scientist | 35735 | HUN | Hungary | Hybrid | HUN | Large | 140 | Europe & Central Asia | Offshore |

# DATA PREPARATION

## Original Data:



| ...1 | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | company_size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2020 | MI | FT | Data Scientist | 70000 | EUR | 79833 | DE | 0 | DE | L |
| 1 | 2020 | SE | FT | Machine Learning Scientist | 260000 | USD | 260000 | JP | 0 | JP | S |
| 2 | 2020 | SE | FT | Big Data Engineer | 85000 | GBP | 109024 | GB | 50 | GB | M |
| 3 | 2020 | MI | FT | Product Data Analyst | 20000 | USD | 20000 | HN | 0 | HN | S |
| 4 | 2020 | SE | FT | Machine Learning Engineer | 150000 | USD | 150000 | US | 50 | US | L |

Convert experience levels, remote status, & company size to ordinal values *(factors)*

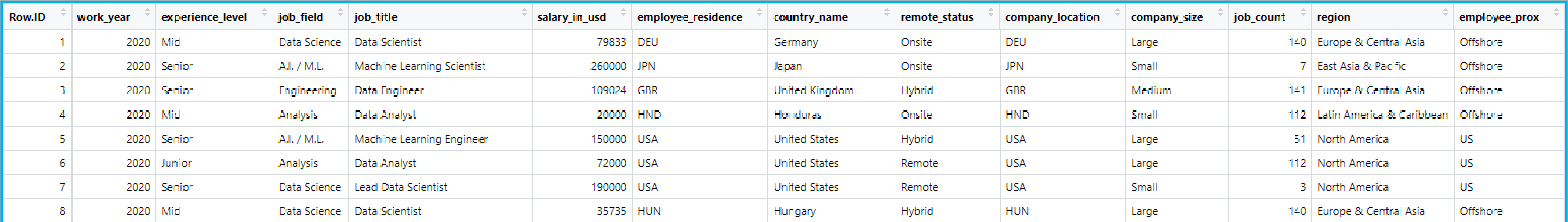
Remove all "Data Consultants" *(not an ideal candidate, and extreme outliers)*

Import ISO 3166-1 Country Code Data
-- Fix NA values
-- Use 3-letter country codes, country name, & region

Add column for total counts of each job title

Add column for filtering between "**US**" and "**Offshore**" employees

## Cleaned Data:



| Row.ID | work_year | experience_level | job_field | job_title | salary_in_usd | employee_residence | country_name | remote_status | company_location | company_size | job_count | region | employee_prox |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2020 | Mid | Data Science | Data Scientist | 79833 | DEU | Germany | Onsite | DEU | Large | 140 | Europe & Central Asia | Offshore |
| 2 | 2020 | Senior | A.I. / M.L. | Machine Learning Scientist | 260000 | JPN | Japan | Onsite | JPN | Small | 7 | East Asia & Pacific | Offshore |
| 3 | 2020 | Senior | Engineering | Data Engineer | 109024 | GBR | United Kingdom | Hybrid | GBR | Medium | 141 | Europe & Central Asia | Offshore |
| 4 | 2020 | Mid | Analysis | Data Analyst | 20000 | HND | Honduras | Onsite | HND | Small | 112 | Latin America & Caribbean | Offshore |
| 5 | 2020 | Senior | A.I. / M.L. | Machine Learning Engineer | 150000 | USA | United States | Hybrid | USA | Large | 51 | North America | US |
| 6 | 2020 | Junior | Analysis | Data Analyst | 72000 | USA | United States | Remote | USA | Large | 112 | North America | US |
| 7 | 2020 | Senior | Data Science | Lead Data Scientist | 190000 | USA | United States | Remote | USA | Small | 3 | North America | US |
| 8 | 2020 | Mid | Data Science | Data Scientist | 35735 | HUN | Hungary | Hybrid | HUN | Large | 140 | Europe & Central Asia | Offshore |

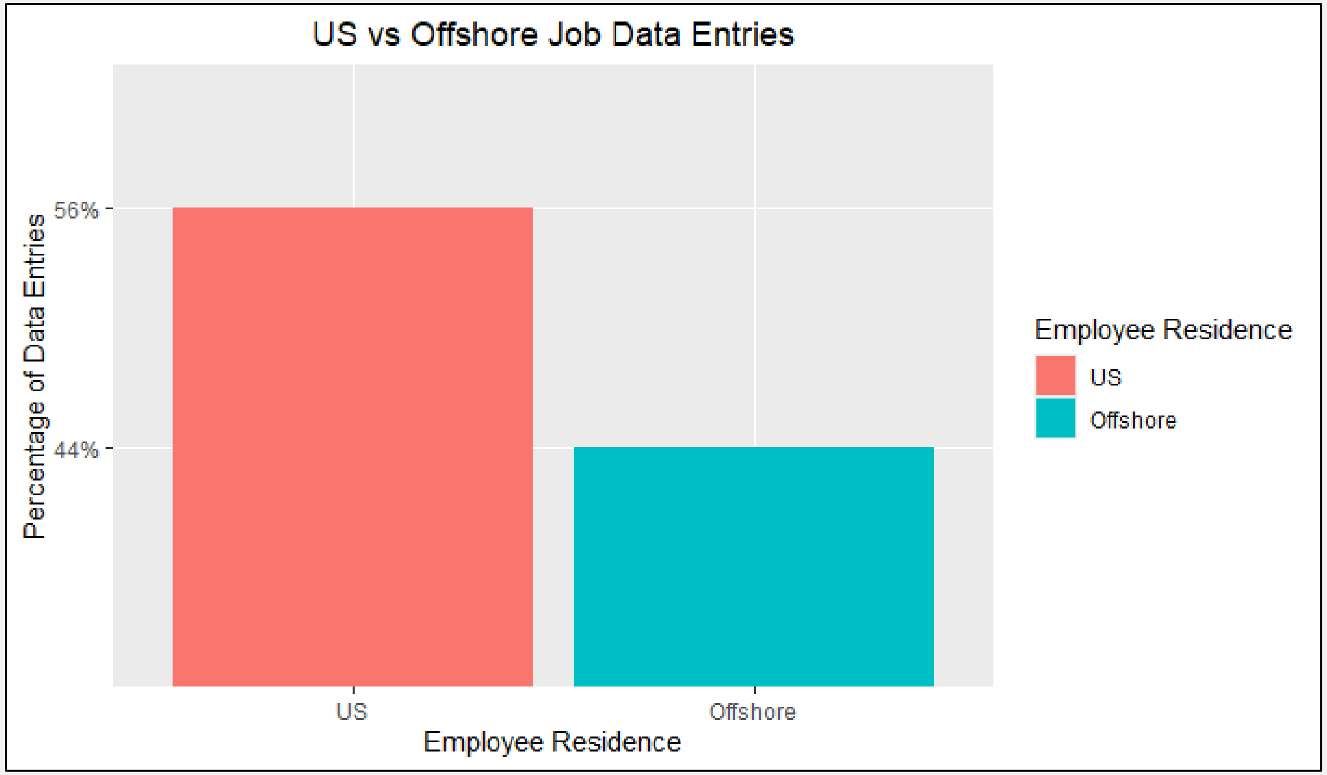
# EMPLOYEE COUNTRY OF RESIDENCE

The dataset can easily be categorized into 2 groups:

- US employee residence
- Offshore employee residence

This approach allows the following steps for Exploratory Data Analysis (EDA):

- Part 1: U.S. Data Analysis
- Part 2: Offshore Data Analysis
- Part 3: Compare U.S. vs Offshore



# PART I
# U.S. DATA ANALYSIS

"Salaries are going up
due to the great recession"

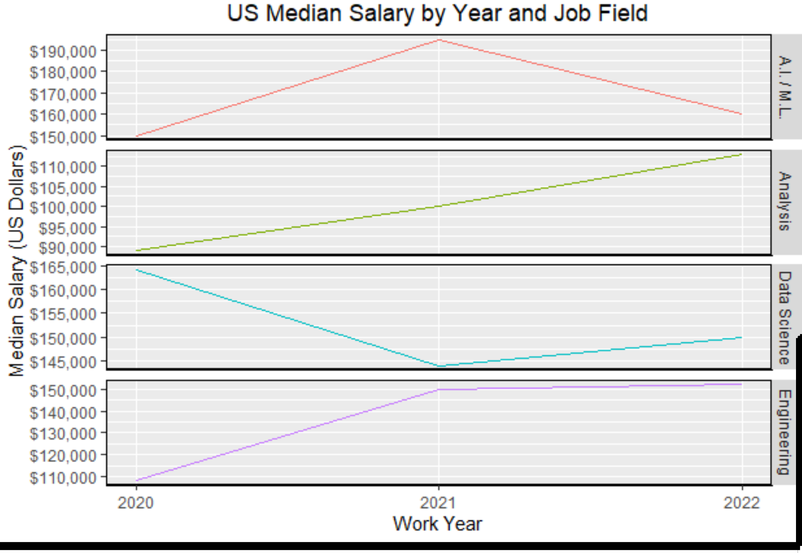
Does the data support the theory of salary increases over recent years? Several data aggregations are compiled and plotted for U.S. salaries:

Top-left plots Job Field trends.

Middle plots Experience trends.

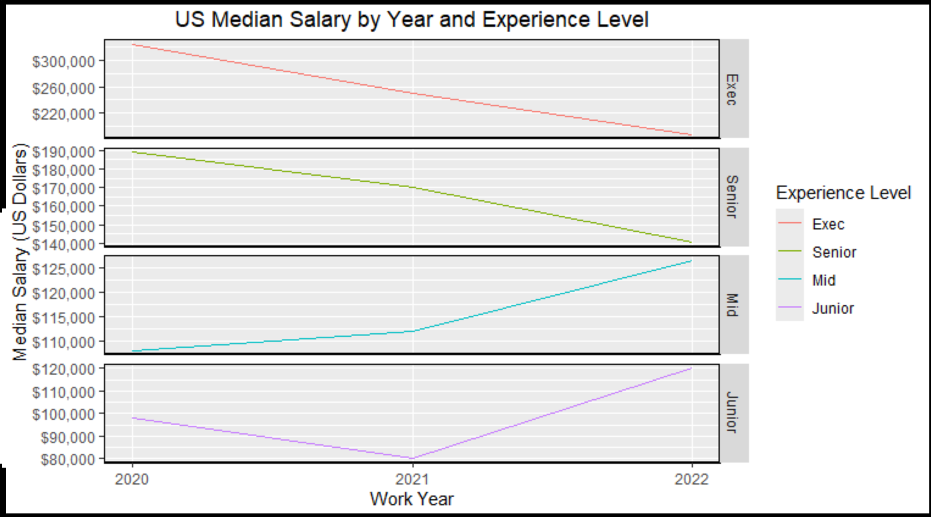
Bottom-left plots lines for each experience level within a Job Field.

See details in the corresponding text for each line plot. Ultimately, there does not seem to be a trend for increasing salaries in Data Science, but our offers should still be competitive.



