Data Exploration Report

Author: Tirumalesh Nagothi

Contents

Introduction

Problem description:
This project explores the trends of IT jobs related to data science and how they are affected with the Covid 19 during the years 2020, 2021 and 2022. There are many aspects to consider, such as job role, location of the company that is providing the job etc. And the salary is wholly based on the situation and the other factors collected. The covid19 and IT salaries data are combined for further exploration and analysis.

Motivation:
Since I've developed an interest in data science, I used to explore the data science field and its exposure in future years. I learned about the demand in various areas that comes under Artificial Intelligence. But, the roles don't stick to only Machine Learning, Deep Learning and Data Science. There were lot more roles derived from those marked fields. As I want to be an entrepreneur in data science, I wanted to know more about the pay of the data science roles and on what conditions they are paid low or high.

Questions:
1. In which trends are the IT role salaries have gone during the period 2020 to 2022. Which job at the entry-level got better to pay each year?
2. Are more jobs changed from office to remote due to covid19? During and after the pandemic, in which country were the more remote or office jobs provided?
3. Did companies pay less to the employees due to the covid19 effect, and what are the jobs that became more each year?

Data Wrangling

Data sources:
There are two datasets where one. One is taken from a website (ai-jobs.net), and another is taken from google API.

1. IT job salaries related to the data science dataset are downloaded from- https://salaries.ai-jobs.net. And this dataset is maintained by a third-party host, and the data is entirely given by the public who are in the IT field. The size of the data in terms of rows increases daily.

   Feature description:

- work_year:- This represents the year as a numeric value type.
- experience_level:- This means the experience of the employee.
- employment_type:- This means the employment type.
- job_title:- Role of the employee
- salary:- The continuous value given as salary paid to that role in their country's currency.
- salary_usd:- Salary for that particular role after converting to US dollars.
- employee_residence:- Country where the employee resides in.
- remote_ratio:- What is the percentage of remote environment involved?
- company_location:- Country where the company is located at.
- company_size:- How big the company is.

2. Covid 19 dataset is taken from google API:- https://storage.googleapis.com/covid19-open-data/v3/location/<country_code>.csv

   If we want to get the data of the USA(United States of America), we need to call the API by including the country code in the link as https://storage.googleapis.com/covid19-open-data/v3/location/US.csv.

   Note:- There are lots of features in this dataset. And we are going to consider only the required columns.

   These two datasets aren't in the exact structure with the same number of rows, but they match the countries with the view of columns. The primary dataset - IT job salaries combines with the covid19 dataset that supports the primary dataset by providing the positive cases of each country.

- Extracted the covid19 data using the country codes in the primary dataset's company_location feature. So, only the data with those countries in the primary dataset appends as a single dataset.
- Only a few features such as country_code, date, new_confirmed, new_deceased and new_tested are included. All these might not be involved doesn't mean they will be used. They are excluded because those were the common data feature that supports flexibility for exploration.

Data Transformations:

- Iterated through the country_code, year and new_confirmed features, the mean for new_confirmed cases is calculated by filtering the data specifically with the country_code and year. These values are appended to lists: mean_values list, country_code and year.

- A new data frame is created that includes those three different lists of values in 3 other columns.

- Merged this new data frame to the primary data by mapping the mean value based on the country_code and the year.

Tools used:
For the above process of cleaning and transforming the data, I've used python 3.9 along with the libraries such as pandas for loading the data from the local machine and calling from the cloud, performing data manipulation and SSL for getting the certificates for calling google API for covid data.

Data Checking:

- It was checked whether there were any null values in the primary and covid19 datasets and found nothing. All the cells were filled with data related to their columns.
- Removed the salary and salary_currency columns in the primary dataset because the salary_usd is common for every country's salary currency conversion. So, Instead of involving those two columns and exploring, it is better to use a standard column that replaces and supports those two columns.
- During the final checking of the data, I found no errors. And all the values are ideally related to each of their columns.

Tools used: Python 3.9 and pandas to remove the unnecessary columns from the dataset.

Data Exploration

Salary trends of IT roles for each experience level

Dataset used:- IT job salaries related to data science.
Tools used:- R programming language for data filtering and plotted graphs with R and ggplot library.

