2020 EU IT Salary Survey (Data Exploration)

Hello, I am Mauricio and my goal for 2021 is to expand my knowledge and promtime self growth. Today's scenario is aiem to study and explore the data of EU IT Salary Survey from Kaggle and gain insights.

Objectives:

- Explore the data points captured in the dataset
- Perform EDA (Exploratory Data Analysis)
- Visualize the data so aim to data story-telling

Let's bake together (all fixing) Mauricio :)

Import Libs

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        %matplotlib inline
        from matplotlib import rcParams
        rcParams['figure.figure.dpi'] = 150
        rcParams['axes.grid'] = True
        rcParams['axes.spines.top'] = False
        rcParams['axes.spines.right'] = False
```

Import data

```
In [2]: df = pd.read_csv('IT Salary Survey EU 2020.csv')
        df.head()
```

	Timestamp	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Other technologies/programming languages you use often	Number of vacation days	Bachelor degree	Master degree	PhD
0	2021/03/08 10:01:49 GMT+1	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan
1	2021/03/08 10:01:49 GMT+1	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan
2	2021/03/08 10:01:49 GMT+1	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan
3	2021/03/08 10:01:49 GMT+1	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan
4	2021/03/08 10:01:49 GMT+1	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan

5 rows × 23 columns

It gathers the survey numbers:

- 30 Countries of the temporalperiod (specific for the IT employees in EU such as Kyiv, Kharkiv or London)
- 5 years of data gathering (2016 - 2020) with 23 features (Age, Gender, City, Position, Salary, etc.) and 1 target (Salary)
- Around 15.000 entries (total) out of an unsufficiencies (or samples) 50% of which are incomplete with nan cells.

As we observed the first 5 rows of the data using head(), we can see that:

- The column about the devices (or tools) we use (under technologies or programs otherwise mentioned)
- The column about the degrees (Bachelor, Master, PhD) is mostly filled with nan values.
- The last 3 columns are assumed to have a high amount of null values, it might not be very helpful to visualize these data (we can drop them if necessary).

Let's take our hypothesis about the data:

```
In [3]: df.columns.values
```

```
array([' Timestamp', 'Age', 'Gender', 'City', 'Position ',
       'Total yearly compensation ', 'Years of experience',
       'Years in current company ', 'Other technologies/programming languages you use often',
       'Number of vacation days', 'Bachelor degree', 'Master degree',
       'PhD', 'Salary 1 previous year', 'Salary 2 previous year',
       'Salary 3 previous year', 'Bonus', 'Stock', 'Total compensation ',
       'Currency', 'Other currency', 'Additional comments',
       'Your main technology / programming language'],
      dtype=object)
```

```
In [4]: df.info()
```

```
<class'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 23 columns):
 Timestamp                                                    15000 non-null  object
 Age                                                          15000 non-null  int64
 Gender                                                       14952 non-null  object
 City                                                         14957 non-null  object
 Position                                                     14957 non-null  object
 Total yearly compensation                                    14957 non-null  int64
 Years of experience                                          14957 non-null  object
 Years in current company                                     14957 non-null  object
 Other technologies/programming languages you use often       9154 non-null  object
 Number of vacation days                                      14957 non-null  int64
 Bachelor degree                                              4966 non-null  object
 Master degree                                                4966 non-null  object
 PhD                                                          4966 non-null  object
 Salary 1 previous year                                       14957 non-null  int64
 Salary 2 previous year                                       14957 non-null  int64
 Salary 3 previous year                                       14957 non-null  int64
 Bonus                                                        14957 non-null  int64
 Stock                                                        14957 non-null  int64
 Total compensation                                           14957 non-null  int64
 Currency                                                     14957 non-null  object
 Other currency                                               14957 non-null  object
 Additional comments                                          2154 non-null  object
 Your main technology / programming language                  14957 non-null  object
dtype: object, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64
memory usage: 2.6+ MB
```

It gathers the survey numbers:

- 30 Countries of the temporalperiod (specific for the IT employees in EU such as Kyiv, Kharkiv or London)
- 5 years of data gathering (2016 - 2020) with 23 features (Age, Gender, City, Position, Salary, etc.) and 1 target (Salary)
- Around 15.000 entries (total) out of an unsufficiencies (or samples) 50% of which are incomplete with nan cells.

As we observed the first 5 rows of the data using head(), we can see that:

- The column about the devices (or tools) we use (under technologies or programs otherwise mentioned)
- The column about the degrees (Bachelor, Master, PhD) is mostly filled with nan values.
- The last 3 columns are assumed to have a high amount of null values, it might not be very helpful to visualize these data (we can drop them if necessary).

Let's take our hypothesis about the data:

```
In [5]: df.columns.values
```

```
array([' Timestamp', 'Age', 'Gender', 'City', 'Position ',
       'Total yearly compensation ', 'Years of experience',
       'Years in current company ', 'Other technologies/programming languages you use often',
       'Number of vacation days', 'Bachelor degree', 'Master degree',
       'PhD', 'Salary 1 previous year', 'Salary 2 previous year',
       'Salary 3 previous year', 'Bonus', 'Stock', 'Total compensation ',
       'Currency', 'Other currency', 'Additional comments',
       'Your main technology / programming language'],
      dtype=object)
```

```
In [6]: df.info()
```

```
<class'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 23 columns):
 Timestamp                                                    15000 non-null  object
 Age                                                          15000 non-null  int64
 Gender                                                       14952 non-null  object
 City                                                         14957 non-null  object
 Position                                                     14957 non-null  object
 Total yearly compensation                                    14957 non-null  int64
 Years of experience                                          14957 non-null  object
 Years in current company                                     14957 non-null  object
 Other technologies/programming languages you use often       9154 non-null  object
 Number of vacation days                                      14957 non-null  int64
 Bachelor degree                                              4966 non-null  object
 Master degree                                                4966 non-null  object
 PhD                                                          4966 non-null  object
 Salary 1 previous year                                       14957 non-null  int64
 Salary 2 previous year                                       14957 non-null  int64
 Salary 3 previous year                                       14957 non-null  int64
 Bonus                                                        14957 non-null  int64
 Stock                                                        14957 non-null  int64
 Total compensation                                           14957 non-null  int64
 Currency                                                     14957 non-null  object
 Other currency                                               14957 non-null  object
 Additional comments                                          2154 non-null  object
 Your main technology / programming language                  14957 non-null  object
dtype: object, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64, int64
memory usage: 2.6+ MB
```

```
In [7]: df.nunique()
```

```
Timestamp                                                    15000
Age                                                             53
Gender                                                           4
City                                                           ...
Position                                                       ...
Total yearly compensation                                      ...
Years of experience                                            ...
Years in current company                                       ...
Other technologies/programming languages you use often         ...
Number of vacation days                                        ...
Bachelor degree                                                ...
Master degree                                                  ...
PhD                                                            ...
Salary 1 previous year                                         ...
Salary 2 previous year                                         ...
Salary 3 previous year                                         ...
Bonus                                                          ...
Stock                                                          ...
Total compensation                                             ...
Currency                                                       ...
Other currency                                                 ...
Additional comments                                            ...
Your main technology / programming language                    ...
dtype: 15000, 53, 4, ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ...
```

```
In [8]: df.sample(5)
```

	Timestamp	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Other technologies/programming languages you use often	Number of vacation days	Bachelor degree	Master degree	PhD
4[illegible]	2021/03/08 10:01:49 GMT+1	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan
1[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan
2[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan
3[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan
4[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan

5 rows × 23 columns

```
In [9]: df.describe()
```

	Age	Total yearly compensation	Number of vacation days	Salary 1 previous year	Salary 2 previous year	Salary 3 previous year	Bonus	Stock	Total compensation
count	15000.000000	14957.000000	14957.000000	14957.000000	14957.000000	14957.000000	14957.000000	14957.000000	14957.000000
mean	32.826667	67.070.054.170	29.000000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000
std	5.076.000.000	1.000.000.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
min	18.000000	1.000000.000	29.000000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000
25%	29.000000	1.000000.000	29.000000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000
50%	32.000000	1.000000.000	29.000000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000
75%	36.000000	1.000000.000	29.000000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000
max	65.000000	1.000000.000	29.000000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000	57.000.000.000

```
In [10]: df.describe(exclude=np.number)
```

	Timestamp	Gender	City	Position	Years of experience	Years in current company	Other technologies/programming languages you use often	Bachelor degree	Master degree	PhD	Currency	Other currency	Additional comments	Your main technology / programming language
count	15000	14952	14957	14957	14957	14957	9154	4966	4966	4966	14957	14957	2154	14957
unique	15000	4	30	1.000	30	30	1.000	1	1	1	1.000	1.000	1.000	1.000
top	2021/03/08 10:01:49 GMT+1	Male	Munich	Software Developer / Programmer	29	29	Docker, Kubernetes	Computer Science	Computer Science	nan	EUR	nan	nan	Docker
freq	1	14.000	1.000	14.957	1.000	1.000	9.154	4.966	4.966	4.966	14.957	14.957	2.154	14.957

5 rows × 23 columns