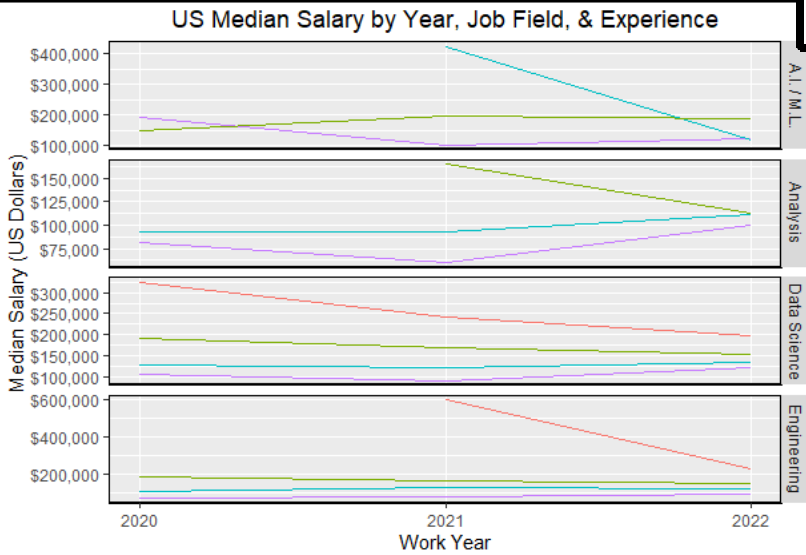
## Top-Left

Analysis and Engineering fields show salary increases over time, whereas A.I./M.L. and Data Science do not show consistent trends.



## Middle

Executive and Senior levels have decreasing salary trends, whereas Mid and Junior levels show increasing salary trends overall.

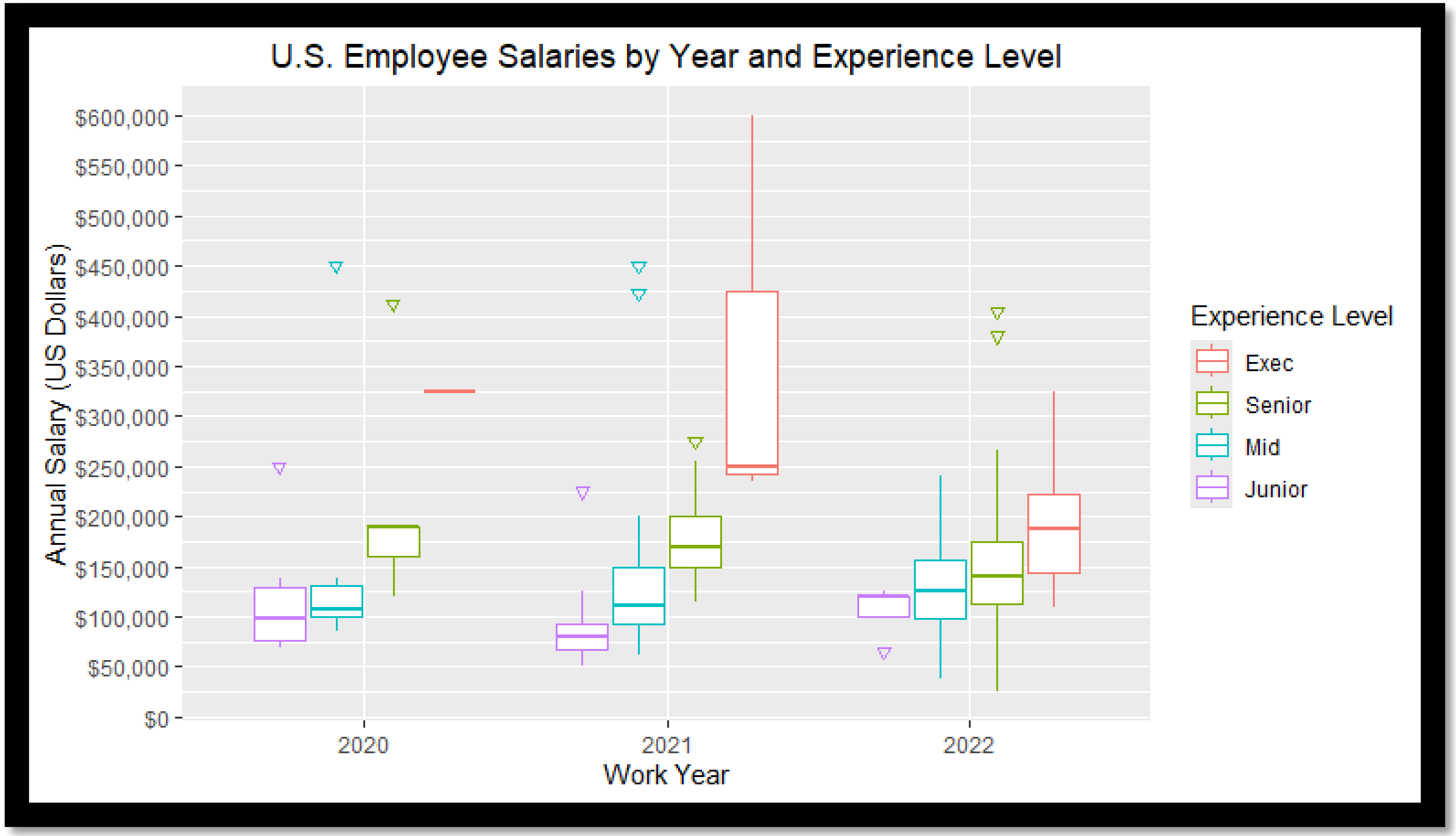


## Bottom-left

The most relevant data here is the Data Science field, which shows a decreasing salary trend for Exec and Senior levels (our area of focus for talent) whereas the Mid and Junior levels have very slight increasing salary trends.

# PART I
# U.S. DATA ANALYSIS

## How do salaries across experience levels match up?



Boxplots show high-level visuals that are useful for preliminary analysis of data distributions.

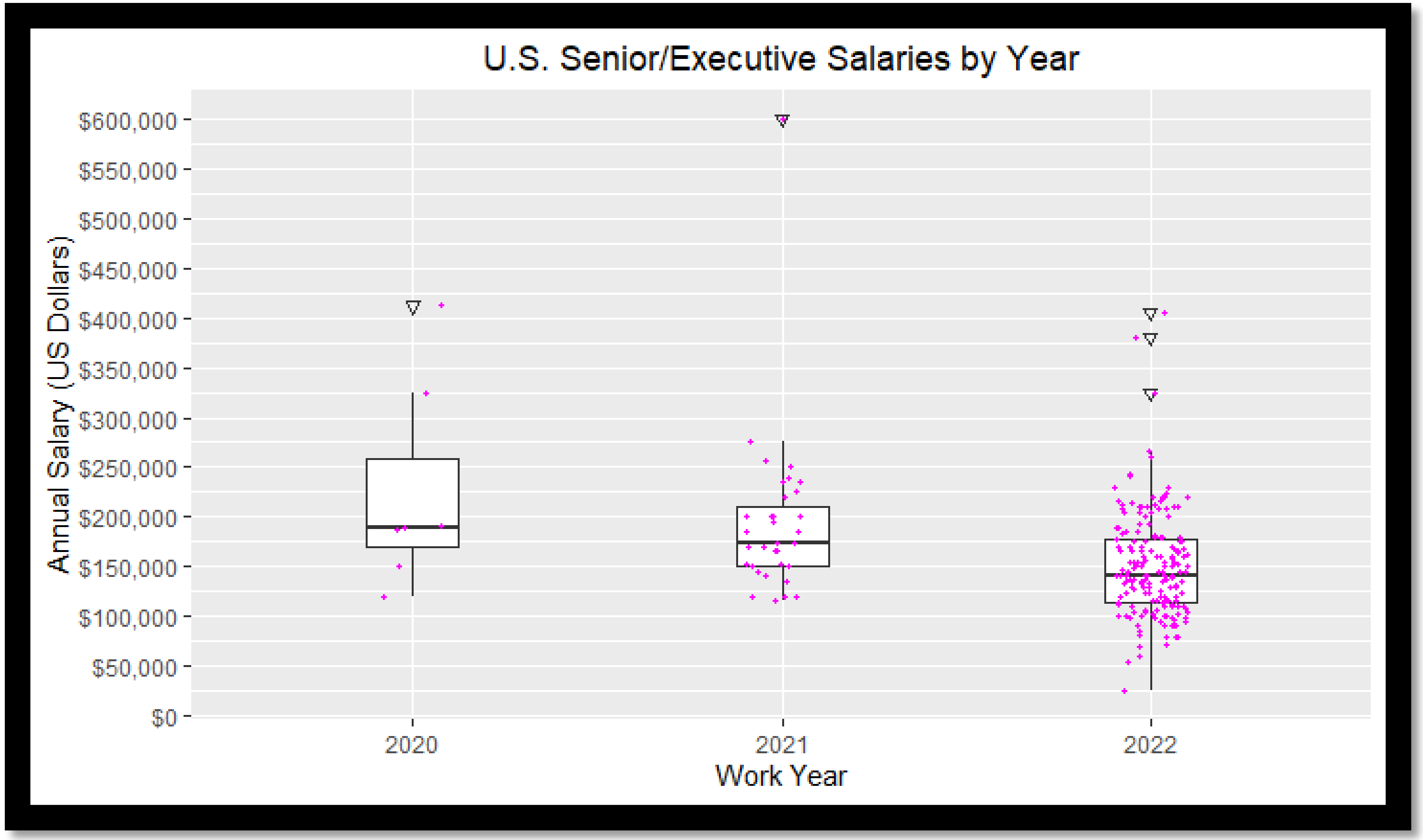
For each “experience level” boxplot, the U.S. salaries from 2020 to 2022 appear to be as expected when compared from Junior up to Executive, showing the Junior pay ranges as lowest and increasing accordingly up through the Executive level.

Based on the desire for “top talent” and a potential team lead, the Junior & Mid experience levels will now be excluded, and the data will focus on the Senior and Executive level salaries.

# PART I
# U.S. DATA ANALYSIS

## *Senior & Executive experience*

## Salaries by Year

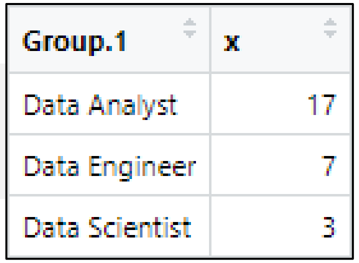


Examining the Senior and Executive salaries in 2022, there appears to be a high count of salaries $100k or less, whereas 2020 and 2021 are all above that number. That does <u>not</u> seem like competitive U.S. salary for top-talent.

