Data Science
Project Document

Team 15

Name	Section	BN
Maryam Salah	2	21
Nadeen Ayman	2	31
Noran Hany	2	34
Hala Hamdy	2	35

Table of contents

Team Member Contributions

Name	Workload
Maryam Salah	Cleaning + Q6+Q10+Q11
Nadeen Ayman	Cleaning +Q7+Q8+Q9
Noran Hany	Cleaning + Q4 +Q5+Q6
Hala Hamdy	Cleaning +Q1,2+Q3

Business Value

IT jobs are incredibly important in today's world as technology continues to play an increasingly crucial role in almost every industry. Here are some reasons why IT jobs are so important:

- Cybersecurity: They are responsible for implementing security measures and ensuring that systems are secure and up-to-date
- Data analytics: With the growing importance of data in decision-making, IT professionals who specialize in data analytics are becoming increasingly valuable. They are responsible for analyzing data and providing insights that can inform business strategy.
- IT jobs are among the fastest-growing occupations, and the demand for IT professionals is expected to continue to grow in the coming years

So It's important to Collect data about IT salaries so that it can provide valuable insights into the job market and help both employers and employees make informed decisions.

Here are some reasons why data collection on IT salaries is important:

1- Salary benchmarking: Employers can use salary data to benchmark their compensation packages against industry standards. This can help them attract and retain top talent by offering competitive salaries

2- Negotiating salaries: Employees can use salary data to negotiate better salaries when they are offered a job or during performance reviews. Having access to data can help them understand what a fair salary range is for their position and experience

3- Identifying trends: Collecting data over time can help identify trends in IT salaries, such as how salaries are changing in response to changes in the economy or the job market. This information can be useful for both employers and employees when making decisions about compensation.

4- Addressing pay equity: Salary data can help identify pay disparities and potential biases in the hiring and compensation process. This information can be used to address pay equity issues and ensure that all employees are paid fairly for their work.

Dataset

An anonymous salary survey has been traditionally conducted annually since 2015 among European IT specialists with a stronger focus on Germany. The purpose of this survey is to learn the competitive value of a skill set for IT specialists depending on years of experience, position, language, etc.

With that in mind , we were given access to 3 consecutive years (2018-2020) in which we have combined the 3 datasets & merged them into one. We have noted down how inconsistent the survey from one year to another. Not only that, but also we have noticed that the dataself has various diverse responses that includes several inconsistencies that we had to deal with first. Combining them all together we have ~ 3000 records.

Data Cleaning

Company Type Cleaning

Normalization

Method	Demonstration
Remove Trailing and Leading spaces	"Bank " → "Bank"
Remove dashes -	E-commerce → Ecommerce
Lowercase all strings	Consulting, Consulting → consulting Ecommerce, eCommerce → ecommerce
Group several consulting categories into one generic consultation type.	Technology Consulting, IT Consultancy, Consulting Company, IT Consultants → consultancy
Group types that contain the word "commerc" into one type named "ecommerce"	Big commercial, ecommerce firm → ecommerce
Group types that contain "research" or "institute" into one type named "research"	Research institute, Research → research
Group types that contain "corborat" into one type named "corborate"	corporate incubator, corporation → corporate

Filter out company types that have occurred less than 5 times.	The instances of such occurrences are of very low frequency, making it unreliable to draw meaningful insights from the analysis based on them. Before filtering: 2918 responders After filtering: 2839 responders By filtering out these specific company types, we observe that we did not lose a significant amount of information.

By enforcing the above mentioned techniques, the number of unique company types was reduced from 102 to 72

Clustering

By visually inspecting and analyzing the 75 different company types, we aim to cluster and group them into more generalized categories. This approach allows us to reduce the overall number of company types while increasing the frequency of occurrence for each category. Rather than simply removing sparse company types with low frequencies, we utilize human judgment to create more meaningful and consolidated clusters. For instance, the words media, publishing and technology and publisher was placed under the category: Media/Publishing.

Doing so, the number of unique company types was reduced from 72 to 40

Seniority Level Cleaning

Spell checking & correction

- By manually checking how many unique values we got from this column, we found there are around 25 different values

which is much more than the expected and the standard (Head-Senior-Mid level-Junior).

- We found that many were mis-spelled and others were not standard levels that maps to one of the standard ones e.g (entry-level = Junior)
- Removing not relevant values like : (No level , no idea, there are no ranges in the firm)
- The final number of categories became 4 (Head-Senior-Mid level-Junior).

Salary Cleaning

No currency standardization or conversion was needed , as all are in Euros.

Outliers removal

After visualizing mean/median/max/min/std of salary, it was clear that there nonsense outliers that affected the salary statistics greatly, As the results were :

Mean salary: 300307.1538352403
Median salary: 70000.0
Standard deviation of salary: 10695071.571778344
Max salary: 500000000.0
Min salary: 6000.0

Value of mean is the indicator that there are outliers that make it extremely high.

After removal the values became more logical :

Mean salary: 69135.70384615385
Median salary: 68375.0
Standard deviation of salary: 15808.153332636677
Max salary: 110000
Min salary: 30000

Years of Experience cleaning

It was expected to have a large number of unique values as it's a continuous value, but not expected to have text. So we started to extract numbers in text e.g('6 (not as a data scientist, but as a lab scientist)) to result in only floating numbers ranging from 0 -40 years of experience .

Contract duration cleaning

The main categories in construct duration we found were ['Unlimited contract' 'Temporary contract' nan 'unlimited' '6 months' 'more than 1 year' '1 year' '3 months']

We only focused on mapping 'Unlimited contract' & 'unlimited' to the same thing

Company Size Cleaning

Due to the inconsistencies in the survey itself from year to year, we figured slightly different ranges by only 1 person difference for the company size which was unified to be all the same as shown:

```
df[company_size]=df[company_size]
        .replace('11-50','1 0-50')
        .replace('51-100','50-100')
        .replace('101-1000','100-1000')
```

Technology Cleaning

To be able to unify all the technologies altogether we have figured several inconsistencies which includes {aggregated values , different lower & upper case letter , leading & trailing spaces, misspelling}

To solve these issues we lowered all the strings , stripped to remove the spaces , then we repeated the same record for the different technologies he had to deal with the aggregate values. To address consistency and misspelling issues, a string matching solution was applied. This was done through the fuzzy wuzzy library which has different tools to use such as best_x and ratio. It uses Levenshtein Distance to calculate the differences between sequences in a simple-to-use package.

City Cleaning

With the same challenges we faced in the technology column we have suffered from similar problems here adding to the unicode (Zürich -> Zurich) which was solved by unidecode. Also , some cities are just not misspelled but written in a different language so replacements have been done as follows:

```
corrections = {
    'Koln': 'Cologne',
    'Duesseldorf': 'Dusseldorf',
    'Dusseldurf': 'Dusseldorf',
    'Saint petersburg': 'Saint-petersburg',
    'Tampere (finland)':' Tampere',
    'Kiev':'Kyiv',
    'Konstanz area': 'Konstanz',
    'Munchen': 'Munich',
    'Nurnberg': 'Nuremberg',
    }
```

Position Cleaning

- Firstly , we put the most common positions in a list [software engineer , backend developer, frontend developer, software architect, fullstack developer,mobile developer,devops,designer(ui/ux), data scientist, ml

engineer', qa engineer, qa, ios developer, product manager, researcher, security engineer, software tester]

- Then we use two ways for cleaning we use Fuzzy library to map all positions to similar ones in the list, then we make a manually cleaning by looking to the unmapped position and replace them with a suitable position in the list , and finally we make `Other` position, this position for any position that can't be mapped manually or by fuzzy library.

Questions

Q1: What is the most used technology in the mid sized company?

Q2: Can we say that the most popular technology for mid sized companies is the most popular for big sized companies ?

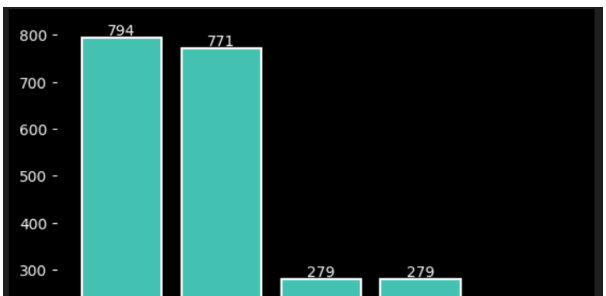
State & Refine QS

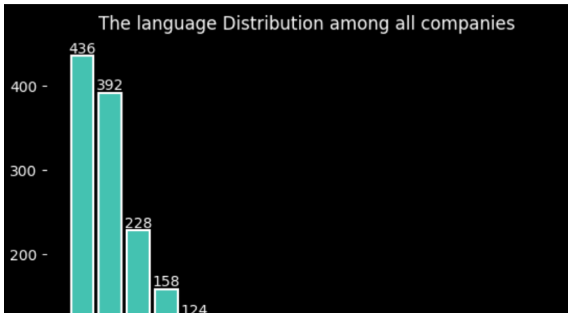
QS: What is the most used technology in Software Engineering for a mid-sized level company?

QS: Can we say the same technology is the most popular for a big-sized company or not?

Set Expectation	Collect Info	Match Results
The Quetsion is Answerable & of interest & Novel	Did some reasearch online and poeple always ask what language do we need to learn and Which coding languages are used by the world's top companies?	*Matches* because people may be interested in this question as the language can influence their job opportunities and career growth.Knowing these languages could potentially increase a programmer's chances of being hired by these companies Such af FAANG companies

Explore Data



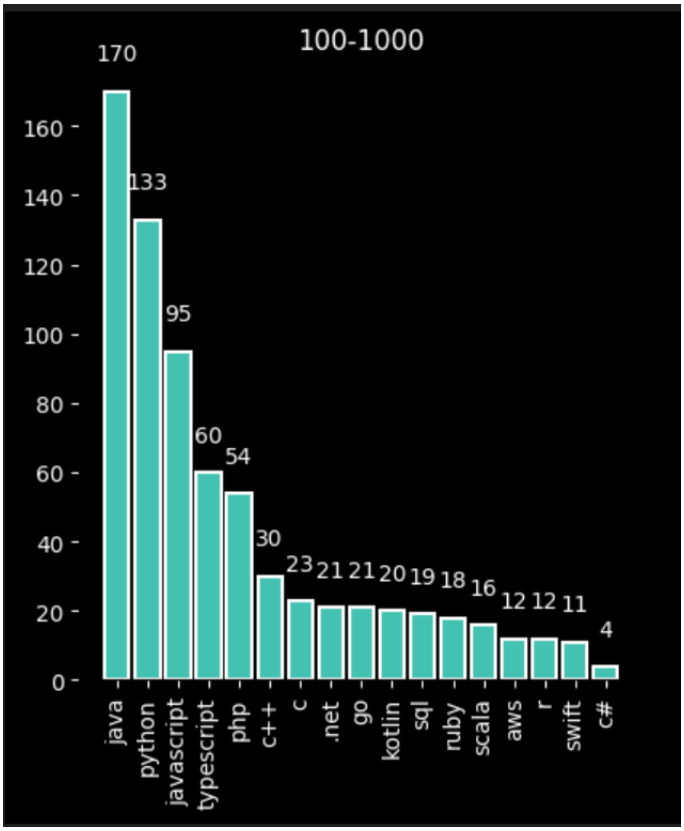


Interpretations:

- *We can see that generally python is the most used technology among the rest of the technologies regardless of the company size*
- *Followd by Java which is still a very big peak*

Interpretations:

- *We can see that the big Sized Companies (1000+) have the highest % in the dataset*
- *Followed by Mid Sized Companies (100-1000)*
- *The number of people in each category of Bigsized and MidSized are quite close to each other*



Set Expectation	Collect Info	Match Results
Python to be the most used langauge in all company sizes	Showed Several plots	*DIDnot Match* looking back at the distribution of langauages used. Yes although python is the most used, still java is frequently used by a minor difference between them

Build Models

Z_test for proportion

- *We need to proof that "Java" which is the most used technology in Midsized companies , is the most used in Medium sized companies or not?*
- *So typically to perform such test, we will need to perform "z test for propotion" for EVERY technology other than "Java"*

