Exploratory Data Analysis

2019 IT salary survey in EU region

Created by

Jakub Smyk

This is dataset on which Exploratory Data Analysis (EDA) was conducted. Attributes were previously prepared for this analysis through process of data cleaning. The data cleaning process as well as EDA of this dataset are located on GitHub. The dataset has 20 columns from a 2019 IT salary survey from EU region. The raw survey data originates from Kaggle. Features describe IT sector employees, that took part in this survey. Goal of this analysis is to characterize a group of respondents, that took part in this survey.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 525 entries, 0 to 527
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             525 non-null    int32
 1   gender                          525 non-null    object
 2   city                            525 non-null    object
 3   seniority                       525 non-null    object
 4   position                        525 non-null    object
 5   experience_years                525 non-null    int64
 6   main_technology                 525 non-null    category
 7   yearly_brutto_salary_eur_2019   525 non-null    int32
 8   yearly_bonus_eur_2019           525 non-null    int32
 9   yearly_stocks                   525 non-null    float64
 10  yearly_brutto_salary_eur_2018   525 non-null    int32
 11  yearly_bonus_eur_2018           525 non-null    int32
 12  yearly_stocks_2018              525 non-null    float64
 13  vacation_days                   525 non-null    int32
 14  home_office_days_monthly        525 non-null    int32
 15  work_language                   525 non-null    object
 16  company_size                    525 non-null    object
 17  company_type                    525 non-null    object
 18  contract_duration               525 non-null    object
 19  business_industry               525 non-null    object
```

The process of EDA consists of 4 segments:

1. Estimates of location
2. Estimates of variability
3. Data distribution
4. Correlation

Estimates of location are central values that best describe the data. Examples of such estimates are mean or median.

Estimates of variability are values which consist information about sparsity of the data. Examples for this estimate are variance or standard deviation.

Data distribution is visualisation of estimates mentioned before. It shows all possible values of data.

Correlation shows relationships between features. Correlation can be positive, negative or equal to zero. Positive correlation means that both attributes are linearly increasing together. Negative correlation means, that when one attribute increases then the second one is decreasing. When correlation coefficient is equal to zero there is no linear relationship between testes variables. Correlation can be weak or strong. The further from zero it is the stronger it becomes.

Goal of this analysis is to characterize data, a group of respondents and gather conclusions about salary.

Key salary conclusions

- **There is a short supply (around 4% of respondents) of advanced skills such as Solidity, Google Cloud, SAP / ABAP, Rust, Kubernetes. Its salary median is 27% higher than median of other technologies.** This means that those technologies are rare skill and are highly paid by companies.
- **Product, Consulting companies, Startups and Banks have 35% higher yearly salary mean than** workers of **Universities and Outsourcing**
- **Consulting employees have the highest additional income mean (stocks and bonuses)** which is around 3000 euros in 2018 and 2019 while **bank employees don't have additional income.**
- **English and German are most desirable and payable languages in IT business in Central Europe** although this conclusion may be rigged by high number of German speaking specialists from German cities (43% of respondents), that took part in this survey.

Respondents group characteristic

- Most respondents are in the **20 – 30 years old age range** with **around 10 years of work experience**
- **Males are majority** of respondents
- The most **popular** technologies are **Java and Python**.
- **Top 3 cites**, in which respondents live in are **Berlin, Munich and Amsterdam**.
- **German and English** are **most popular languages** at work.
- **The largest group of respondents are senior positions** followed by midds and juniors.
- Majority of respondents have **unlimited time contracts**.
- **Around 50%** of respondents have **30 days of vacation per year**
- **Most common roles** are **Backend Developers, Data Scientists and Fullstack Developers.**
- Majority works at **large companies** with 100 – 1000+ employees
- Majority **works in Commerce, Finance** and **Transport.**

Estimates of location

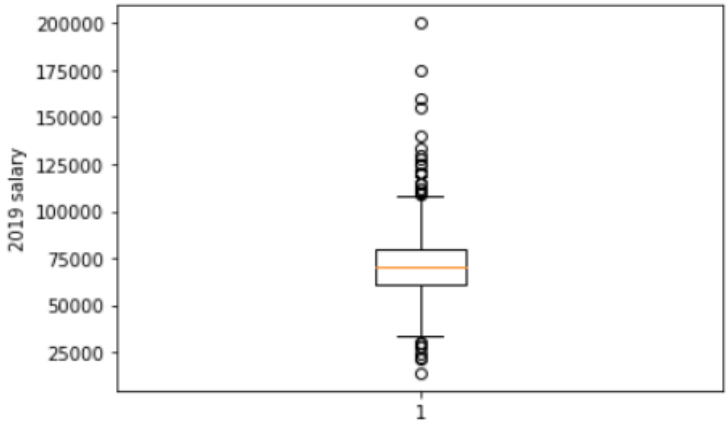
Measure of estimates of location takes into account continuous variables. Using tools such as median, average and boxplots. Performed calculations mean following information:

- Average experience years in a job among respondents is around 10 years
- Median of yearly brutto salary earned in 2019 is 70 000 EUR
- Median of yearly brutto salary in 2018 is 65 000 EUR
- Median of sum of yearly bonuses in 2019 is equal to zero. Same goes for bonuses in 2018 and number of stocks in both years held by respondents.
- Median of vacation days per year is 28
- Median of days in home office per month is equal to 4.

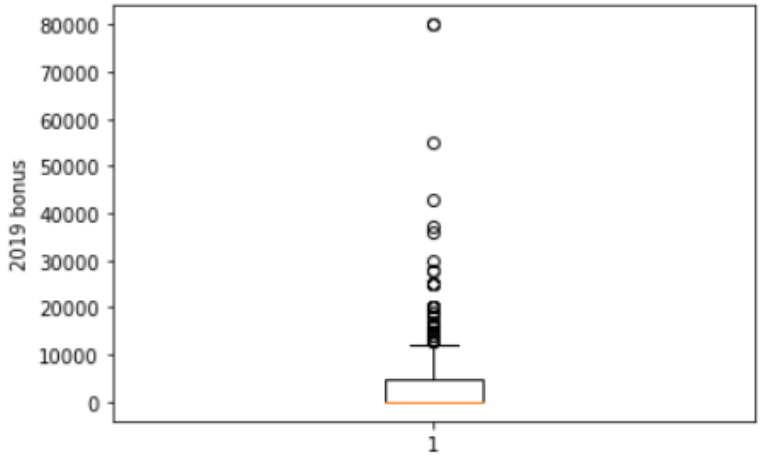
```
In [69]: np.average(df['experience_years'], weights=df['yearly_brutto_salary_eur_2019'])
Out[69]: 9.79304874048867