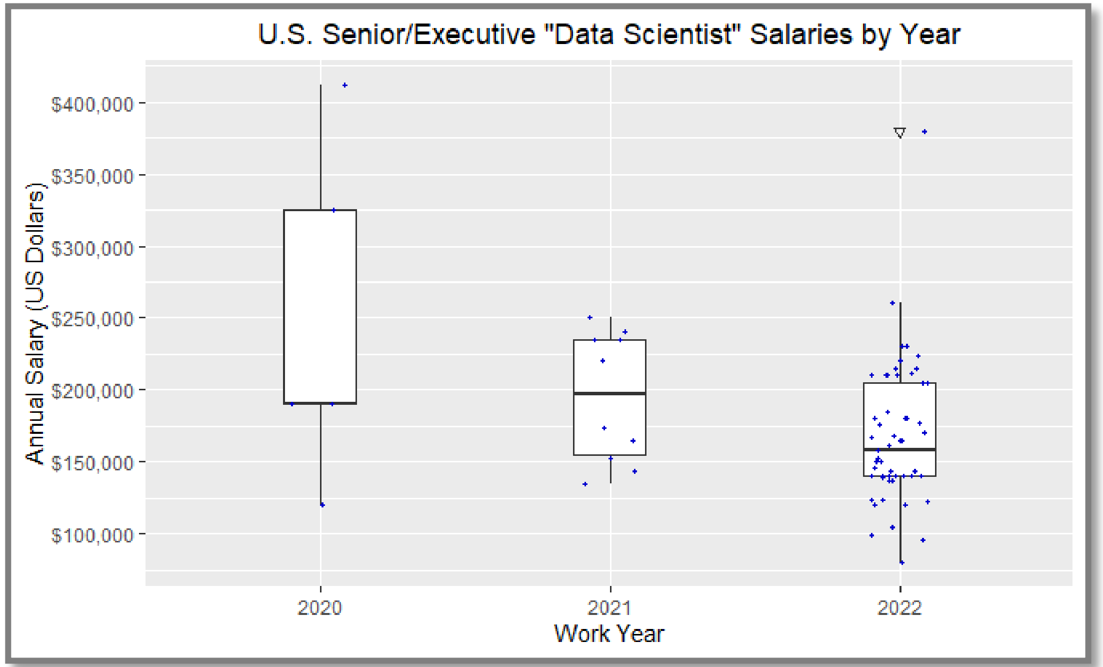
To investigate, data is aggregated for counts of each Job Title with senior or executive experience level that has a salary of $100k or less.

The aggregated data (bottom table) illustrates how Data Analysts and Data Engineers account for 24 of the 27 low-paying salaries (89%). This can be addressed by filtering the Job Fields.

Job Title counts:
Sr/Exec salaries $100k or less



| Group.1 | x |
|---|---|
| Data Analyst | 17 |
| Data Engineer | 7 |
| Data Scientist | 3 |



After filtering the data to the "Data Science" field, the median salary for 2022 increased from less than $150k to more than $150k.

# PART I
# U.S. DATA ANALYSIS

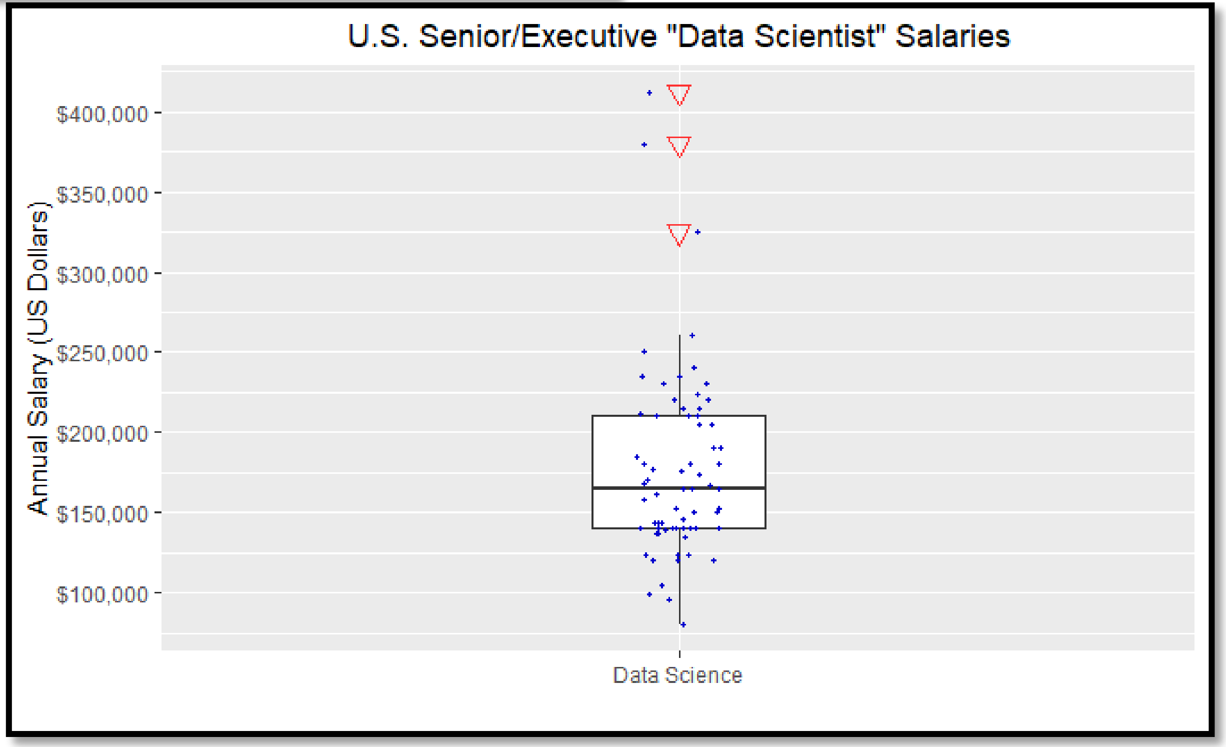
*Senior & Executive experience*

Salaries for Data Science field only

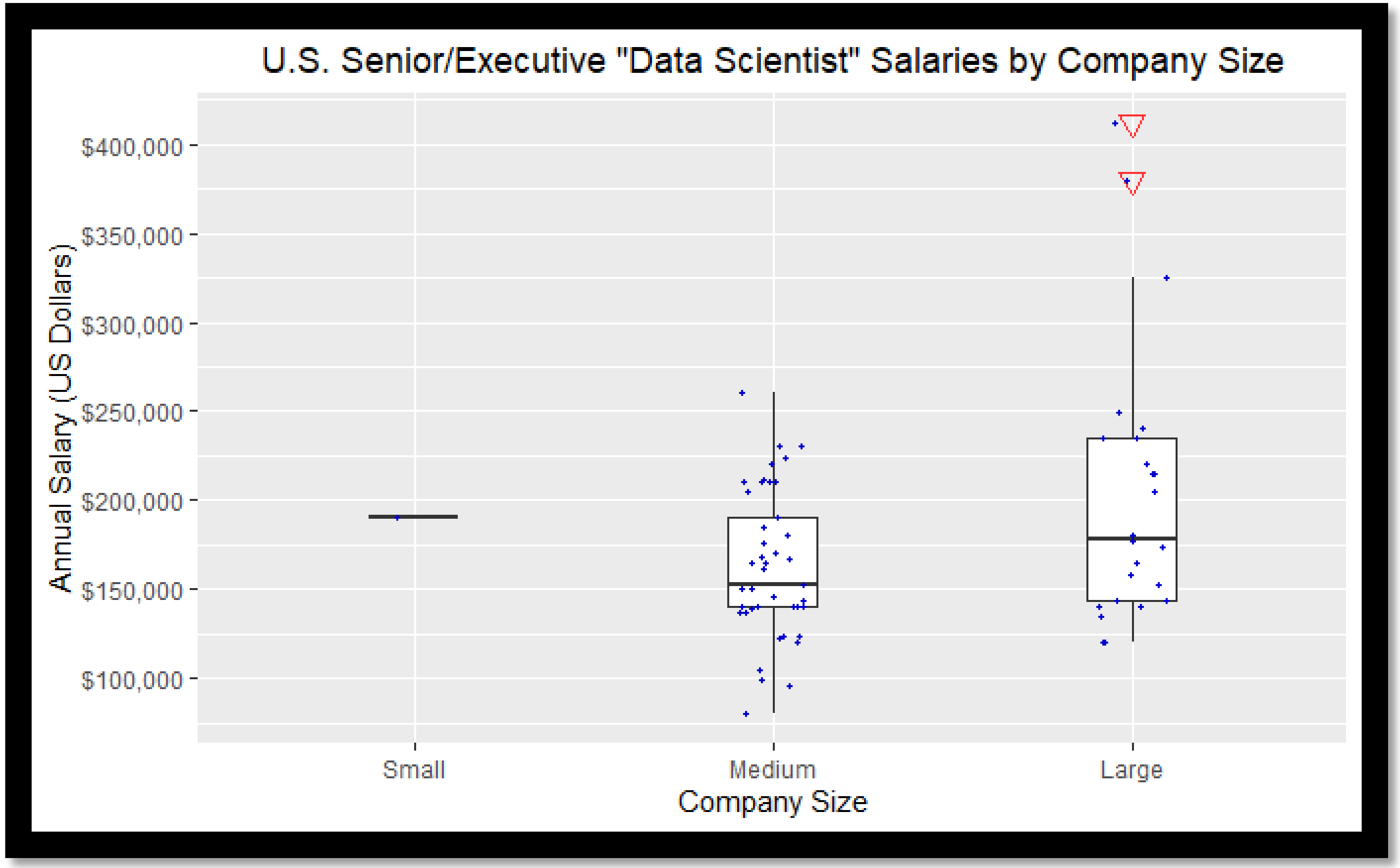
"Data Science" is the most relevant Job Field based on the requirements. By focusing on that field, the data reflects the salaries of our targeted talent more accurately and is no longer skewed by other fields like Analyst or Data Engineering.

A.I. / M.L. jobs will also be filtered out, but this is acceptable given that we are specifically looking to analyze salaries for a Data Science position. Also worth noting, in the U.S. there is no AI/ML data at the Executive level.

Now we will address the red outliers.



Combining all 3 years, a few outliers (in red) appear to be salaries over $300k, which is most likely out of our pay range considering our company size.