$H_o : \%Java <= \%technology$

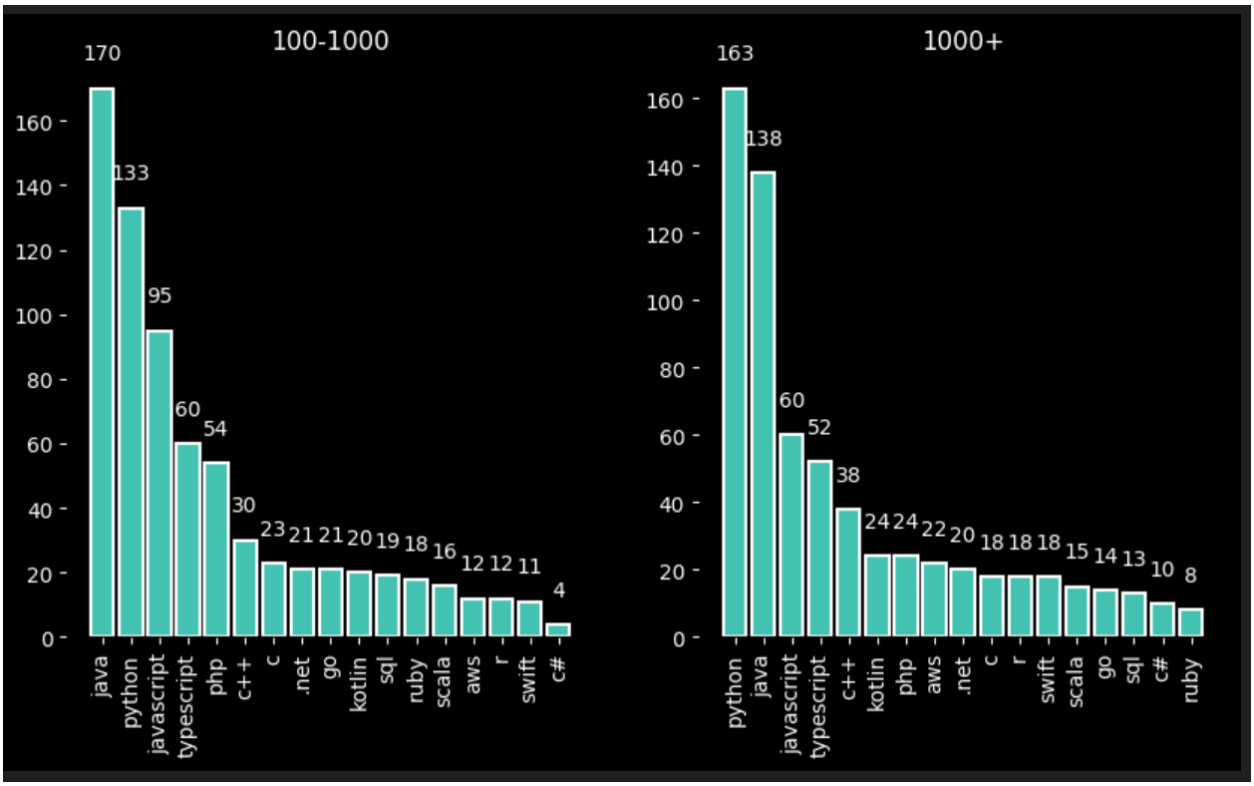
where technology is all other technologies in Big sized companies other than Java

$H_a : \%Java > \%technology$

Ztest for proportion

- *Assumes that we have atleast 10 records to perform it*
- *Here it is right tailed test , so we must have z_values >0*
- *the restured Pvalue should be divide by 2*

	typescript	php	javascript	python	c#	scala	go	kotlin	swift	c++	.net	r	sql	c	aws
0	True	True	True	False	True	True	True	True	True	True	True	True	True	True	True



Set Expectation	Collect Info	Match Results
Java to be the most used langauge in all Big sized companies	Did z hypothesis testing for proprotion	*DIDnot Match* looking back at the distribution of langauages , we found out that Python is the most used language in Big sized companies

Interpret Results

Interpretations:

- *Java is the most used technology in the midsized companies exculding python*
- *It MAYBE the case that java < python in midsized companies but we doesnot have evidence to suggest that over the population*

Communicate Results

Suggestions:

- *It seems that Java & python are in a very competitive case. Although no certain case that shows that one language is more dominant than the other , it is still a good potential if you are a new learner and want to join the field to start with one of these 2 language.* -So we are just narrowing down the space of searching to only 2 language & you might be condifent that one of these is top in Big sized companies if you wish to join the big companies

---

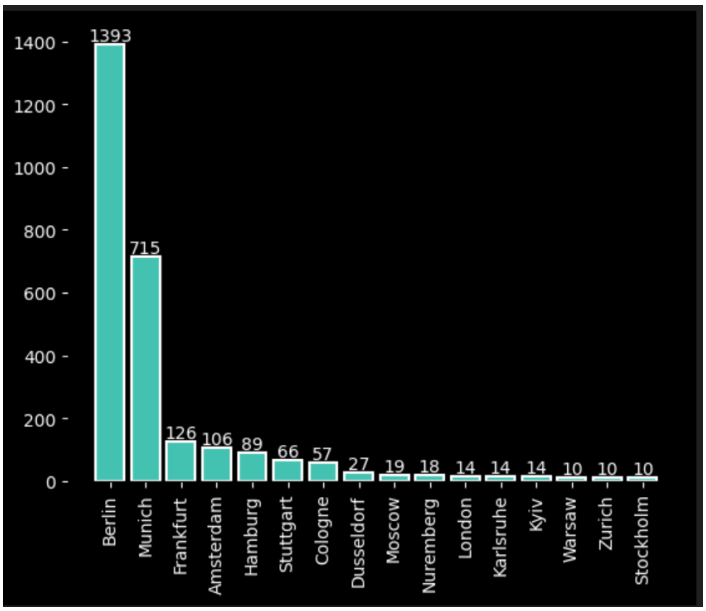
Q3: How does the salary for the same position differ from city to another?

State & Refine QS

QS:How the salary for the same position varies from city to another

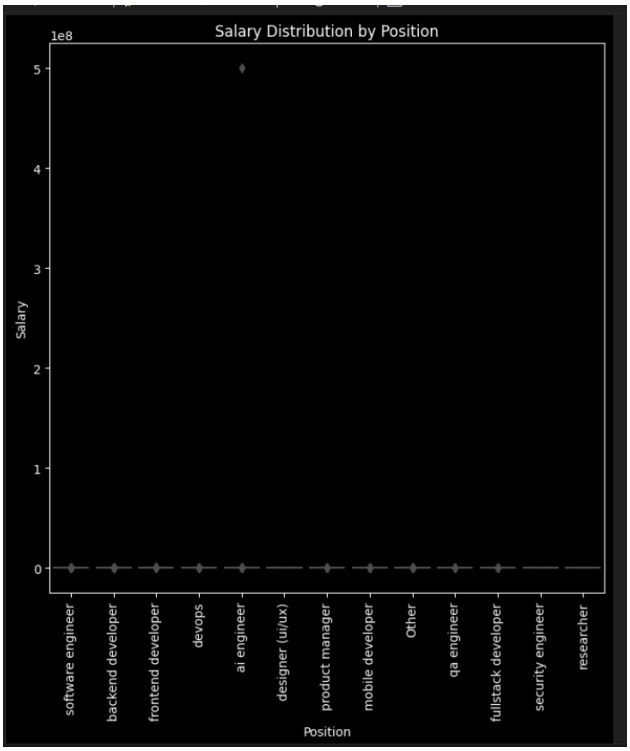
Set Expectation	Collect Info	Match Results
There exists different cities in EU for a valid question	See the Columns in our dataset	*Matches* as there are different Cities & positions in our dataset & the salary is provided

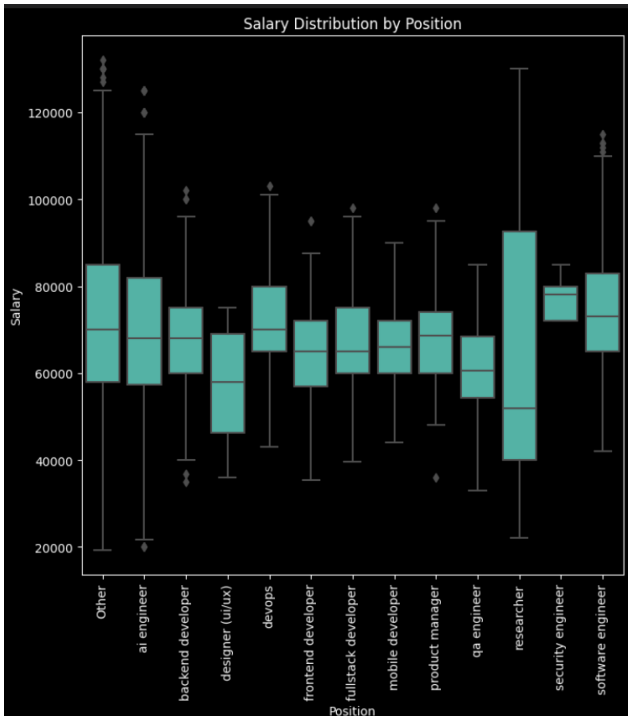
Explore Data



Interpretations:

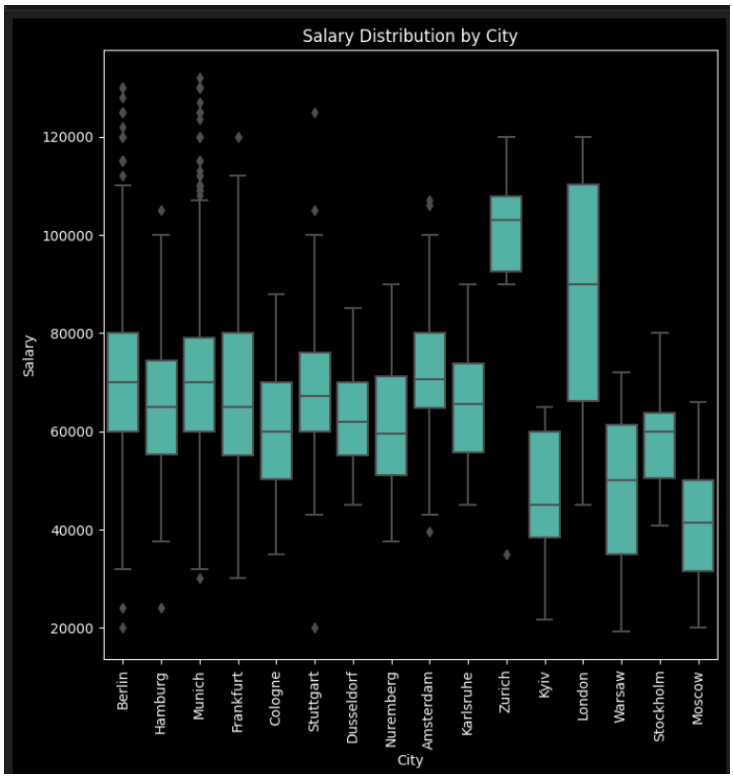
- *Most of the cities lie in Germany as states in the dataset collection*
- *After filtering , all the cities are in EU continetnt*
- *Berlin , which is the capital of Germany, has the most records which does make sense*
- *Some cities ,which has very minor count,could be of less use especially after grouping by the position*





Interpretations:

- *It seems that there is a ckear outlier is salary in ai engineer, so it was removed by IQR*
- *It seems that most of the positions have median salary mostly between 60-80k*
- *The Researcher seemed to come less than this predefined common median working arounf 50k*
- _Some positions such as {ai Engineer ,researcher} have high variance showing diverse payment in the position in general_