In [70]: plt.boxplot(df['yearly_brutto_salary_eur_2019'])
         plt.ylabel('2019 salary')
Out[70]: Text(0, 0.5, '2019 salary')

In [72]: np.median(df['yearly_brutto_salary_eur_2019'])
Out[72]: 70000.0

In [75]: np.median(df['yearly_bonus_eur_2019'])
Out[75]: 0.0

In [77]: np.median(df['yearly_stocks'])
Out[77]: 0.0

In [78]: np.median(df['yearly_brutto_salary_eur_2018'])
Out[78]: 65000.0

In [79]: np.median(df['yearly_bonus_eur_2018'])
Out[79]: 0.0

In [80]: np.median(df['yearly_stocks_2018'])
Out[80]: 0.0

In [82]: np.median(df['vacation_days'])
Out[82]: 28.0

In [84]: np.median(df['home_office_days_monthly'])
Out[84]: 4.0
```

```
In [69]: np.median(df['age'])
Out[69]: 32.0
```

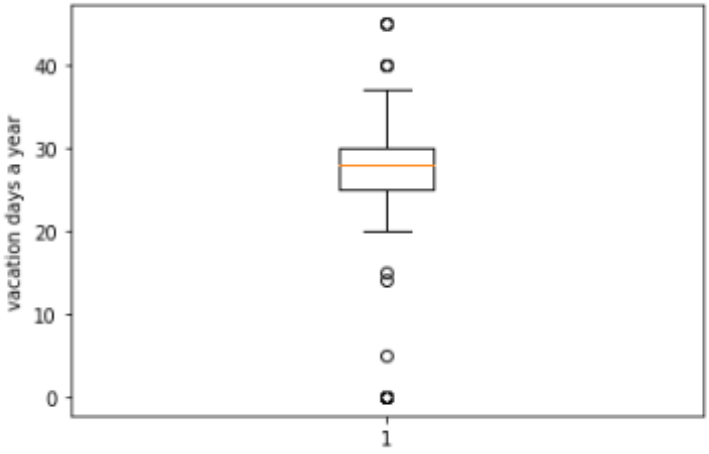
Visualisations expand on information from calculations:



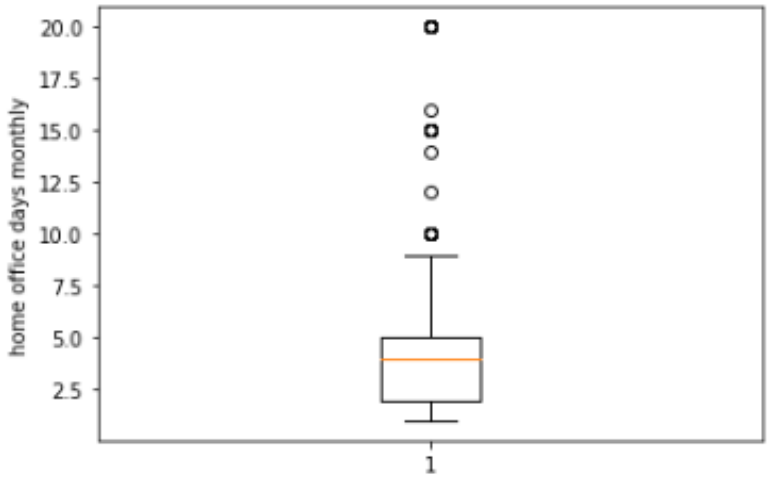
- It can be inferred from boxplot with 2019 salary, that around 50% of respondents had a yearly salary in the year 2019 between 55 000 – 80 000 euros. Outliers above 75th percentile and upper earn way more than 100 000 euros. There is a group of people earning below 40 000 euros. It lies below 25th percentile and bottom whisker.



- Boxplot with 2019 yearly bonus shows that there is a lot of outliers above 10 000 euros of a bonus value. Still most people - which is around 50% - have bonuses with range 0 – 10 000 euros. Most people have 0 bonuses which is derived from median being equal to 0.



- Respondents mostly have around 20 – 30 vacation days per year. Although because of a few outliers, respondents in overall have fallen in range of 0 to 50 vacation days a year.



- Approximately 50% of people spend around 2 – 6 days in a home office monthly. Outliers show that there are employees, that spend most of the month in home office ranging from 10 to 20 days.

Estimates of variability

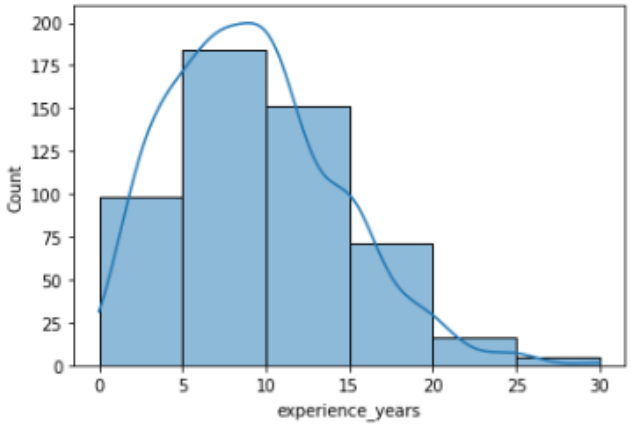
Estimates of variability were measured using mean absolute distribution. MAD is the average distance between each data value and the mean of the data set. The greater MAD is the bigger the sparsity of data.

The highest sparsity among variables is around 14 000 MAD and belongs to yearly brutto salaries from 2018 and 2019. Variables with the lowest values around 2 – 4 MAD are age, experience years, vacation days a year and home office days in a month. Sparsity's of bonuses and number of stocks from both years are close to each other and in between highest and lowest values of MAD with sparsity around 3000/4000.

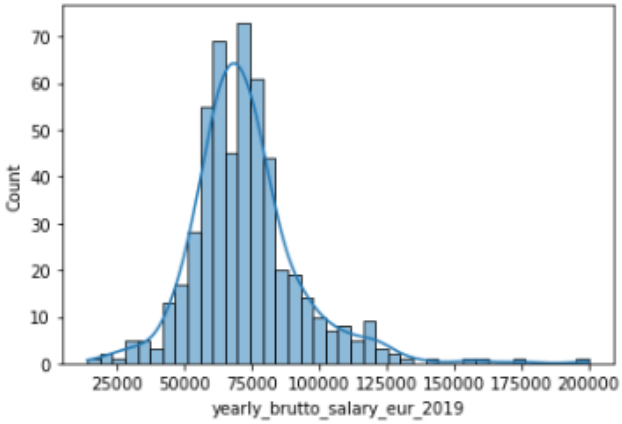
```
In [89]: df.mad() # mean absolute distribution

