BAIS:3200
Final Project

Alex Bender, Nathan Kurtz, Reilly Tabares, Teja Jetty, Jack Wilson

Introduction

Every data scientist wants to earn a higher salary, but how can they attain a higher salary? In our project, we applied data-driven analysis to identify key factors such as experience, work setting, location, and more that contribute to salaries in the popular field of data science. Our database application and analysis may be useful for data scientists looking to increase their salary, or serve as a tool for data scientists and aspiring data scientists to examine their earnings potential in the data science field.

Data

Our project uses data collected from a 2024 Kaggle dataset about the data science industry across the world (Job and Salaries in Data Science). The original dataset contained 9,355 rows across 12 items. Our dataset was reduced to 10 items relevant to our analysis, and we reduced the size of the dataset to only include data from 2022-2023, leaving us with 9,087 rows across the 10 items. To uniquely identify records in our dataset, four primary keys were created, one for each relation. Table 1 displays a description of the data.

Table 1 Data Dictionary

Field	Type	Description
employment_id	Text	Unique ID for each employment
company_id	Text	Unique ID for each company
job_id	Text	Unique ID for each job
salary_id	Text	Unique ID for each salary
work_year	Numeric	Indicates the year the data was recorded
job_title	Text	Specifies the specific job role
job_category	Text	Classifies the job role into broader categories
salary_in_usd	Numeric	Presents the annual gross salary converted to United States Dollars
employee_residence	Text	Specifies the employee's country of residence
experience_level	Text	Classifies the professional experience level
employment_type	Text	Specifies the type of employment (e.g., full-time, part-time, contract)
work_setting	Text	Describes the work setting (e.g., hybrid, remote, in-person)
company_location	Text	Indicates the country where the company is located
company_size	Text	Represents the size of the employer company

The dataset features three strong entities. First is EMPLOYMENT, identified by employment_id, second is COMPANY, identified by company_id, and third is JOB, identified by job_id. All attributes in each entity are required, and there are no multivalued attributes. Additionally, we created an associative entity, SALARY, identified by salary_id, to connect the three strong entities. Each strong entity has a one-to-many relationship with the SALARY entity, and the primary keys of the strong entities are stored as foreign keys in the SALARY entity. Figure 1 displays the ERD for our data.

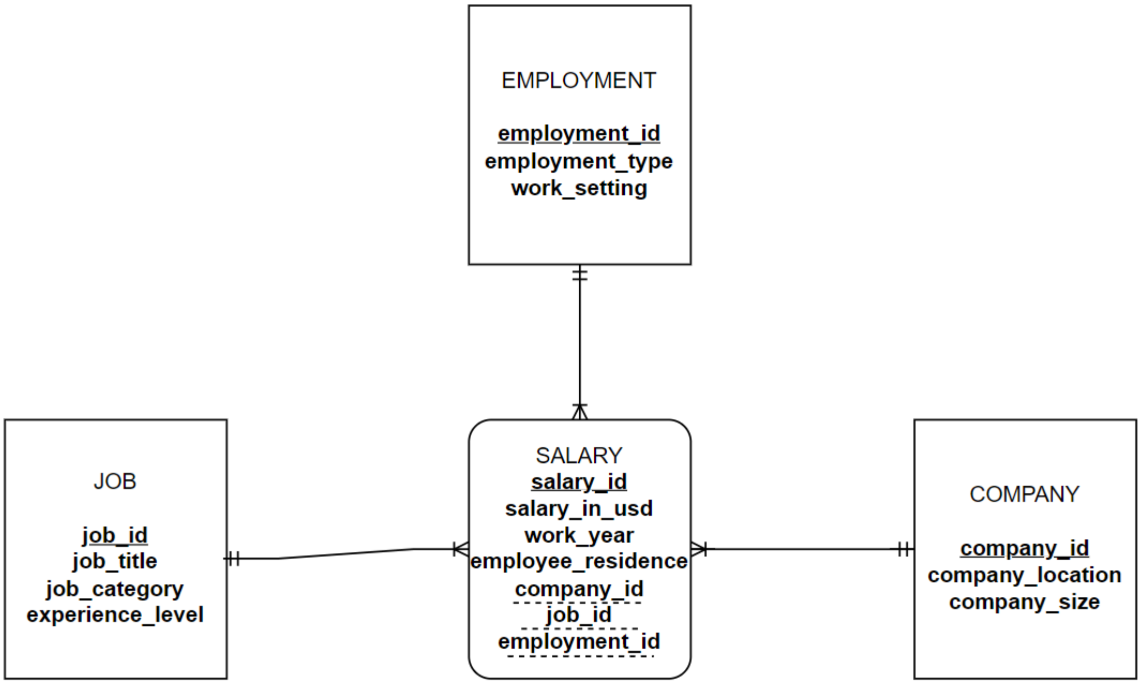
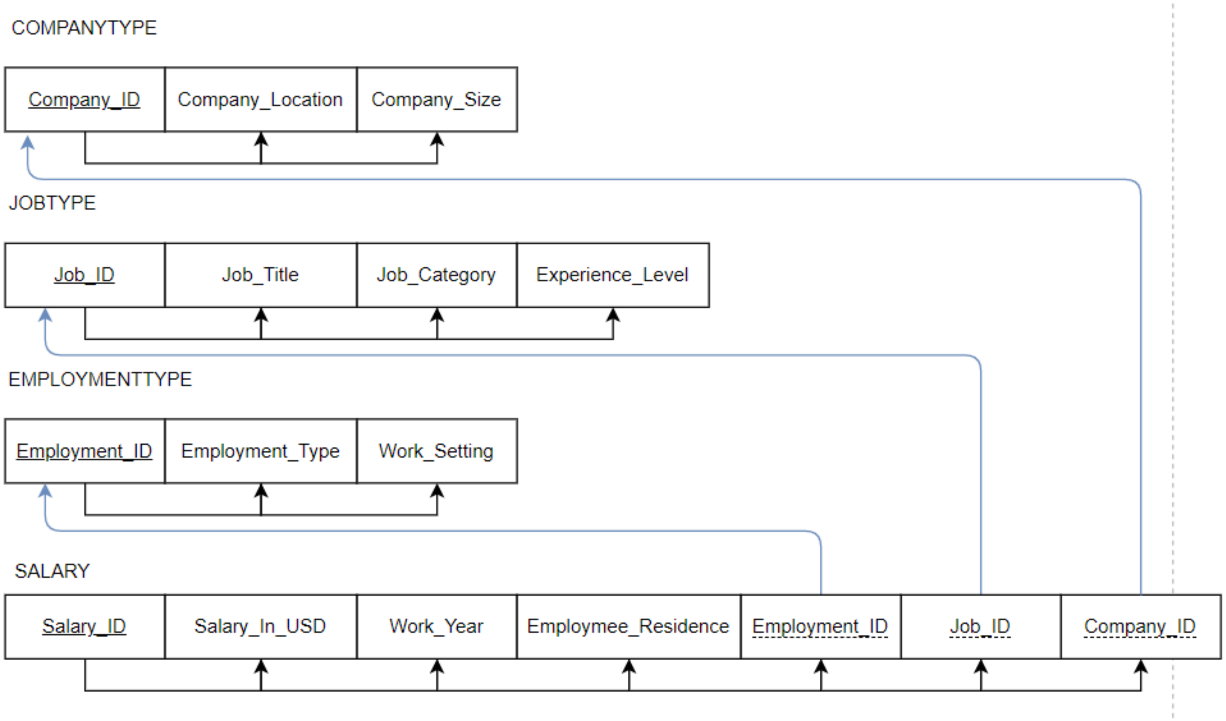