- Ultimately there are four common types of experiences in the salary data Entry level, Mid-level, Senior level, and Executive level. For each experience level, the salary is varied each year.
- The figures below were plotted in a bar chart to get the trends for limited years(3). There were many plots plotted and saved in a pdf file. The screenshots were taken for a few features due to the high number. And some figures have the executive level with a single category of the year 2022 because, in today's scenario of jobs, there are more entry-level to mid-level positions than at the executive level, where it takes many years to reach.

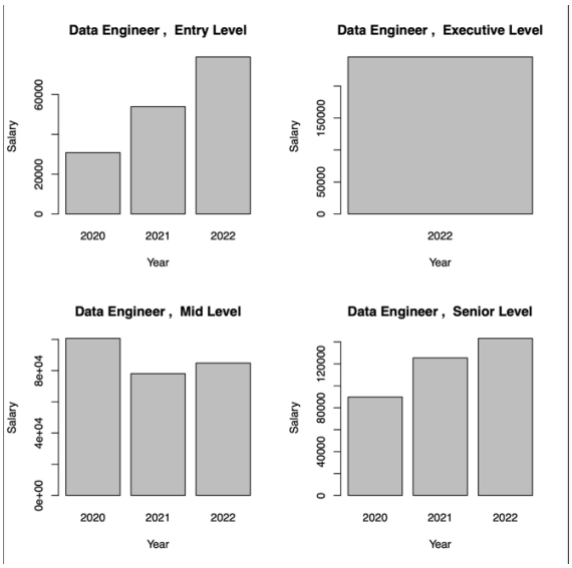

Figure 1.1

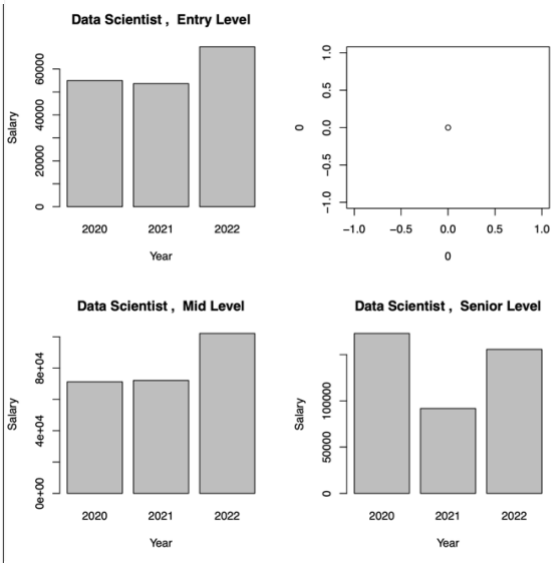

Figure 1.2

The average salary for each job experience was considered by filtering and creating a new dataset. I plotted these paragraphs using simple R bar graphs to combine all the charts of each job role, and also, these bar graphs clearly represent each year as a category and give the trend path too.

Salary paid for each fresher job.

Dataset(s) used:- IT job salaries related to data science
Tool(s) used:- R, ggplot2

Filtered the data based by considering the average salary of each entry-level job role. Plotted three different plots for each year using ggplot to get clear and notifiable bar graphs.