Out[89]: age                              3.680907
         experience_years                 4.024882
         yearly_brutto_salary_eur_2019    14209.229257
         yearly_bonus_eur_2019            4867.626667
         yearly_stocks                    3123.291151
         yearly_brutto_salary_eur_2018    14017.338282
         yearly_bonus_eur_2018            4867.626667
         yearly_stocks_2018               1496.242103
         vacation_days                    2.830476
         home_office_days_monthly         3.468575
         dtype: float64
```

Data distribution

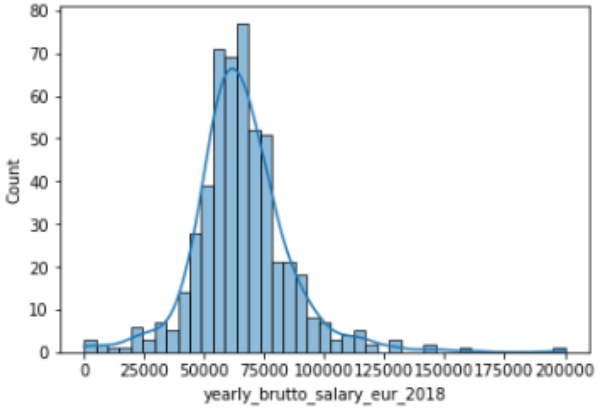
Data distribution is a graph, that shows every value of a variable and it's frequency. These plots show central tendency and sparsity of data.



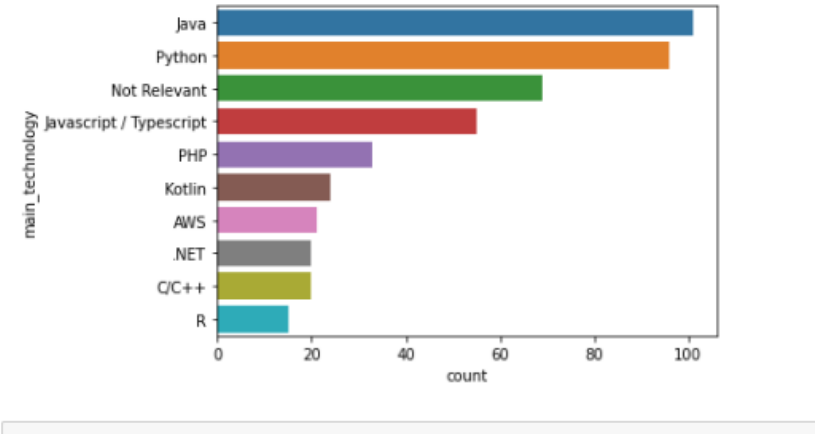
The first plot shows distribution of job experience years provided by respondents. It shows that the most common number of years is around 10 with around 180 respondents. This is reflected by mean of this variable. The data ranges from 0 to 30 years of experience. Distribution is right-tailed.



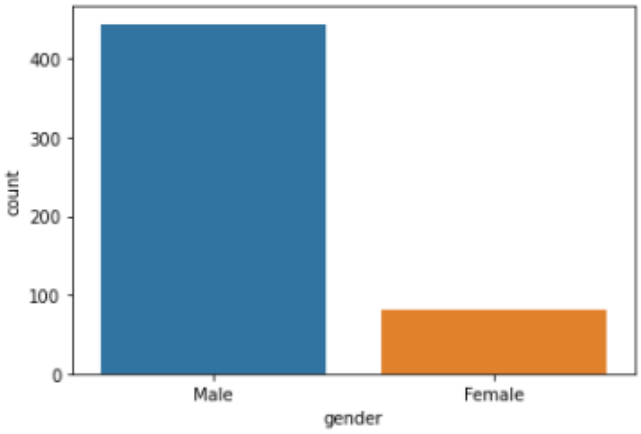
Yearly brutto salary in 2019 has a median of 70 000 euros. The data ranges between around 25 000 to 200 000 euros. Data distribution is normal.



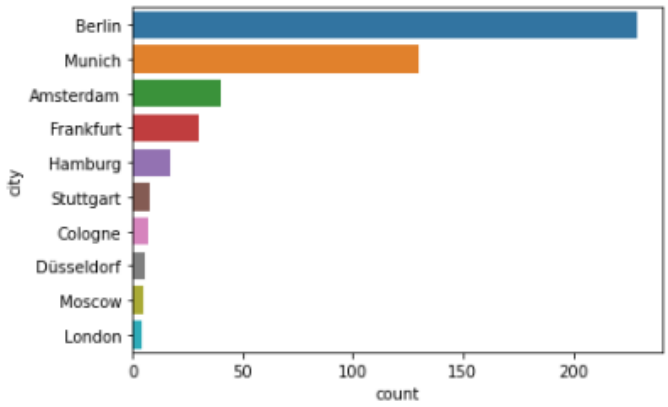
Yearly brutto salary in 2018 has a median of 65 000 euros. The data ranges between 0 to 200 000 euros. The changes to distribution of salary from 2018 to 2019 are probably an outcome of a group of respondents who have 1 year of experience and didn't work in 2018. Data distribution is normal.



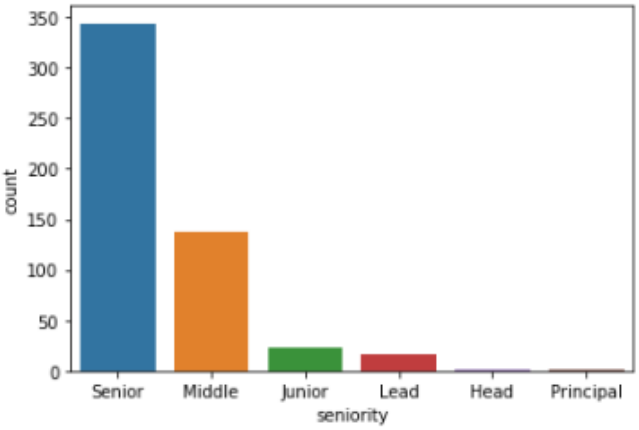
Distribution of main technology used by respondents at work shows, that most of them uses Java, which is closely followed by a group of Python users. Least number of respondents used C/C++, .NET, AWS, Kotlin and R.



