Portfolio Data Analytics
3nd

Salary Project

TAKEN FROM CASE STUDY
RevoU Mini Course - Data Analytics

DHEA AMALIA LUTFIANI
(25 JULY – 5 AUGUST)

Case Study Instructions

QUESTION

Table of interest :

Salary Dataset

1. Look at this data and start thinking. List down 3 trends/points that you want to show.
2. From here, try to explore the data and make changes, filter, and prepare the data that you need.
3. Create some visualizations or dashboard with the best type of chart you have learned.

   The easiest is with Google Data Studio or Google Sheets.
4. Then, make 1-2 slides from the Graphs with the insights you got to present your findingsto the stakeholders

   (read this article from HBR)

Preview Data

Info from Dataset

```
salary = pd.read_csv("salary_dataset.csv")
salary.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           607 non-null    int64
 1   experience_level    607 non-null    object
 2   employment_type     607 non-null    object
 3   job_title           607 non-null    object
 4   salary              607 non-null    int64
 5   salary_currency     607 non-null    object
 6   salary_in_usd       607 non-null    int64
 7   employee_residence  607 non-null    object
 8   remote_ratio        607 non-null    int64
 9   company_location    607 non-null    object
 10  company_size        607 non-null    object
dtypes: int64(4), object(7)
memory usage: 52.3+ KB
```

From this information, it can be seen that:

- There are 11 columns and the amount of data varies is 607 data
- The data type of the salary dataset are int64 and object

Preview Data

Preview the Dataset before data cleaning:

salary

	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 Developer	260000	USD	260000	JP	0	JP	S
2	2020	SE	FT	Data Engineer	85000	GBP	109024	GB	50	GB	M
3	2020	MI	FT	Data Analyst	20000	USD	20000	HN	0	HN	S
4	2020	SE	FT	Machine Learning Developer	150000	USD	150000	US	50	US	L
...	...	...	...	...	...	...	...	...	...	...	...
602	2022	SE	FT	Data Engineer	154000	USD	154000	US	100	US	M
603	2022	SE	FT	Data Engineer	126000	USD	126000	US	100	US	M
604	2022	SE	FT	Data Analyst	129000	USD	129000	US	0	US	M
605	2022	SE	FT	Data Analyst	150000	USD	150000	US	100	US	M
606	2022	MI	FT	Data Scientist	200000	USD	200000	IN	100	US	L

607 rows × 11 columns

Python and Data Cleaning

Data Cleaning

Messy data is a common problem you'd likely face when you have data from sources like spreadsheets. It is important to clean your data before doing any analysis.

Things to do in data cleaning :

1. Change data type
2. Remove duplicated data : from 607 data to 562 data, where there are 45 duplicate data, which is 7.41% of the total data.
3. Remove empty data
4. Remove outliers : Delete data that has a value that exceeds the upper and lower limits (data outliers) in the 'salary_in_usd' column, so that the total data becomes 552.
5. Remove unnecessary data

Link Google Colab :
https://colab.research.google.com/drive/1I2zP1P0Rt_ThnIZ7aHZOMfDXjs602xsS?usp=sharing

Python and Data Cleaning

Preview the Dataset after data cleaning:

salary

	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 Developer	260000	USD	260000	JP	0	JP	S
2	2020	SE	FT	Data Engineer	85000	GBP	109024	GB	50	GB	M
3	2020	MI	FT	Data Analyst	20000	USD	20000	HN	0	HN	S
4	2020	SE	FT	Machine Learning Developer	150000	USD	150000	US	50	US	L
...	...	...	...	...	...	...	...	...	...	...	...
602	2022	SE	FT	Data Engineer	154000	USD	154000	US	100	US	M
603	2022	SE	FT	Data Engineer	126000	USD	126000	US	100	US	M
604	2022	SE	FT	Data Analyst	129000	USD	129000	US	0	US	M
605	2022	SE	FT	Data Analyst	150000	USD	150000	US	100	US	M
606	2022	MI	FT	Data Scientist	200000	USD	200000	IN	100	US	L

552 rows × 11 columns

Preview Data

After cleaning the data in python on Google Colab, then we make a query from the data with BigQuery

Schema from the Dataset :