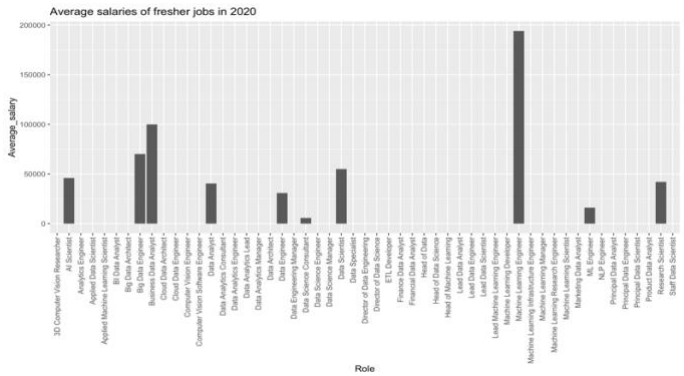

Figure 2.1

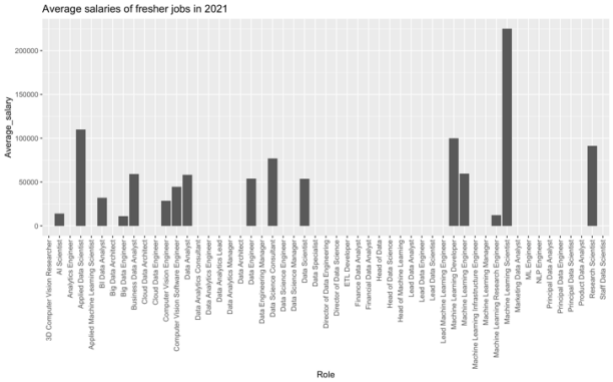

Figure 2.2

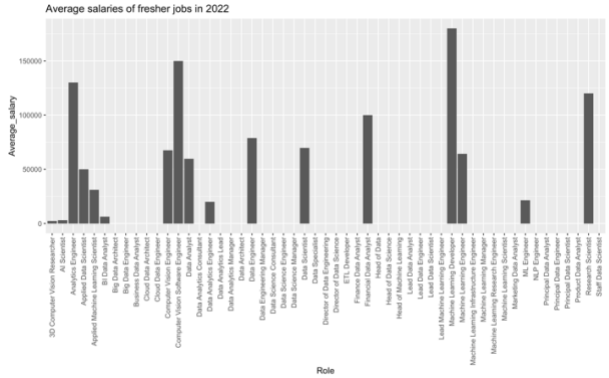

Figure 2.3

Trends of type of work

Dataset(s) used:- IT job salaries related to data science
Tool(s) used:- R, ggplot2

Filtered the data based on the ratio of remote, which has a discrete numeric value that consists of only 0, 50 and 100, where 0 represents complete office work, 50 represents a hybrid, and 100 represents fully remote work.
The data is extracted by getting the count of remote, hybrid and office works and converted to percentages to understand correctly.

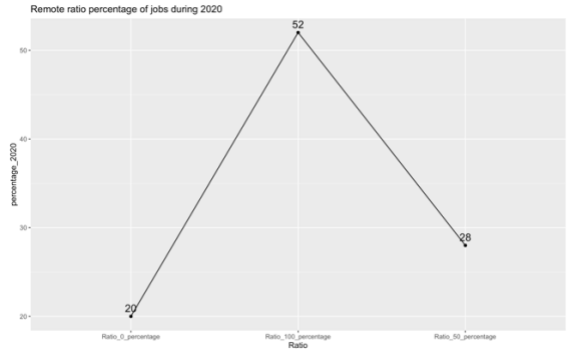

Figure 3.1

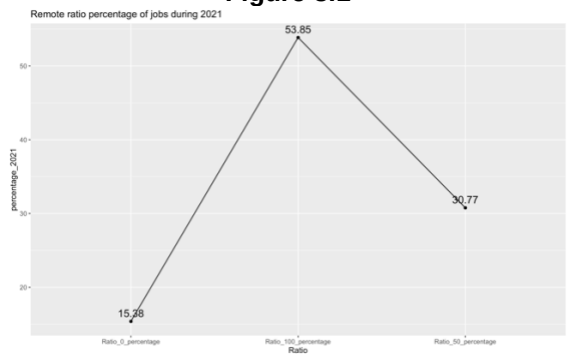

Figure 3.2

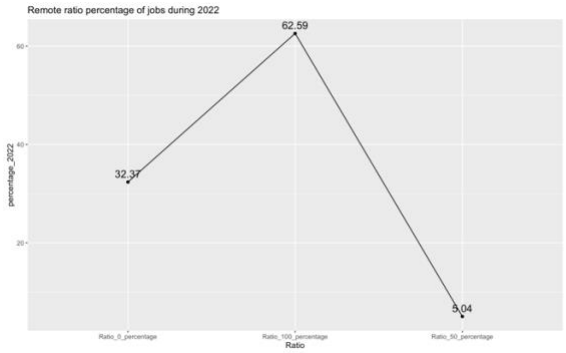

Figure 3.3

Number of jobs in each country during and post-pandemic

Dataset(s) used:- IT job salaries related to data science, Covid19 dataset
Tool(s) used:- R, Tableau

- Calculated the number of jobs provided by each country each year and calculated the average daily cases in that country.
- Saved all these data with country names in a CSV file.
- Loaded the data into Tableau.
- The colour saturation represents the severity of covid cases counted in a country.
- The countries are represented with their country code, and the count is mentioned below.

The job counts represent the mapping of count to high scale number of jobs.

