LE BRUIT

Salaries
in AI, ML and Big Data

LE BRUIT

INDEX

01.

LE BRUIT

PROJECT PRESENTATION

LE BRUIT

INTRODUCTION

Le Bruit *is a small size consulting company*

We were contacted by Glasswindow, a community for workplace conversations, support, and resources to make the most out of work life

They're adapting to the growing IT market and want to develop strategies to ***help workers maximizing their work choices and opportunities***

Alongside this main goal, they're also working on becoming a ***more reliable source of information****, using solid data (not hypotheses) to corroborate the information they allow their users to share between themselves*

LE BRUIT

Year | Job Title | Experience Level | Employment Type | Company Size | Salary

IT MARKET SALARY DATASET

Year	Job Title	Experience Level	Employment Type	Work Model	Company Size	Salary
2024	Data Scientist	Senior	Full-time	Onsite	Medium	324000
2024	Data Developer	Senior	Full-time	Onsite	Medium	109800
2024	Machine Learning Scientist	Senior	Full-time	Remote	Medium	245400
2024	Machine Learning Scientist	Senior	Full-time	Remote	Medium	139000
2024	BI Specialist	Intermediate	Full-time	Onsite	Medium	115000
2024	BI Specialist	Intermediate	Full-time	Onsite	Medium	93000
2024	Data Analyst	Junior	Full-time	Remote	Medium	62000
2024	Data Analyst	Junior	Full-time	Remote	Medium	55000
2024	Machine Learning Engineer	Senior	Full-time	Onsite	Medium	181600
2024	Machine Learning Engineer	Senior	Full-time	Onsite	Medium	125100
2024	Research Engineer	Intermediate	Full-time	Onsite	Medium	203000
2024	Research Engineer	Intermediate	Full-time	Onsite	Medium	133000
2024	Research Engineer	Senior	Full-time	Onsite	Medium	235680
2024	Research Engineer	Senior	Full-time	Onsite	Medium	153040

1 - 20 / 9152

(*) The data is collected through anonymous salary surveys filled out by professionals in the AI, ML, and data science fields. The dataset is published in the public domain under the CC0 license, ensuring it can be freely used and distributed.

02.