Fig. 1 Entity Relationship Diagram



COMPANYTYPE

Company_ID	Company_Location	Company_Size

JOBTYPE

Job_ID	Job_Title	Job_Category	Experience_Level

EMPLOYMENTTYPE

Employment_ID	Employment_Type	Work_Setting

SALARY

Salary_ID	Salary_In_USD	Work_Year	Employee_Residence	Employment_ID	Job_ID	Company_ID

Fig. 2 Graphical Relational Schema

Database Implementation

COMPANY Table:

```
CREATE TABLE COMPANY (
Company_ID VARCHAR(12) NOT NULL,
Company_Location VARCHAR(25) NOT NULL,
Company_Size CHAR(1) NOT NULL,
CONSTRAINT COMPANY_PK PRIMARY KEY (Company_ID)
);
```

JOB Table:

```
CREATE TABLE JOB (
job_id varchar(25) not null,
job_title varchar(50) not null,
job_category varchar(50) not null,
experience_level varchar(25) not null,
CONSTRAINT job_pk PRIMARY KEY (job_id)
);
```

EMPLOYMENT Table:

```
CREATE TABLE EMPLOYMENT (
EMPLOYMENT_ID VARCHAR(20) NOT NULL,
EMPLOYMENT_TYPE VARCHAR(20) NOT NULL,
WORK_SETTING VARCHAR(20) NOT NULL,
CONSTRAINT EMPLOYMENT_PK PRIMARY KEY (EMPLOYMENT_ID)
);
```

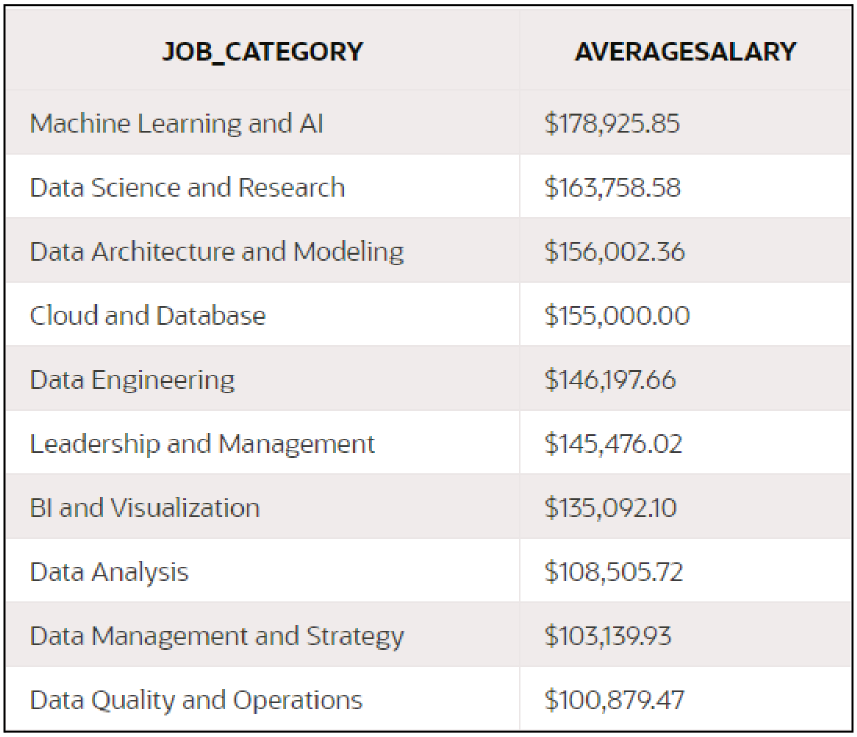
SALARY Table:

```
CREATE TABLE SALARY (
SALARY_ID VARCHAR2(10) NOT NULL,
COMPANY_ID VARCHAR2(25) NOT NULL,
JOB_ID VARCHAR2(10) NOT NULL,
EMPLOYMENT_ID VARCHAR2(20) NOT NULL,
SALARY_IN_USD NUMBER(7,0) NOT NULL,
WORK_YEAR NUMBER(4,0) NOT NULL,
EMPLOYEE_RESIDENCE VARCHAR2(25) NOT NULL,
CONSTRAINT SALARY_PK PRIMARY KEY (SALARY_ID),
CONSTRAINT COMPANY_FK FOREIGN KEY (COMPANY_ID) REFERENCES COMPANY
(COMPANY_ID),
CONSTRAINT EMPLOYMENT_FK FOREIGN KEY (EMPLOYMENT_ID) REFERENCES EMPLOYMENT
(EMPLOYMENT_ID),
CONSTRAINT JOB_FK FOREIGN KEY (JOB_ID) REFERENCES JOB (JOB_ID)
);
```

Analysis

Question 1: What are the differences in pay for different job categories?

```
SELECT Job_Category, TO_CHAR(AVG(Salary_In_USD), '$999,999.00') AS
AverageSalary
FROM JOB
    JOIN SALARY ON JOB.Job_ID = SALARY.Job_ID
GROUP BY Job_Category
ORDER BY AverageSalary DESC;
```



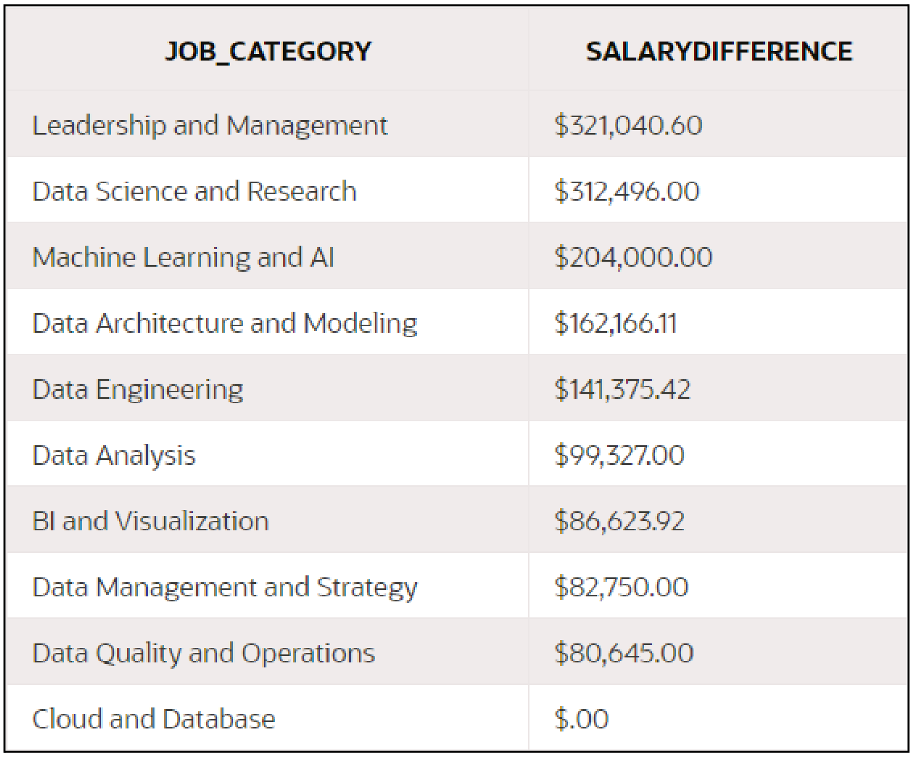
JOB_CATEGORY	AVERAGESALARY
Machine Learning and AI	$178,925.85
Data Science and Research	$163,758.58
Data Architecture and Modeling	$156,002.36
Cloud and Database	$155,000.00
Data Engineering	$146,197.66
Leadership and Management	$145,476.02
BI and Visualization	$135,092.10
Data Analysis	$108,505.72
Data Management and Strategy	$103,139.93
Data Quality and Operations	$100,879.47