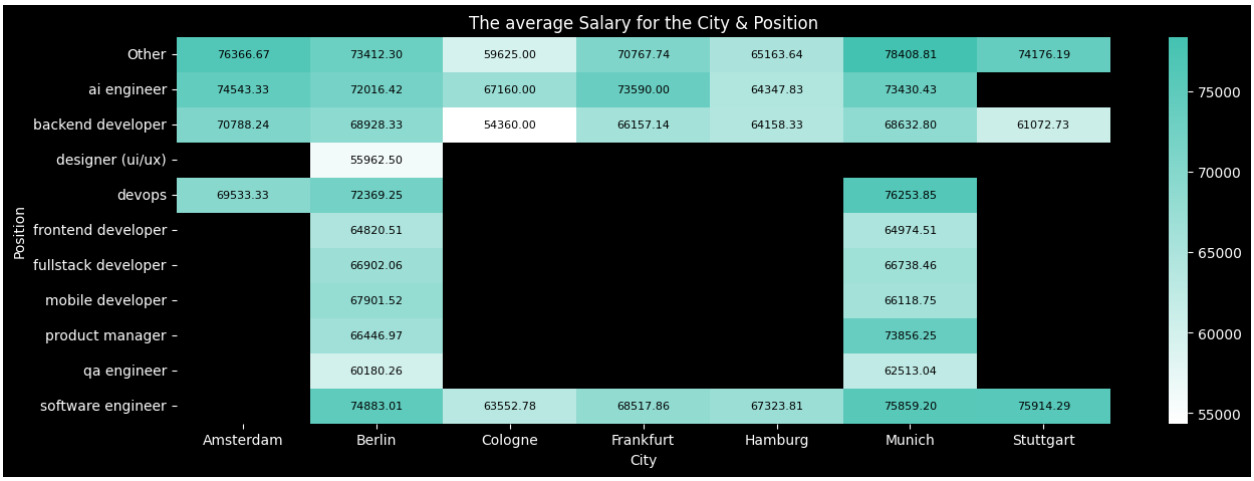


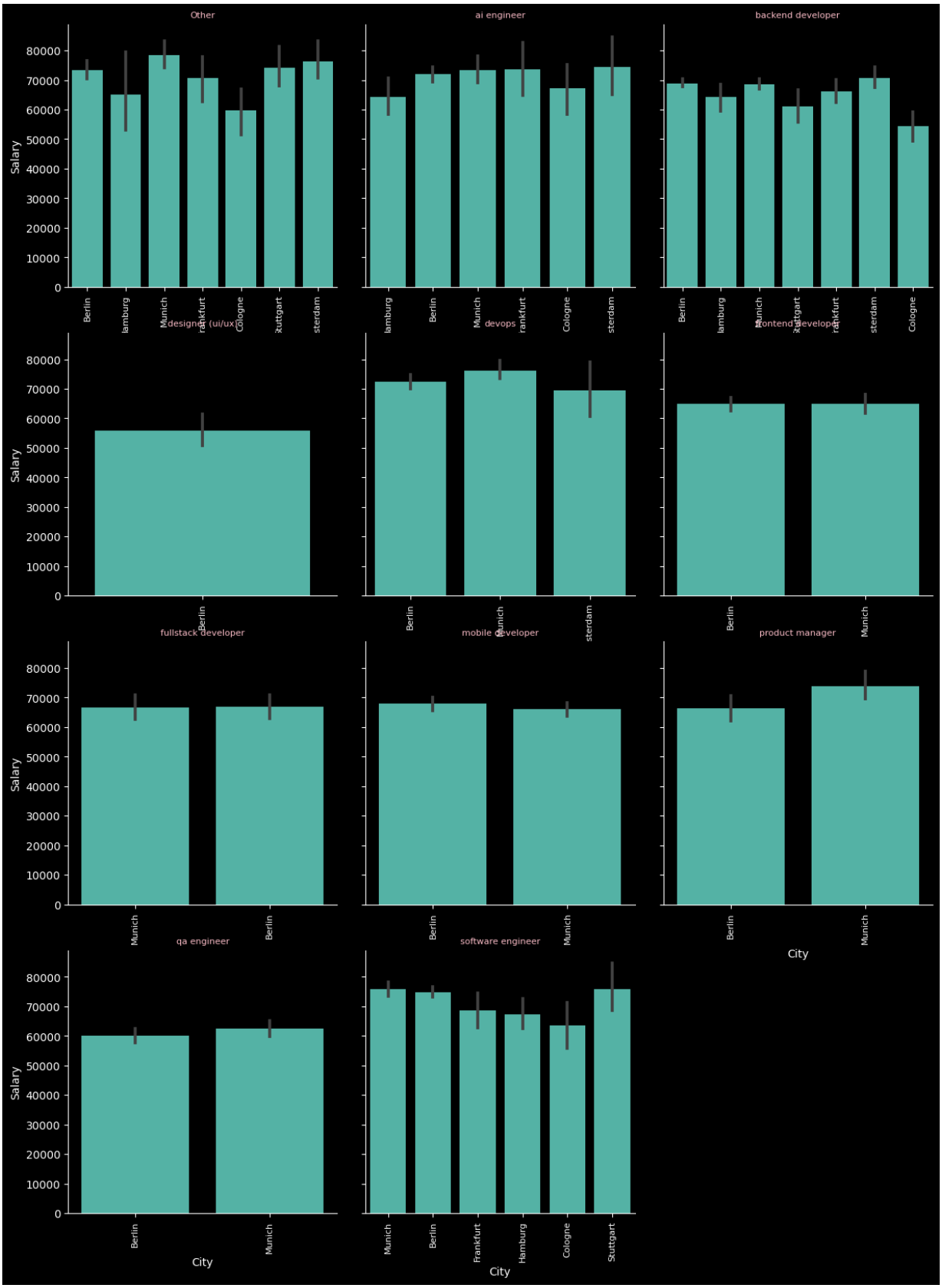
Interpretations:

- *As seen from graph , zurich & London seems to have the highest median salaries*
- *With Zurich the highest, it does make sense as it is well known how it has high Living standards and with that comes high paying salaries*
- *Meanwhile Moscow & kyiv seems to have the least Salary*

Interpretations:

- *It is seen that the frequency after grouping the position & city is niche in some combinations e.g London seems to have 2,6,1,3 people in its available positions which would reflect a non statistcal significnce for the enitre population*
- *We can see that that happened for cities that has a very low frequency in the 1st histogram we drew. E.g: Zurich and London were below 20 people*
- *Should drop those that will cause statistically insignificant results*




Other
ai engineer
backend developer
(ui/ux) designer
devops
frontend developer
fullstack developer
mobile developer
product manager
qa engineer
software engineer
Salary
City
Berlin
Hamburg
Munich
Frankfurt
Cologne
Stuttgart

Interpretations:

- *It seems that For some positions , indeed, the salary varies from a city to another For example:*
  - *Software engineer seemed to be payed highest in Stuggart*
  - *Backend developers seemed to vary fhe most between Amesterdam & Cologne*
- *However the barcharts shows that the difference is very minor*
- *Now it is time to prove that statistically*

City

Set Expectation	Collect Info	Match Results
Enough Cities with frequent records in each city for the question to be answerable	Several plots about the City Column	*Matches* the expectations after dropping certain niche values this is due to the variety in cities that were provided

Build Models

$\eta^2 = \frac{\sigma_{\bar{y}}^2}{\sigma_y^2}$ where:

- η^2 : correlation ratio
- $\sigma_{\bar{y}}^2$: is standard deviation between the groups e.g {Munich, Berlin..etc}
- σ_y^2 : is the standrd deviation within the group {salary in Berlin itself has a lot of variance} + between groups

In other simple words it is : $correlationRatio = \frac{betweenGroups}{betweenGroups+withinGroup}$

* The bigger the correlation ratio ~1 the more we can say that the variation is due to difference in difference between groups.In our case it will be due the differenece in salary between cities such as Munich & Berlin ..

* The smaller the correlation ratio the more we can say it is difference within the groups. In our case, the Variation in salaries in the city itself is big rather than having significant difference between the cities

Interpretations:

- *Overall it seems that there is weak relation between the city & the salary, So mostly all the cities for the same position are quite close to each other in terms of salary*
- *Designer having correlation coeff=0 means that there is NO difference in the mean of the salary between the different cities. That was expected because as we see that these 3 positions have only 1 city so the difference is all due to within_group term*
- *Product Manager seems to have the highest correlation among them all , which was seen from the heatmap & the histogram graph it had the highest value of difference*

Set Expectation	Collect Info	Match Results
Finding minor variation in the payment/salary for the same position from city to another	Used the correlation ratio to give me the degree of between_groups variation	*Match* When inspecting we can say that ALL the cities except Amesterdam are in Germany and All in the EU region which could have a close salary for the same position. If it was in different continent maybe it would be a significant result. After searching I can say that the difference is very insignificant For example, Frankfurt am Main accounts for the largest average gross salary at €66,529, followed by Stuttgart (€66,174) and Munich (€65,164) . So it is supported by the Grid of histograms better as it shows the minor difference

Interpret Results

Interpretations:

- *There is a very minor difference in the payment salary in Eu Region (Mainly in Germany) for the same position*

Communicate Results

Summary:

- *Who ever is interested in relocation in Germany and is confused where to head to , the salary factor should not be considered as the main decison upon which he/she build his/her decison*

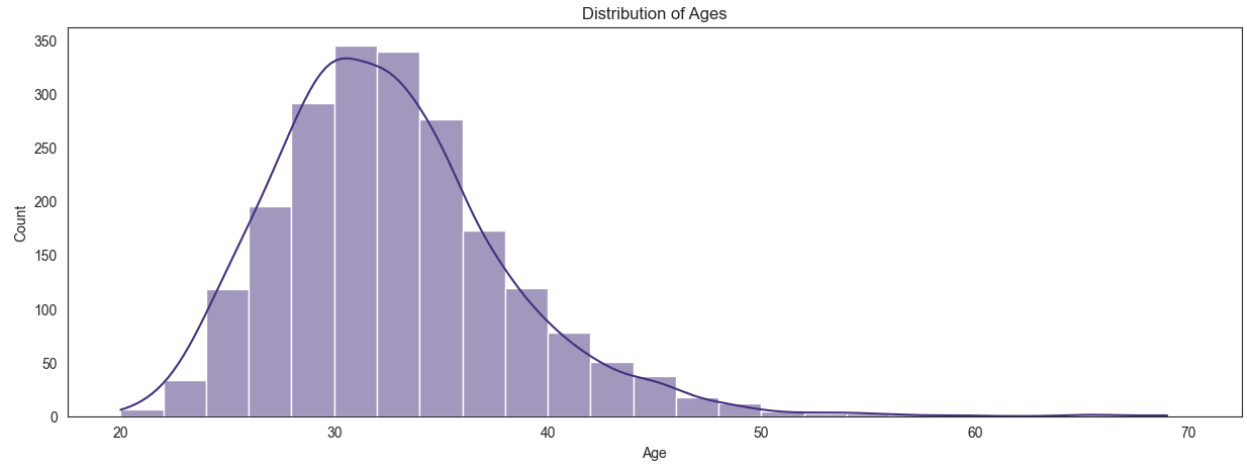
Q4: What is the most preferred language/technology for each age segment?

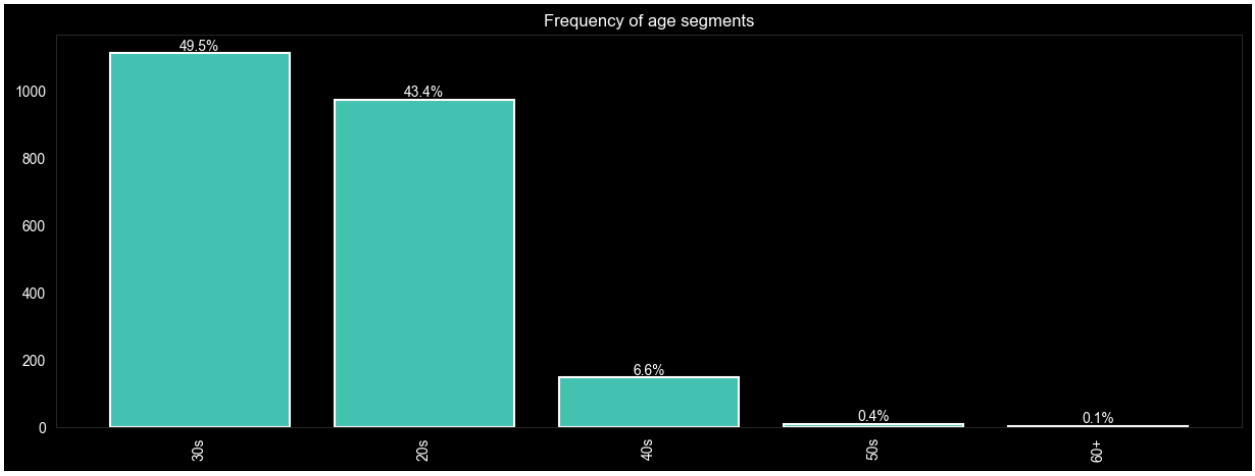
State & Refine QS

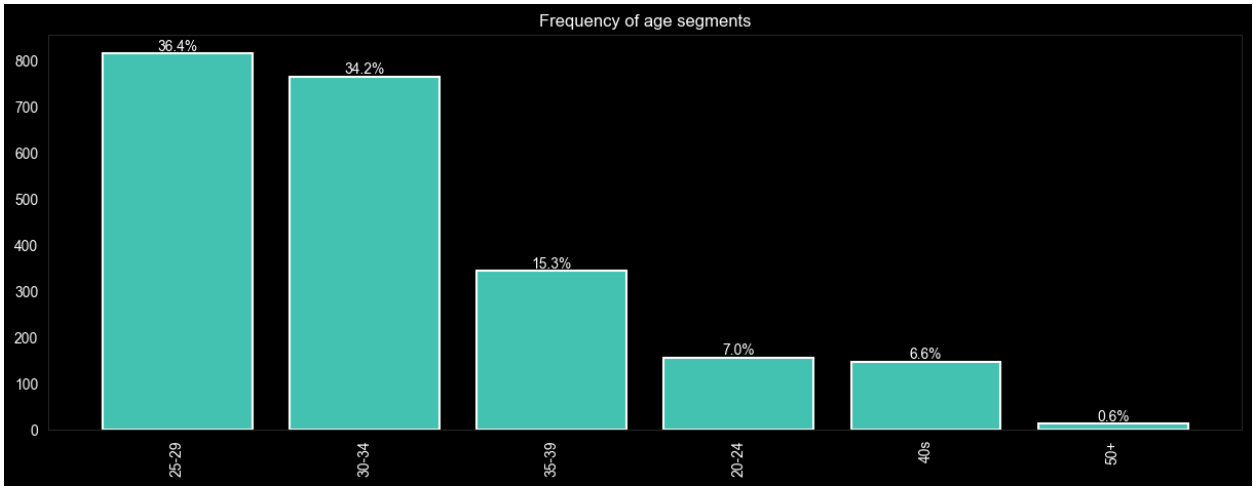
QS: What is the most preferred language/technology for each age segment?