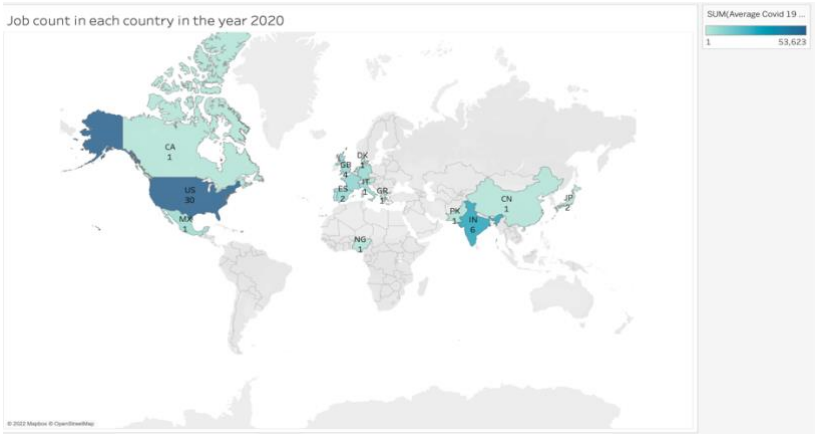

Figure 4.1

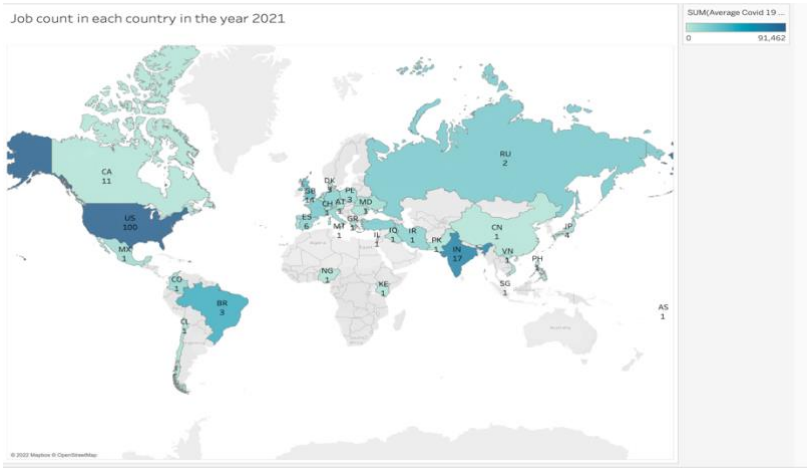

Figure 4.2

Figure 4.3

Salaries paid to the employees

Dataset(s) used:- IT job salaries related to data science, Covid-19 dataset.
Tool(s) used:- R, Tableau

- Created a new dataset after extracting the average pay in a country for each year from the primary dataset. And getting the moderate covid19 cases in that year in each country from the covid19 dataset.
- Loaded the data into a tableau.
- The saturation represents the average covid 19 cases in that country map.
- The salary is represented as continuous values, and the country is represented with the country code in its respective location.