LE BRUIT

OVERALL ANALYSIS

LE BRUIT

Year

Experience Level

Work Model

Employment Type

HIGHEST SALARY

Median Salary

126.7K

In Time Range

Experience Level Salary

140.9K

Senior

Work Model Salary

144.0K

Onsite

Employment Type Salary

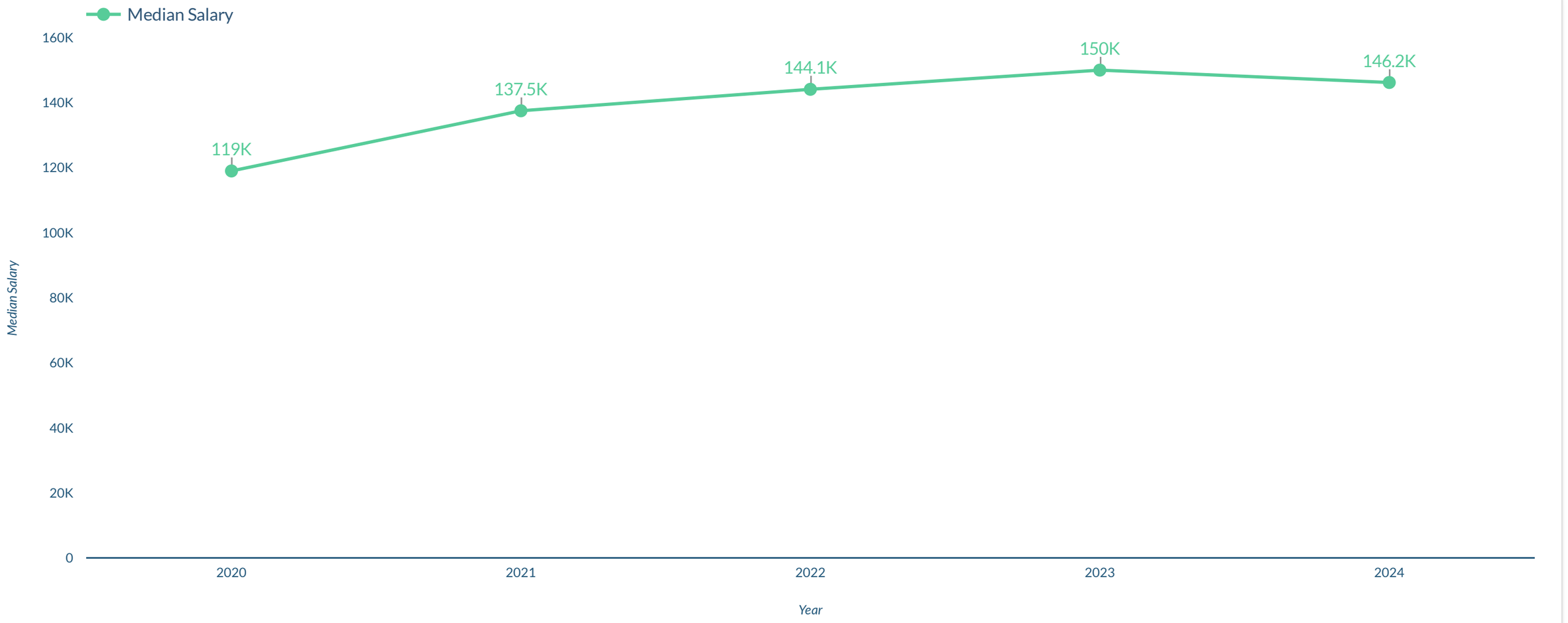
115.4K

Full-time

