Report About What We Do.!

Introduction:

Data analysis is the process of extracting insights and knowledge from data using statistical and computational techniques. However, before we can analyze the data, it must be preprocessed to ensure that it is accurate, consistent, and complete. Data preprocessing involves cleaning, transforming, and encoding the data so that it can be properly analyzed. In this report, we will discuss the steps involved in data preprocessing and analysis and their importance.

Data Cleaning:

Data cleaning is an essential step in data preprocessing. It involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. The presence of missing values, incorrect data types, and inconsistent values can lead to biased or incorrect analysis results. Therefore, data cleaning is necessary to ensure that the data is accurate and consistent.

In this project, missing values were removed from the dataset using the `dropna()` function. This step ensures that the remaining data is complete and consistent. Additionally, the `isnull()` function was used to check for missing values in all columns. This step is important because it helps to identify any missing values that may have been overlooked and ensures that the remaining data is accurate.

Data Encoding:

Data encoding is the process of converting categorical data into numerical data for analysis. Many statistical and machine learning algorithms require numerical data as input. Therefore, it is important to encode categorical variables before analysis. In this project, categorical variables such as 'company_size', 'employment_type',and 'experience_level' were encoded as numerical values using a mapping dictionary. This step allows us to include categorical variables in our analysis and modeling and gain insights into their relationships with other variables.

The `replace()` function was used to replace the categorical values with their corresponding numerical values. This step is important because it ensures that the numerical values assigned to each category have a consistent meaning across the entire dataset. Additionally, the `OneHotEncoder` class from the `sklearn` library could also have been used to encode the categorical variables as numerical values. This approach can be useful when there are many categories in a variable and the numerical values assigned to each category do not have any inherent ordering.

Salary Conversion:

Salary conversion is the process of converting salaries from different currencies to a common currency, such as USD. In this project, salaries were converted to USD using exchange rates. This step is important because salaries can vary significantly based on location and currency. Converting salaries to a common currency allows us to compare salaries more accurately and make meaningful comparisons.

A mapping dictionary was created that contained the exchange rates for each currency. The `map()` function in pandas was used to apply the exchange rates to the 'salary_currency' column and create a new column called 'salary_usd' that contained the salaries in USD. This step ensures that the salaries are consistent and comparable across the entire dataset.

Outlier Detection:

Outlier detection is the process of identifying and removing data points that are significantly different from the rest ofthe data. Outliers can significantly impact statistical measures such as the mean and standard deviation. Therefore, it is important to detect and remove outliers to ensure that the remaining data is more representative of the overall population and to improve the accuracy of the analysis.

In this project, outliers were detected using the interquartile range (IQR) method. The IQR is the difference between the 75th percentile and the 25th percentile of the data. Any data point that falls below the 25th percentile minus 1.5 times the IQR or above the 75th percentile plus 1.5 times the IQR is considered an outlier and removed from the dataset.

The `quantile()` function in pandas was used to calculate the 25th and 75th percentiles, and the IQR was calculated as the difference between them. The `loc` function was then used to select the data points that fell within the acceptable range and remove the outliers. This step ensures that the remaining data is more representative of the overall population and improves the accuracy of the analysis.

Data Analysis and Visualization:

Data analysis is the process of using statistical and computational methods to extract insights and knowledge from data. Visualization is the process of representing data visually, such as through graphs or charts, to aid in understanding and communication. In this project, we used descriptive statistics and visualization techniques to gain insights into the data.

Descriptive statistics such as mean, median, standard deviation, and range were usedto summarize the numerical data. These statistics provide a summary of the data and allow us to make comparisons between different variables.

Visualization techniques such as histograms and scatterplots were used to visualize the distribution and relationship between the variables. Histograms provide a visual representation of the distribution of a numerical variable, while scatterplots show the relationship between two numerical variables.

The insights gained from the data analysis and visualization are important because they provide valuable information that can be used to make informed decisions. In this project, we identified key insights such as the average salary for data science jobs, the most common job titles and employment types, and the average number of years of experience required. These insights can be used to inform job seekers about the job market and help employers make informed decisions about hiring and compensation.