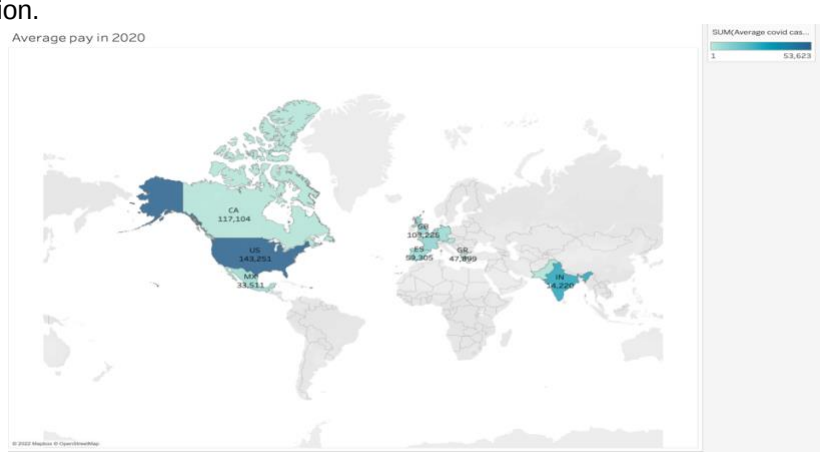

Figure 5.1

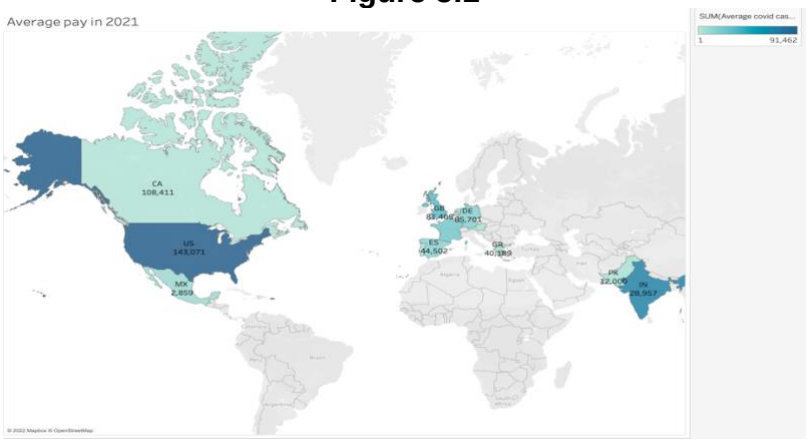

Figure 5.2

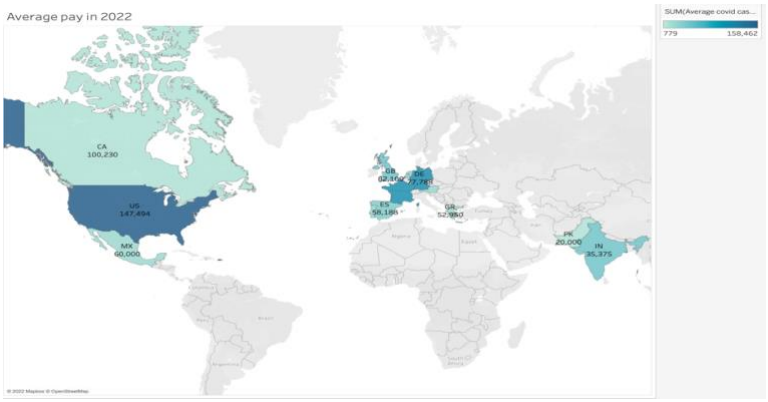

Figure 5.3

Count of different job roles in each year

Dataset(s) used:- IT job salaries related to data science
Tool(s) used:- R, Tableau

- Filtered the data from the primary dataset(IT role salaries) by extracting the count of each job role in a specific year. According to the year, the average cases of covid 19 are removed from the covid 19 dataset.
- Saved the filtered data into a CSV file.
- Loaded into Tableau.
- Plotted the data using bar graphs.
- This represents the count of each job role each year and shows the trends of each.

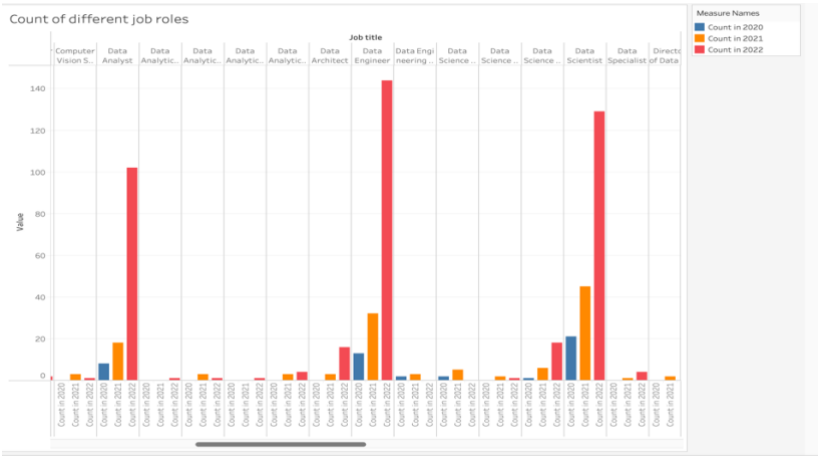

Figure 6.1

Conclusions

For Question 1:

In figure 1.1, firstly, the entry-level for data engineers went up balanced. It almost seems equal a rise each year. Whereas senior level salary also increased but not in the equal share compared to the previous year. And the mid-level salary was high in 2020, went low in 2021 and had a slight angle in 2022.