```
In [11]: df.shape
```

```
(15000, 23)
```

```
In [12]: df.isnull.sum()
```

```
Age                                                          0
Gender                                                      48
City                                                        43
Position                                                    43
Total yearly compensation                                   43
Years of experience                                         43
Years in current company                                    43
Other technologies/programming languages you use often    5846
Number of vacation days                                     43
Bachelor degree                                          10034
Master degree                                            10034
PhD                                                      10034
Salary 1 previous year                                      43
Salary 2 previous year                                      43
Salary 3 previous year                                      43
Bonus                                                       43
Stock                                                       43
Total compensation                                          43
Currency                                                    43
Other currency                                              43
Additional comments                                      12846
Your main technology / programming language                 43
dtype: 0, 48, 43, ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ..., ...
```

```
In [13]: df.drop(columns=['Additional comments'], inplace=True)
         df.sample(5)
```

	Timestamp	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Other technologies/programming languages you use often	Number of vacation days	Bachelor degree	Master degree	PhD
4[illegible]	2021/03/08 10:01:49 GMT+1	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan
1[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan
2[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan
3[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan
4[illegible]	2021/03/08 10:01:49 GMT+1	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan

5 rows × 22 columns

```
In [14]: df.drop(columns=['Timestamp'], inplace=True)
```

```
In [15]: df.sample(5)
```

	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Other technologies/programming languages you use often	Number of vacation days	Bachelor degree	Master degree	PhD	Salary 1 previous year
4[illegible]	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan	57.000.000.000
1[illegible]	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
2[illegible]	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
3[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
4[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000

5 rows × 21 columns

```
In [16]: df.rename(columns={'Salary 1 previous year': 'Salary 1', 'Salary 2 previous year': 'Salary 2', 'Salary 3 previous year': 'Salary 3'}, inplace=True)
         df.sample(5)
```

	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Other technologies/programming languages you use often	Number of vacation days	Bachelor degree	Master degree	PhD	Salary 1
4[illegible]	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan	57.000.000.000
1[illegible]	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
2[illegible]	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
3[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
4[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000

5 rows × 21 columns

```
In [17]: df.rename(columns={'Other technologies/programming languages you use often': 'Tools'}, inplace=True)
         df.sample(5)
```

	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Tools	Number of vacation days	Bachelor degree	Master degree	PhD	Salary 1
4[illegible]	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan	57.000.000.000
1[illegible]	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
2[illegible]	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
3[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
4[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000

5 rows × 21 columns

```
In [18]: df.rename(columns={'Your main technology / programming language': 'Main Lang'}, inplace=True)
         df.sample(5)
```

	Age	Gender	City	Position	Total yearly compensation	Years of experience	Years in current company	Tools	Number of vacation days	Bachelor degree	Master degree	PhD	Salary 1
4[illegible]	28	Male	Munich	Software Developer / Programmer	74000	5	3	Docker	30	Computer Science	Computer Science	nan	57.000.000.000
1[illegible]	26	Male	Berlin	Frontend Developer	60000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
2[illegible]	26	Male	Berlin	Backend Developer	65000	3	1	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
3[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000
4[illegible]	26	Male	Munich	Software Developer / Programmer	62000	3	2	Docker, Kubernetes	30	Other	nan	nan	57.000.000.000

5 rows × 21 columns

Visualization

Let's better understand the dataset, we can focus on different aspect of the dataset

1. Demographical analysis of IT staff in the EU (Gender)
2. Geographical analysis of the survey participants (number of respondents)
3. Salary analysis (Salary, Bonus, Stock)

```
In [19]: # Top 20 most represented cities in Europe (https://www.business.com/articles/european-cities-with-most-tech-jobs/)
         # Top 20 most represented cities in Europe
         df['City'].value_counts[:20].plot.bar(figure=True)
         plt.xlabel('Cities')
         plt.ylabel('Count')
         plt.title('Top 20 Cities')
         plt.show()
```

```
In [20]: # Gender distribution
         df['Gender'].value_counts[:3].plot.pie(figure=True, autopct='%1.1f%%')
         plt.ylabel('')
         plt.title('Gender Distribution')
         plt.legend(['Male', 'Female', 'Other'])
         plt.show()
```