LE BRUIT

See details

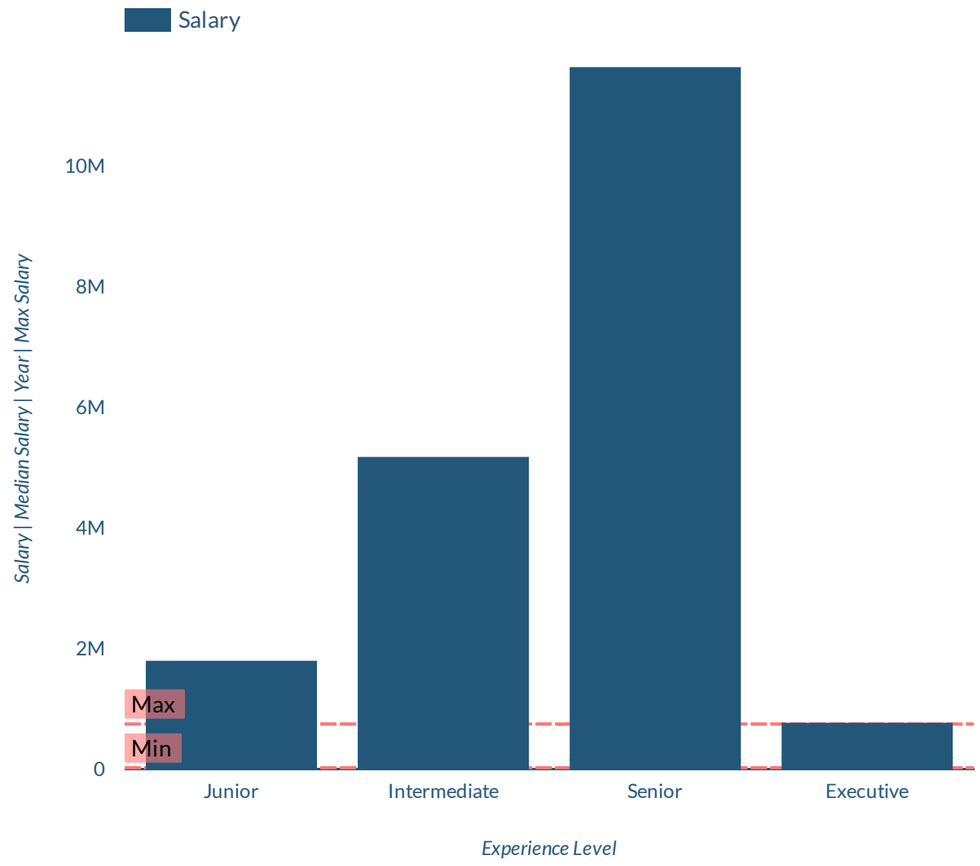
IT MARKET SALARY OVER THE LAST YEARS



LE BRUIT

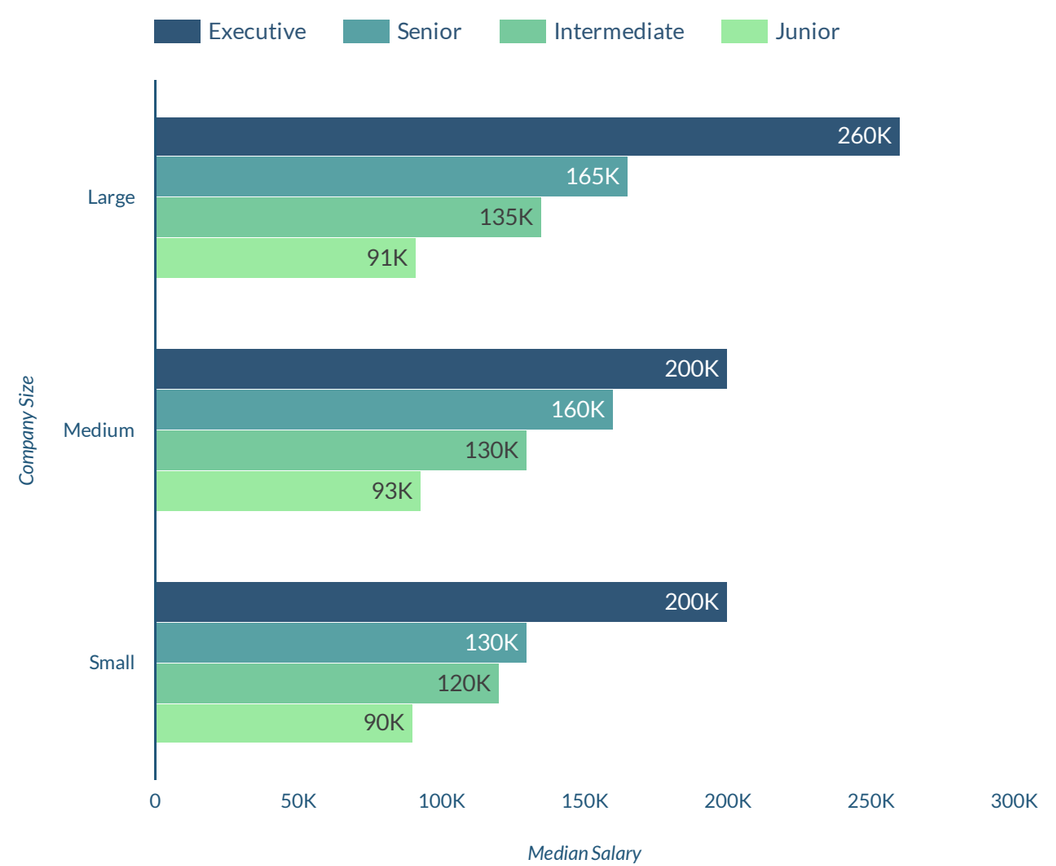
Year

SALARY BY EXPERIENCE LEVEL



Executive positions tend to offer the highest median salaries

SALARY BY COMPANY SIZE

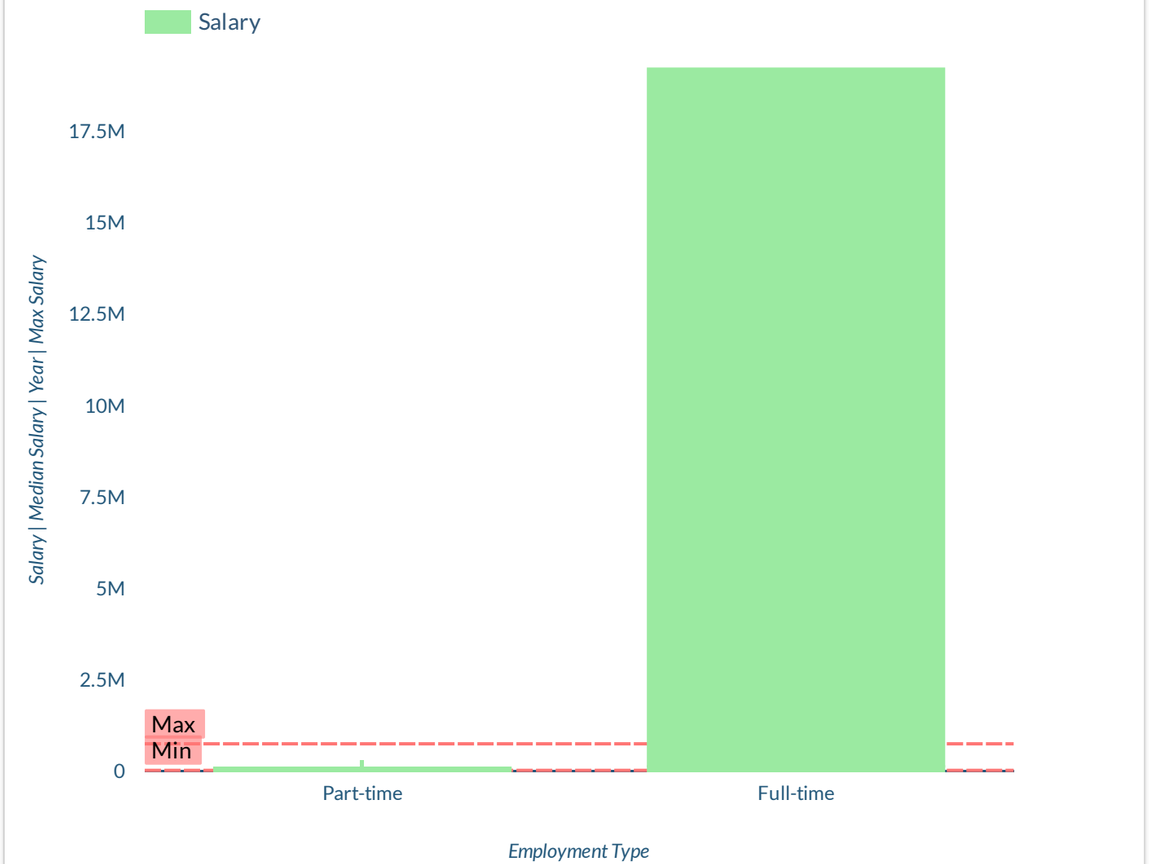