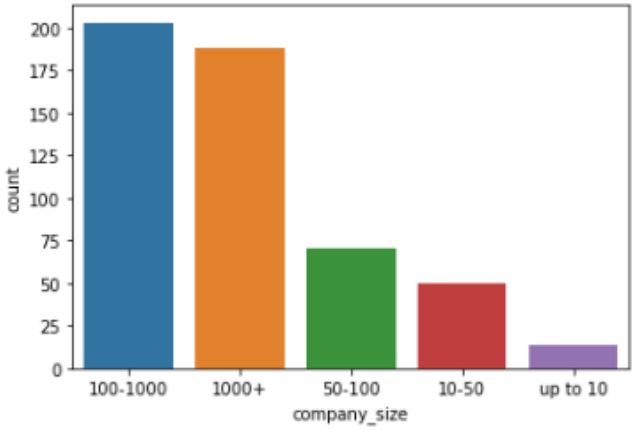
Gender of respondents show, that most of them are males. There is around 350 male respondents more than female respondents. There is around 100 female respondents and around 400 male respondents.



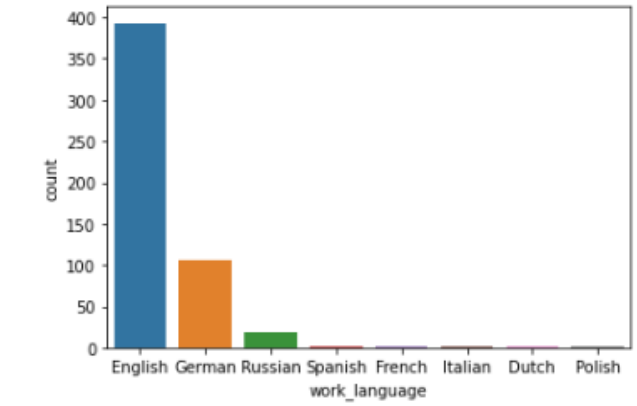
Top 10 City distribution shows that most of respondents works in Berlin with second large group of respondents working in Munich. Amsterdam, Frankfurt and Hamburg are in the middle of a list. Least number of respondents works in London, Moscow, Düsseldorf, Cologne and Stuttgart.



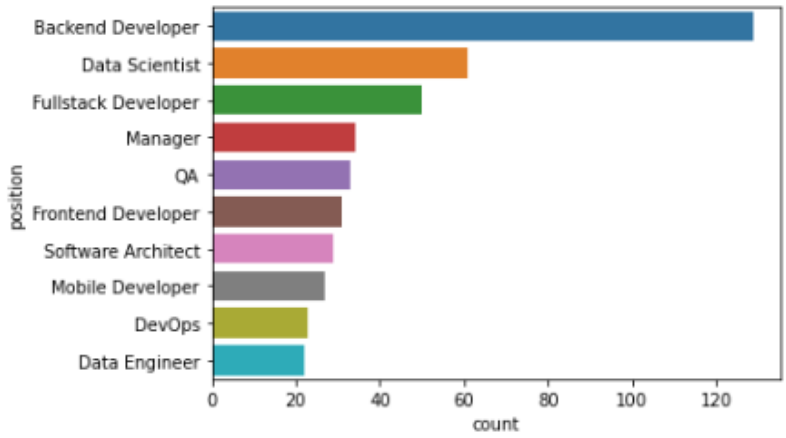
Most respondents are seniors. Distribution runner-up are a group of middle positions, which is followed by juniors and team leaders. Least number of respondents are head and principals.



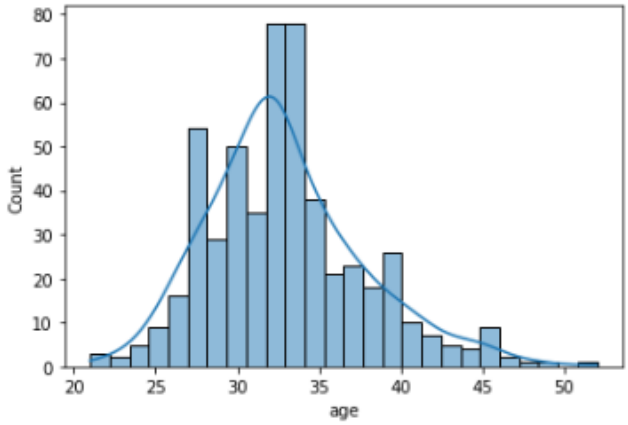
Most respondents represented companies within 100-1000 and 1000+ people employed range. The least number of respondents works at small companies with 10-100 employees.



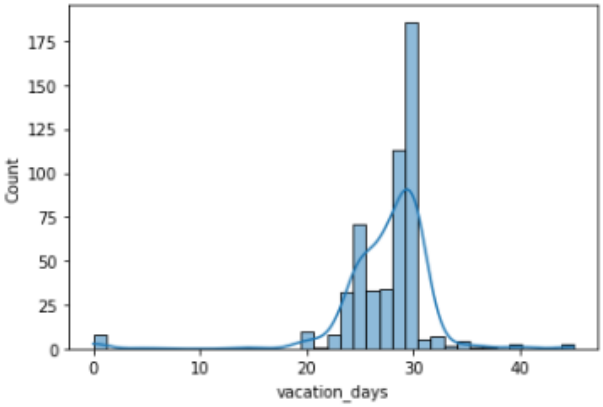
The top 3 languages that employees speak in is English, German and Russian. Small group of respondents speak Spanish, French, Italian, Dutch and Polish.



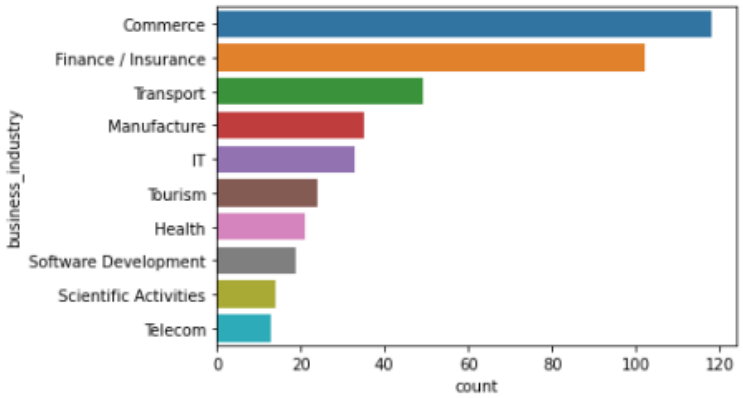
Largest group of respondents are Backend Developers who are followed in numbers by Data Scientists and Fullstack Developers. Least number of people are Data Engineers, DevOps, Mobile Developers.



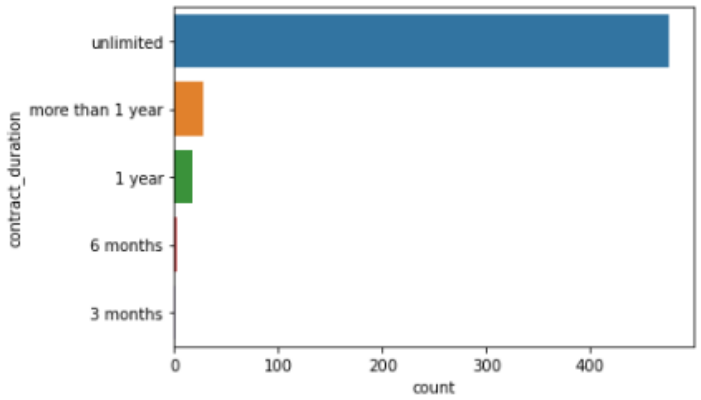
Age is distributed normally with central tendency, median equal to 32 years of age. Workers age ranges from around 20 to 50. Lots of people are in their 20s. There is less respondents in ages between 35 and 50 compared to younger people.