Fig. 3 Average Salary by Category

Our data suggests that the job category has a large impact on average salary. Salaries range from an average of $100,879 in Data Quality and Operations to $178,925 in Machine Learning & AI. A difference of over $70,000 from the minimum average salary to the maximum average salary demonstrates that a large difference in average salary based on job category exists.

Question 2: Are there major differences in the salaries for different job titles in the same category?

```
SELECT JOB.Job_Category, TO_CHAR(MAX(s.AverageSalary) - MIN(s.AverageSalary),
'$999,999.00') AS SalaryDifference
FROM JOB
    JOIN SALARY ON JOB.Job_ID = SALARY.Job_ID
    JOIN (SELECT Job_Category, Job_Title, TO_CHAR(AVG(Salary_In_USD),
‘$999,999.00”) AS AverageSalary
    FROM JOB
        JOIN SALARY ON JOB.Job_ID = SALARY.Job_ID
    GROUP BY Job_Category, Job_Title) s ON JOB.Job_Category = s.Job_Category
GROUP BY JOB.Job_Category
ORDER BY AverageSalary DESC;
```



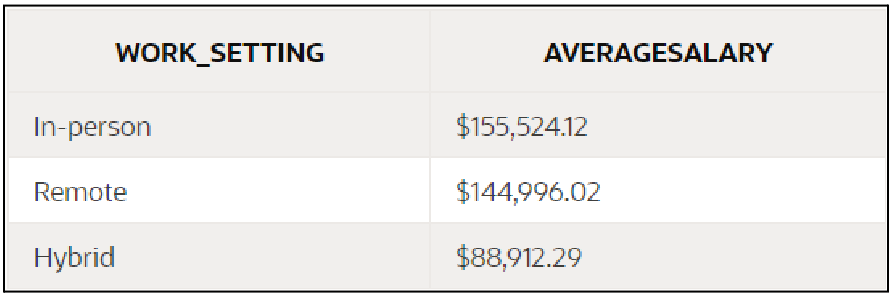
JOB_CATEGORY	SALARYDIFFERENCE
Leadership and Management	$321,040.60
Data Science and Research	$312,496.00
Machine Learning and AI	$204,000.00
Data Architecture and Modeling	$162,166.11
Data Engineering	$141,375.42
Data Analysis	$99,327.00
BI and Visualization	$86,623.92
Data Management and Strategy	$82,750.00
Data Quality and Operations	$80,645.00
Cloud and Database	$.00

Fig. 4 Max Difference in Average Salary for Job Titles in the Same Category

Our data shows that different jobs within the same category can come with major salary differences. For example, in the Machine Learning and AI category, salaries can differ by as much as 204,000 dollars based on job title. The Cloud and Database category has a $0 difference because the category only has one job title in our dataset, Cloud Database Engineer.

<u>Question 3</u>: What work setting pays the highest salary?

```
SELECT Work_Setting, TO_CHAR(AVG(Salary_In_USD), '$999,999.00') AS
AverageSalary
FROM EMPLOYMENT
    JOIN SALARY ON EMPLOYMENT.Employment_ID = SALARY.Employment_ID
GROUP BY Work_Setting
ORDER BY AverageSalary DESC;
```



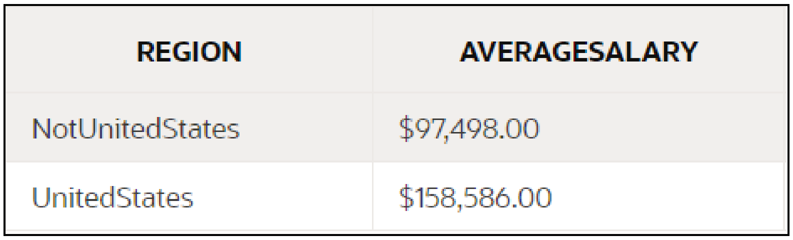
WORK_SETTING	AVERAGESALARY
In-person	$155,524.12
Remote	$144,996.02
Hybrid	$88,912.29

Fig. 5 Average Salary by Work Setting

The “In-person” work setting pays the highest average salary based on our data, with “Remote” coming in second, and “Hybrid” in last. “In-person” has the highest average, which makes logical sense, but it is interesting that “Hybrid” is the lowest, as a “Hybrid” working arrangement features both “In-person” and “Remote” components.

Question 4: Do Data Scientists get paid more when working outside of the United States? What proportion of Data Scientists work in the United States?

```
select CASE
    when employee_residence != 'United States' then 'NotUnitedStates'
    else 'UnitedStates'
    end as Region,
    to_char(avg(salary_in_usd), '$999,999.00') as AverageSalary
from salary
group by CASE
    when employee_residence != 'United States' then 'NotUnitedStates'
    else 'UnitedStates'
    end;
```

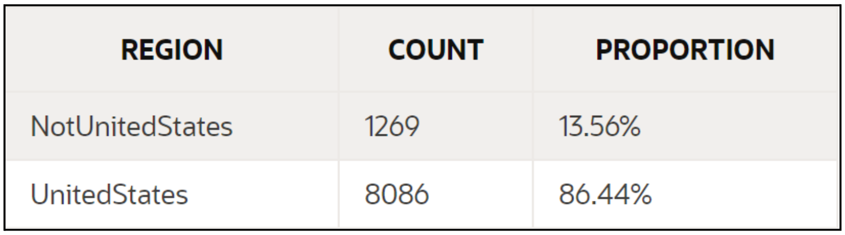


REGION	AVERAGESALARY
NotUnitedStates	$97,498.00
UnitedStates	$158,586.00

Fig. 6 Average Salary by Work Location

According to our data, data scientists working in the United States make about $61,088 more per year on average than their counterparts outside the United States.

```
SELECT CASE
    WHEN Employee_Residence != 'United States' THEN 'NotUnitedStates'
    ELSE 'UnitedStates'
    END AS Region, COUNT(Employment_ID) AS COUNT,
    ROUND((COUNT(Employment_ID) / (SELECT COUNT(Employment_ID) FROM SALARY))
* 100, 2) || '%' AS Proportion
FROM SALARY
GROUP BY CASE
    WHEN Employee_Residence != 'United States' THEN 'NotUnitedStates'
    ELSE 'UnitedStates'
    END;
```