Executive salaries show the biggest difference between companies

LE BRUIT

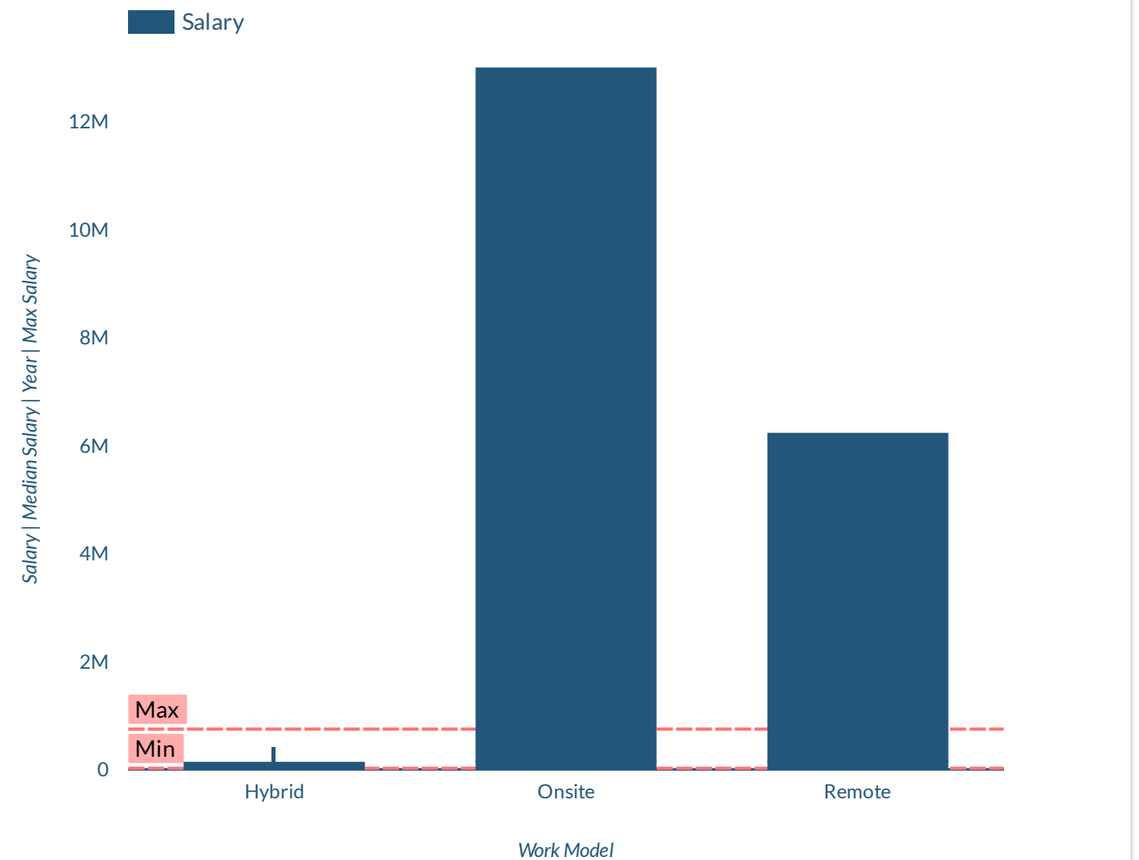
Year

SALARY BY EMPLOYMENT TYPE



Full-time positions offer the highest salaries

SALARY BY WORK MODEL

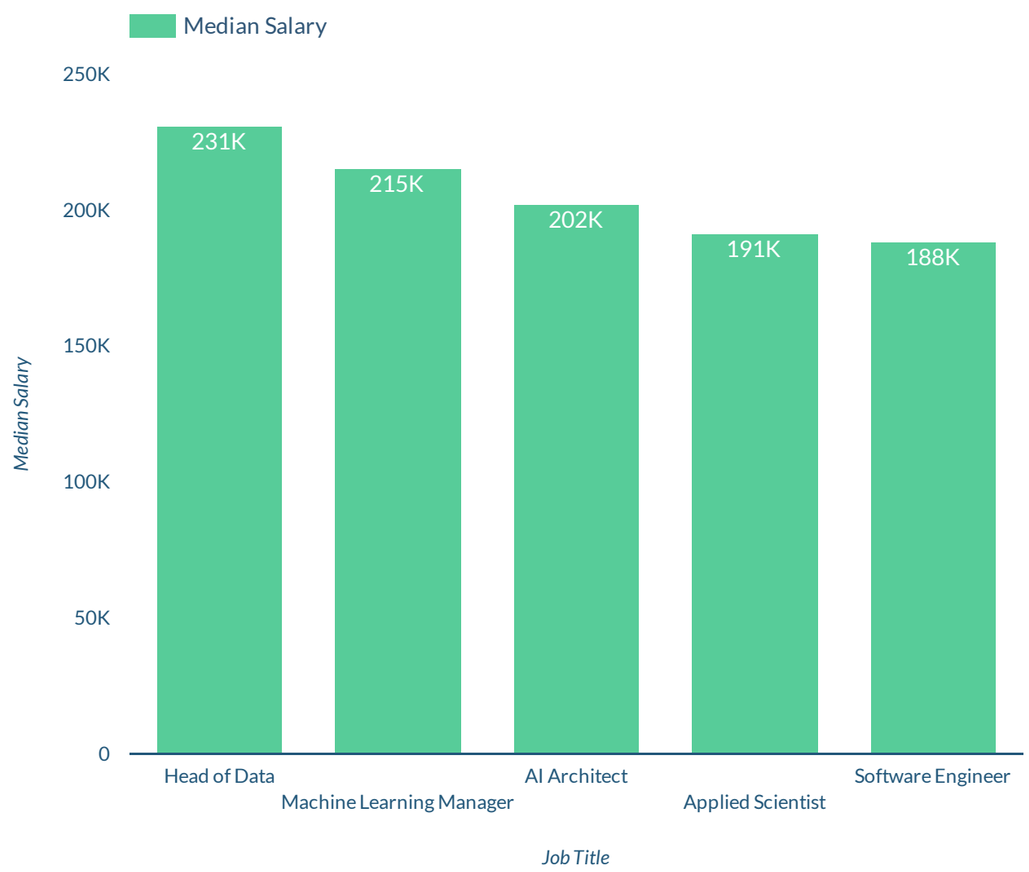