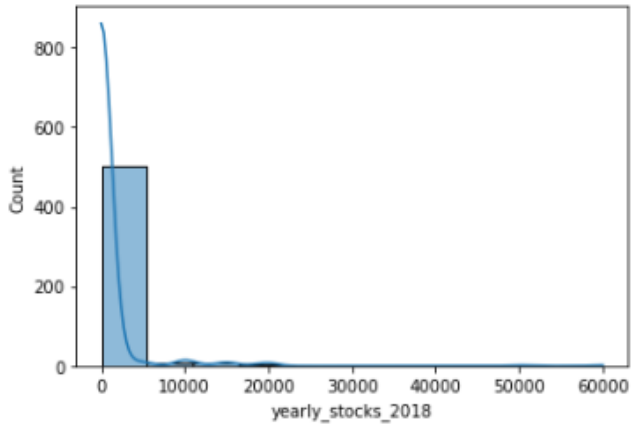
Central tendency – median of vacation days in a year is equal to 28. Data ranges from 0 to around 50 days. There is a large group of respondents (Around 275 people) who have around 30 days off in a year. Distribution shows, that most people fall in range of 20 – 30 days off a year.



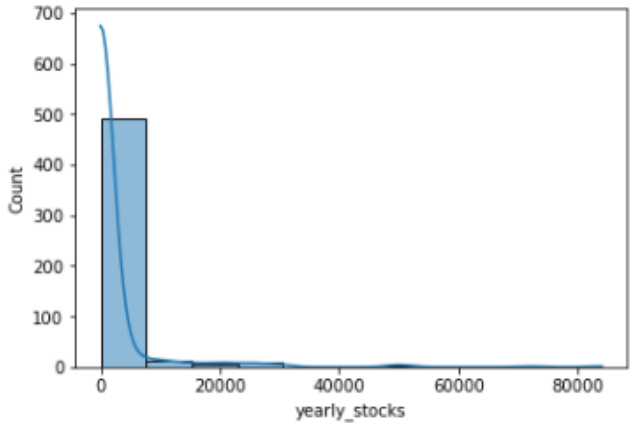
Count plot of respondents belonging to specific business industries shows, that most people are employed in Commerce, Finances / Insurance and Transport. Least amount of people is employed in Telecommunication, Scientific field and Software Development.



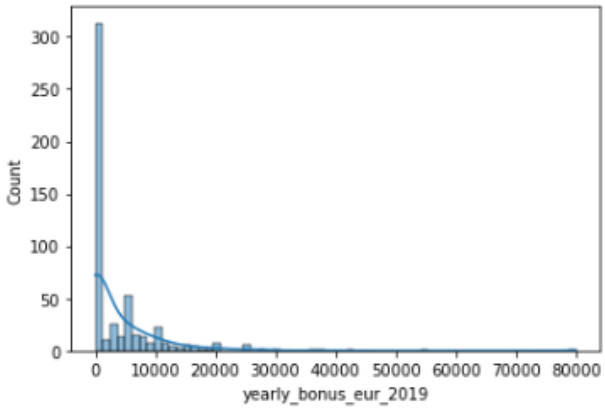
A lot of respondents are on a unlimited contract. Minority has limited time of contract (1 year+, 1 year, 6 months). There is 1 person with 3 months contract.



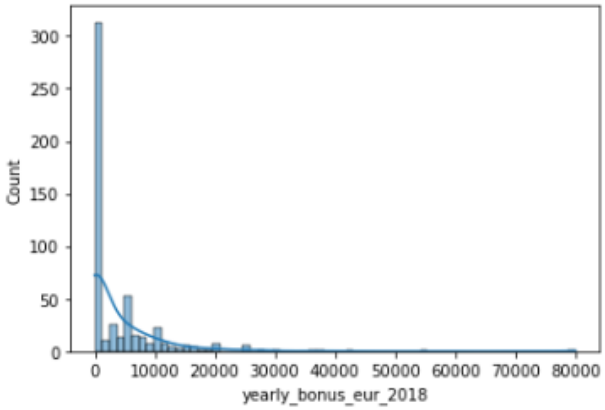
Yearly stocks from 2018 is skewed right. Central tendency – median is equal to 0. Lots of people didn't have stocks.



Yearly stocks from 2019 is skewed right. Central tendency – median is equal to 0. Lots of people didn't have stocks.

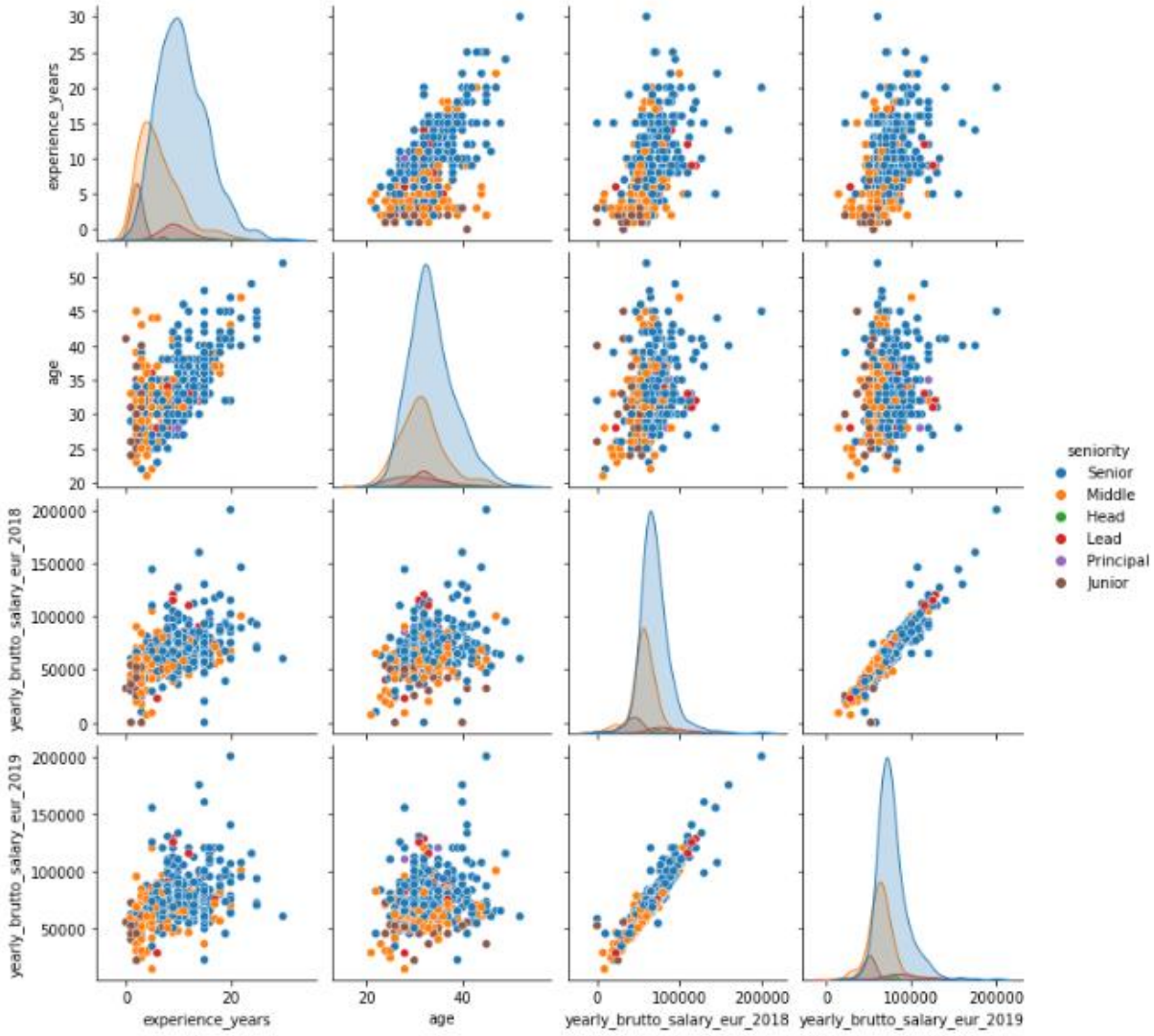


Yearly bonus from 2019 is skewed right. Central tendency – median is equal to 0. Lots of people didn't have bonuses. Minority had received bonuses in 2019.



Yearly bonus from 2018 is skewed right. Central tendency – median is equal to 0. Lots of people didn't have bonuses. Minority had received bonuses in 2018.

Correlation



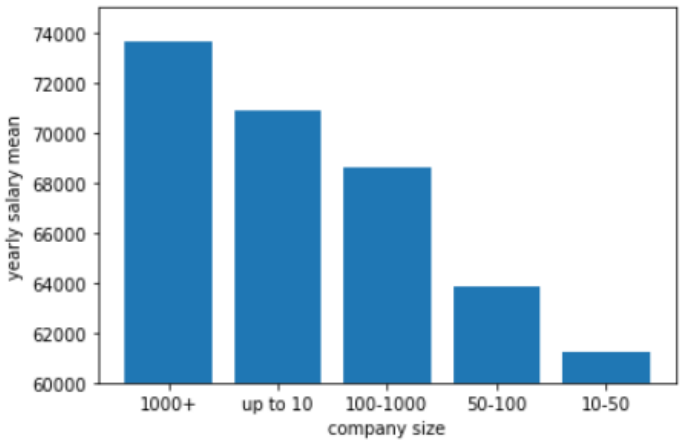
I've chosen variables which have the strongest relationships with each other based on correlation matrix and plotted them as is seen above. Those variables represent experience years, age of respondents, yearly salaries in 2018 and 2019. I've classified the data points based on seniority level. Conclusions from relationships between variables:

- The older a person is the more experience a person tends to have
- Older people tend to have higher salary than youngsters
- Yearly salary is somewhat dependent on experience years and age.

Business insight

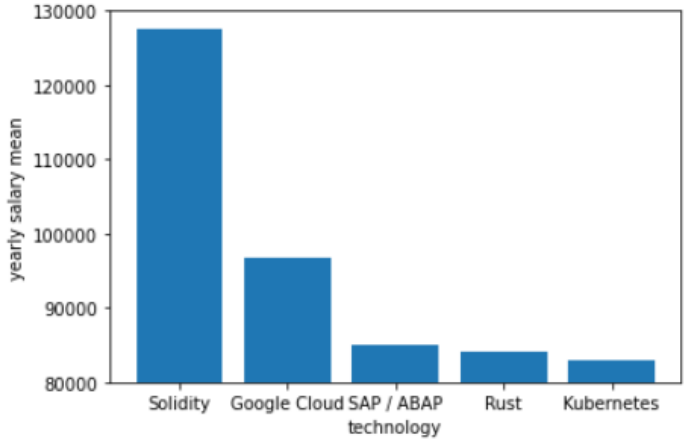
I’ve categorized yearly salary mean (sum of 2018 and 2019 salaries by employee was averaged) and additional income (averaged sum of bonus and stocks) by categories of some variables. Data provided below can be used to draw useful information about characteristics of employees and companies, that foster the best salary.

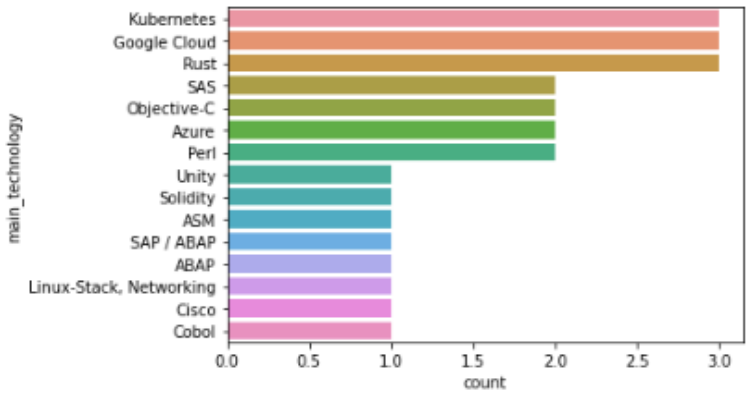
company_size	yearly_salary_mean
1000+	73662.646277
up to 10	70896.428571
100-1000	68631.591133
50-100	63850.714286
10-50	61252.250000



The greater company size, the bigger salary tends to be. Companies up to 10 employees are exception to this statement as they have second largest yearly salary mean. This may be accounted to successes of some companies like this. These type of firms can have a lot of money to divide between small amount of employees.

main_technology	yearly_salary_mean
Solidity	127500.000000
Google Cloud	96666.666667
SAP / ABAP	85000.000000
Rust	84000.000000
Kubernetes	82833.333333



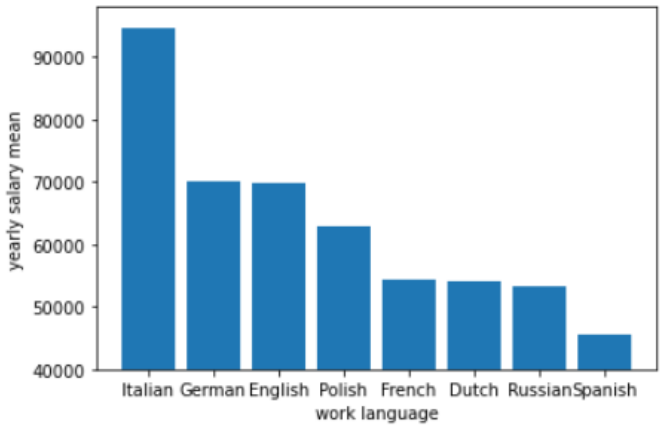


Technologies with highest yearly salary mean showed in a table have been made for advanced purposes. There is small amount of people, that uses them. Small supply for such skills is a reason for high salaries for people who know them. Count plot above shows, that technologies earning most are used by small amount of people (ex. Solidity, highest yearly salary mean is used by 1 respondent).