Set Expectation	Collect Info	Match Results
Question is answerbale, there's a suffiecient number of rows having both age and technology as not nons.	Only 136 rows out of 2244 rows for the age is missing. 141 rows out of 2244 rows for the technology is missing.	*Matches* There are a few rows that have missing values for both age and technology, which indicates that we still have sufficient data available to answer the question.

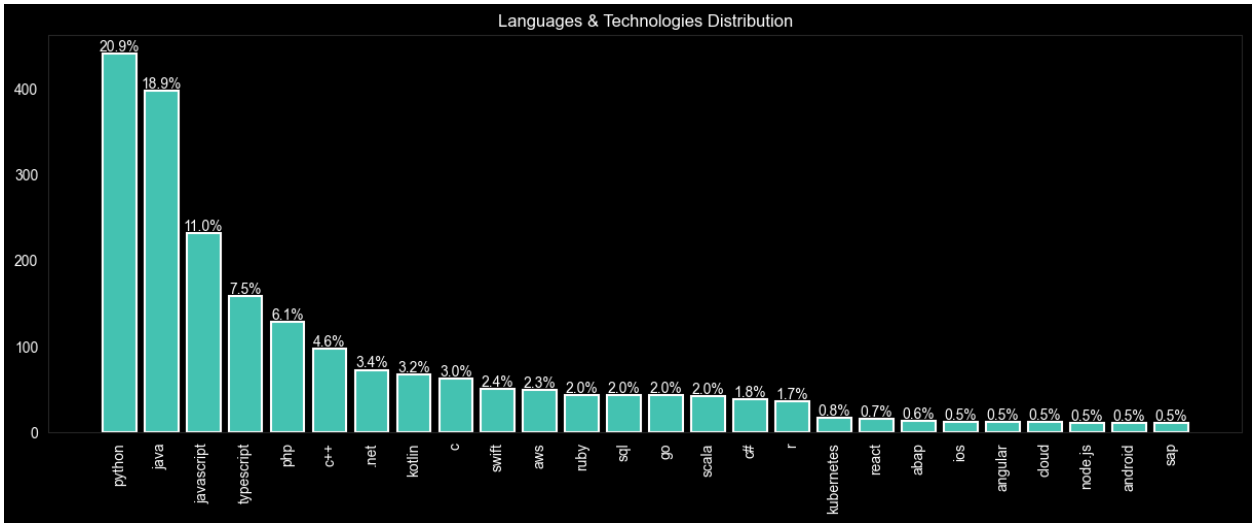
Explore Data





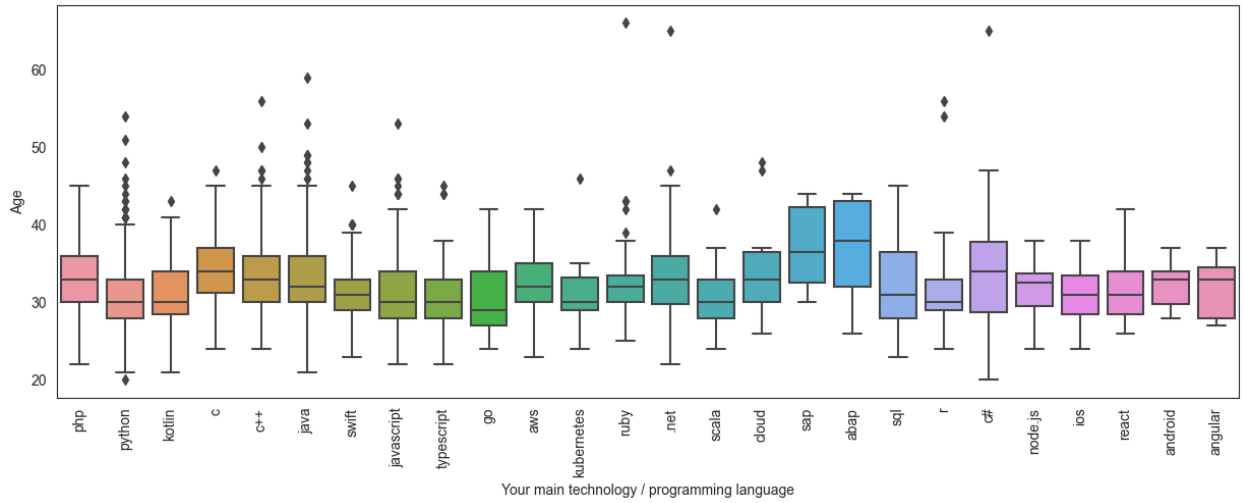


Expectation	Collected	Match?
The question is feasible to answer and the data exhibits a balanced distribution across the age groups of 20s, 30s, 40s. Therefore, we can analyze and provide insights for each age group individually.	92.8% of the responders are within their 20s and 30s, only 6.6% are at their 40s.	The 40s age segment is not represented suffieciently. It seems that people at their 40s don't engage much in such surveys. To address this issue and ensure more reliable analysis, we will divide the survey responders into more specific age segments by grouping them into intervals of five years each. This will allow us to have more groups to study the question on. ❌
Suffiecent rows for employees in their 50s	Only 0.4% of the respondents fall into the age group of 50s.	This low percentage collected makes sense, because employees in their 50s may not be as familiar with or inclined to fill out online surveys. ❌
People above 60 years are more likely to be retired.	Retirment age in germany is 65 (and 7 months)	*Matches*
There's should be no teen-ager or child because this is an IT salary, and people below 20 very few of them are hired in the IT industry.	Youngest age was 20, no year below.	*Matches*



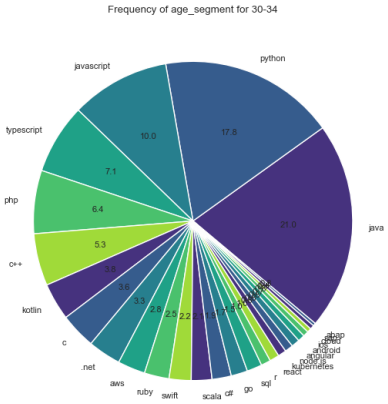
Expectation	Collected	Match?
Based on a study conducted in May 2019 regarding the most popular coding languages among unicorns, the distribution of languages in the survey is expected to align with the findings of the study. The study can be found at this link: https://flyaps.com/blog/top-10-coding-languages-used-by-global-companies/	Computing a histogram of the langauges filled by survey responders	*Matches*

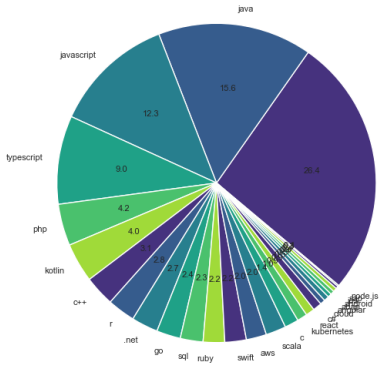
Study link:
https://flyaps.com/blog/top-10-coding-languages-used-by-global-companies/

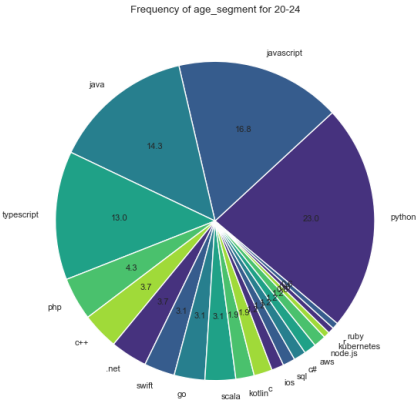


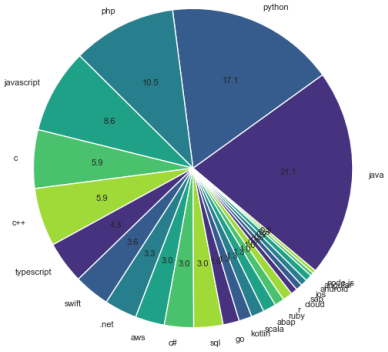
Interpretations:

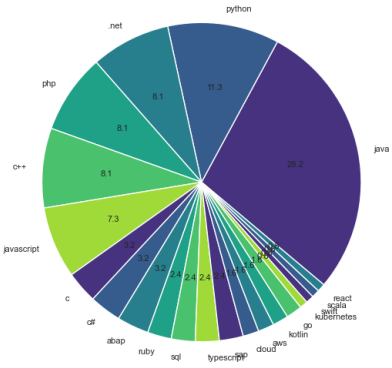
- *The median of most technologies lies within the 30s, this is obvious since most survey repsonders are at their 30s*
- *The IQR for SAP and ABAP is around the 35-40s, this makes sense since SAP was started in 1972 and SAP ABAP was created in 1983 for the development of business applications in the SAP environment. Thus more elder employees knows and uses SAP ABAP*

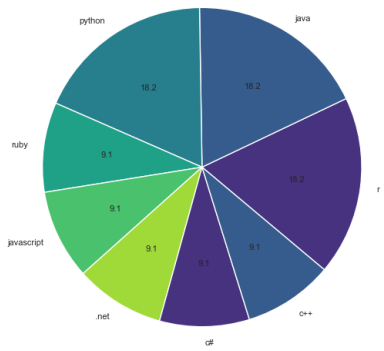












Interpretations:

- *Python and Java are popular among all age segments.*
- *Type script is popular among 20s, less popular in the ages from 30-35, and Even very less popular for 35+, this makes sense since typescript was released in 2012. Thus eldery age won't be much familiar with it.*
- *Based on the data provided, we can observe that the usage of PHP and .Net varies across different age groups. The usage of PHP shows a pattern of 8.1% in the 40s, 10.5% in the 35-39 age group, 6.4% in the 30-34 age group, 4.3% in the 25-29 age group, and 4.2% in the 20-24 age group. Similarly, the usage of .Net exhibits a trend of 8.1% in the 40s, 3.3% in both the 35-39 and 30-34 age groups, 2.7% in the 25-29 age group, and 3.7% in the 20-24 age group. From this analysis, it can be concluded that PHP and .Net are relatively more popular among individuals in their late 30s and 40s compared to other age segments.*

Build Models

	age_segment	Most Popular Technology
0	20-24	[python]
1	25-29	[python]
2	30-34	[java]
3	35-39	[java]
4	40s	[java]
5	50+	[java, python, r]

Interpret Results

Interpretations:

- *PHP, Python, etc existed with a significant amount within each group segment, due to the popularity and different usages of such langauages.*
- *Analyzing the usage of TypeScript, a language that emerged relatively recently, we observe that it is more prevalent among older individuals who may be less up-to-date with newer languages. This suggests that older individuals might be inclined to stick with older languages they are familiar with, rather than adopting newer technologies.*
- *This observation can be supported by the notion that younger individuals are generally more adaptable and open to learning new technologies. Their flexibility and ability to quickly grasp new concepts may make them more inclined to explore and embrace emerging languages.*
- *The above analysis demonstrates that the distribution of langauges within each age segment differs, however, the **most popular langauage** within each age segment doesn't depend on a specific age segment. The most popular are nearly the same across all age segments.*

Communicate Results

- *Python and Java emerge as the most popular technologies based on the analysis. For individuals venturing into the IT field, these languages serve as excellent starting points due to their high demand and widespread usage.*
- *The most popular langauage that you are encouraged to start learning in the field will not differ according to your age very much. They most popular are nearly the same across all age segments.*

Q5: Is there a relation between the age of employee and the business type of the company he/she is joining?

State & Refine QS

QS: How does age affect the type of company that a person is joining? In other words, do startups tend to have people with a smaller age?

Set Expectation	Collect Info	Match Results
Question is answerable, there exist enough company types to answer it.	There is a variety of 102 company types	*Matches*
Younger people are more likely to work in startups. People above 35 are more likely to work for big sized corporates	63.5% and 20.04% of the company types are `product` and `startup`, respectivley. Only 10 corporate records. However, it is important to note that the `product` based companies does not necessarily indicate the size of the company. Instead, it typically refers to a company that primarily focuses on developing and manufacturing tangible goods or physical products, irrespective of its scale.	❌ Our initial expectation does not align with the data we have collected. Therefore, we need to modify our question from focusing on the company size to focusing on the company business type, to incorporate the insights gained from the data.