# PART I
# U.S. DATA ANALYSIS

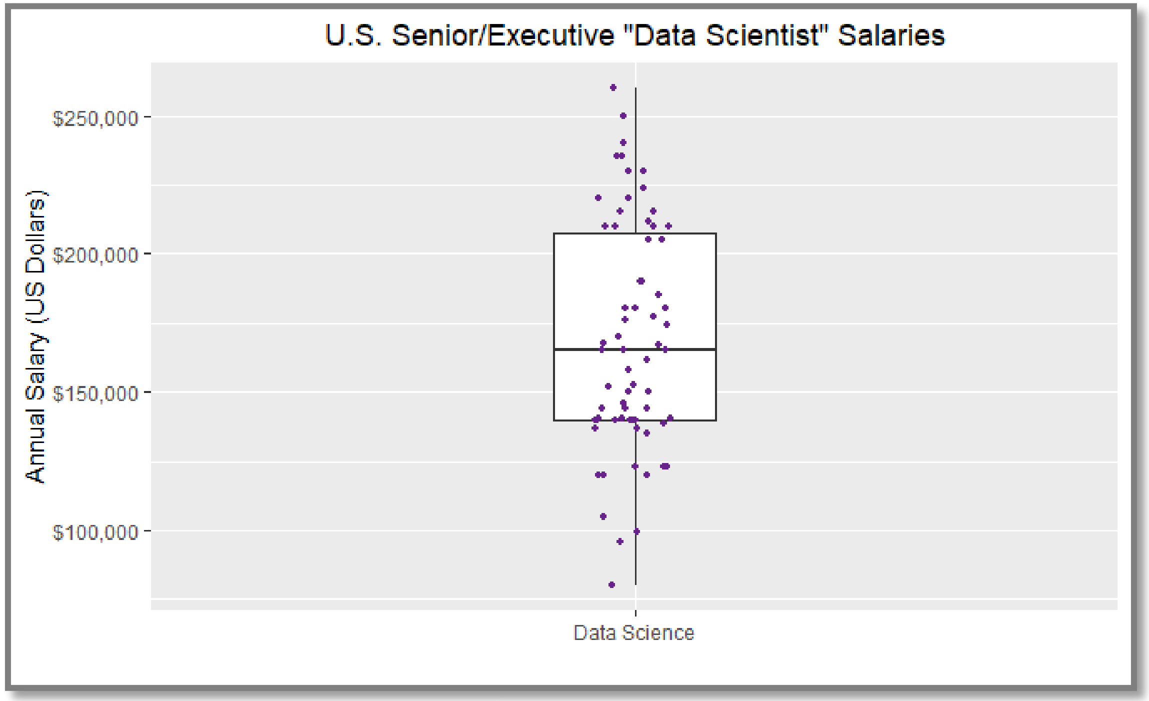
*Senior & Executive experience*

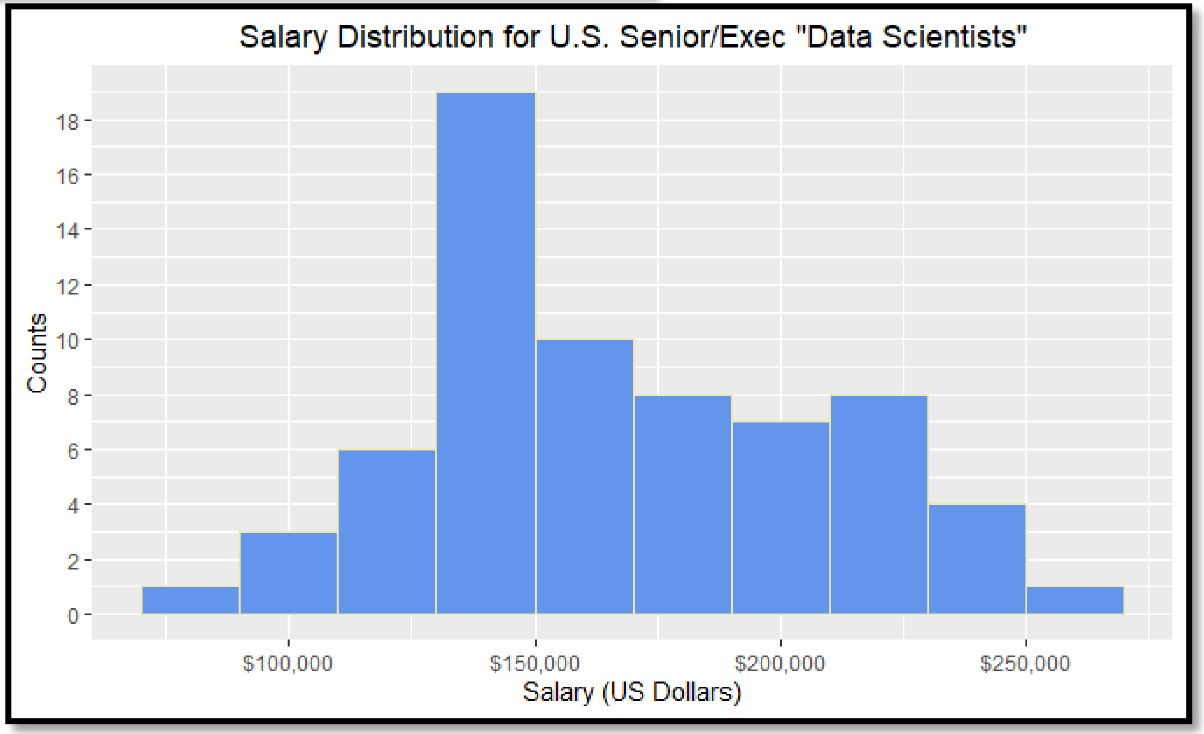
## Salaries for Data Science by Company Size

A look at salary distribution by company size offers some insight on the outlying values: **Large** companies.

It is understandable that larger companies can afford to overpay a few high-level employees. However, as a small-to-mid-sized company, we are not financially positioned to compete with these outliers.

It is best to filter out salaries that exceed $300k and focus on data that is close to financially viable.





# PART I
# U.S. DATA ANALYSIS

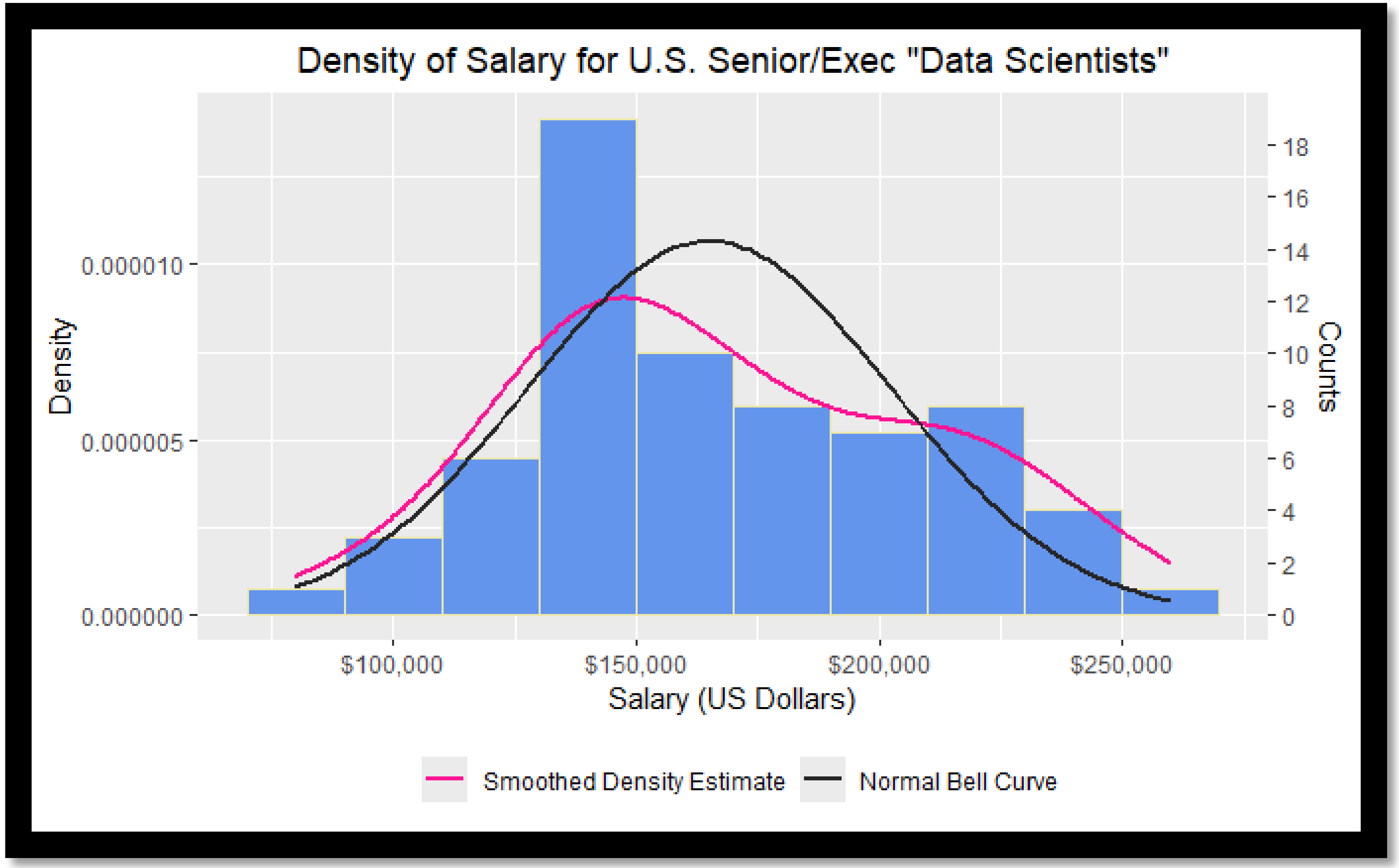
*Senior & Executive experience*

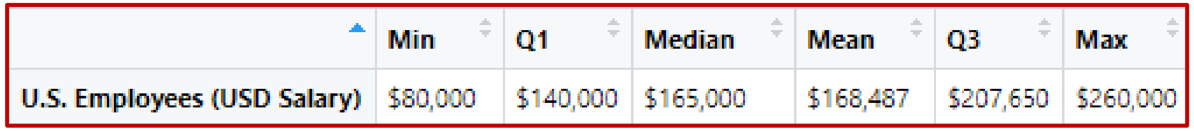
## Data Science Salary Distributions

These are the salary distribution plots for 67 U.S. Data Science jobs at the senior & executive level, excluding outliers from the large companies.

Although not exact, the data does appear to be *relatively* normal.