Field name	Type	Mode
int64_field_0	INTEGER	NULLABLE
work_year	INTEGER	NULLABLE
experience_level	STRING	NULLABLE
employment_type	STRING	NULLABLE
job_title	STRING	NULLABLE
salary	INTEGER	NULLABLE
salary_currency	STRING	NULLABLE
salary_in_usd	INTEGER	NULLABLE
employee_residence	STRING	NULLABLE
remote_ratio	INTEGER	NULLABLE
company_location	STRING	NULLABLE
company_size	STRING	NULLABLE

Description :

Column	Description
work_year	The year the salary was paid.
experience_level	The experience level in the job during the year with the following possible values: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director
employment_type	The type of employement for the role: PT = Part-time / FT = Full-time / CT = Contract / FL = Freelance
job_title	The role worked in during the year.
salary	The total gross salary amount paid.
salary_currency	The currency of the salary paid as an ISO 4217 currency code.
salary*in*usd	The salary in USD
employee_residence	Employee's primary country of residence in during the work year as an ISO 3166 country code.
remote_ratio	The overall amount of work done remotely, possible values are as follows: 0 No remote work (less than 20%) 50 Partially remote 100 Fully remote (more than 80%)
company_location	The country of the employer's main office or contracting branch as an ISO 3166 country code.
company_size	The average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large)

From this data preview, we can find out the description of each column and know which columns can be used to answer problems or which columns can be useful for finding new insights.

Preview Data

Preview the Data :

SCHEMA DETAILS PREVIEW

Row	int64_field_0	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
1	130	2021	EN	FT	Machine Learning Developer	100000	USD	100000	IQ	50	IQ	S
2	153	2021	EN	FT	Data Scientist	13400	USD	13400	UA	100	UA	L
3	191	2021	EN	FT	Machine Learning Developer	21844	USD	21844	CO	50	CO	M
4	196	2021	EN	FT	Data Analyst	9272	USD	9272	KE	100	KE	S
5	279	2021	EN	FT	Data Analyst	50000	EUR	59102	LU	100	LU	L
6	281	2021	EN	FT	Data Scientist	100000	USD	100000	JE	0	CN	L
7	487	2022	EN	PT	Data Scientist	100000	USD	100000	DZ	50	DZ	M
8	489	2022	EN	CT	Machine Learning Developer	29000	EUR	31875	TN	100	CZ	M
9	505	2022	EN	FT	Data Scientist	120000	AUD	86703	AU	50	AU	M
10	510	2022	EN	FT	Data Engineer	150000	USD	150000	AU	100	AU	S
11	96	2021	EN	PT	Data Scientist	12000	USD	12000	BR	100	US	S
12	499	2022	EN	FT	Data Scientist	66500	CAD	52396	CA	100	CA	L
13	600	2022	EN	FT	Data Analyst	67000	USD	67000	CA	0	CA	M
14	601	2022	EN	FT	Data Analyst	52000	USD	52000	CA	0	CA	M
15	45	2020	EN	PT	Data Engineer	14000	EUR	15966	DE	100	DE	S

Defining Question

List down trends/points that you want to show :

1. Average salary_in_usd on experience level every year
2. Average salary_in_usd on job title every year
3. What jobs have the highest and lowest salaries?
4. What types of employment_type pay the most?
5. Located where most of the workers work?