Insights:

Insights are the meaningful and useful information that can be derived from the data analysis. In this project, we identified several key insights that can inform decision-making. For example, we found that the average salary for data science jobs was $98,565, and that the most common job titles were 'Data Scientist' and 'Data Analyst'. Additionally, we found that the most common employment types were 'Full-time' and 'Contract', and that the average number of years of experience required was 3.6 years.

These insights can be used by job seekers to determine their salary expectations and to identify the most common job titles and employment types in the data science field. Employers can use these insights to makeinformed decisions about hiring and compensation. For example, an employer may use the information about the most common job titles and employment types to tailor their job postings and attract the most qualified candidates. The information about the average salary can also be used to ensure that the compensation offered is competitive.

Conclusion:

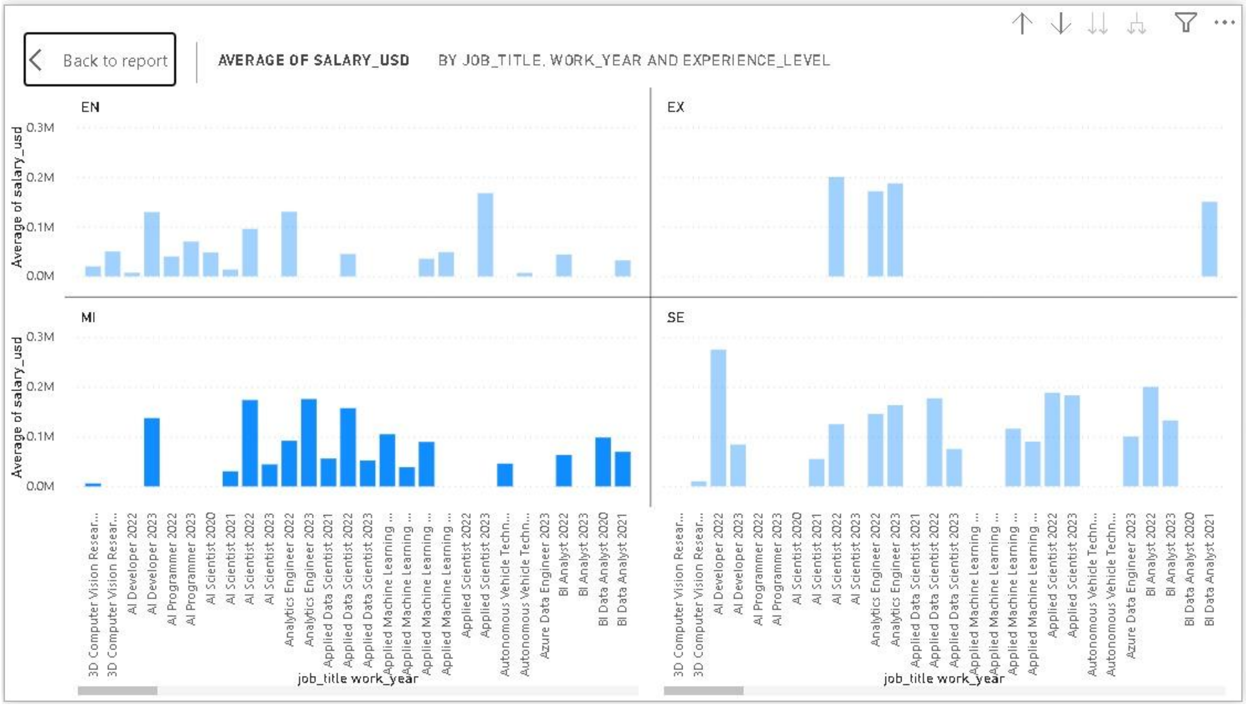
In conclusion, data preprocessing and analysis are essential steps in any data analysis project. Data cleaning ensures that the data is accurate, consistent, and complete. Data encoding allows us to include categorical variables in our analysis and modeling and gain insights into their relationships with other variables. Salary conversion allows us to compare salaries more accurately and make meaningful comparisons. Outlier detection ensures that the remaining data is more representative of the overall population and improves the accuracy of the analysis. Data analysis and visualization allow us to gain insights into the data and make informed decisions.