In figure 1.2, the salaries of data scientists with entry-level had no drastic change, i.e. only a slight difference (54000 USD - 55000 USD) between the first two years(2020 and 2021), and in 2022, the salaries went more than 60000 USD. Coming to the mid-level salary, the scenario is the same as the fresher trend but on a colossal scale. While in senior level pay, it went up-down-up manner, i.e. in 2020, the salary is more than 150000 USD; in 2021, it went down below 100000 USD; in 2022, it nearly reached 150000 USD.

Figure 2.1. In 2020, Machine Learning Engineer was paid to a margin nearer to 120000 USD, next comes to Business Data Analyst's salary, around 100,000 USD, and Big Data Engineer paid nearly 74,000 USD.

Figure 2.2. In the year 2021, a Machine learning scientist's salary is 250000, Applied data scientists are above 100000, and a Machine learning developer is around 100,000.

Figure 2.3. In the year 2022, the salary of a Machine learning Developer went high to above 175,000, Computer Vision Software Engineer got around 150,000, and Analytics Engineer's salary slightly above 125000.

For Question 2:

Figure 3.1. In 2020, the ratio of fully remote jobs was 52% more than the hybrid(50% remote) and the office-work, which has 28% and 20%.

Figure 3.2. In 2021, the ratio of complete remote jobs increased slightly to 53.85%(1.85% than in 2020), and hybrid work increased by 2.77% and is 30.77%. So, the complete offline work is reduced to 15.38, which is lower than 2020.

Figure 3.3. In 2022, the ratio of fully remote jobs is again but in more percentage than in 2021, i.e. 53.85% to 62.59%. Post the pandemic; the hybrid jobs count went entirely down to 5.04 from 30.77, which is an enormous change. And the in-office jobs went to 32.37%, which increased more than half the percentage of office jobs in 2021.

Figure 4.1. In 2020, the US provided more jobs during the first year of the pandemic, even when the covid 19 cases were more than the rest of the world. And next comes India, offering jobs with covid19 patients more than 30000, and then Great Britain provided more jobs.

Figure 4.2. In 2021, the scenario was same as in 2020 in providing jobs, but there is an increase in covid19 cases and also an increase in employment.

Figure 4.3. In 2022, there will be changes in job offerings as the USA usually stands in front of providing the jobs. But, Canada became 2nd in giving the number of positions, and India and Great Britain ended up at third highest to offer employment.

Data maps in fig 4.1, 4.2 and 4.3 would look more apparent if the percentages were included instead of the jobs count. That makes information more technical.

For Question 3:

Figure 5.1. In 2020, the US paid more to the employees. Secondly, Canada has spent better but lesser than the US. Great Britain stands in third place for salary payments. But, the covid cases in Canada and Great Britain isn't as severe as in the US.

Figure 5.2. In 2021, the US again remained at the top for offering more pay, and there was not much increase in salary pay. Whereas Canada stands in second place and Great Britain has a notified decrease, Denmark comes third highest paying.

Figure 5.3. In 2022, the US's salary went up but not too high. And the wages of Canada went down again but not least than the other countries except the US. Also, the pay in Denmark went low, and Great Britain's salary was slightly above. In addition, India and Pakistan had no fall when comparing their average pay with the previous year. There is some increase each year.

Figure 6.1. The job roles such as Data Analyst, Data Engineer and Data Scientist have a notable rise in their count. The count of Data Scientist jobs is highest in 2020 and 2021, whereas the count of Data Engineer roles is the highest in 2022.

Reflection

- I learned how to merge the datasets to create questions that give broad conclusions, which help in diving into deep exploration techniques.
- Lots of learning by getting familiar with R programming and maintaining the structure to analyse or plot the graphs.
- Using Tableau is handy for me, where I can plot data after cleaning them using R or any tools.
- Getting through the data analysis tasks gave me more ideas on what to explore beyond the questions I got.
- There are many changes, and exploration needs to be done with the data to answer many questions.
- I got to know how we can spread our analysis by merging it with another dataset.

Bibliography

[1] Javier Canales, Luna (2022), *Article shows 'Data science salary expectations in 2022*, 18.05.2022. URL: https://www.datacamp.com/blog/data-science-salaries [Accessed on: 07.08.2022].