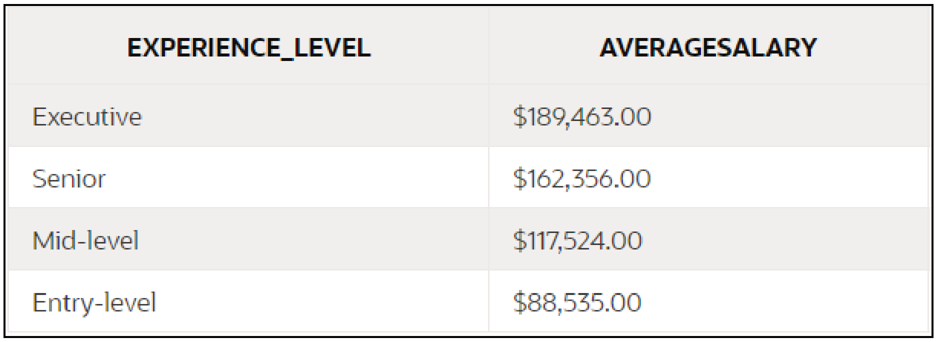
REGION	COUNT	PROPORTION
NotUnitedStates	1269	13.56%
UnitedStates	8086	86.44%

Fig. 7 Proportion of Workers by Work Location

The vast majority – over 85% – of the Data Scientists in our dataset work in the United States, which is higher than we expected as our dataset is composed of global data. A possible explanation for the high proportion of American data scientists is that the data may have been collected more heavily on U.S. based jobs, or the U.S. is simply leading the Data Science industry.

Question 5: How much does experience level affect salary?

```
SELECT JOB.experience_level, TO_CHAR(AVG(SALARY.salary_in_usd),
'$999,999.00') as AverageSalary
FROM SALARY
FULL OUTER JOIN JOB on JOB.job_id = SALARY.job_id
GROUP BY JOB.experience_level
ORDER BY AverageSalary DESC;
```



EXPERIENCE_LEVEL	AVERAGESALARY
Executive	$189,463.00
Senior	$162,356.00
Mid-level	$117,524.00
Entry-level	$88,535.00

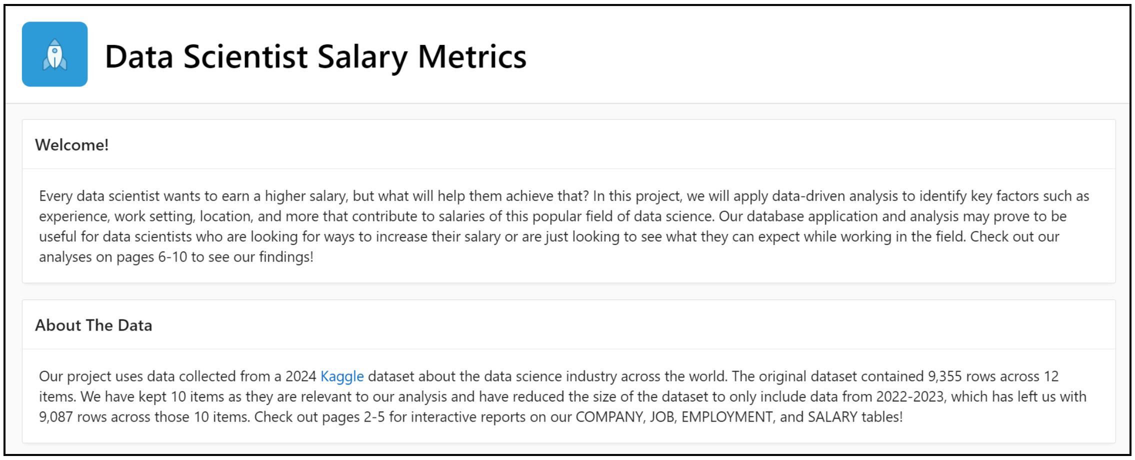
Fig. 8 Average Salary by Work Experience

The analysis of our dataset demonstrates that a data scientist's experience level leads to major differences in average salary. On average, a data scientist with more experience will be paid more than a data scientist with less experience. For example, if a data scientist progresses from entry-level to executive level, their annual salary can nearly double based on the trends in the dataset.

Web Application

[Database Management Final Project Application]

Home Page: This page gives an introduction to our application as well as a short description of the dataset we used for our analyses.



Data Scientist Salary Metrics

Welcome!

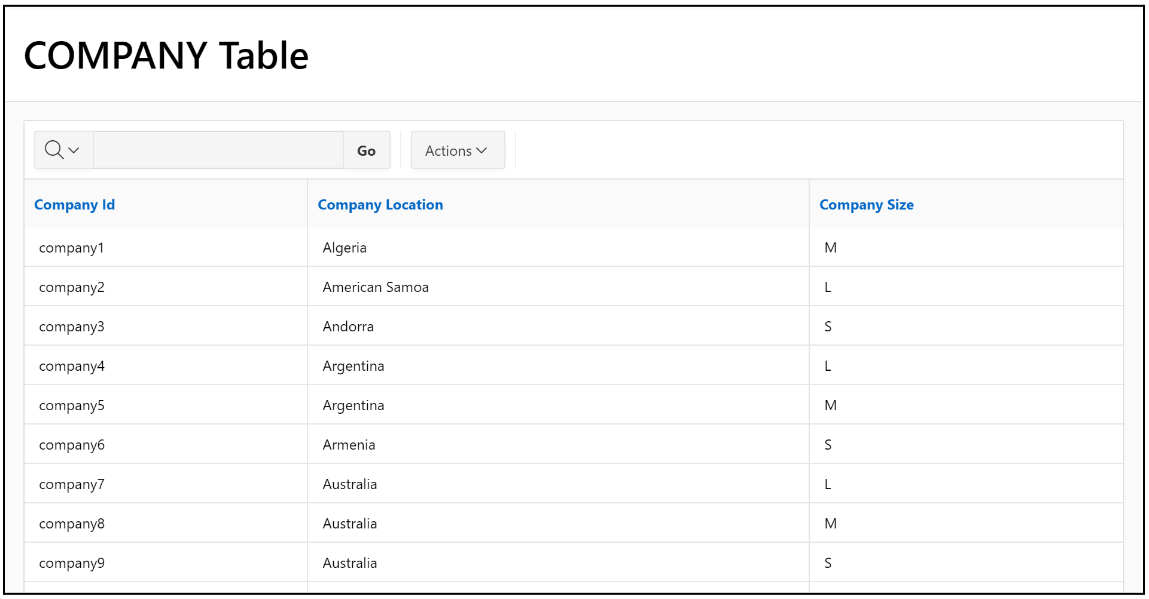
Every data scientist wants to earn a higher salary, but what will help them achieve that? In this project, we will apply data-driven analysis to identify key factors such as experience, work setting, location, and more that contribute to salaries of this popular field of data science. Our database application and analysis may prove to be useful for data scientists who are looking for ways to increase their salary or are just looking to see what they can expect while working in the field. Check out our analyses on pages 6-10 to see our findings!