To confirm, a density histogram can be plotted with a smoothed estimation curve, then compared to a bell curve.





| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| U.S. Employees (USD Salary) | $80,000 | $140,000 | $165,000 | $168,487 | $207,650 | $260,000 |

# PART I
# U.S. DATA ANALYSIS

*Senior & Executive experience*

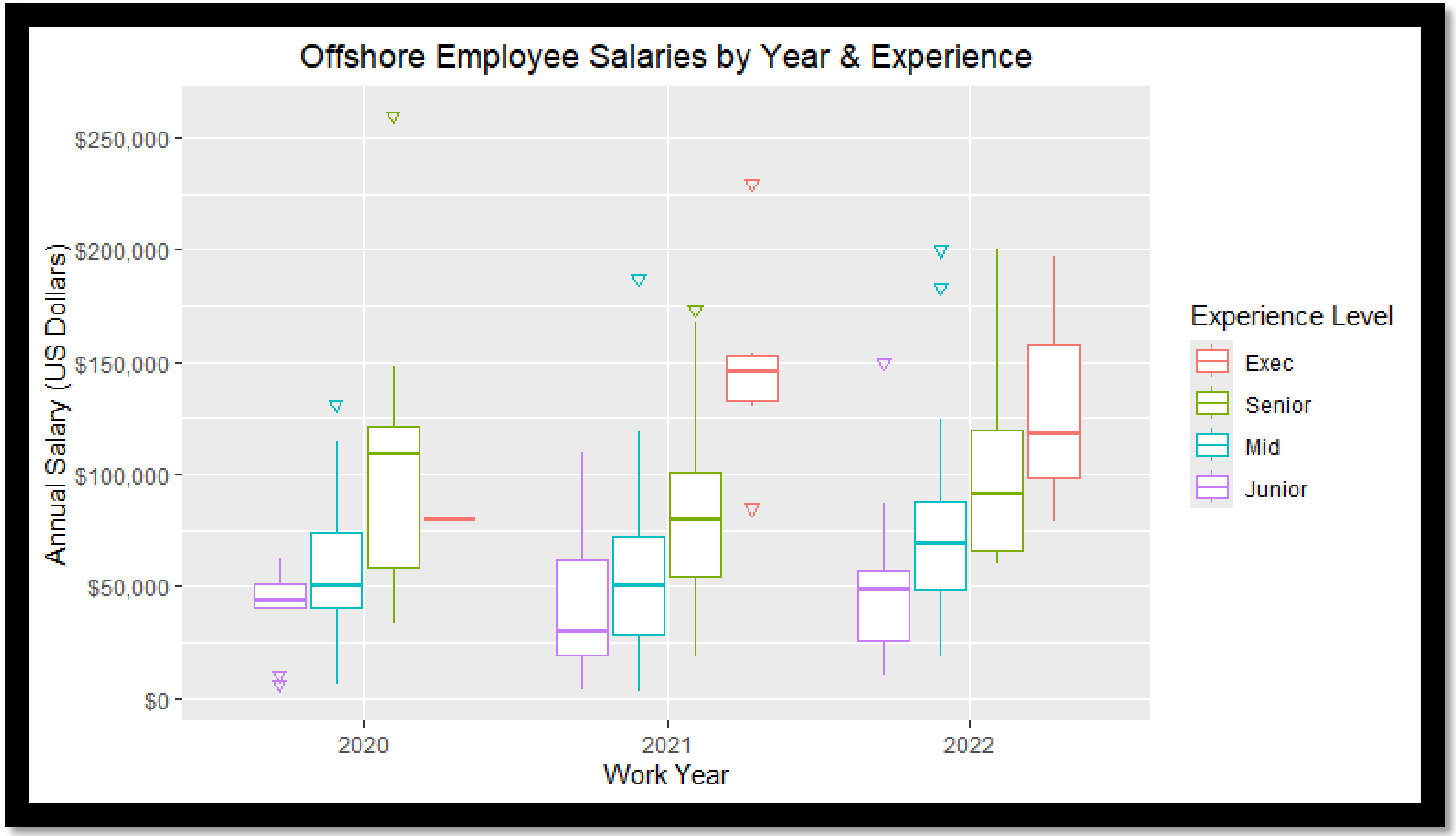
## Data Science Salary Distributions and Statistics

The black "bell" curve illustrates what a normal distribution would be.
The pink density curve estimates the actual salary distribution.

Although the peak of the curves are slightly different, the shapes appear to be quite similar, enough so that this can be treated as normal distribution.

A table of important distribution statistics are displayed at the bottom, which will be used to compare against offshore analysis and for final recommendations.

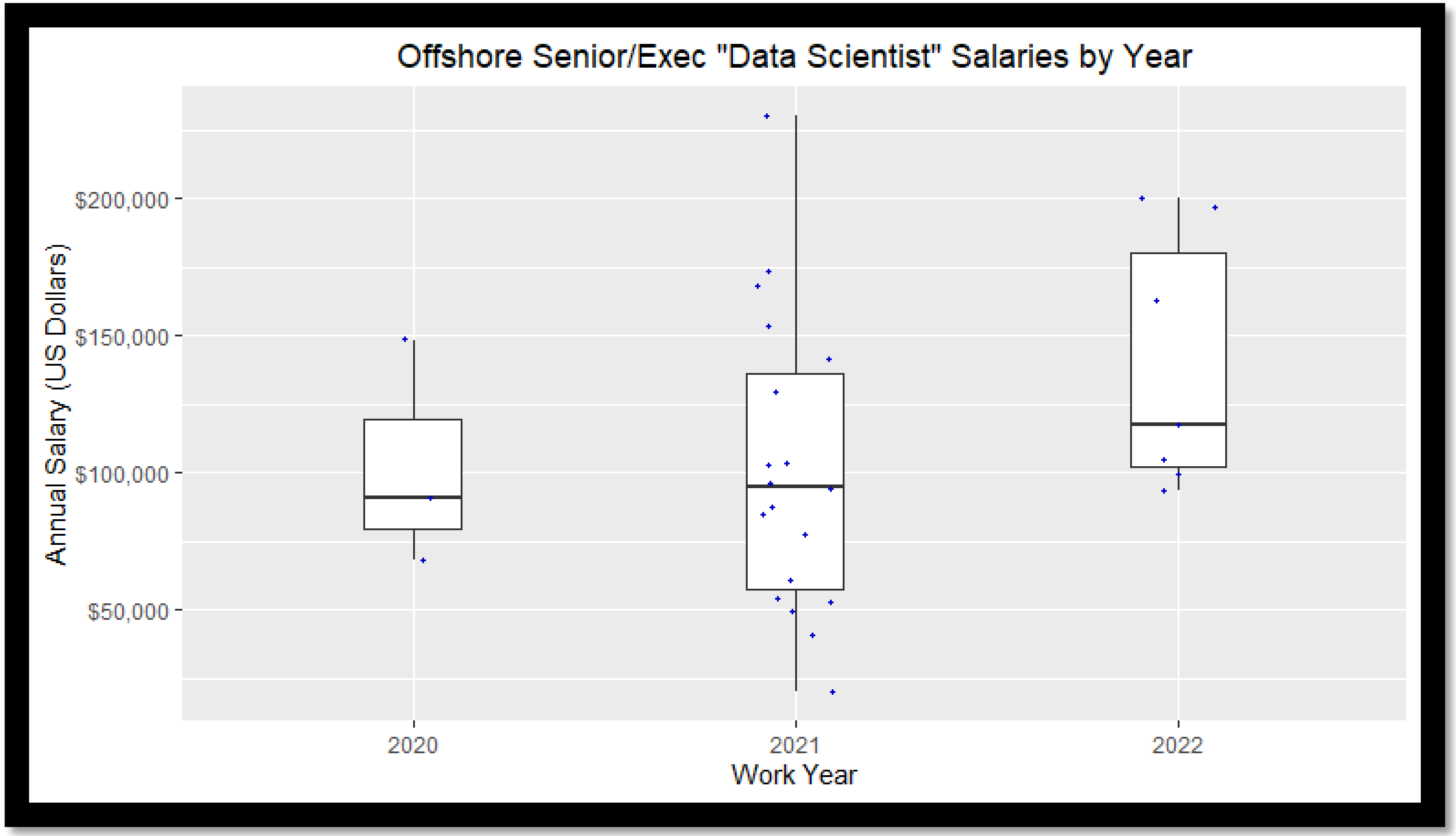
# PART II
# OFFSHORE DATA ANALYSIS



## How do salaries across experience levels match up?

Similar to the U.S. analysis, for each Offshore "experience level" boxplot the salaries from 2020 to 2022 appear to be as expected when compared from Junior up to Executive.

Based on the desire for "top talent" and a potential team lead, the data will now focus only on the Senior and Executive level salaries.