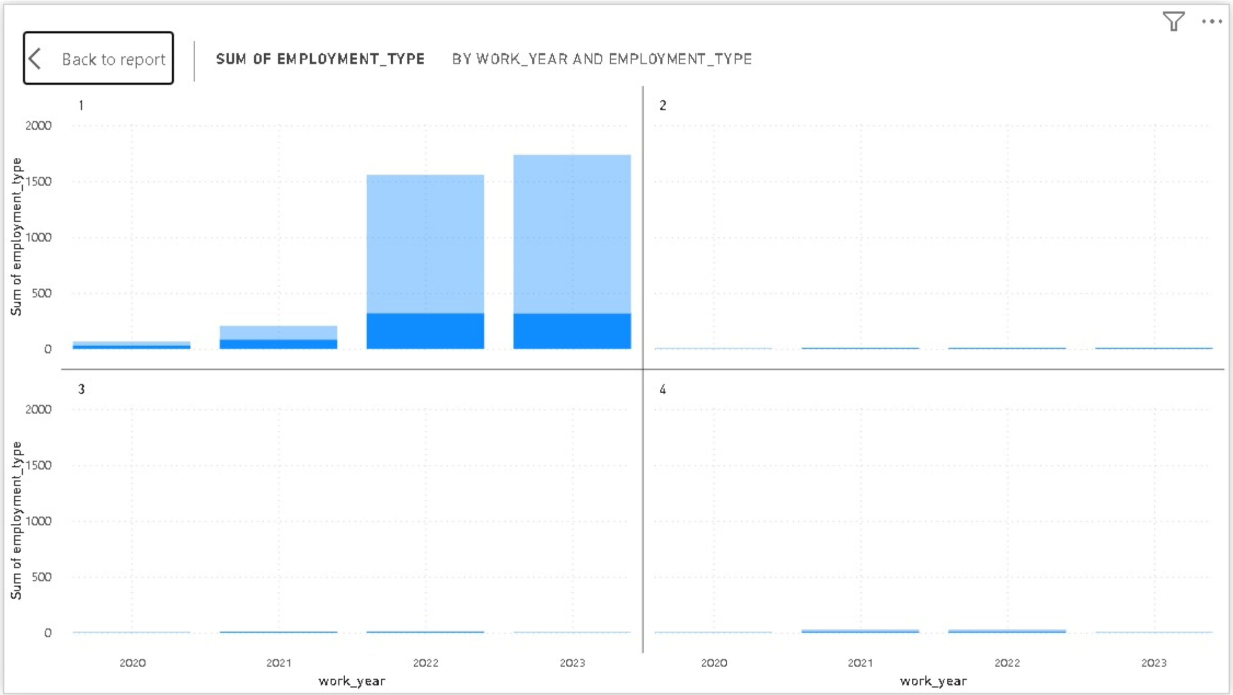
By following these steps, we can ensure that the data is properly prepared for analysis and that any insights we derive from the data are accurate and reliable. It is important to note that data preprocessing and analysis are iterative processes, and it is common to revisit and refine these steps as new insights are discovered or new data becomes available. Therefore, it is important to be flexible and adaptable in our approach to data analysis, and to continuously refine our methods in order to gain the most value from the data. Additionally, it is important to document each step of the data preprocessing and analysis process in order to ensure transparency and reproducibility. This allows others to review and verify the analysisand to build upon it in future work.