In 2024 the Hybrid model is becoming the highest paid

LE BRUIT

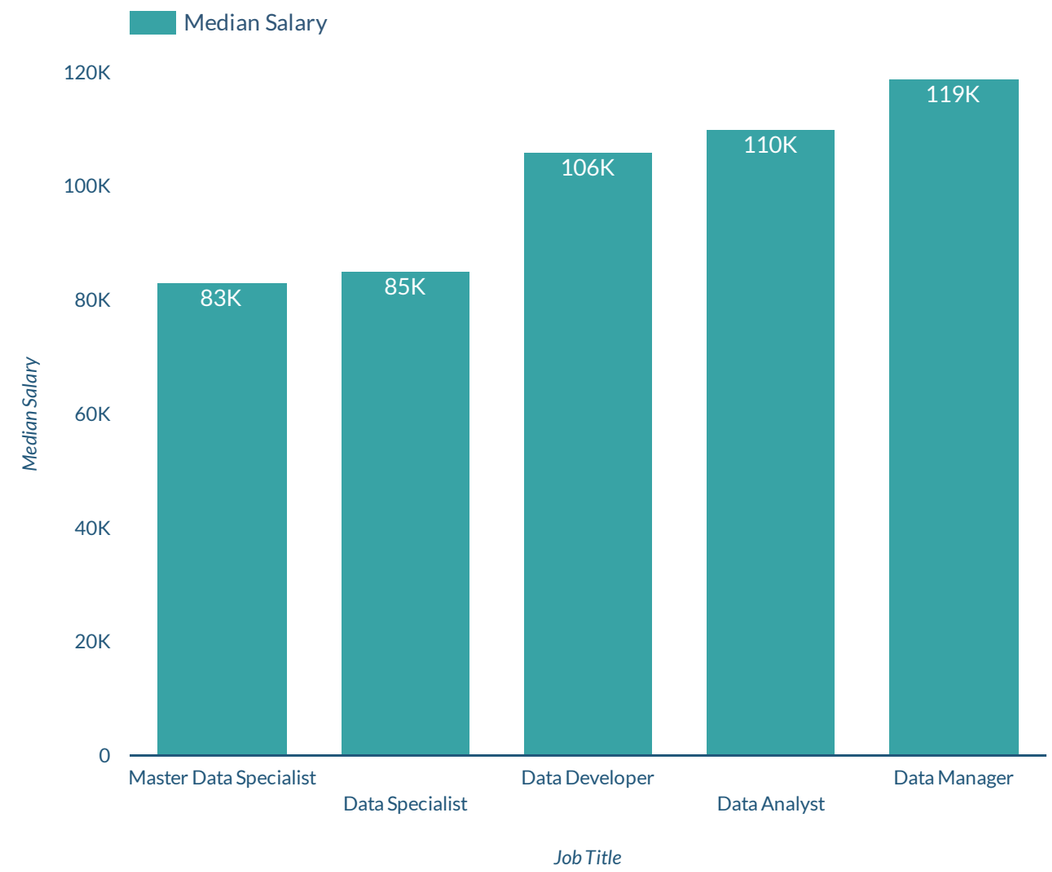
See details

HIGHEST 5 EARNER JOB TITLES



Over the years ML roles have always been in the 5 highest earners

LOWEST 5 EARNER JOB TITLES



Data-field related jobs have always been the majority of the 5 lowest

03.