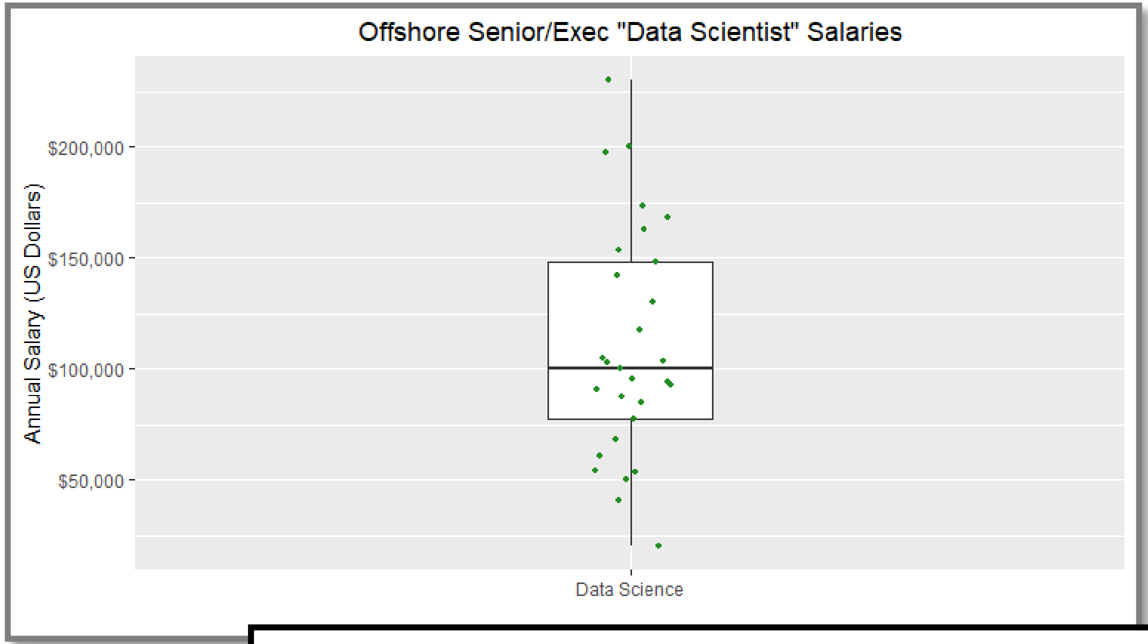
## PART II
## OFFSHORE DATA ANALYSIS

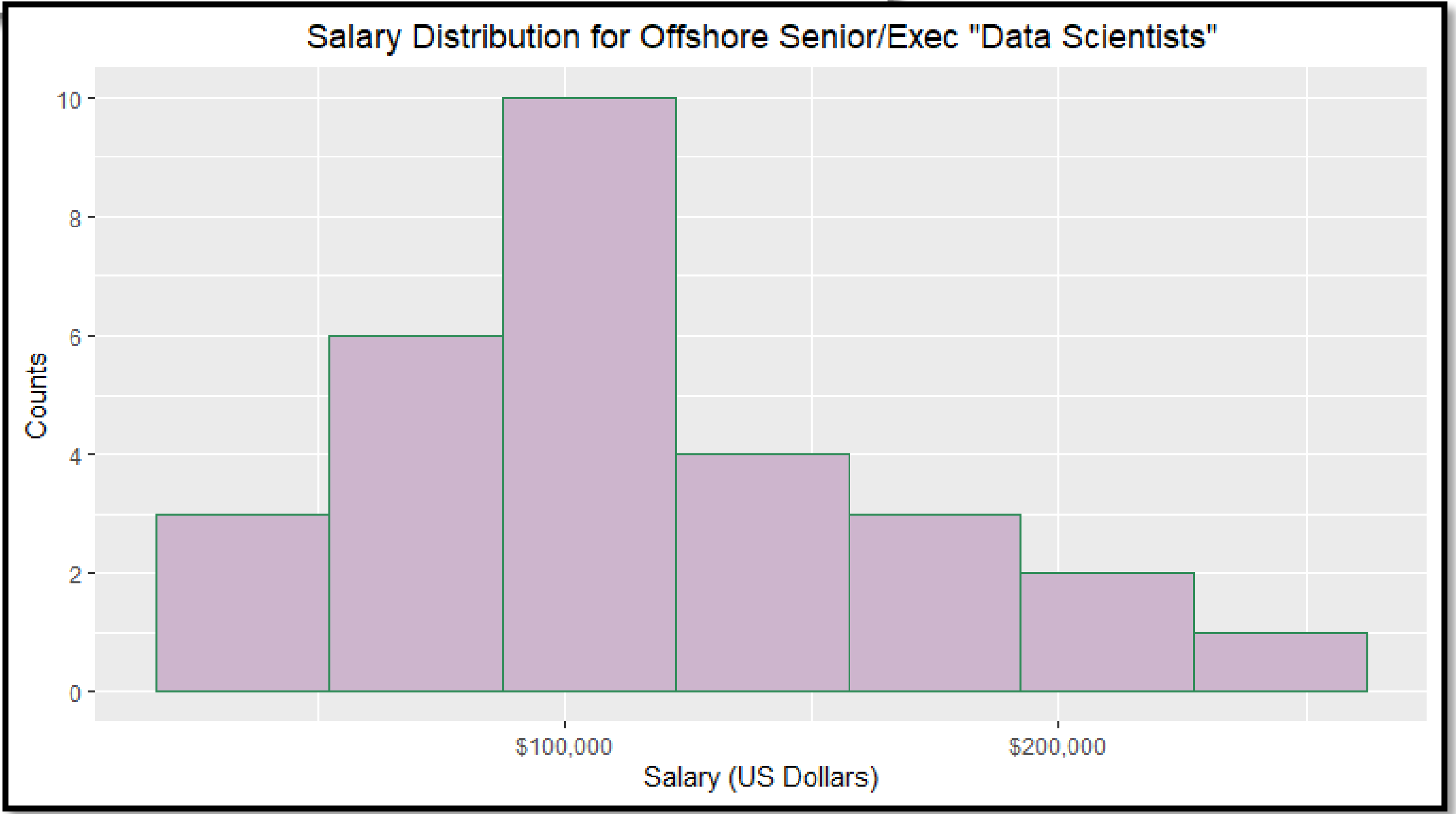
### *Senior & Executive experience*

### Salaries by Year

Although there appear to be some significantly low-paying salaries for "top-talent", it is difficult to make assumptions about what qualifies as competitive pay throughout other foreign countries.

However, to ensure that the Offshore statistics can be appropriately compared to the U.S. statistics, the job field must now be filtered to only include the Data Science field.





# PART II
# OFFSHORE DATA ANALYSIS

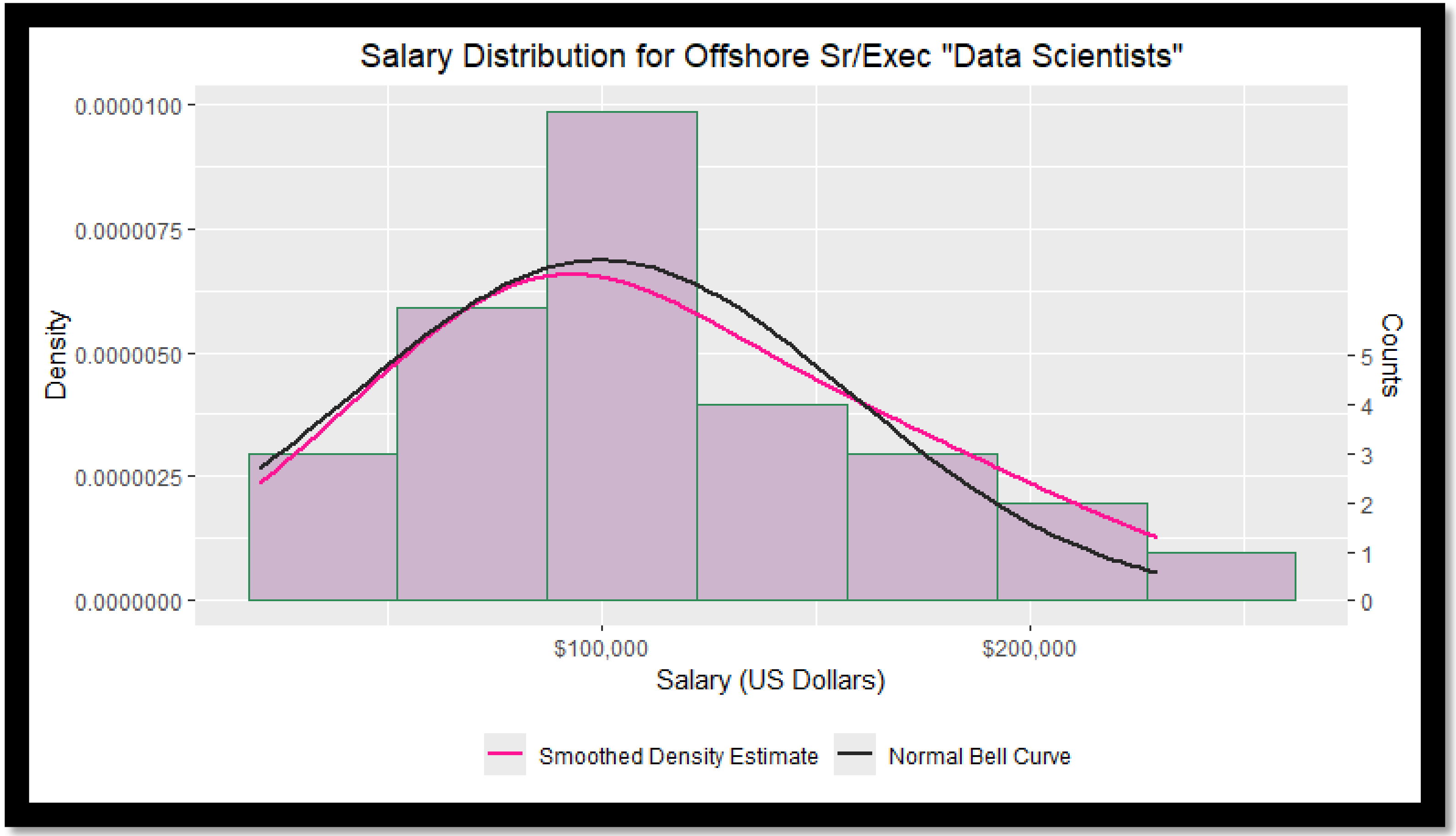
## *Senior & Executive experience*

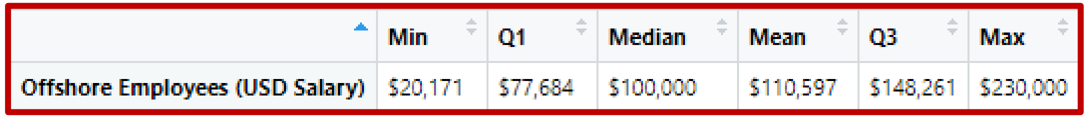
## Data Science Salary Distributions

These are the salary distribution plots for 29 offshore Data Science jobs at the senior & executive level.

Although not exact, the data does appear to be *relatively* normal.

To confirm, a density histogram can be plotted with a smoothed estimation curve, then compared to a bell curve.





| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| Offshore Employees (USD Salary) | $20,171 | $77,684 | $100,000 | $110,597 | $148,261 | $230,000 |

# PART II
# OFFSHORE DATA ANALYSIS

### *Senior & Executive experience*

## Data Science Salary Distributions and Statistics

The black "bell" curve illustrates what a normal distribution would be.
The pink density curve estimates the actual salary distribution.

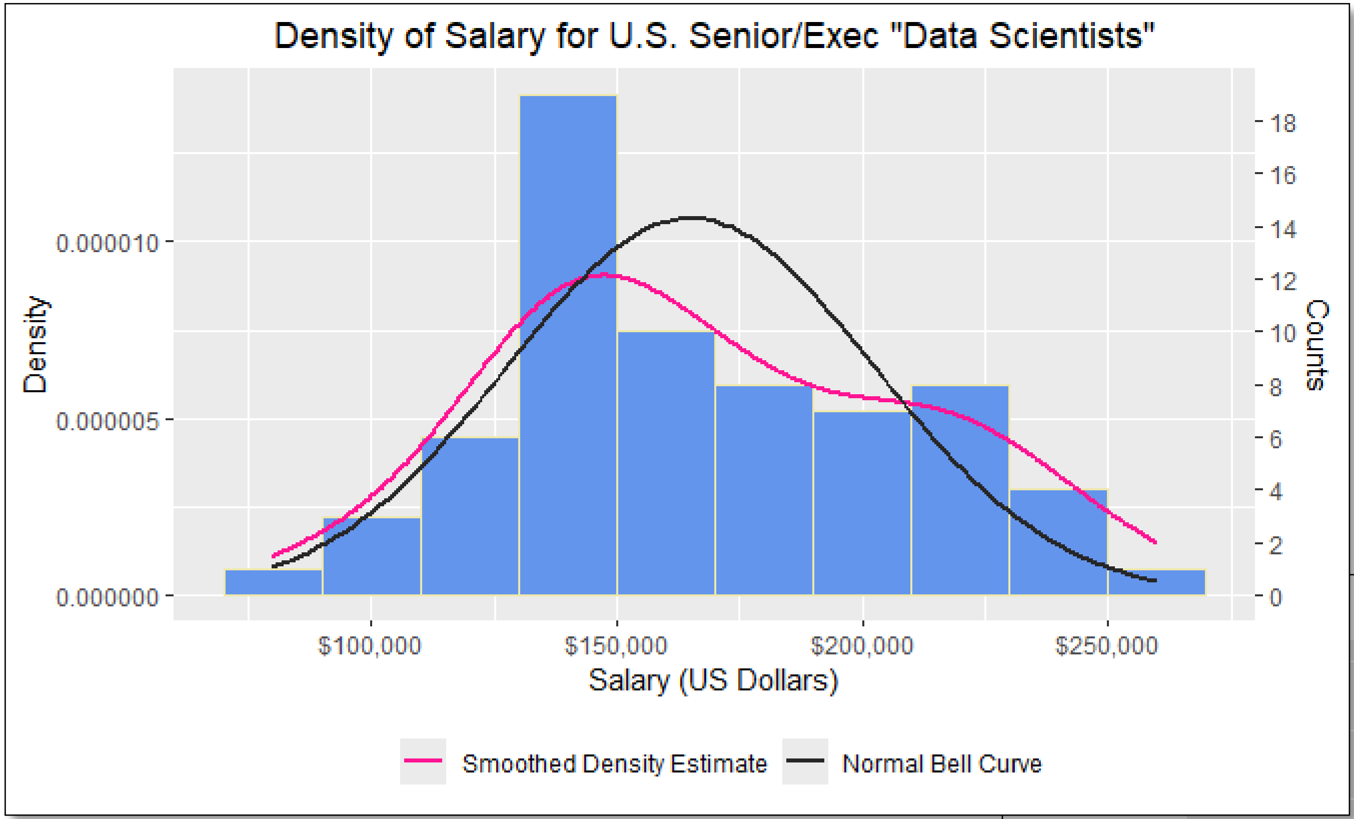
The curves and shapes appear to be very similar. This can be treated as a normal distribution.