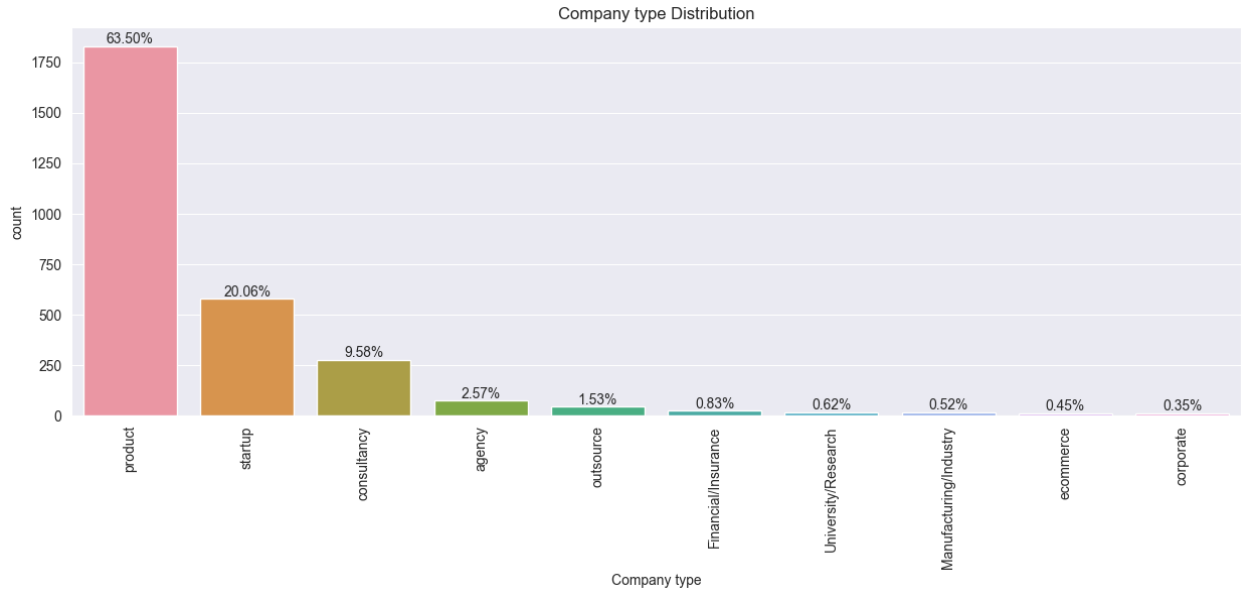
New QS: Is there a relation between the age of employee and the type of company he/she is joining?

Explore Data

Set Expectation	Collect Info	Match Results
The company type is divided into a few categories, such as `Product` or `Startup`.	There is a variety of 102 company types	❌ This is because it's a survey, where people either choose or write in a text field the type of the company they work at. In 2018 and 2020, there are many variations in the company types filled by survey responders, including "Technology Consulting," "IT Consultants," "IT Consulting," and more. Since the column primarily focuses on the company type rather than the specific business sector, it is possible to group several categories into one generic company type. For example, any type of company related to consulting can be considered as a generic consulting company. Also, group the types with the same stem together, and unify the capitalization of the text.

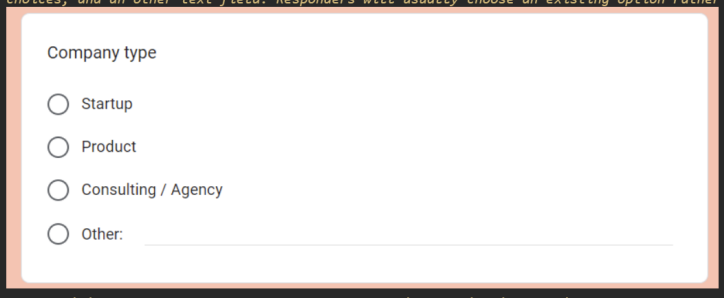
Univariate Analysis: Company type Distribution

After applying Cleaning to the Company type as mentioned in the clean effort section.



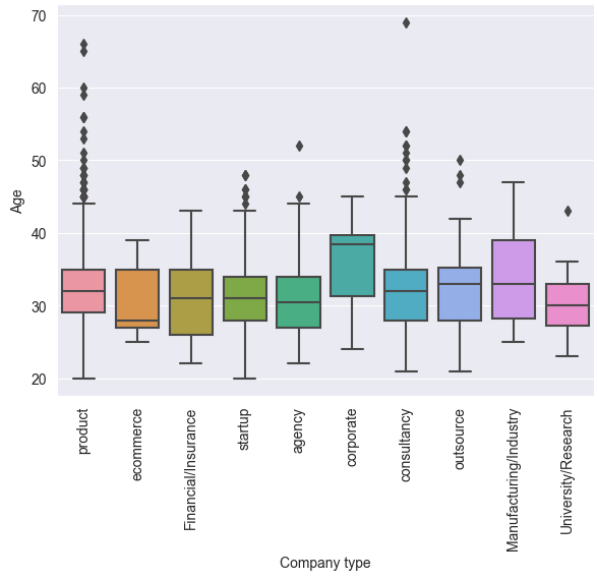
Interpretations:

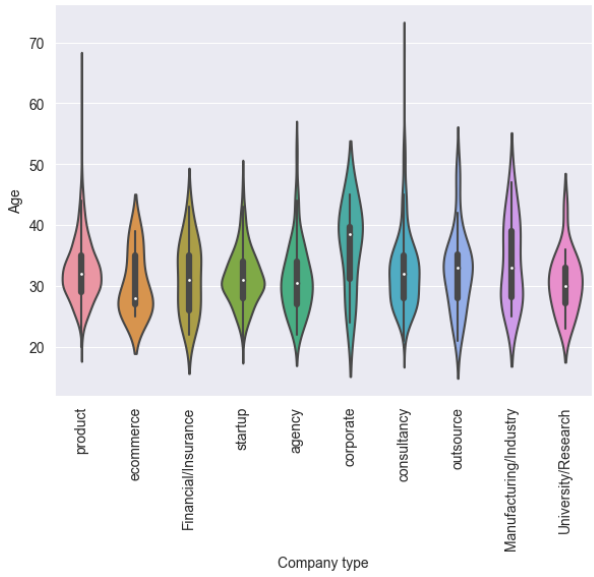
- *Product company type is dominating.*
- *Most Dominating company types are prodcut, startup, consultancy and agency, this is beacuse the survey the data was collected through have them as choices, and an other text field. Responders will usually choose an existing option rather than writing in the others text field.*



- *The remaining company types have very low frequencies, making it unreliable to conduct analysis using them. Therefore, we will filter out company types that have fewer than 10 respondents.*

Bivariate Analysis: Distribution of Age within each Company Business Type

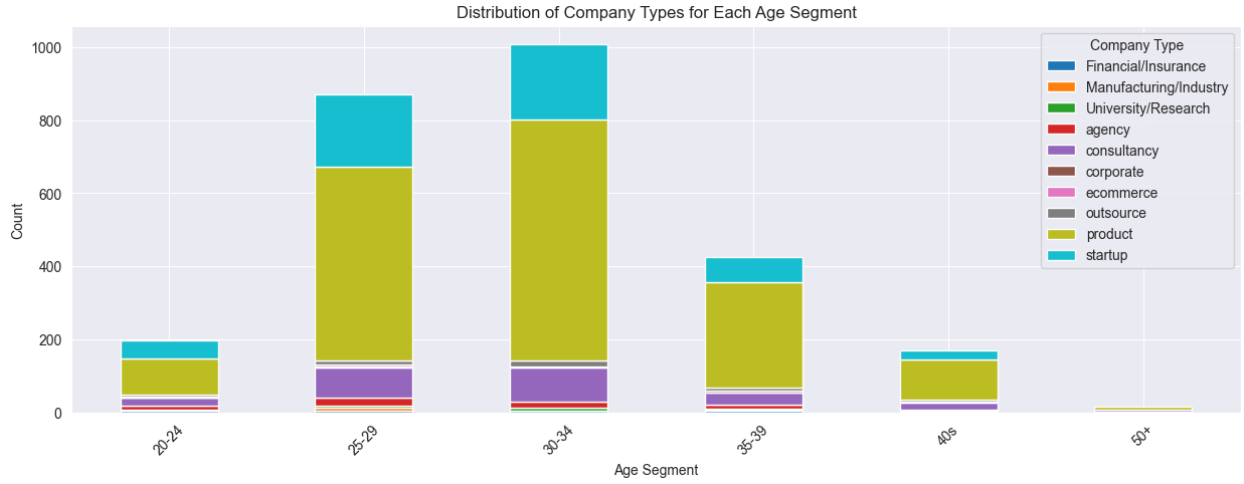




Interpretations:

- *Most company types have most of their employees around the age of 30 because the majority of this survey's responders are around the age of 30.*
- *In the consultancy company type, we observe the presence of outliers or higher age values. This can be attributed to the fact that older employees, who have accumulated significant experience over the years, are more likely to work as consultants and provide valuable advice within a consultancy company.*
- *Most people working in ecommerce are below 30 years unlike other company types. This observation can be attributed to the fact that younger individuals are more familiar with technology, online platforms, and the digital landscape. Given the nature of ecommerce businesses that heavily rely on online transactions and digital marketing, younger employees may possess the necessary skills, adaptability, and understanding of modern consumer behavior, making them well-suited for this industry.*
- *Most people who works in corporate, companies characterized by their hierarchical organizational structure, are around 40 years*

Distribution of Company Types for Each Age Segment



- The distribution of different types of companies within each age group appears to be similar to the overall distribution of company types across all age groups. This suggests that there may not be a noticeable relationship between age and the type of company. In order to further validate the absence of this relationship, a correlation ratio will be calculated.

Build Models

To determine does changing a company type leds to change in the age of its employees, a correlation ratio will be calculated.

$\eta^2 = \frac{\sigma^2_{\bar{y}}}{\sigma^2_y}$ where:

- η^2 : correlation ratio
- $\sigma^2_{\bar{y}}$: is standard deviation between the groups e.g {Startup, Product, etc.}
- σ^2_y : is the standrd deviation within the group + between groups

In other simple words it is : $correlationRatio = \frac{betweenGroups}{betweenGroups + withinGroup}$

Correlation ratio = 0.112495001968

- *Since the correlation ratio is smaller than 1, it indicates that there is a significant variation in age within each company type separately. This implies that company types do not have a specific age range associated with them.*
- *Age is not a strong determinant or defining factor for the type of company a person is associated with. The distribution of company types is similar across different age groups, indicating that people of various ages can be found in each type of company.*

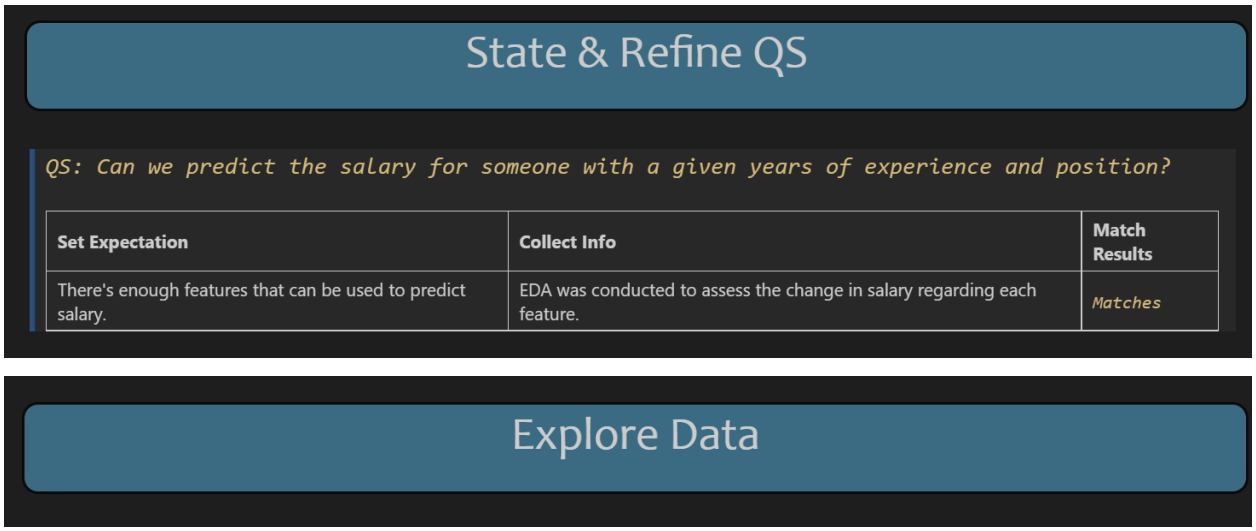
Interpret Results

- *Both the correlation ratio and the visualization provide compelling evidence that there is no discernible link between the type of company and the age of its employees, contradicting our initial expectation.*
- *After conducting thorough analysis and seeking guidance from a senior data scientist, it became apparent that the absence of this relationship can be attributed to deliberate efforts by companies to cultivate age diversity within their workforce. By actively hiring individuals from different age groups, companies aim to create a well-rounded blend of experience and fresh perspectives.*
- *Maintaining a diverse range of ages among their employees not only facilitates the transfer of knowledge and expertise but also enables companies to mitigate the potential impact of losing older employees. This is achieved by ensuring a pool of younger individuals who can be mentored and educated, thereby ensuring continuity and the preservation of institutional knowledge.*