LE BRUIT

DATA ANALYTICS *VS.* DATA SCIENCE

LE BRUIT

Job Title | Experience Level | Work Model | Employment Type

HIGHEST SALARY 2023 | 2024

Median Salary

109K

Until April

Job Title Salary

128K

Data Scientist

Work Model Salary

111K

Onsite

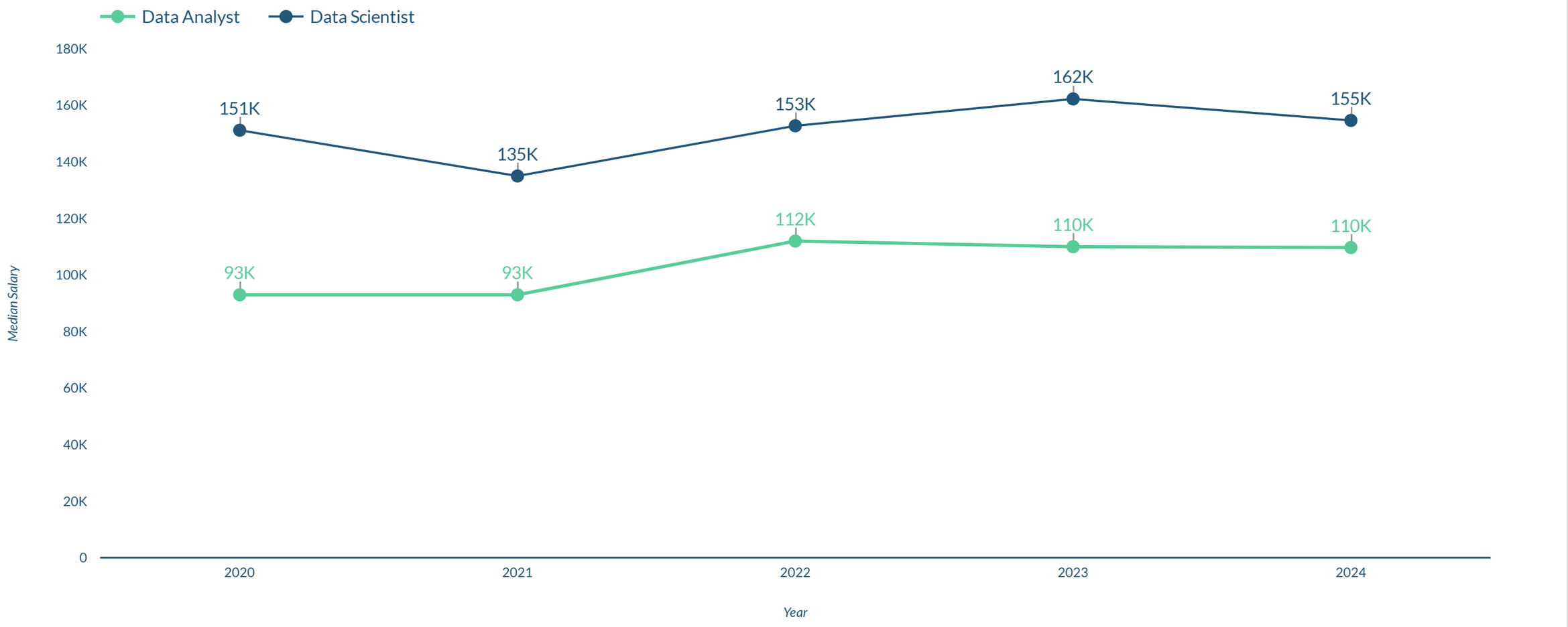
Employment Type Salary

103K

Full-time

LE BRUIT

SALARIES OVER THE LAST YEARS



04.

LE BRUIT

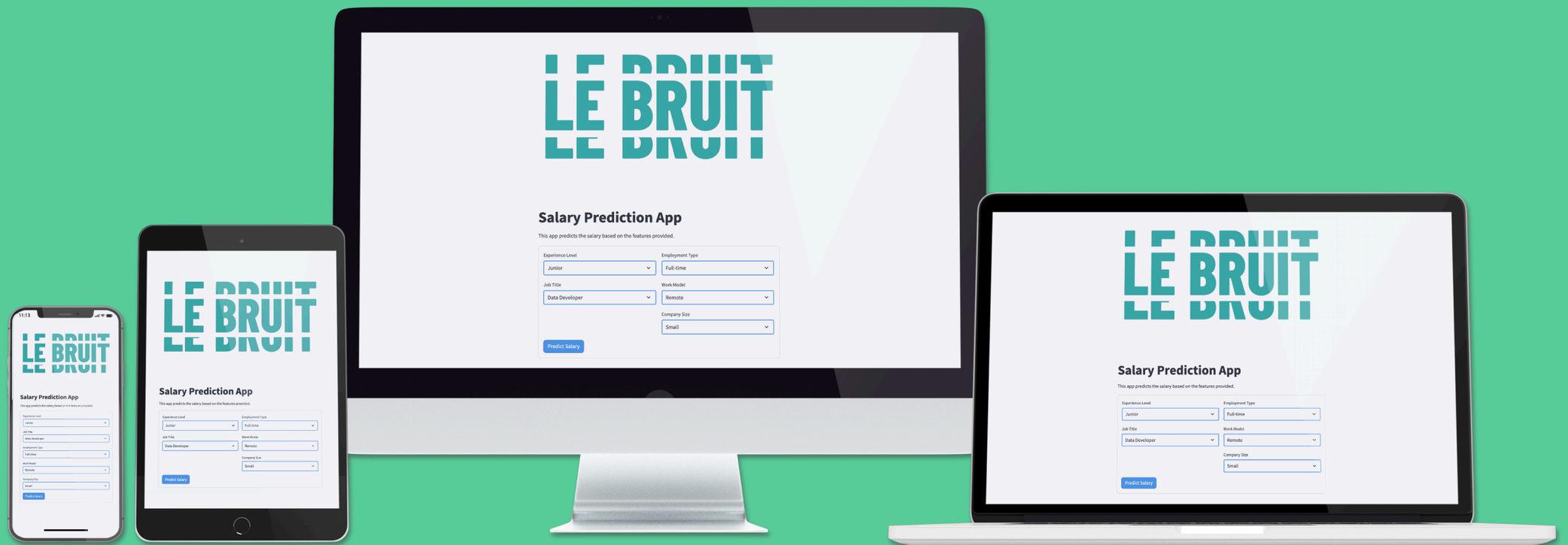
CONCLUSION

WHAT SOLUTION DID

LE BRUIT

LE BRUIT

CAME UP TO?



AN APP WHERE YOU CAN FIND YOUR ANSWERS

LET'S GO!

TRY IT FOR YOURSELF!