About The Data

Our project uses data collected from a 2024 Kaggle dataset about the data science industry across the world. The original dataset contained 9,355 rows across 12 items. We have kept 10 items as they are relevant to our analysis and have reduced the size of the dataset to only include data from 2022-2023, which has left us with 9,087 rows across those 10 items. Check out pages 2-5 for interactive reports on our COMPANY, JOB, EMPLOYMENT, and SALARY tables!

Fig. 9 Home Page (Page 1)

COMPANY Table: This page displays an interactive report for our COMPANY table.

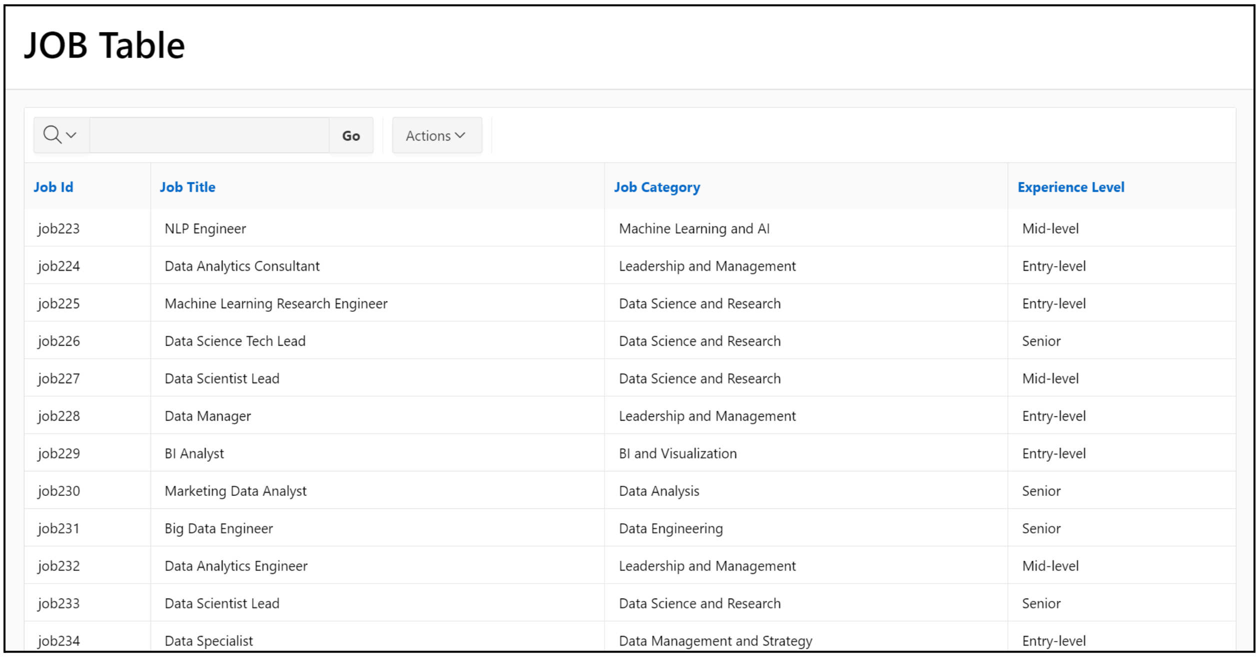


COMPANY Table

Company Id	Company Location	Company Size
company1	Algeria	M
company2	American Samoa	L
company3	Andorra	S
company4	Argentina	L
company5	Argentina	M
company6	Armenia	S
company7	Australia	L
company8	Australia	M
company9	Australia	S

Fig. 10 COMPANY (Page 2)

JOB Table: This page displays an interactive report for our JOB table.

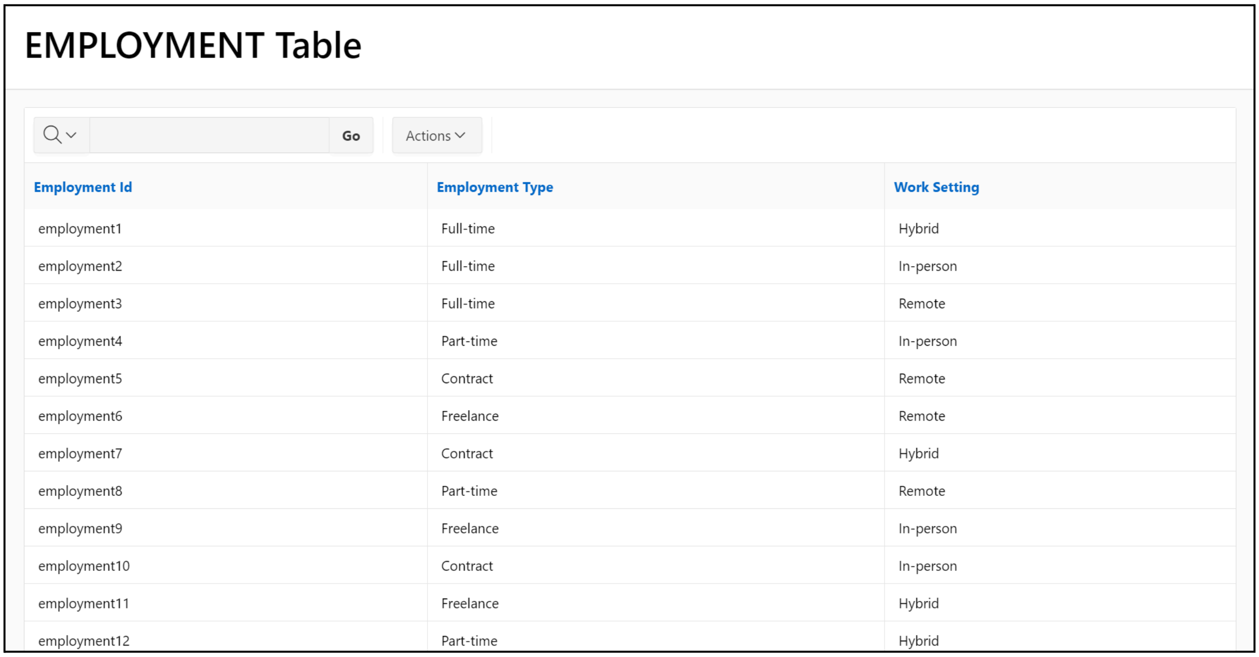


JOB Table

Job Id	Job Title	Job Category	Experience Level
job223	NLP Engineer	Machine Learning and AI	Mid-level
job224	Data Analytics Consultant	Leadership and Management	Entry-level
job225	Machine Learning Research Engineer	Data Science and Research	Entry-level
job226	Data Science Tech Lead	Data Science and Research	Senior
job227	Data Scientist Lead	Data Science and Research	Mid-level
job228	Data Manager	Leadership and Management	Entry-level
job229	BI Analyst	BI and Visualization	Entry-level
job230	Marketing Data Analyst	Data Analysis	Senior
job231	Big Data Engineer	Data Engineering	Senior
job232	Data Analytics Engineer	Leadership and Management	Mid-level
job233	Data Scientist Lead	Data Science and Research	Senior
job234	Data Specialist	Data Management and Strategy	Entry-level

Fig. 11 JOB (Page 3)

EMPLOYMENT Table: This page displays an interactive report for our EMPLOYMENT table.