Exploring Data

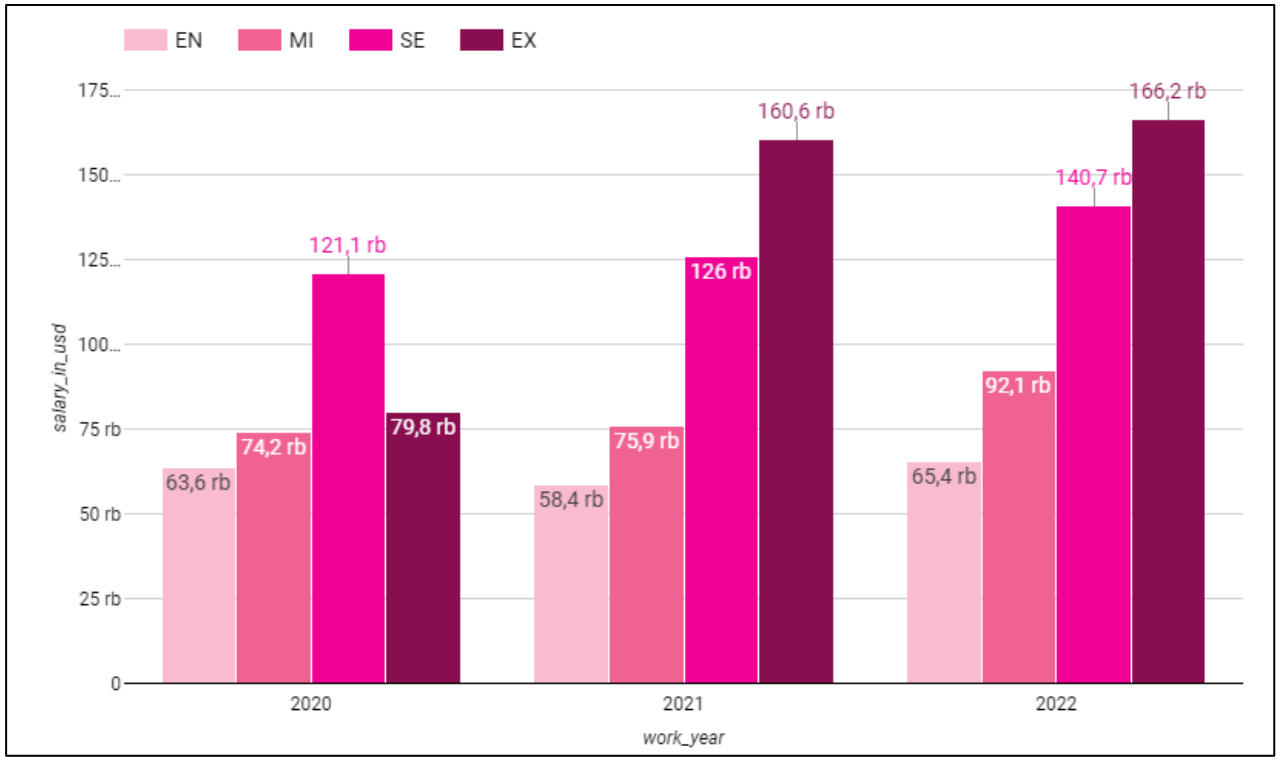
Include SQL with BigQuery

- The following is a ready-to-use query for data visualization in Google Data Studio :

```sql
SELECT
  work_year,
  experience_level,
  employment_type,
  job_title,
  salary,
  salary_currency,
  salary_in_usd,
  employee_residence,
  remote_ratio,
  company_location,
  company_size
FROM `assignment-mini-course-da.salary_dataset.salary` LIMIT 1000
```

Visualization with Insight

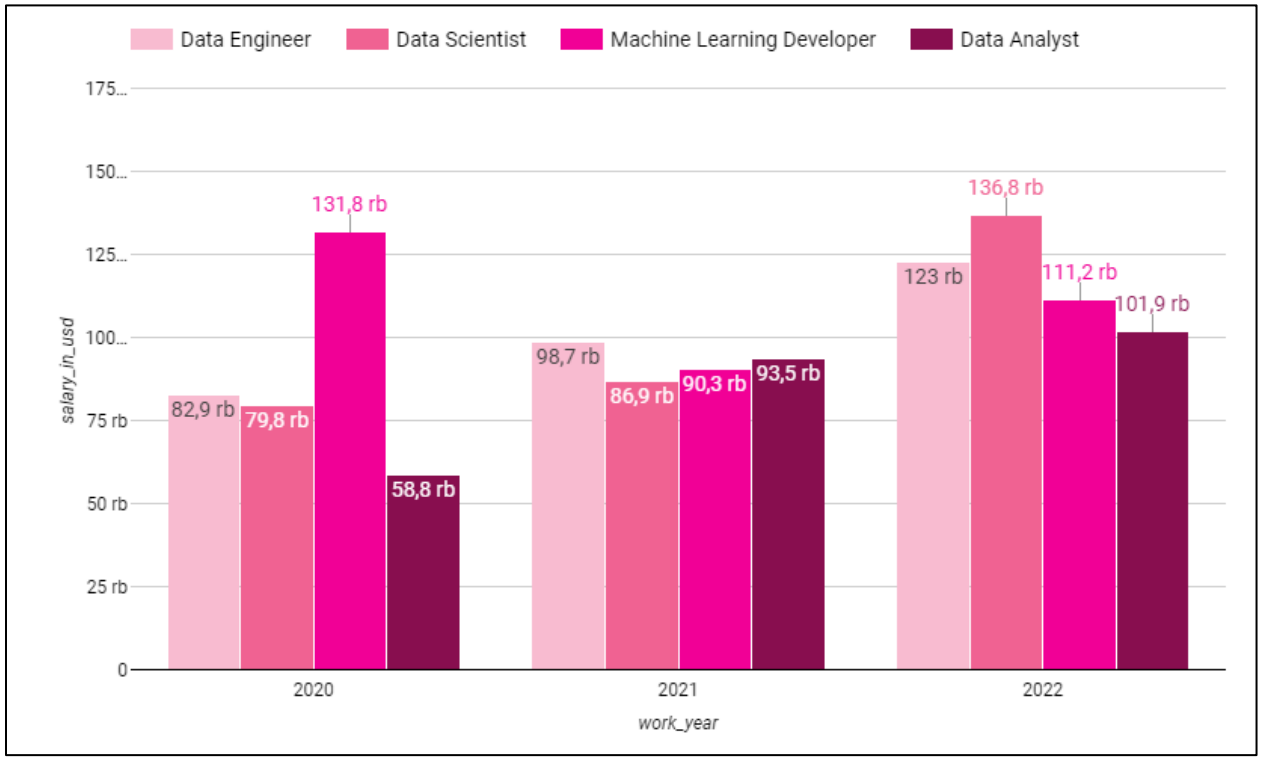
1. Average salary_in_usd on experience level every year