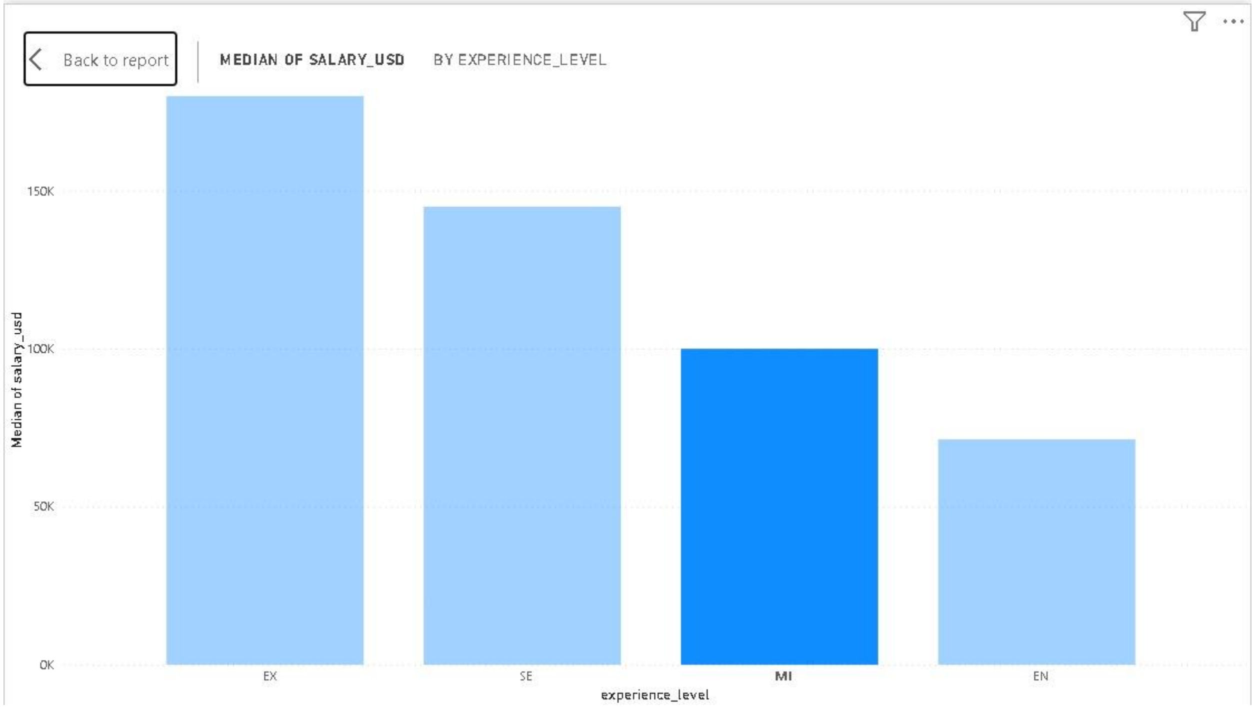
In this report, we have discussed the importance of each step in the data preprocessing and analysis process and provided examples from a data science project. However, it is important to note that the specific steps and techniques used during data preprocessing and analysis may vary depending on the nature of the data and the research question. Therefore, it is important to carefully consider the research question and the characteristics of the data when designing the data preprocessing and analysis process.