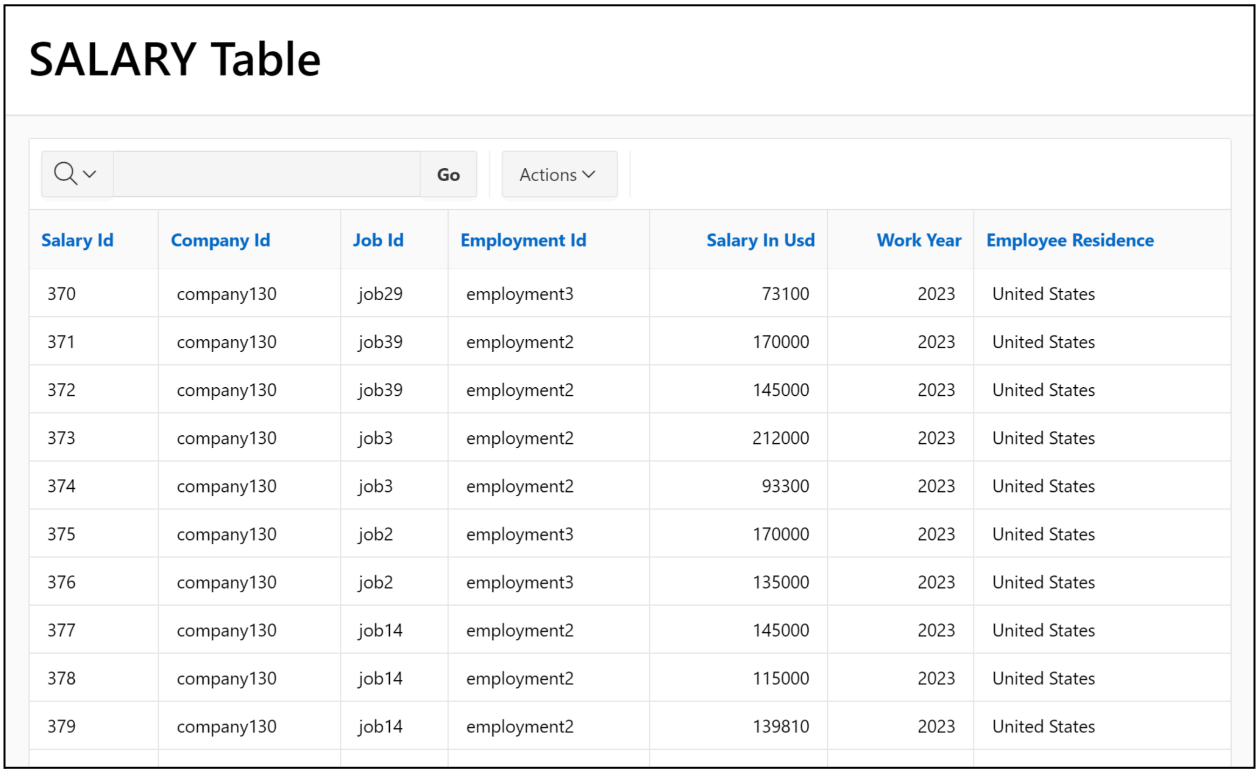


EMPLOYMENT Table

Employment Id	Employment Type	Work Setting
employment1	Full-time	Hybrid
employment2	Full-time	In-person
employment3	Full-time	Remote
employment4	Part-time	In-person
employment5	Contract	Remote
employment6	Freelance	Remote
employment7	Contract	Hybrid
employment8	Part-time	Remote
employment9	Freelance	In-person
employment10	Contract	In-person
employment11	Freelance	Hybrid
employment12	Part-time	Hybrid

Fig. 12 EMPLOYMENT (Page 4)

SALARY Table: This page displays an interactive report for our SALARY table.

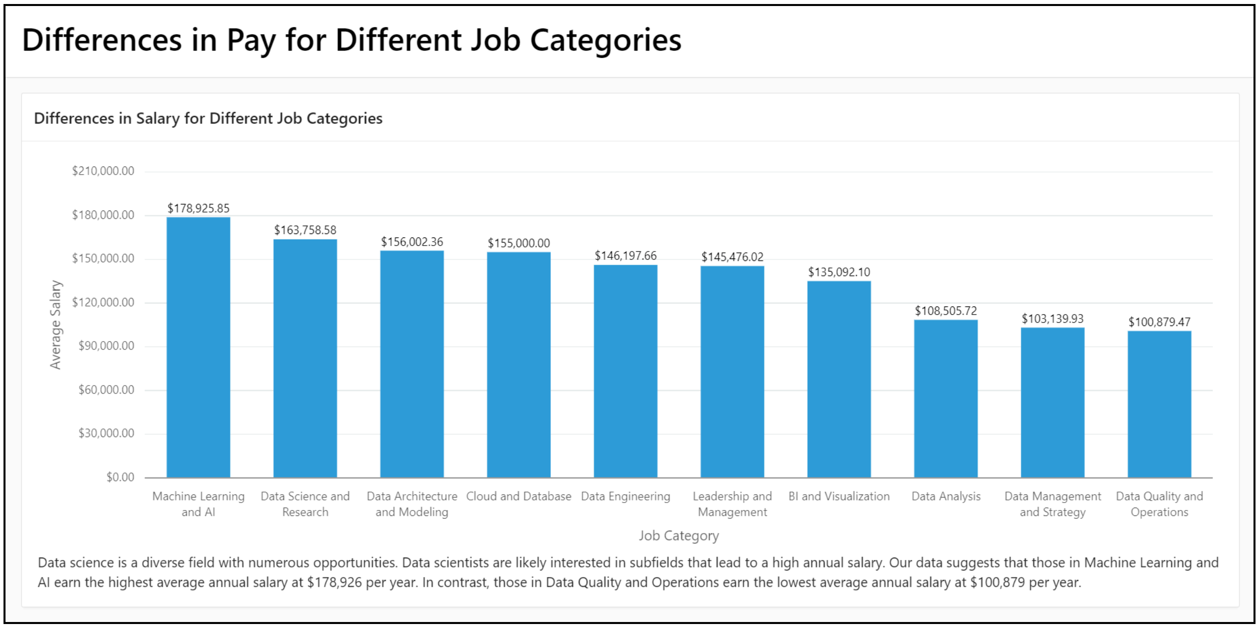


SALARY Table

Salary Id	Company Id	Job Id	Employment Id	Salary In Usd	Work Year	Employee Residence
370	company130	job29	employment3	73100	2023	United States
371	company130	job39	employment2	170000	2023	United States
372	company130	job39	employment2	145000	2023	United States
373	company130	job3	employment2	212000	2023	United States
374	company130	job3	employment2	93300	2023	United States
375	company130	job2	employment3	170000	2023	United States
376	company130	job2	employment3	135000	2023	United States
377	company130	job14	employment2	145000	2023	United States
378	company130	job14	employment2	115000	2023	United States
379	company130	job14	employment2	139810	2023	United States

Fig. 13 SALARY (Page 5)

Differences in Salary for Different Job Categories: Page 6 displays a bar chart comparing average salaries across different data science job categories.