- From the diagram, it can be seen that the average salary in each year tends to be constant, even though the salary at the EX experience level has increased by 2 times in 2021.
- And if you look at the average salary given according to the level of work experience. Where if the worker has higher experience then the salary given is also higher.
- The average salary given to EX in 2020 is lower than that given to SE in that year. This can be caused by the number of EX workers more than SE and with a different nominal for each company.

Visualization with Insight

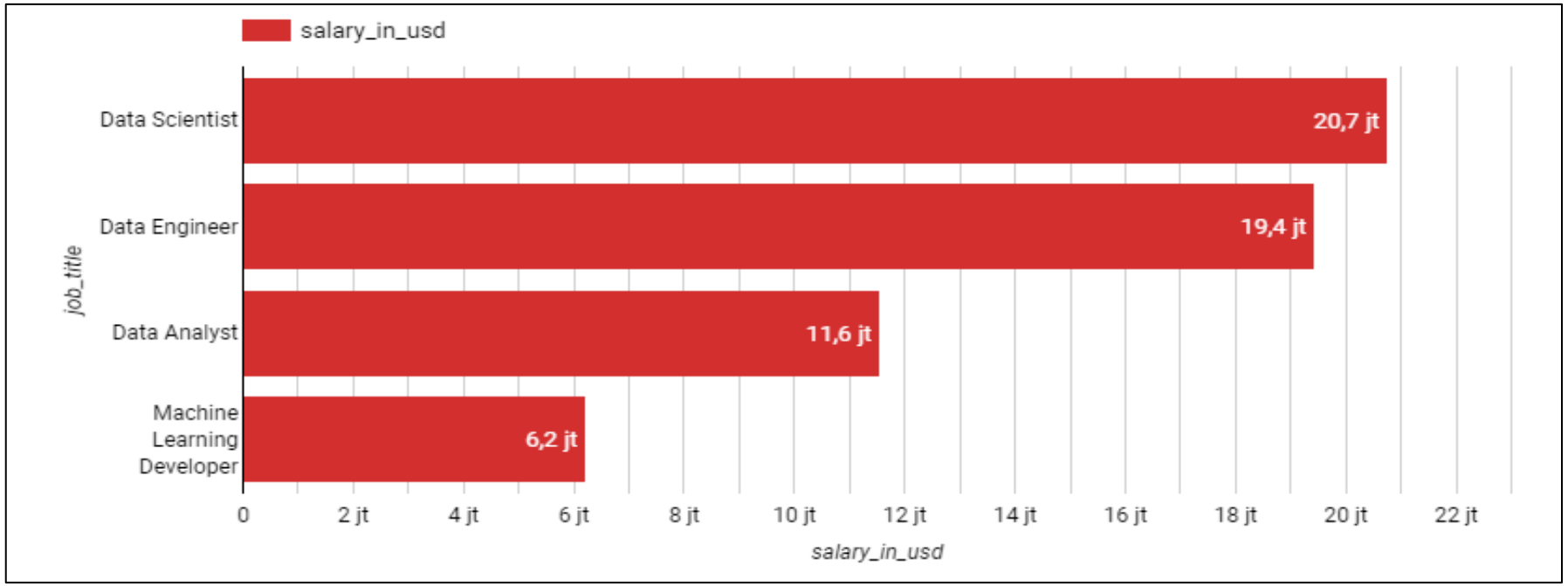
2. Average salary_in_usd on job title every year