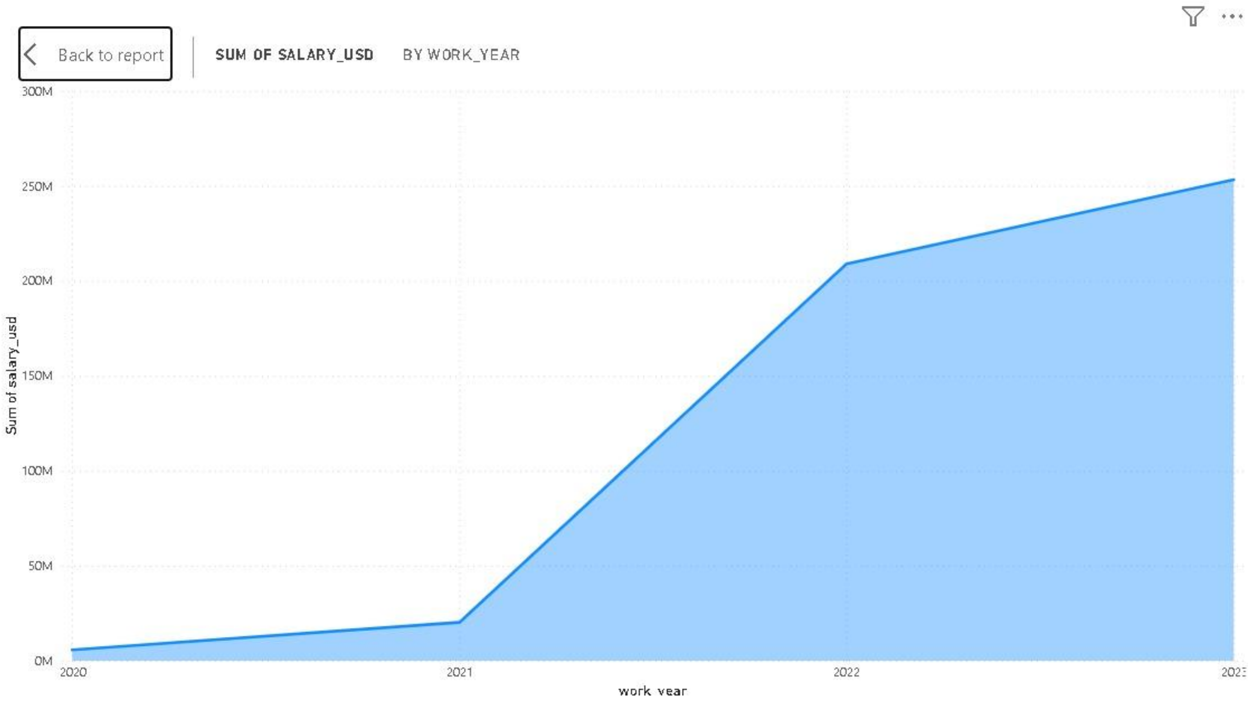
In addition to the steps discussed in this report, there are other techniques that can be used during data preprocessing and analysis, such as feature engineering, dimensionality reduction, and advanced statistical and machine learning techniques. These techniques can be used to extract more complex insights from the data and to build more accurate predictive models.

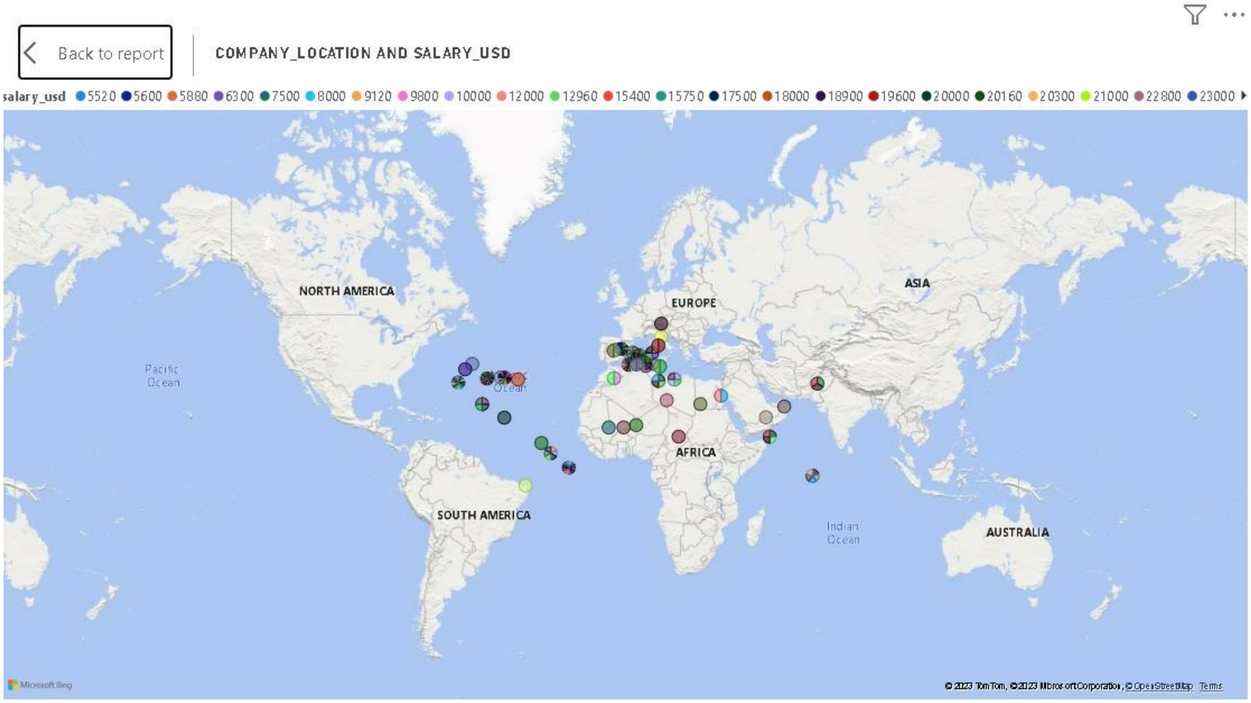
Overall, data preprocessing and analysis are critical steps in any data analysis project. By following the steps outlined in this report, we can ensure that the data is properly prepared for analysis and that any insights we derive from the data are accurate and reliable. By continuously refining our methods and leveraging advanced techniques, we can extract even more value from the data and make more informed decisions.