LE BRUIT

Key Insights

LE BRUIT

(*) The data is EXCLUSIVELY from the US

DATASET ANALYSIS

- Null values weren't found
- Duplicates were found, and removed from the dataset
- Outliers were found, and therefore removed from the dataset
- The time range runs from 2020 to April of 2024
- The year appearing the most is 2024 (50.9%), followed by 2023 (39.3%), 2022 (8.6%), 2021 (0.9%), and the one that has the lowest number of entries is 2020 (0.3%)
- Experience levels are divided in four categories: Junior, Intermediate, Senior, and Executive
- Employment type has four categories: Full-time, Part-time, Contract, and Freelance
- Work Models are divided in three categories: Onsite, Hybrid and Remote
- Company size has three categories: Small, Medium and Large

OVERALL ANALYSIS

EVOLUTION:

- The IT term was first published in 1958 in the Harvard Business Review, however we have witnessed a boom in this sector's roles the past years
- Breakthroughs in artificial intelligence, cloud computing, blockchain… have propelled the industry forward, resulting in increased job opportunities across various domains, including software development, data science, cybersecurity, and project management. Skilled professionals are in high demand, leading to competitive salaries and benefits

LE BRUIT

(*) The data is EXCLUSIVELY from the US

SALARIES:

- The minimum salary is $24k and the maximum is $750k, with a median of $134k.
- The metrics we analysed such as experience levels, employment types, company sizes and work models do influence the salaries.

GEOGRAPHICAL LOCATION:

- On the first approach to the data analysis we realised that, although we had entries from 85 countries regarding the employees and 77 countries regarding companies, the majority of the entries was from the US, with 83.4%, which creates a major geographical disparity.
- For that reason we decided to analyse only data from the US.

EXPERIENCE LEVEL:

- The experience level represented the most is Senior (60.2%) followed by Intermediate (26.7%), Junior (9.2%), and Executive (3.9%), although the Intermediate and Junior levels have been consistently growing in recent years.
- The Executive level naturally offers the highest salaries, even though it has the fewest active employees.

EMPLOYMENT TYPE:

- Full-time employment is massive (99.6%), compared to part-time (0.24%), contract (0.26%) and freelance (0.19%).
- Full-time positions consistently offer better wages, regardless of the employee's experience.

LE BRUIT

(*) The data is EXCLUSIVELY from the US

WORK MODEL:

- Between 2020 and 2022 the number of employees in remote work was bigger than the onsite work. That may be attributed to Covid-19 lockdown, once in the following years onsite overtakes the remote.
- From the last months of 2020 until the middle of 2021, the hybrid work model had a higher salary than onsite and remote. After mid 2021 the hybrid model was surpassed by both onsite and remote although in the last few months it became the highest salary again.
- The work environment tends to favor the onsite (67.3%) model, followed by remote (32.3%), and finally hybrid (0.43%).
- Until 2024, the employees that work onsite had a higher salary, but the hybrid type has stood out as the best paid in the recent months, despite being the least adopted work model.
- The remote work model offers lower, but very similar salaries to the onsite model, having the experience level as a factor of influence.

COMPANY SIZE:

- The Medium size companies represent almost the total of the dataset entries (95.4%), followed by the Large companies (4.1%), and finally the Small ones (0.57%).
- Since 2022, Small and Large companies tend to offer higher salaries for Executive roles. However, when it comes to Junior positions, both Large and Medium sized companies offer similar top-tier salaries.
- There's a noticeable difference for Intermediate and Senior roles, once Small companies pay less compared to Medium and Large ones, showing a 20% salary gap between both Intermediate and Senior levels.

LE BRUIT

(*) The data is EXCLUSIVELY from the US

COMPARATIVE INSIGHTS (2023):

- As a Junior employee, the highest paid role is Applied Scientist while Data Analyst has the lowest salary.
- In the Intermediate experience category, Machine Learning Scientist is the highest paid role while Data Specialist is the least paid role.
- In the Senior level, the highest paying role is Head of Data while Data Specialist is the least paid role.
- Finally, the Executive level has the Head of Data as the highest paying role, and Data Analyst as the lowest.
- The top 5 highest earner titles are Head of Data, AI Architect, Applied Scientist, AI Scientist and Machine Learning Engineer, belonging the lowest earner role to the Data Specialist.

COMPARATIVE INSIGHTS (2024):

- As a Junior employee, the highest paid role is Research Scientist while AI Engineer has the lowest salary.
- In the Intermediate experience category, the highest paid role is Data Specialist while Applied Scientist has the lowest salary.
- In the Senior level, the highest paying role is Machine Learning Manager while BI Engineer is the least paid role.
- Finally, the Executive level has the Machine Learning Engineer as the highest paying role, and BI Engineer as the lowest.
- The top 5 highest earner titles are Machine Learning Manager, Head of Data, AI Architect, Software Engineer, and Applied Scientist, belonging the lowest earner role to the Master Data Specialist.

LE BRUIT

Data Cleaning

LE BRUIT

DATA COLLECTION

```
#importing pandas library and reading the dataset
import pandas as pd

df = pd.read_csv('/content/drive/MyDrive/LeWagon/7. Project/Colab Notebooks/global_ai_ml_data_salaries.csv')
df
```