Communicate Results

- *If you are an IT company, regardless of your business type (whether it's a startup, product-based, or service-based), it is highly recommended to consider hiring employees from diverse age groups. This approach promotes the transfer of knowledge and expertise within your organization, fostering a dynamic environment that benefits from a well-rounded blend of experience and fresh perspectives. This will also mitigate the potential impact of losing older employees, through ensuring a pool of younger individuals who can be mentored and educated, thereby ensuring continuity and the preservation of institutional knowledge.*
- *As an individual seeking to enter the field of IT, you have the opportunity to join any company, irrespective of its business type. The specific business type is unlikely to significantly affect the acceptance of your job application. Therefore, you can explore opportunities in various IT companies without being restricted by the specific nature of their business.*

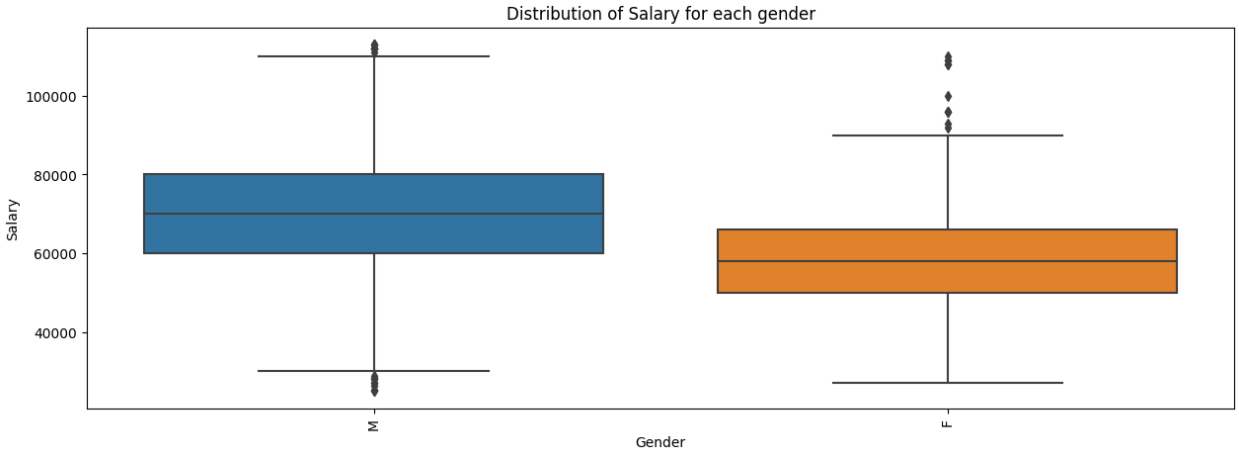
Q6: Can we predict the salary for someone with a given years of experience and position?



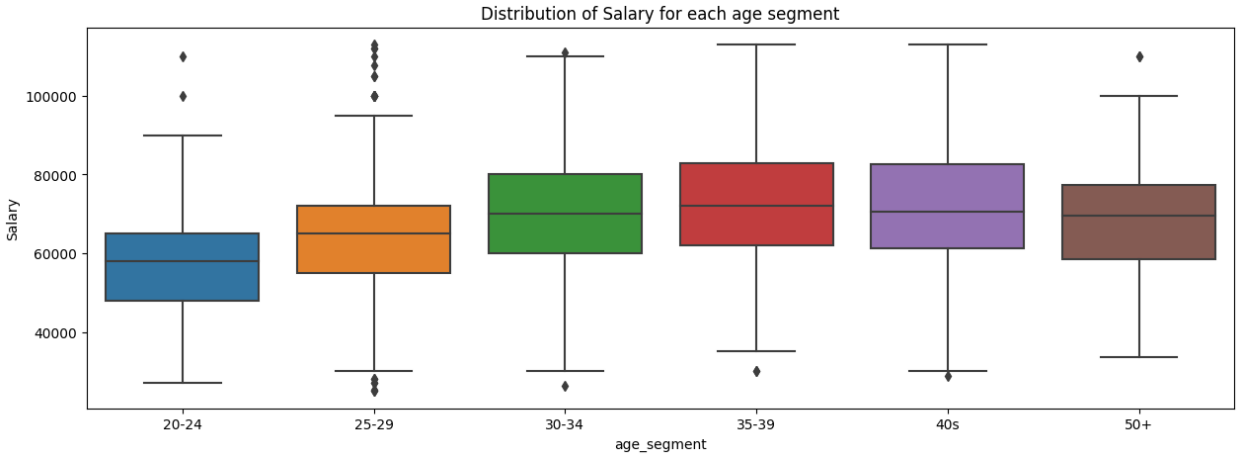
Predictive features used:

- Company size
- age
- gender
- city
- position
- Years of experience
- Seniority level

Assessing the distribution of the salary with each predicting feature:

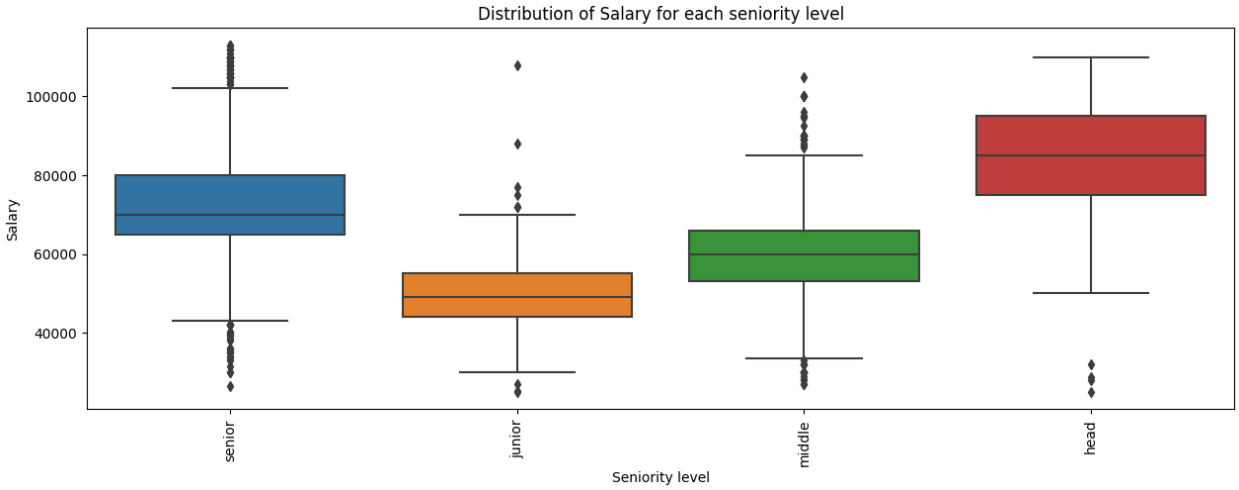


Set Expectation	Collect Info	Match Results
Men recieve higher salaries than women, due to what's called the gender pay gap in tech.	Upon plotting the box plot, it's obvious that the median salaries for males are higher than females. With 70000.0 Salary median for men and 58000.0 Salary median for women. Also, based upon codeacademy's survey, men were offered higher salaries than women for the same job title at the same company 59% of the time, according to a 2021 survey from Hired.	*Matches*



Set Expectation	Collect Info	Match Results
Fresh graduates and younger employees recieve smaller salary than older ages who have been there for a while in the field.	Upon plotting the box plot, it's obvious that the median salaries for the ages from 20-24 is lower than other ages segments.	*Matches*

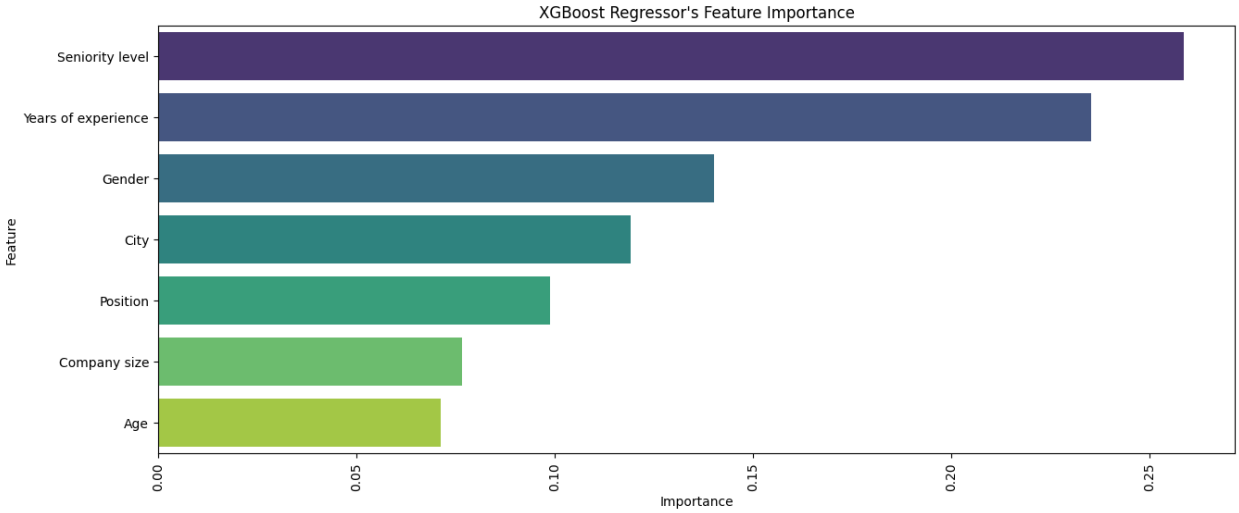
Set Expectation	Collect Info	Match Results
Small company size will pay less salaries than large company size.	Upon plotting the box plot, it's obvious that the median salaries for the up to 10 and 10-50 company sizes are lower than bigger company sizes.	*Matches*



Set Expectation	Collect Info	Match Results
As the expirence and seniority level increases, the employee influence on the company increases, thus should be his/her salary.	Upon plotting the box plot, it's obvious that the median salaries increases for the junior then middle then senior then head.	*Matches* The expectation matches the data collected with the same chronical order.

Cleaning was then performed on each feature.

Importance of each feature in predicting:



- *XGBoost regressor, linear regression and svr has been built.*
- *Holdout validation was used to assess the performance in predicting the salaries for such different models.*
- *The XGBoost regressor proved to be the best regressor, with 2898.08497, 9688.511969 mae for train and test respectively.*
- *Grid search was then used to find the best hyperparameters for the xgboost regressor.*
- *The xgboost regressor produced the importance of features ordered from the most important to use till the least.*

Communicate Results

- *As an individual working in the field of IT, you can estimate the salary worth your knowledge and expirence. This can help you when targeting a new job, to estimate and negotiate your expected salary based on the understanding of your worth.*
- *As an IT Company, this can help you decide the salaries of your employees. Not all employees working at the same title or team, should recieve the same sallary. There's an on-avergae salary for each position, then you can use the model to customize the avergae salary for each employee based on his/her experience.*

Q7: What is the relationship between contract duration & salary?

Analysis on the Salary

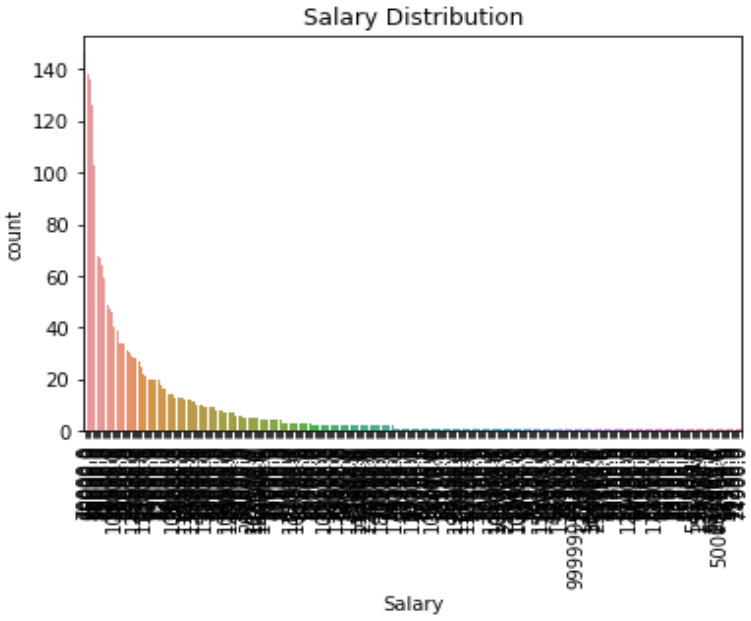
Firstly I started to question: What is the AVG / MIN /MAX/ of the salary?