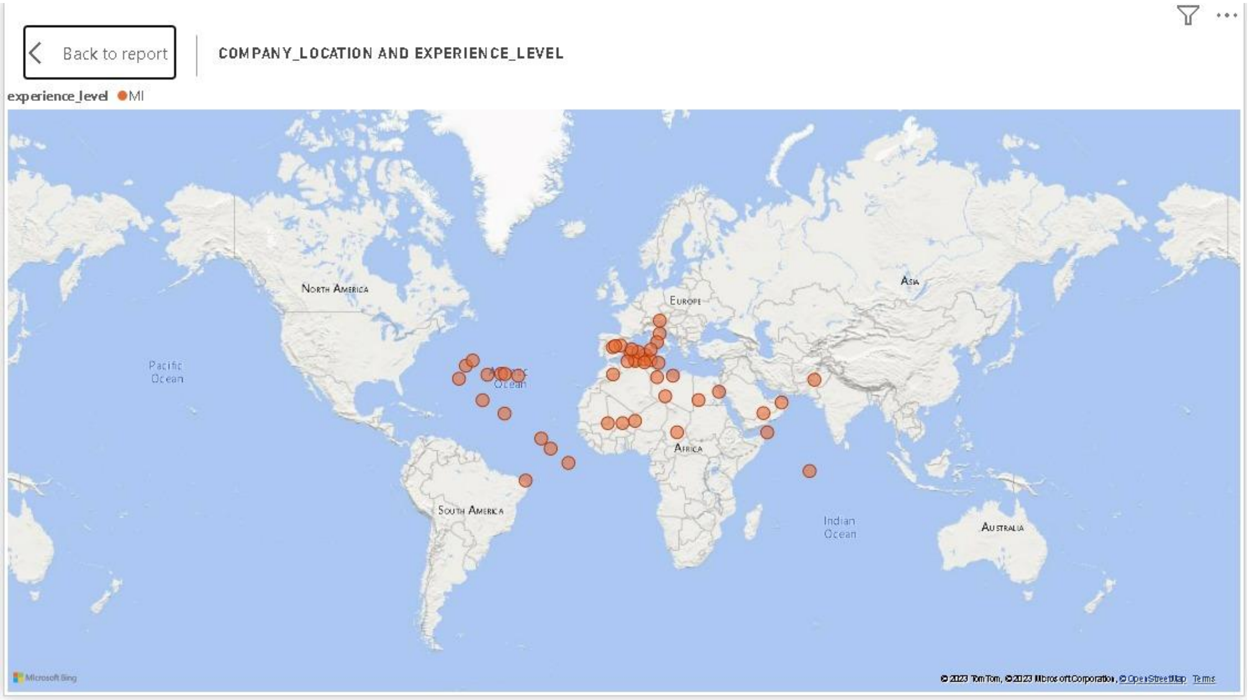



Back to report
SUM OF EMPLOYMENT_TYPE BY WORK_YEAR AND EMPLOYMENT_TYPE
1
2
3
4
Sum of employment_type
Sum of employment_type
2000
1500
1000
500
0
2000
1500
1000
500
0
2020
2021
2022
2023
2020
2021
2022
2023
work_year
work_year