DATA STRUCTURING

```
#see the head of the table
df.head()

#exploring data types
df.dtypes

#searching null values
df.isnull().sum()

#searching duplicates
df.dulpicated().sum()

#cleaning duplicates in the original table
df.drop_duplicates(inplace=True)

#search for unique values
df.nunique()

#check column names
df.columns

#cleaning columns that aren't needed
df.drop(columns=['salary', 'salary_currency'], inplace=True)

#in column remote_ratio, change the values 0, 50 and 100 to Onsite, Hybrid and Remote
df['remote_ratio'].replace({0: 'Onsite', 50: 'Hybrid', 100: 'Remote'}, inplace = True)

#in the column company_size, change the values S, M, L to Small, Medium and Large
df['company_size'].replace({'S': 'Small', 'M': 'Medium', 'L': 'Large'}, inplace = True)

#in the column experience_level, change the values EN, MI, SE and EX to Junior, Intermediate, Senior, and Executive
df['experience_level'].replace({'EN': 'Junior', 'MI': 'Intermediate', 'SE': 'Senior', 'EX': 'Executive'}, inplace = True)

#in the column employment_type, change the values PT, FT, CT, FL to Part-time, Full-time, Contract, and Freelance
df['employment_type'].replace({'PT': 'Part-time', 'FT': 'Full-time', 'CT': 'Contract', 'FL': 'Freelance'}, inplace = True)

#display all the job titles to be able to analyse them
value_counts = df['job_title'].value_counts().reset_index()
print(value_counts)

### realized there were 148 job titles, but many belonged to the same function and decided to group them by these
functions to obtain a more organized data structure ###

#search for 'Business Inteligence' job title (specifying to ignore case variants)
contains_word = df['job_title'].str.contains('Business Inteligence', case=False)

#filter the column to show only the 'Business Inteligence' job title
job_name_df = df[contains_word]
job_name_df

#replace the job title 'Business Intelligence' with a new job title 'BI'
df['job_title'] = df['job_title'].str.replace('Business Intelligence', 'BI', case=False)

#do the same replacement procedure to all job titles that represented the same role
df['job_title'] = df['job_title'].str.replace(r'.*\bData Scientist\b.*', 'Data Scientist', case=False, regex=True)
df['job_title'] = df['job_title'].str.replace(r'.*\bData Analyst\b.*', 'Data Analyst', case=False, regex=True)
df['job_title'] = df['job_title'].str.replace(r'.*\bData Engineer\b.*', 'Data Engineer', case=False, regex=True)
df['job_title'] = df['job_title'].str.replace(r'.*\bMachine Learning Engineer\b.*', 'Machine Learning Engineer',
case=False, regex=True)
df['job_title'] = df['job_title'].str.replace(r'.*\bMachine Learning Scientist\b.*', 'Machine Learning Scientist',
case=False, regex=True)
df['job_title'] = df['job_title'].str.replace(r'.*\bMachine Learning Specialist\b.*', 'Machine Learning Specialist',
case=False, regex=True)
df['job_title'] = df['job_title'].str.replace('Machine Learning Infrastructure Engineer', 'Machine Learning Engineer',
case=False, regex=True)

(...)
```

LE BRUIT

```
#counting the new unique values of the column job_title
df['job_title'].nunique()

### started with 148 job titles and ended up with only 26 after organizing the data ###

#changing column names to make them more understandable
df.rename(columns={'work_year': 'Year', 'experience_level': 'Experience Level',
                   'employment_type': 'Employment Type', 'job_title': 'Job Title',
                   'salary_in_usd': 'Salary', 'employee_residence':'Employee Location',
                   'remote_ratio':'Work Model', 'company_location': 'Company Location',
                   'company_size':'Company Size'}, inplace = True)

#checking the percentage of unique values in each column
df['Year'].value_counts(normalize=True) * 100
df['Experience Level'].value_counts(normalize=True) * 100
df['Employment Type'].value_counts(normalize=True) * 100
df['Job Title'].value_counts(normalize=True) * 100
df['Work Model'].value_counts(normalize=True) * 100
df['Company Size'].value_counts(normalize=True) * 100
df['Company Location'].value_counts(normalize=True) * 100
df['Employee Location'].value_counts(normalize=True) * 100

### noticed the majority of the Employee Location and Company Location entries were from US and decided to
change the focus to these entries, cleaning the table from all the rows that didn't had US as the location ###

#strip any leading/trailing whitespace in the Employee Location and Company Location columns
df['Employee Location'] = df['Employee Location'].str.strip()
df['Company Location'] = df['Company Location'].str.strip()

#get the index of the rows to exclude
e_index_to_exclude = df[df['Employee Location'] != 'US'].index
c_index_to_exclude = df[df['Company Location'] != 'US'].index

#drop the rows with the specified indexes
df = df.drop(e_index_to_exclude)
df = df.drop(c_index_to_exclude)

#check if there are still values different from 'US'
c_other_df = df[df['Company Location'] != 'US']
e_other_df = df[df['Employee Location'] != 'US']
c_other_df
e_other_df

#check Full-time entries in Employment Type column
(df['Employment Type'] == 'Full-time').value_counts()

#check Medium entries in Company Size column
(df['Company Size'] == 'Medium').value_counts()

#checking the Salary range and average
df['Salary'].min()
df['Salary'].max()
df['Salary'].mean()

### noticed there was an outlier in the AI Architect role that was unbalancing all the measures, once it earned 800k
being the real average salary 143k ###

#trying to find the outlier in the table
filtered_df = df[df['Job Title'].str.contains('AI Architect', case=False)]

#confirming this AI Architect's index is 3963
df.iloc[3963]

#clean that AI Architect from the table
df.drop(df.index[3963], inplace=True)
```

DATA EXPORT

```
#creating a .csv file with all the cleaning we did
df.to_csv('le_bruit_US.csv', index=False)

#downloading the .csv file
from google.colab import files
files.download('le_bruit_US.csv')
```