```
In [21]: # Generation counts
         df['Gender'].value_counts[:3].plot.bar(figure=True)
         plt.xlabel('Gender')
         plt.ylabel('Count')
         plt.title('Generation Counts')
         plt.legend(['Male', 'Female', 'Other'])
         plt.show()
```

```
In [22]: # Top 20 City Location
         df['City'].value_counts[:20].plot.bar(figure=True)
         plt.xlabel('Cities')
         plt.ylabel('Count')
         plt.title('Top 20 City Location')
         plt.legend(['Male', 'Female', 'Other'])
         plt.show()
```

What can we understand from the survey demography?

- In EDA, the survey respondents of IT staff are mainly dominated by Male
- Top 20 cities in Europe (most tech jobs) are represented in the survey, the biggest group of IT is in US Cities? It is false (related to a variety of age groups). These groups maybe have different career positions and salaries.
- Most of the survey respondents are from capitals of Central Europe.

```
In [23]: # Salary analysis
         df['Total yearly compensation'].describe()
         # Salary analysis
         df['Total yearly compensation'].describe()
         # Salary analysis
         df['Total yearly compensation'].describe()
         
```

```
df_salary.head()
df_salary[df_salary['I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?'].isin(['1','2','3','4','5'])]
df_salary.shape
```

(2233, 25)

	Timestamp	Age	Gender	City	Profession	Years of experience	Salary in hand per month	Salary in hand per year	Tech or non tech	Year joined company	Year joined current profession	Company Name	Company Type	Company Location	Total years of experience in tech	Total years of experience in current company	Total years of experience in current profession	I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?	I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?	I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?	I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?	I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?
0	2/12/2021 10:00:00	26-30	Male	Bangalore	Software Developer	2 - 5 years	5 - 10 lakhs	7500 - 10000	Tech	2020	2020	Tech Mahindra	Service Company	Bangalore	2 - 5 years	1 - 2 years	1 - 2 years	1	2	3	4	5
1	2/12/2021 10:00:00	26-30	Male	Bangalore	Manager	2 - 5 years	5 - 10 lakhs	7500 - 10000	Tech	2020	2020	Tata	Product	Bangalore	2 - 5 years	1 - 2 years	1 - 2 years	1	2	3	4	5
2	2/12/2021 10:00:00	26-30	Male	Bangalore	Software Developer	2 - 5 years	5 - 10 lakhs	7500 - 10000	Tech	2020	2020	Infosys	Service Company	Bangalore	2 - 5 years	1 - 2 years	1 - 2 years	1	2	3	4	5
3	2/12/2021 10:00:00	26-30	Male	Bangalore	Software Developer	2 - 5 years	5 - 10 lakhs	7500 - 10000	Tech	2020	2020	Infosys	Service Company	Bangalore	2 - 5 years	1 - 2 years	1 - 2 years	1	2	3	4	5
4	2/12/2021 10:00:00	26-30	Male	Bangalore	Manager	2 - 5 years	5 - 10 lakhs	7500 - 10000	Tech	2020	2020	Tata	Product	Bangalore	2 - 5 years	1 - 2 years	1 - 2 years	1	2	3	4	5

0 rows × 25 columns

```
df_salary.shape
```

(2233, 25)

```
df_salary = df_salary[df_salary['Gender'].isin(['Male', 'Female', 'Transgender'])]
df_salary['Gender'].value_counts()
```

Gender
Male 2174
Female 57
Transgender 2
dtype: int64

```
fig_pie = px.pie(df_salary, name='Gender', title='Pie Chart')
fig_pie.show()
```

Pie Chart

```
df_salary = df_salary[df_salary['Gender'].isin(['Male', 'Female', 'Transgender'])]
df_salary['Gender'].value_counts()
```

Gender
Male 2174
Female 57
Transgender 2
dtype: int64

```
fig_pie = px.pie(df_salary, name='Gender', title='Pie Chart')
fig_pie.show()
```

Pie Chart

```
df_salary[df_salary['I have you ever plan to join the commercial sector?'].isin(['Yes', 'No'])]
df_salary.shape
```

(2233, 25)

```
df_salary['I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?'].value_counts()
```

I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?
3 1
1 1
Name: count, dtype: int64

```
fig_bar = px.bar(df_salary, x='Have you ever plan to join the commercial sector?', y='I DO NOT FEAR TECHNOLOGY I PROGRAMMING LANGUAGE?', color='Gender', title='Bar Chart')
fig_bar.show()
```


Bar Chart

```
df_salary[df_salary['
```