- For the average salary in accordance with the role of work each year, this has increased every year for the roles of Data Engineer, Data Scientist, and Data Analyst.
- Meanwhile, the Machine Learning Developer role experienced fluctuations where the highest average salary was in 2020 which then fell by 31.5%, and then increased again in 2022.
- And the largest average salary is for the Data Scientist role in 2022.

Visualization with Insight

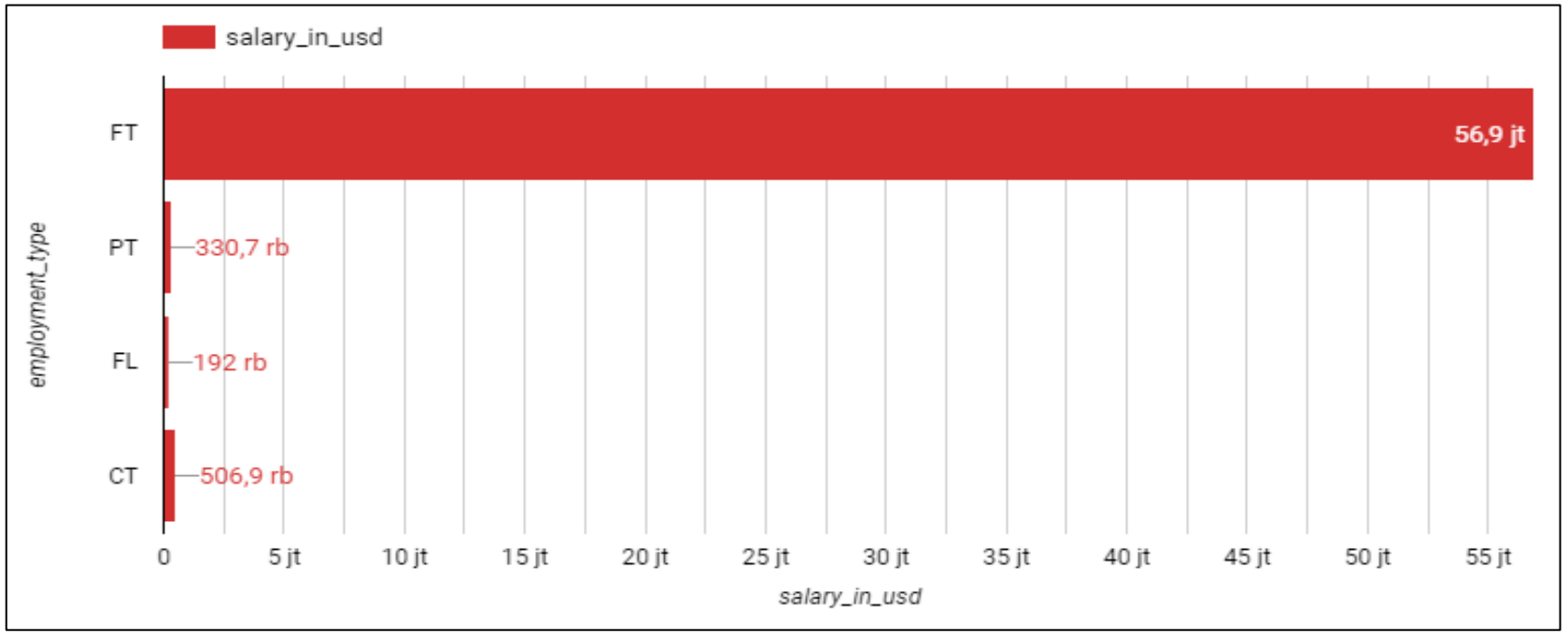
3. What jobs have the highest and lowest salaries?