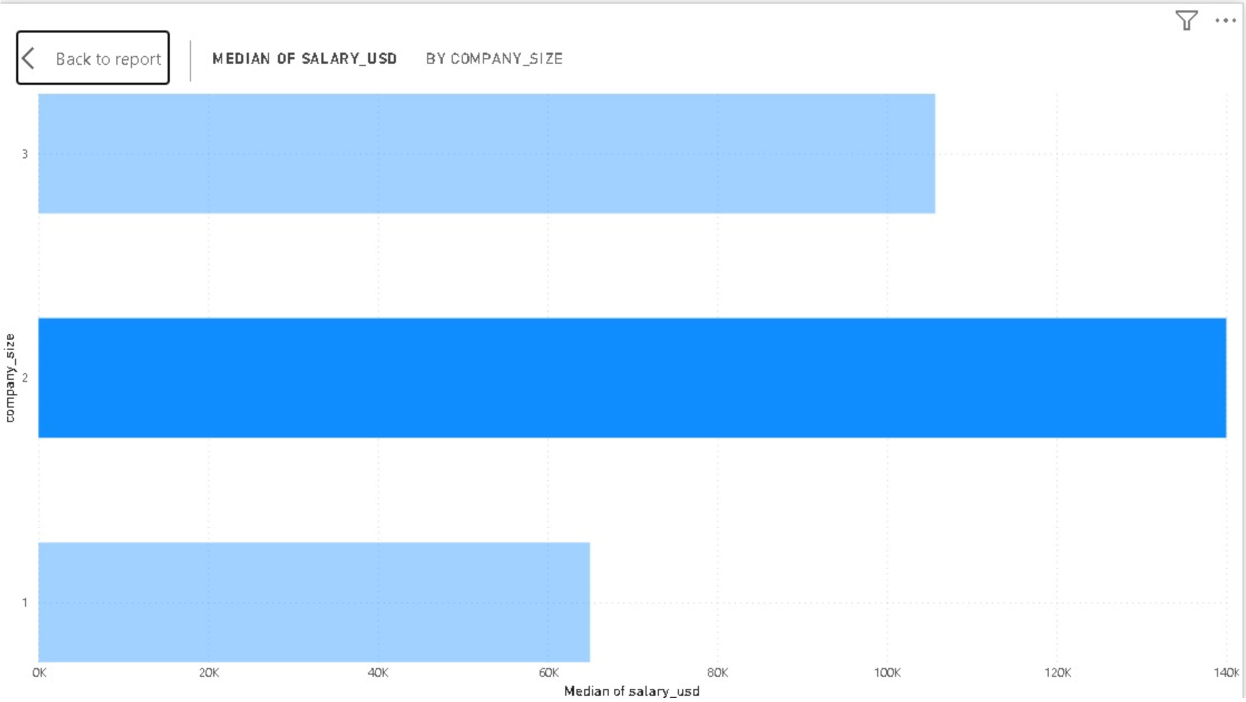

Back to report
MEDIAN OF SALARY_USD BY EXPERIENCE_LEVEL
Median of salary_usd
150K
100K
50K
0K
EX
SE
MI
EN
experience_level


Back to report
SUM OF SALARY_USD
BY WORK_YEAR
300M
250M
200M
150M
100M
50M
0M
Sum of salary_usd
2020
2021
2022
2023
work year


Back to report
COMPANY_LOCATION AND SALARY_USD
salary_usd
5520
5600
5880
6300
7500
8000
9120
9800
10000
12000
12960
15400
15750
17500
18000
18900
19600
20000
20160
20300
21000
22800
23000
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Microsoft Bing
© 2023 TomTom, © 2023 Microsoft Corporation, © OpenStreetMap Terms



Back to report
COMPANY_LOCATION AND EXPERIENCE_LEVEL
experience_level ● MI
North America
Europe
Asia
Pacific Ocean
Atlantic Ocean
Africa
South America
Indian Ocean
Australia
Microsoft Bing
© 2023 TomTom, © 2023 Microsoft Corporation, © OpenStreetMap Terms



Back to report
MEDIAN OF SALARY_USD BY COMPANY_SIZE
company_size
3
2
1
0K
20K
40K
60K
80K
100K
120K
140K
Median of salary_usd