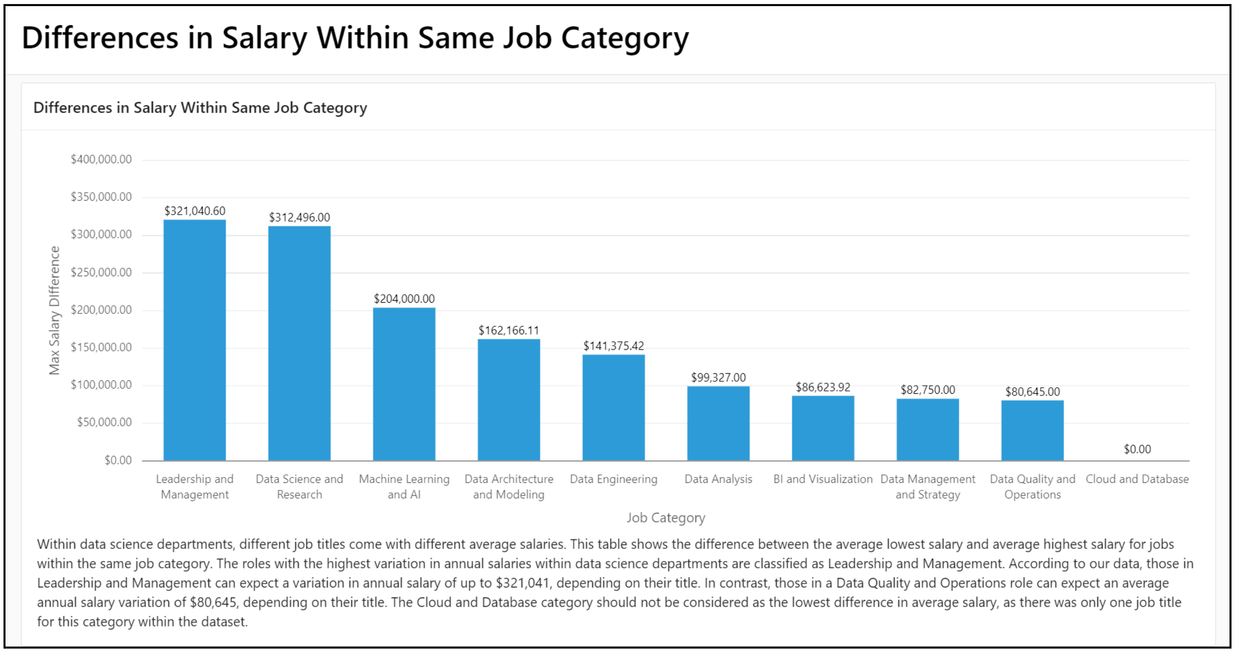
Differences in Pay for Different Job Categories

Differences in Salary for Different Job Categories

Data science is a diverse field with numerous opportunities. Data scientists are likely interested in subfields that lead to a high annual salary. Our data suggests that those in Machine Learning and AI earn the highest average annual salary at $178,926 per year. In contrast, those in Data Quality and Operations earn the lowest average annual salary at $100,879 per year.

Fig. 14 Average Salary by Job Category (Page 6)

Differences in Salary Within Same Job Category: Page 7 displays a bar chart comparing the highest differences in salaries within the same job category.



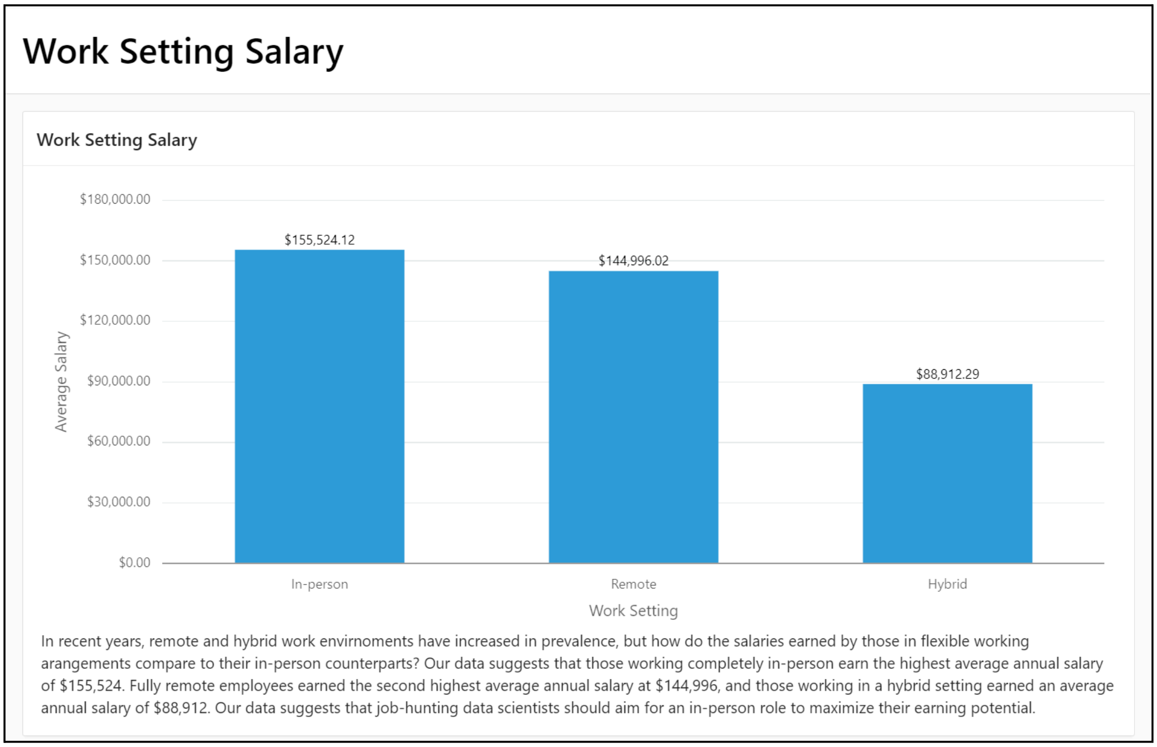
Differences in Salary Within Same Job Category

Differences in Salary Within Same Job Category

Within data science departments, different job titles come with different average salaries. This table shows the difference between the average lowest salary and average highest salary for jobs within the same job category. The roles with the highest variation in annual salaries within data science departments are classified as Leadership and Management. According to our data, those in Leadership and Management can expect a variation in annual salary of up to $321,041, depending on their title. In contrast, those in a Data Quality and Operations role can expect an average annual salary variation of $80,645, depending on their title. The Cloud and Database category should not be considered as the lowest difference in average salary, as there was only one job title for this category within the dataset.

Fig. 15 Max Difference in Average Salary for Job Titles in the same Category (Page 7)

Work Setting Salary: Page 8 displays a bar chart comparing the average salary by work setting (i.e., in-person, remote, hybrid).