- The biggest salary is the salary for Data Scientist jobs and the smallest salary is for Machine Learning Developer jobs

Visualization with Insight

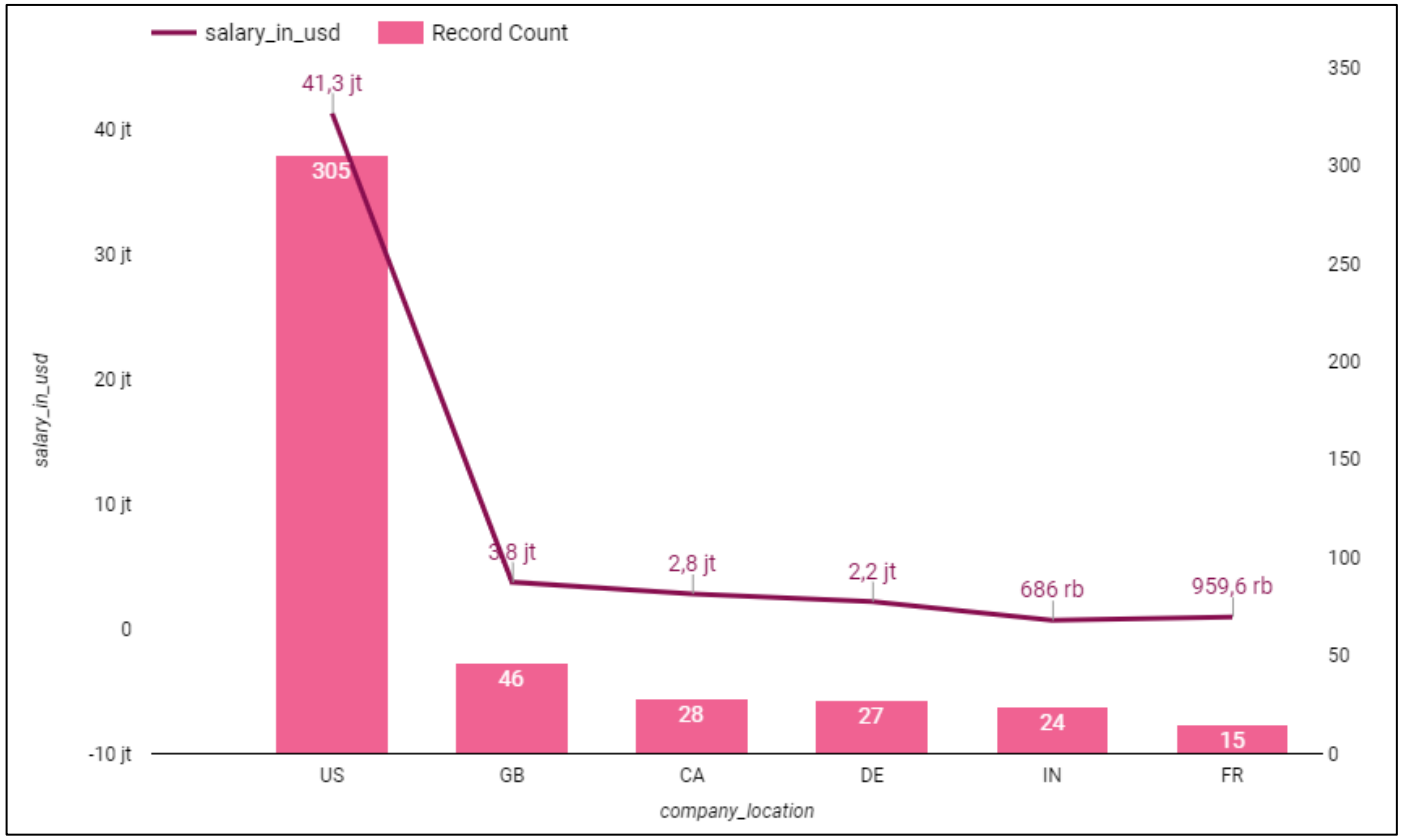
4. What types of employment_type pay the most?



- The employment_type that gets the largest salary is the FT or Full Time type, which means that the majority of the workers in the salary_dataset are Full Time workers.

Visualization with Insight

5. Located where most of the workers work?



- From all the locations in the data, only the top 6 locations for working people were taken.
- It can be seen that the US became the most favorite location among the others, more than 6 times from the second location called GB.
- And the US spends about 41.3 million dollars on its workers.

THANK YOU

DHEA AMALIA LUTFIANI
(25 JULY – 5 AUGUST)