```
In [306]: median_top / median_top_excluded # median of top 5
Out[306]: 1.2708898267648037
```

Yearly salary median of highest earning technologies is 27% higher than median of the rest technologies excluding the top 5.

work_language	yearly_salary_mean
Italian	94500.000000
German	70022.551402
English	69710.270992
Polish	63000.000000
French	54250.000000
Dutch	54000.000000
Russian	53164.083333
Spanish	45500.000000



Italian is a language with highest yearly salary mean. Runner ups are German and English with approx. 70000 euros of yearly salary mean. Languages with the lowest salary mean up to 55000 euros yearly are French, Dutch, Russian and Spanish. Polish found itself behind English and ahead of French.

	city	work_language	yearly_brutto_salary_eur_2018	yearly_brutto_salary_eur_2019	yearly_salary_mean
242	Milan	Italian	95000	115000	105000.0
335	Lugano	Italian	84000	84000	84000.0

```
In [166]: df['city'].unique()

Out[166]: array(['Berlin', 'Frankfurt', 'Munich', 'Hamburg', 'Leipzig', 'Nuremberg',
                 'Cologne', 'Krakow', 'Stuttgart', 'London', 'Karlsruhe', 'Bern',
                 'Düsseldorf', 'Kyiv', 'Amsterdam', 'Pforzheim', 'Kassel', 'Madrid',
                 'Vienna', 'Moscow', 'Warsaw', 'Hannover', 'Milan', 'Odesa', 'Cork',
                 'Heidelberg', 'Bielefeld', 'Dublin', 'Jyvaskyla', 'Toulouse',
                 'Dubai', 'Lingen', 'Dresden', 'Lugano', 'Schleswig-Holstein',
                 'Kaiserslautern', 'Saint Petersburg', 'Leeuwarden', 'Eindhoven',
                 'Hilversum', 'Gdańsk', 'Wroclaw', 'Limassol', 'Würzburg', 'Bremen',
                 'Lausanne', 'Stockholm', 'Rotterdam', 'Minsk', 'Utrecht', 'Kiev'],
                dtype=object)
```

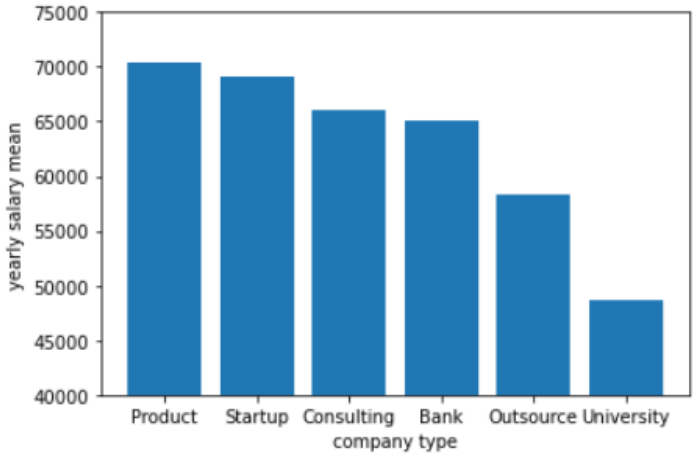
```
In [169]: len(df[df['city'] == ('Berlin' or 'Frankfurt' or 'Munich' or 'Hamburg' or 'Leipzig' or 'Nuremberg' or 'Cologne'
                                or 'Stuttgart' or 'Karlsruhe' or
                               'Bern' or 'Düsseldorf' or 'Pforzheim'
                                 or 'Kassel' or 'Vienna' or'Hannover' or 'Heidelberg'
                                or 'Bielefeld' or 'Lingen' or 'Dresden' or 'Schleswig-Holstein'
                                or 'Kaiserslautern' or 'Würzburg' or 'Bremen')])