1-Expectation : I guess the mean of the annaual salary of an IT engineer in Germany is around 50-60k in euors as Google says so.

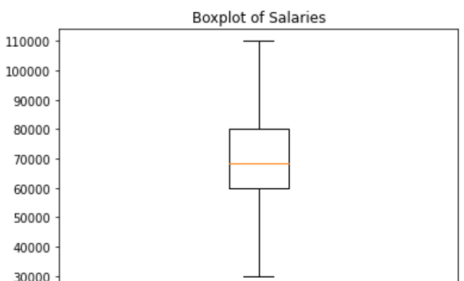
2-After collecting: The data says that the mean of salaries is 4487k, but median is 70k

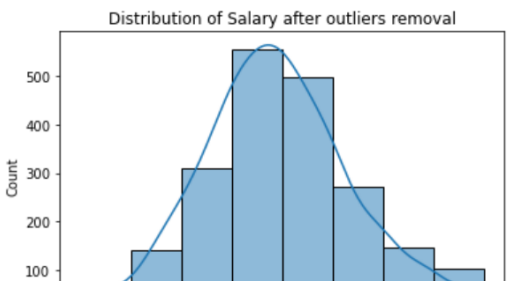
3-Comparing Data and expectations: the median is not much far from the expected mean, but the mean is much more larger than the expected. As we know it, mean is sensitive to outliers .Let's remove outliers and double-check.

Histogram of salary to show its distribution over data regardless of other features.



After outliers removal using IQR the distribution of salary is:





Analysis on the Contract duration

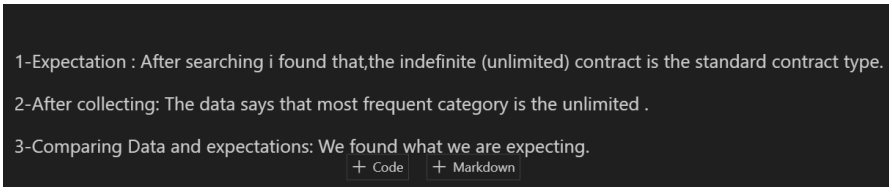
After exploring column values, we found that:

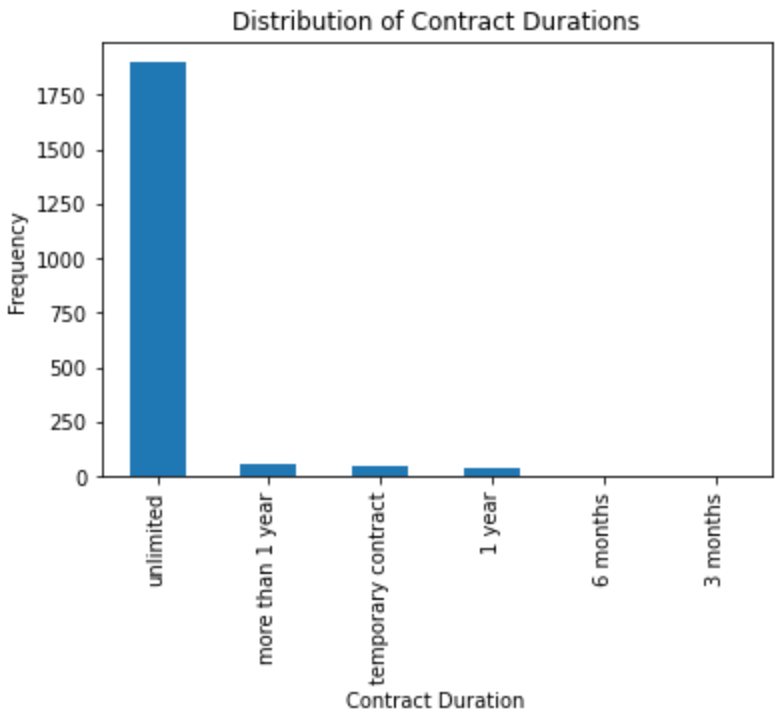
- There is 8 unique value
- The categories are: ['Unlimited contract' 'Temporary contract' nan 'unlimited' '6 months' 'more than 1 year' '1 year' '3 months']
- There are 38 missing values in duration

I decided to drop nan values rather than imputing as they are relatively small

After Cleaning we have only 6 categories : ['Temporary contract' nan 'unlimited' '6 months' 'more than 1 year' '1 year' '3 months']

To Answer: What is the most common (mode) contract duration in our data?





Can we deduce a relationship between the contract duration and salary?

Expectations :

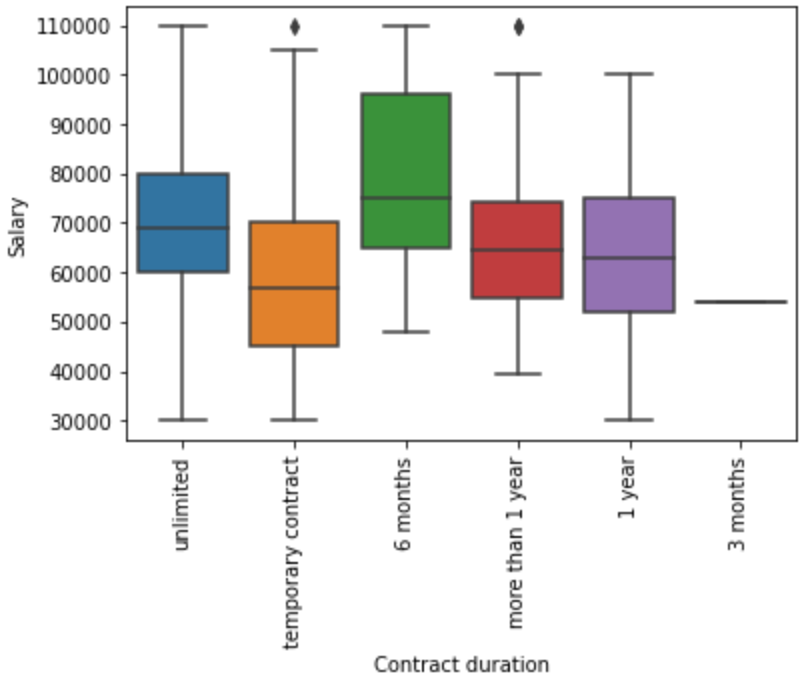
In general, longer contract durations may be associated with higher salaries, as they can provide more job security and stability for employees . However, this relationship can vary depending on factors such as job role, industry sector, and company size. For example, some companies may offer higher salaries for shorter contract durations to attract top talent or fill specific project needs

Collecting information

The graph shows that there is variation in the average salary in different contract categories. For example, 6 months category has the highest average of salaries. Average salaries in the unlimited category is close to the mean & median (expected as it's the majority).

Comparing Data and Expectations

I believe Expectations match collected information



Results : By applying the Kruskal-Wallis hypothesis test

```
Kruskal-Wallis H test
Test statistic: 27.608562189681496
p-value: 4.3405931086896265e-05
```

The Kruskal-Wallis H test is a non-parametric statistical test that compares the median values of the independent groups to determine if there is a significant difference between them. Specifically, it tests the null hypothesis that the median values of all groups are equal against the alternative hypothesis that at least one group has a different median value.

Expectations :

I expect from the graph and other search i did that, the salaries varies in different contract durations based on different point of views

Collected Data:

The hypothesis test kruskal resulted in a p value that is smaller enough than 0.05 to conclude that there is enough evidence to suggest a significant difference in salary among the contract duration categories.

Comparing Data and Expectations

Collected data meets our expectation and we can say that there is some relation between the contract duration and salary, indeed it's not the only affecting factor.However , on average it has some effect

Interpretaion

Based on the results of the Kruskal-Wallis H test with a p-value of 4.3405931086896265e-05, we can conclude that there is a statistically significant difference in salary between at least two of the categories of contract duration. This suggests that contract duration may be a factor that influences salary for IT engineers in Germany.

However, the Kruskal-Wallis H test does not identify which specific groups have significantly different median salaries. To determine which groups are significantly different, we should perform post-hoc tests such as the Mann-Whitney U test or Dunn's test.

It's important to note that statistical significance does not necessarily imply practical significance. Therefore, it's important to interpret the results of statistical tests in the context of the data and consider the practical implications of the findings. Additionally, correlation does not necessarily imply causation, so it's important to consider other factors that may influence salary for IT engineers in Germany, such as experience, education, skills, job role, and company size.

Business value

Recruitment and Hiring: Understanding the relationship between contract duration and salary can help businesses attract and retain top talent by offering competitive compensation packages that align with industry standards. For example, if there is a positive correlation between contract duration and salary, businesses may need to adjust their salary offers for temporary or short-term contract positions to remain competitive. In a positive correlation between contract duration and salary, it's possible that shorter contract durations may be associated with lower salaries. In this case, businesses may need to adjust their salary offers for temporary or short-term contract positions to remain competitive.

One reason to do this is to attract top talent in a competitive job market. Even for short-term or temporary positions, businesses may need to offer competitive salaries to attract qualified candidates who have multiple job offers or options. Offering a lower salary for short-term positions may make the position less attractive to potential candidates, which can make it harder for the business to fill the position and complete the project on time and within budget.

Additionally, offering fair and competitive compensation packages can help businesses build a positive reputation as an employer of choice. This can lead to higher employee satisfaction, better retention rates, and increased productivity and profitability in the long run.

Therefore, even if shorter contract durations are associated with lower salaries, businesses may need to consider offering competitive salaries to remain competitive in the job market and attract top talent. Ultimately, the decision on how to adjust salary offers for different contract durations should be based on a thorough analysis of the data and the specific context of the business and industry.

Q8: How does the position (backend/ Machine learning..etc) affect the required years of experience needed in order to be an official senior?

1)

Exploring position column

Expectaions :

To have aronud 10-15 different position

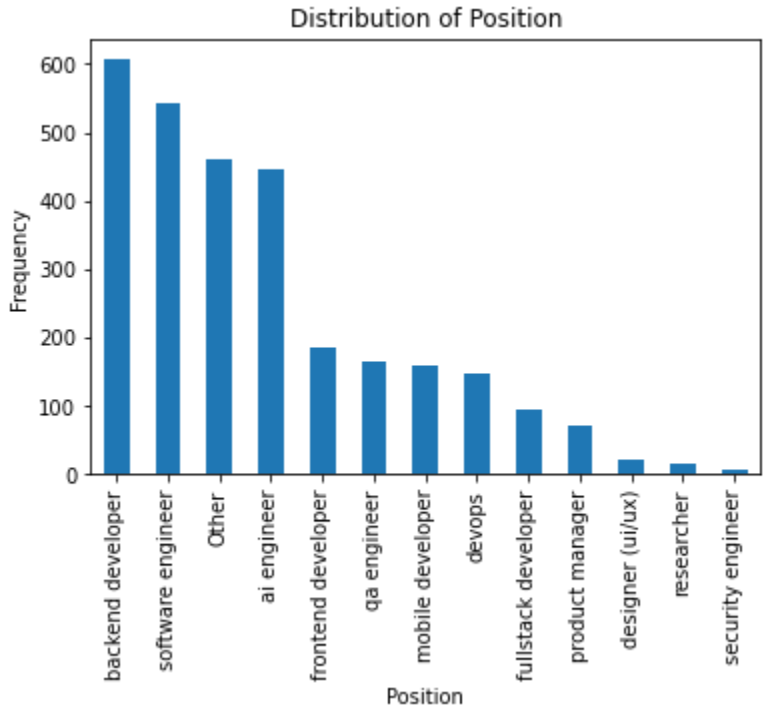
Collected data

we have 511 different position

Comparing Expectation with Collection:

Not matching.The number of uniqe positions is huge, it needs cleaning. Let's unify similar positions to the same category

2) Data Cleaning to match expectations: categories were reduced to 13 categories of interest.



3)

Exploring Years of Experience column

Expectations:

To have only numerical values with ranges from 0-40 at max.