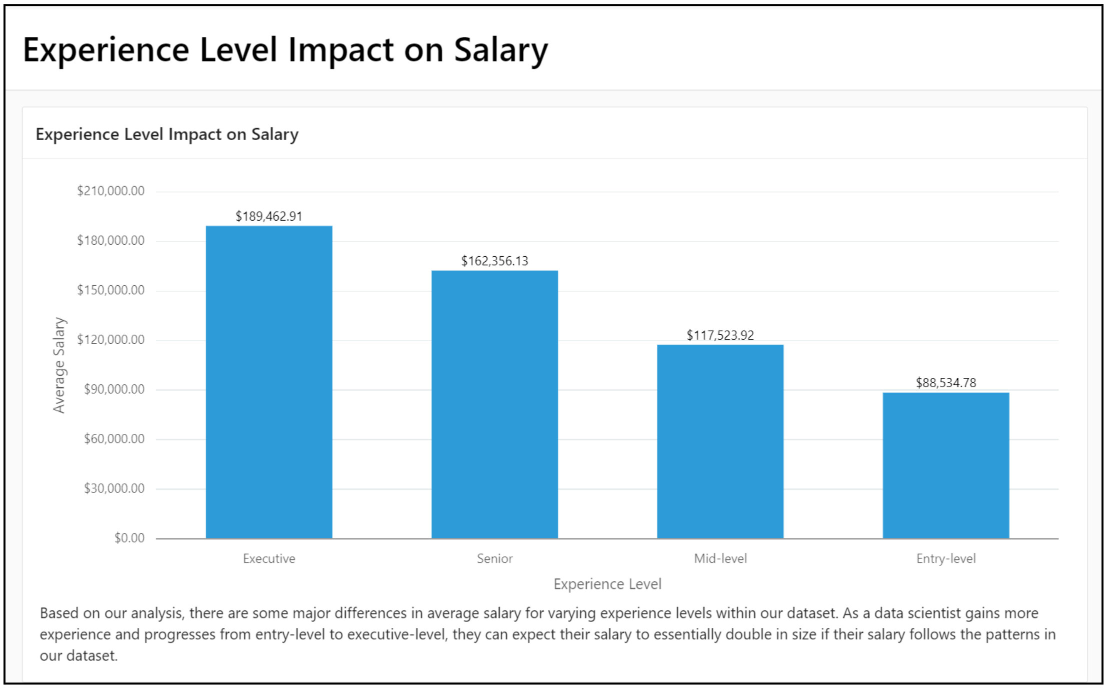
Work Setting Salary

Work Setting Salary

In recent years, remote and hybrid work envirnoments have increased in prevalence, but how do the salaries earned by those in flexible working arangements compare to their in-person counterparts? Our data suggests that those working completely in-person earn the highest average annual salary of $155,524. Fully remote employees earned the second highest average annual salary at $144,996, and those working in a hybrid setting earned an average annual salary of $88,912. Our data suggests that job-hunting data scientists should aim for an in-person role to maximize their earning potential.

Fig. 16 Average Salary by Work Setting (Page 8)

Experience Level Impact on Salary: Page 9 displays a bar chart comparing the difference in average salary across different experience levels.



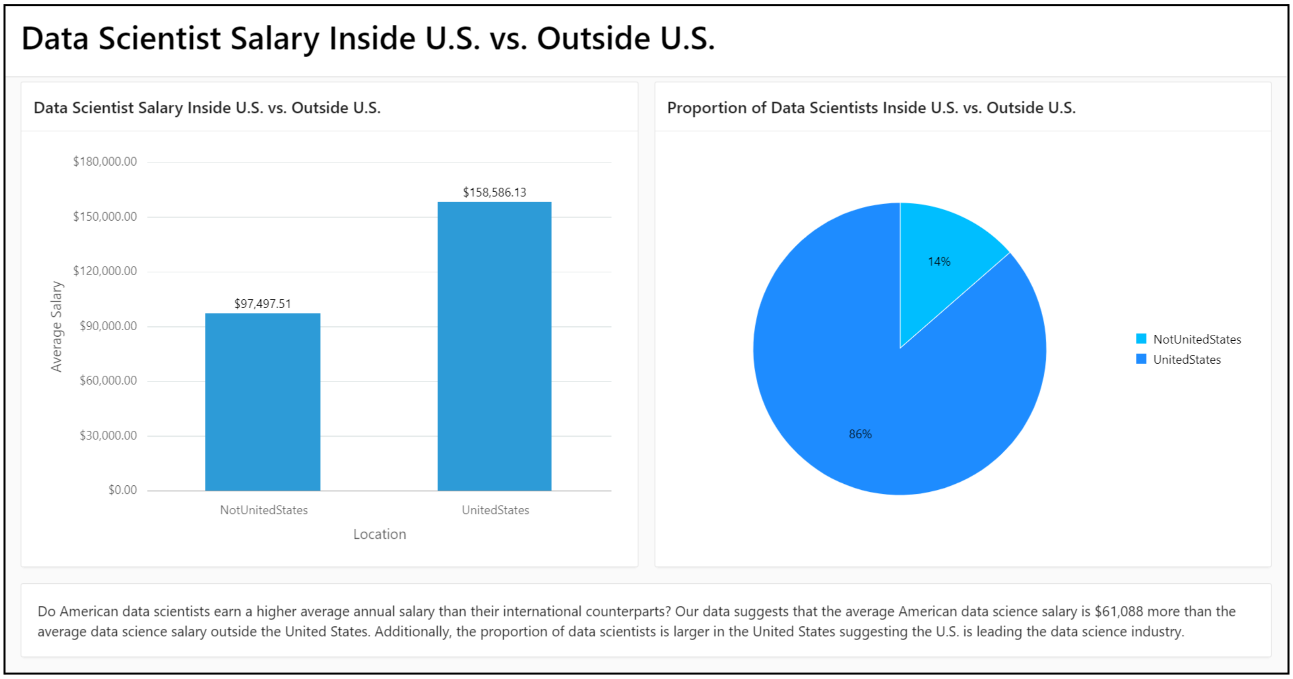
Experience Level Impact on Salary

Experience Level Impact on Salary

Based on our analysis, there are some major differences in average salary for varying experience levels within our dataset. As a data scientist gains more experience and progresses from entry-level to executive-level, they can expect their salary to essentially double in size if their salary follows the patterns in our dataset.

Fig. 17 Average Salary by Experience Level (Page 9)

Data Scientist Salary Inside U.S. vs. Outside U.S.: Page 10 displays a bar chart and a pie chart comparing average data science salaries within the U.S. to average data science salaries outside the U.S. and the proportion of data scientists inside vs. outside the U.S.



Data Scientist Salary Inside U.S. vs. Outside U.S.

Data Scientist Salary Inside U.S. vs. Outside U.S.

Proportion of Data Scientists Inside U.S. vs. Outside U.S.

Do American data scientists earn a higher average annual salary than their international counterparts? Our data suggests that the average American data science salary is $61,088 more than the average data science salary outside the United States. Additionally, the proportion of data scientists is larger in the United States suggesting the U.S. is leading the data science industry.

Fig. 18 Average Salary and Proportion Inside/Outside the U.S. (Page 10)