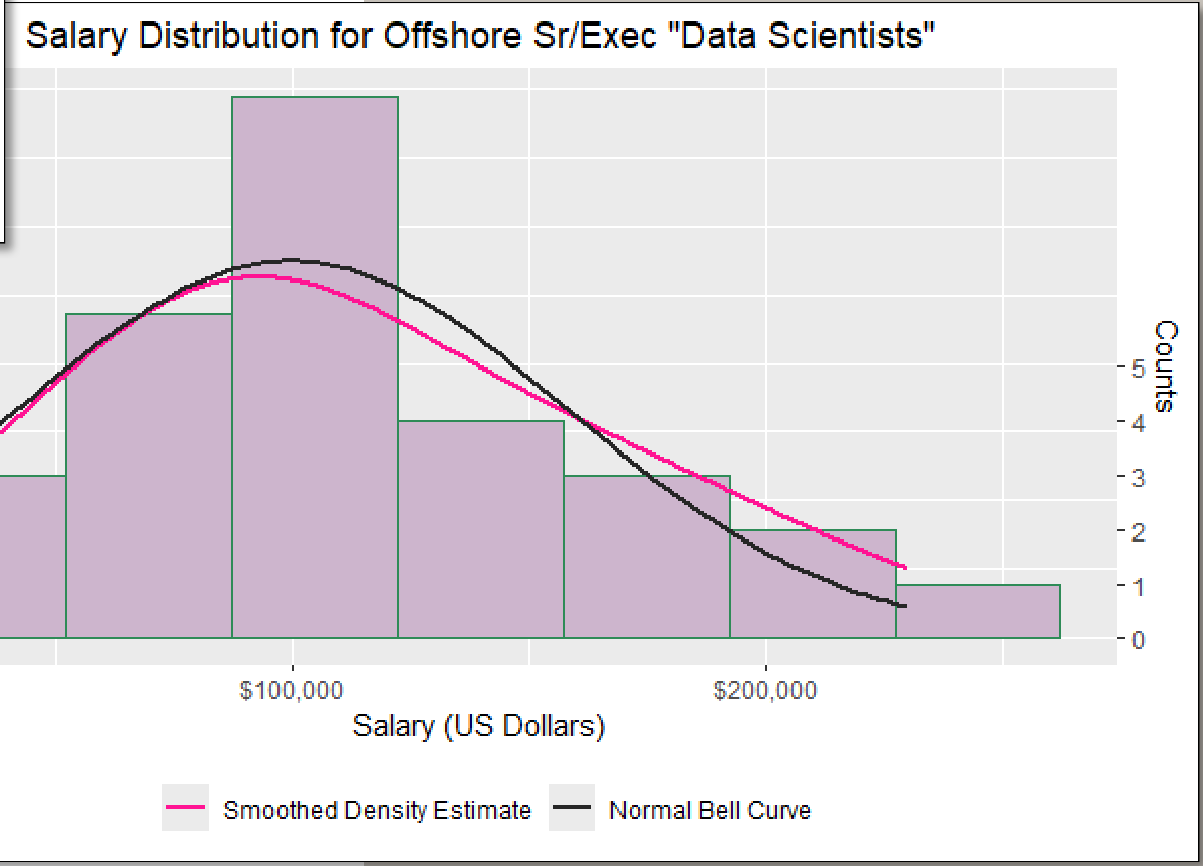
A table of important distribution statistics are displayed at the bottom, which will be used to compare against U.S. statistics and recommendations.

# PART III
# COMPARE U.S. / OFFSHORE

## Histograms and Distribution Curves

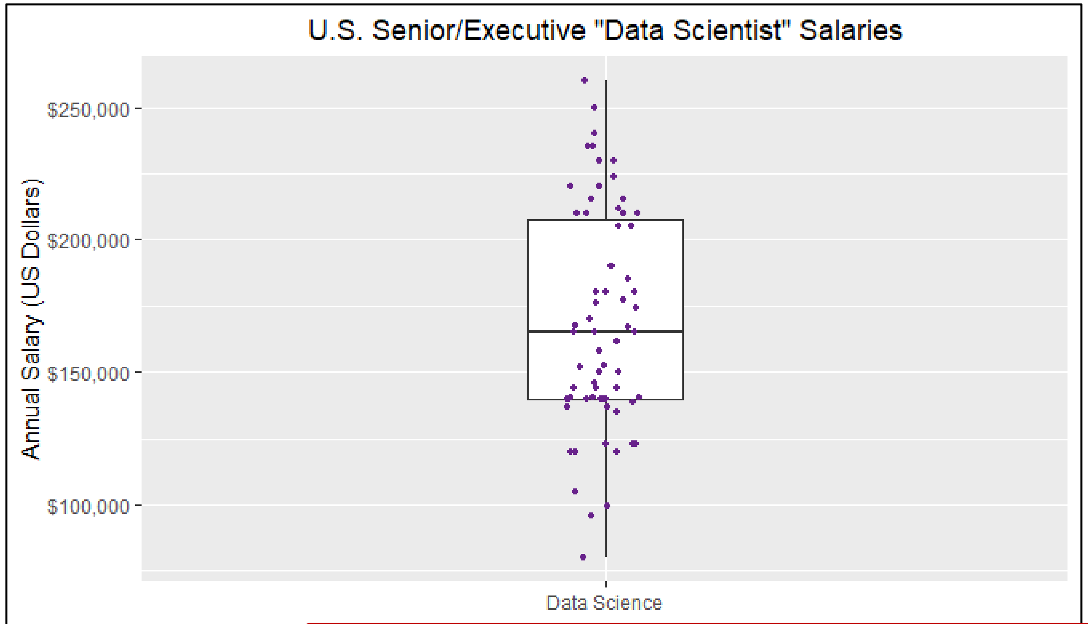
Both normal distributions, but U.S. shifted right by about $50k-$60k.

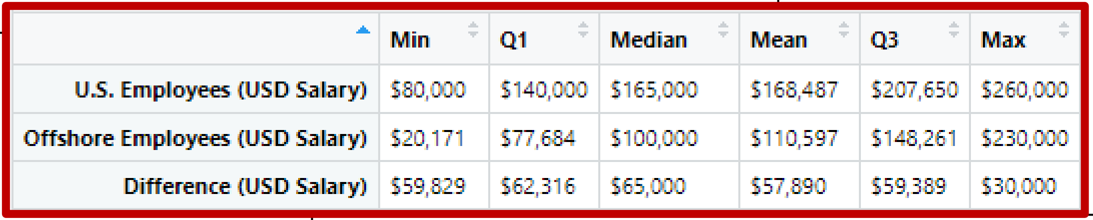




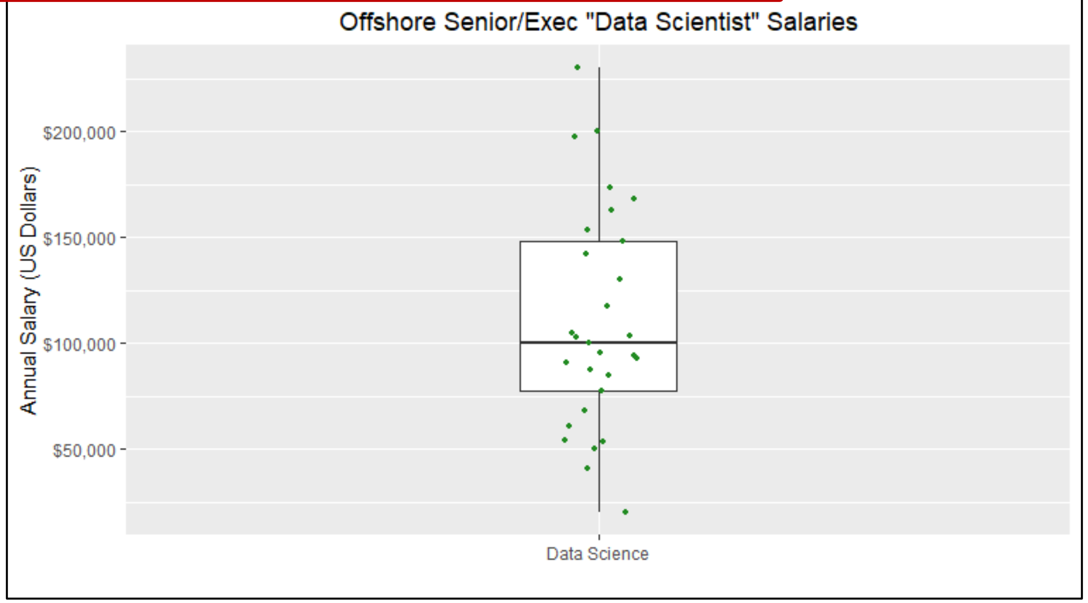
# PART III
# COMPARE U.S. / OFFSHORE

## Boxplots and Statistics





| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| U.S. Employees (USD Salary) | $80,000 | $140,000 | $165,000 | $168,487 | $207,650 | $260,000 |
| Offshore Employees (USD Salary) | $20,171 | $77,684 | $100,000 | $110,597 | $148,261 | $230,000 |
| Difference (USD Salary) | $59,829 | $62,316 | $65,000 | $57,890 | $59,389 | $30,000 |



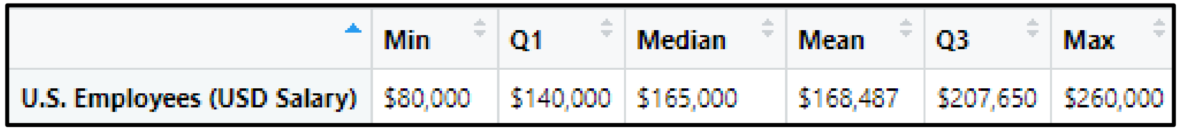
Boxplots show significant visual differences in U.S. vs. offshore distribution of salaries, especially looking at Q1 (bottom of the box which is 25th percentile), the Median (center line, 50th percentile), and at Q3 (top of the box which is 75th percentile).

The statistics clearly support the plots.

The offshore median is $65k less than the U.S., and the offshore mean is about $58k less than the U.S.

Q3 (75th percentile) offshore salaries are about $60k less than the U.S.