Out[169]: 229
```

```
In [158]: len(df[df['work_language'] == 'German'])

Out[158]: 107
```

```
In [159]: len(df[df['work_language'] == 'English'])

Out[159]: 393
```

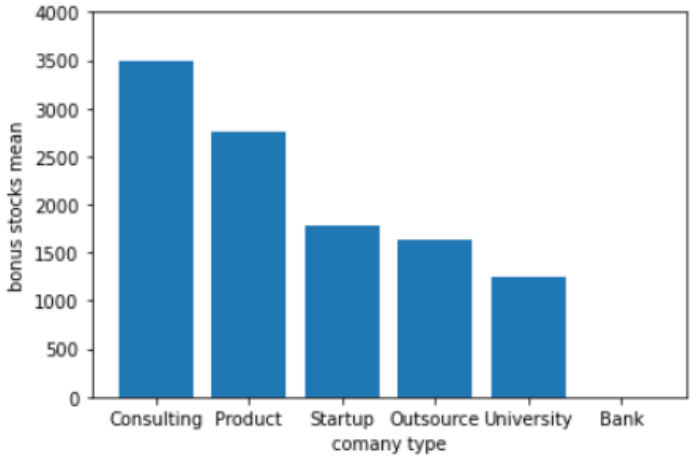
Italian is on top of the earning list because in a survey there are two people speaking Italian who work in Italy and have high salaries, which is not very representative group of respondents. English and German are more representative as there is 107 people speaking German and 393 people speaking English. 500 people out of 525 (95% of respondents) is speaking English and German which may mean that those languages are most desirable to know in IT workspace in Central Europe. This shows that there is only 25 respondents (5%) who speak at work in other languages. High score for German language may be caused by huge number of respondents from German speaking cities (229 people, 44% of respondents). This means, that there is 271 respondents who speak English at work (51%).

company_type	yearly_salary_mean
Product	70348.466387
Startup	69080.113636
Consulting / Agency	65968.341667
Bank	65000.000000
Bodyshop / Outsource	58270.833333
University	48675.000000



Yearly salary mean is highest in Product companies and Startups with around 70 000 euros. Second are Consulting and Banking companies with around 65 000 euros. At the back of the grid there are Outsourcing companies and Universities. In Outsourcing yearly salary mean is around 58 000 euros. Universities yearly salary mean is 48 000 euros.

company_type	bonus_stocks_mean
Consulting / Agency	3485.495875
Product	2749.113725
Startup	1787.224716
Bodyshop / Outsource	1637.500000
University	1250.312500
Bank	0.000000




```
df[['bonus_stocks_mean', 'company_type', 'position']][df['company_type'] == 'Bank']
```

	bonus_stocks_mean	company_type	position
371	0.0	Bank	Backend Developer
410	0.0	Bank	Modelling Specialist
430	0.0	Bank	Backend Developer
525	0.0	Bank	Security Engineer

Consulting employees have highest additional income mean (stocks and bonuses) which is around 3500 euro. Second are employees in Product companies with around 2800 euro mean. Then are Startups, Outsourcing and Universities in a range 1000 – 2000 euros range. It is interesting that bank employees don't have any bonuses and stocks. However this is a small of a group to conclude that banks generally don't give any bonuses or stocks to their employees

```
In [295]: np.mean(outsource_university_excluded['yearly_salary_mean']) / np.mean(university_outsource['yearly_salary_mean'])
Out[295]: 1.3504581613679227
```

Mean of yearly salaries of Product, Consulting and Startups companies is 35% higher than Outsourcing and University yearly salary mean.