# RECOMMENDATION
## FOR
## U.S. SALARY RANGE



| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| U.S. Employees (USD Salary) | $80,000 | $140,000 | $165,000 | $168,487 | $207,650 | $260,000 |

The goal is to hire a top-talent data scientist to help drive data science within the organization, and potentially lead a team in the future.

The analysis of U.S. Data Science salaries for senior & executive level employees has produced the above statistics, which follow a near-normal distribution.

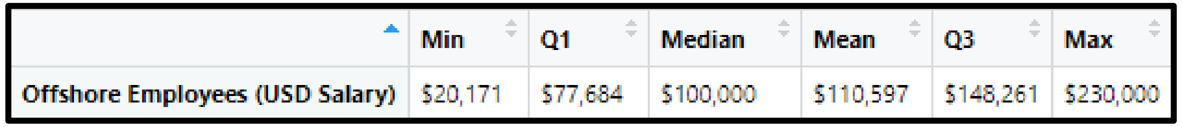
The recommended salary range takes these statistics into consideration, as well as the "competitive" mindset for obtaining our goal.

It is my formal recommendation to use the Median value as the low-end of the salary range, and to use the Q3 value as the high-end of the salary range. This produces a competitive salary range of $42,650:

$165,000 (low) - $207,650 (high)

With that in mind, if the candidate proves to be extremely desirable and the Q3 value is not enough, then the "Max" listed in the statistics ($260,000) can be considered as an **absolute max**. Anything over that number may put our company at financial risk.

# RECOMMENDATION
## FOR
## OFFSHORE SALARY RANGE



| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| Offshore Employees (USD Salary) | $20,171 | $77,684 | $100,000 | $110,597 | $148,261 | $230,000 |

The offshore statistics are complicated, and additional data analysis is likely needed.

For now, my formal recommendation is to follow the same logic as the U.S. range. Use the Median value as the low-end of the salary range and use the Q3 value as the high-end of the salary range, yielding a competitive salary range of $48,261:

$100,000 (low) - $148,261 (high)

Again, the “Max” listed in the statistics ($230,000) can be considered an **absolute max**. Anything over that number may put our company at financial risk.

However, there are more complications to consider with an offshore hire, some of which are not available in the current data. For example, language differences can potentially cause communication issues, and time zone differences can introduce scheduling problems.

Because of this, I **highly recommend** hiring a U.S. employee for the initial Data Science position. This will allow us more time to further analyze the data and consider some of the complexities.

Additional analysis of salaries by specific countries or regions may provide better clarity on the drastic salary ranges observed in the current data set.

# ARE THERE EXTRACTABLE INSIGHTS TO HELP DEVELOP A FUTURE DATA SCIENCE TEAM?

This is the last question from the initial list of research questions.

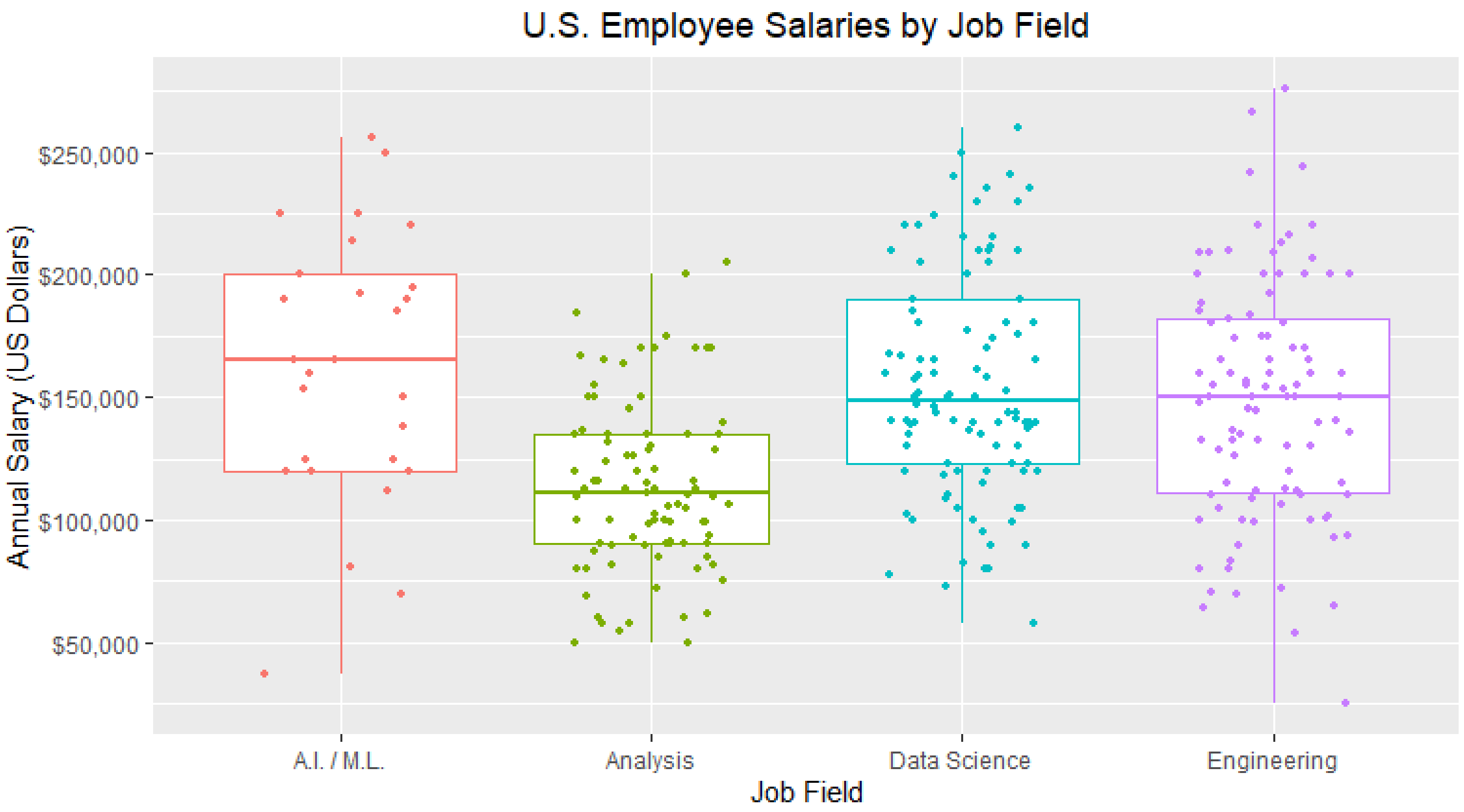
Full analysis on this topic is not within the scope of this recommendation, but a quick high-level look at some data across job fields in the U.S. may help keep things in perspective in terms of a current data science hire and a future data science team.

A future data science team may require additional data scientists, as well as machine learning engineers, data engineers/architects, and data analysts.

The following slides demonstrate some basic insights available that may help with future considerations.

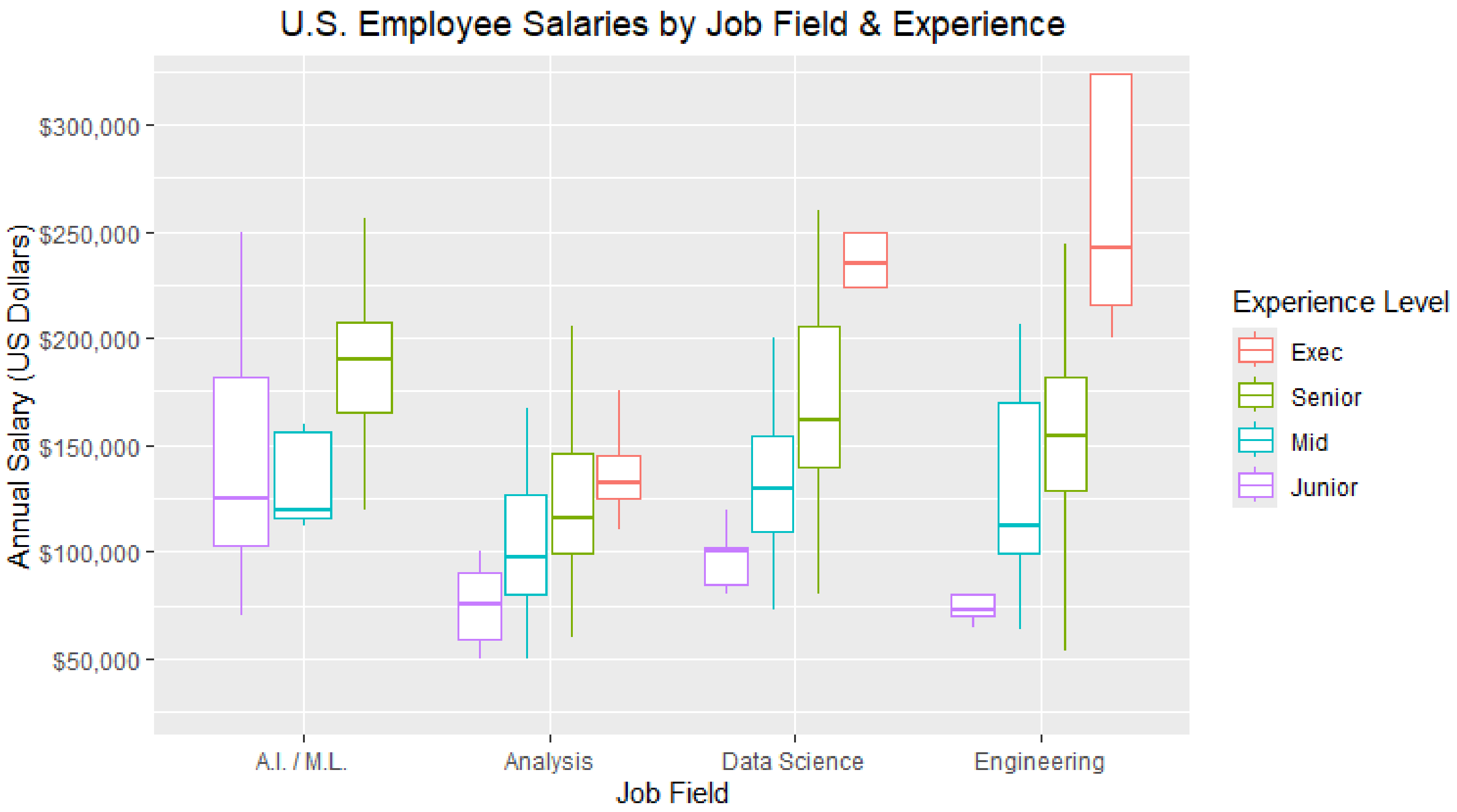
# FUTURE CONSIDERATIONS

DATA SCIENCE TEAM



# FUTURE CONSIDERATIONS

DATA SCIENCE TEAM


U.S. Employee Salaries by Job Field & Experience
Annual Salary (US Dollars)
$300,000
$250,000
$200,000
$150,000
$100,000
$50,000
A.I. / M.L.
Analysis
Data Science
Engineering
Job Field
Experience Level
Exec
Senior
Mid
Junior