Collection data :

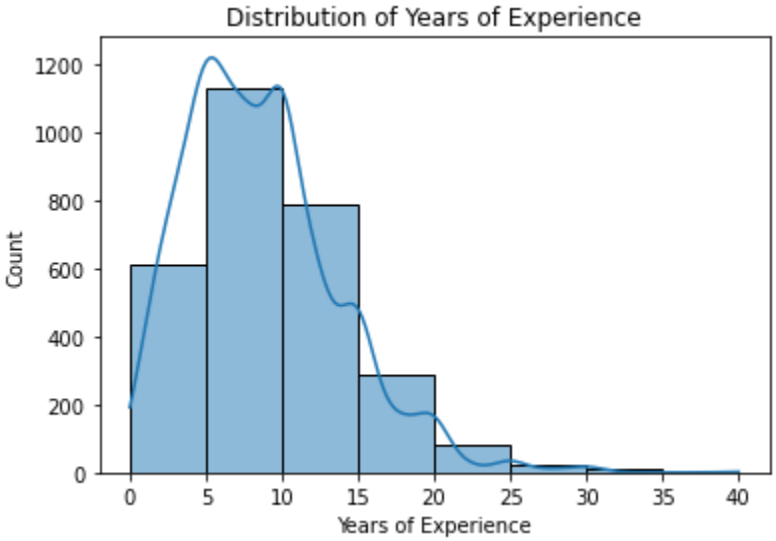
It shows that we have text in the years column ,out of range values (343) ,mis-written values(1,5 instead of 1.5)

Comparison:

Not Matching. Cleaning is needed

...Needs cleaning.

4) visualization:

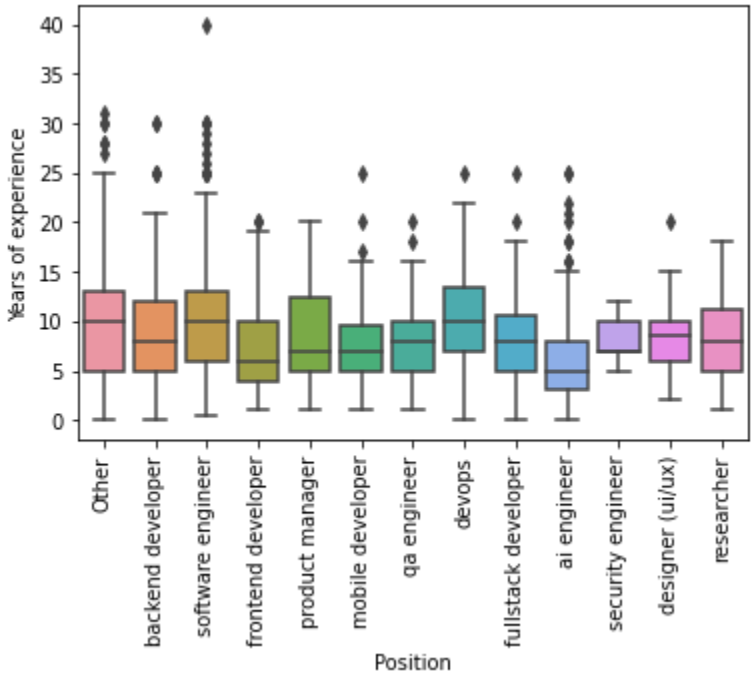


1-Expectation : I thins the mean will be around 10-15 years of experience, is the IT field has started growing in the last 15 years.

2-After collecting: It's clear that the data mean of years of experience is between 5-10 years.

3-Comparing Data and expectations: not exact but close to what is expected. ✓

5) Visualizing distribution of years of experience in different positions

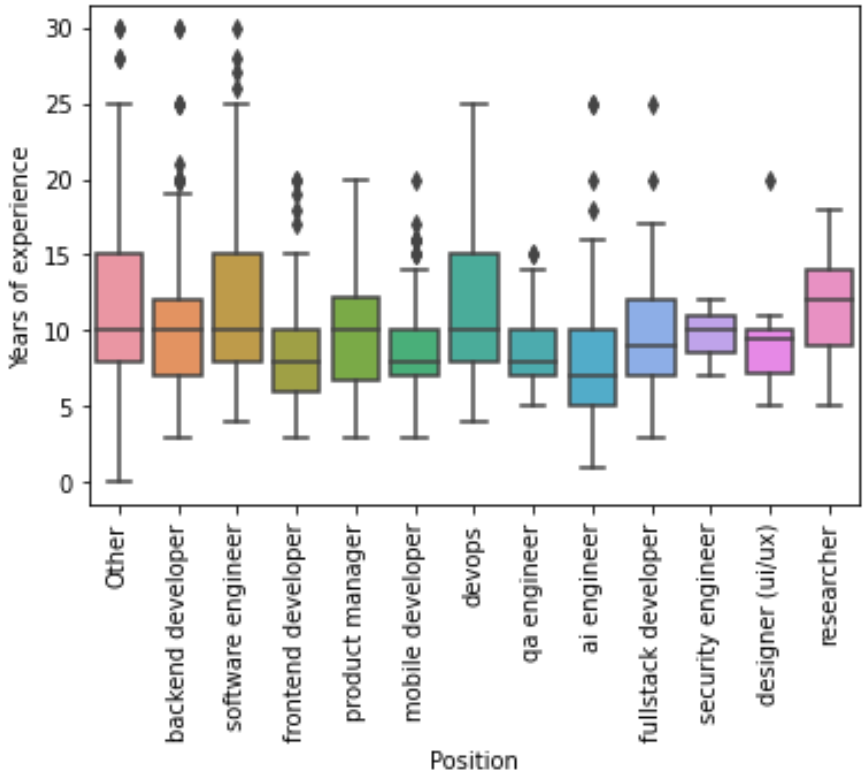


6) Now we will filter on Seniority level, to focus on the needed years of experience for the seniors only

['Senior' 'Middle' 'Junior' 'Head' 'Lead' 'Principal' 'No level' 'VP' 'Manager' 'Work Center Manager' 'CTO' 'No level ' 'Director' 'Key' 'C-level executive manager' 'intern' 'Student' 'no idea, there are no ranges in the firm ' 'C-Level' 'Working Student' 'Entry level' 'Intern' 'student' 'Self employed']

Into : 4 levels only [Senior, Mid level, Junior, Head]

7) To answer if there is a relationship between the position and the required years , we need to visualize



We can see that there is disparity in the needed years of experience to become a senior in each position,

For example, the average needed years for the QA is around 5-7 years,Unlike for the researcher position where you need at least 10-11 years and on average it is around 13 years! That's interesting.

We can categorize the positions into two types:

1. 10-15 Needed years which includes: [SW, BE, DevOps, Product Manager, Others, IT, Researcher, Embedded]
2. 5-10 Needed years which includes: [FE, UI/UX designer, Mobile Development, Data Scientist, Software Testing, ML, QA]

After applying hypothesis test, the results of p-value are:

	sum_sq	df	F	PR(>F)
Position	2757.050046	12.0	12.498375	9.858017e-25
Residual	29780.012110	1620.0	NaN	NaN

Reject null hypothesis as there is a significant difference in the means of years of experience across position.

Q9: What is the most paying position in Berlin?

Set Expectations ❓

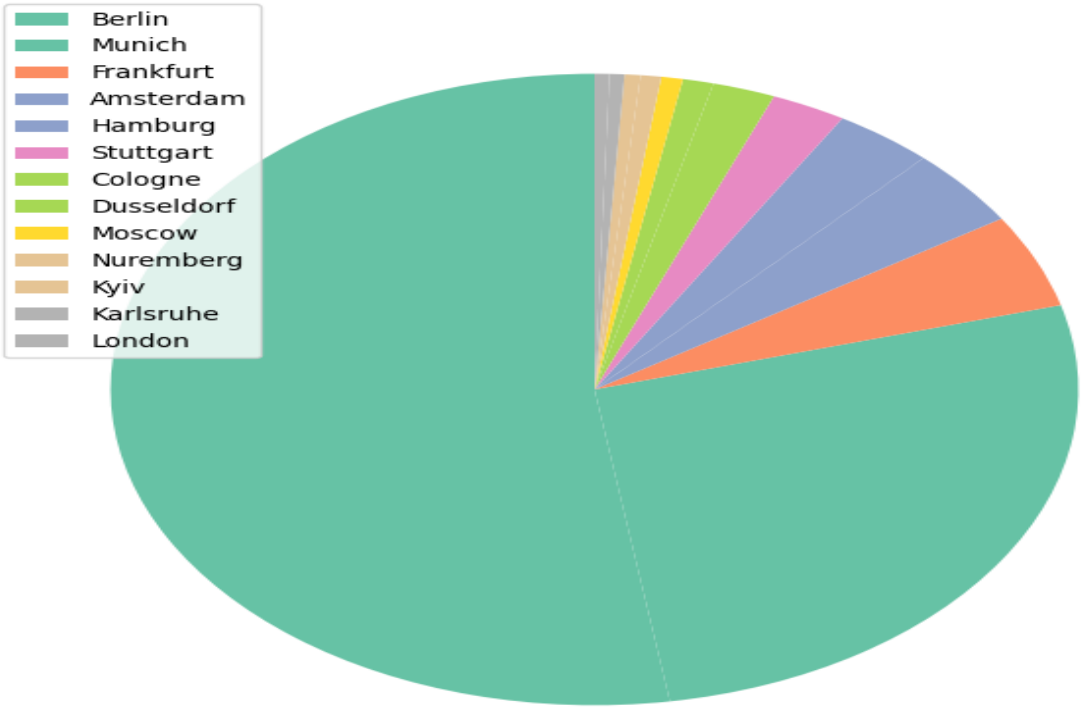
1. ai engineers and software engineers with high Seniority levels have a high salary
2. The junior's distribution for average salaries may be different from the seniors and Middle as in IT Positions, some positions do not commonly to need juniors such as Machine learning, for example, compared to frontend developers or software developers in general

Data Collection

1. Filter the data to get positions in Berlin only
2. Get the average Salary For each position in the senior level only and sort them
3. Plot the average for each position in Berlin
4. Look at the top 5 positions if they contain Machine Learning

Explore Data

Distribution of the cities in the data:



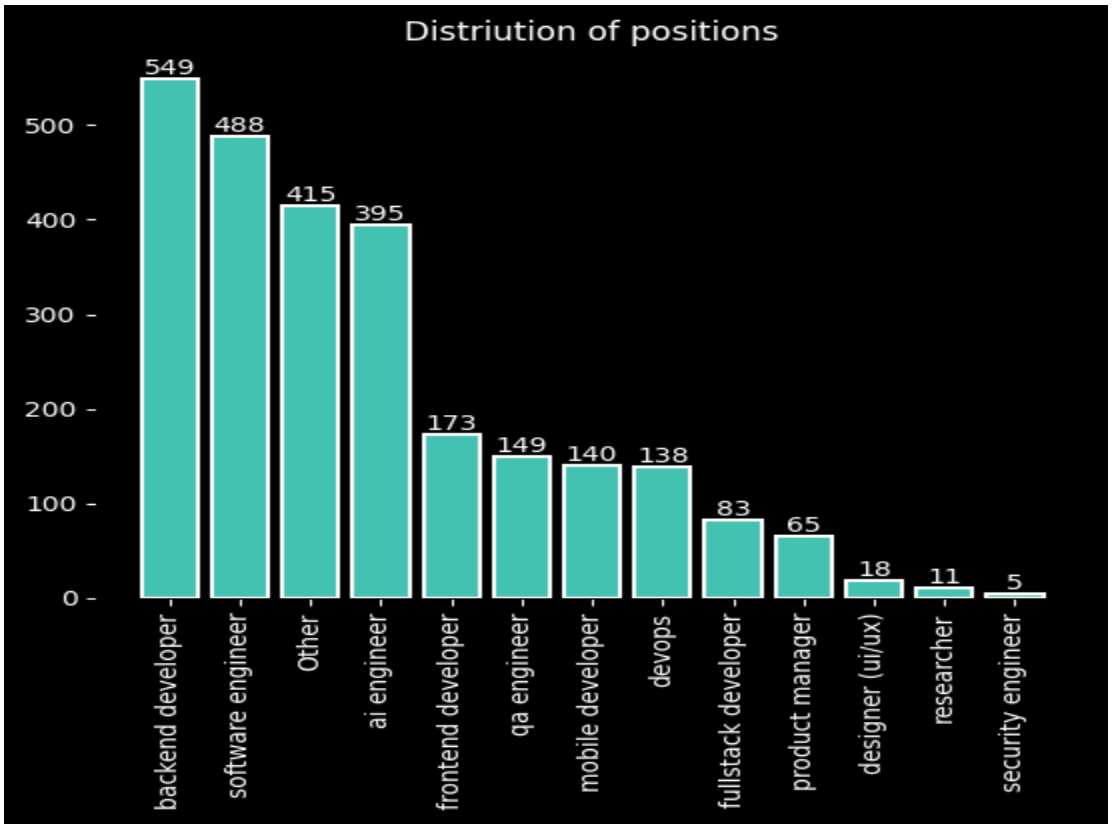
Interpretations ⚡

- It seems that the most frequent city is Berlin, then the cities' frequency is ordered as follows: Berlin, Munich,

Frankfurt, Hamburg, Amsterdam, Stuttgart, Cologne, dusseldorf.

- Match the expectations we expect that as the capital of Germany is Berlin

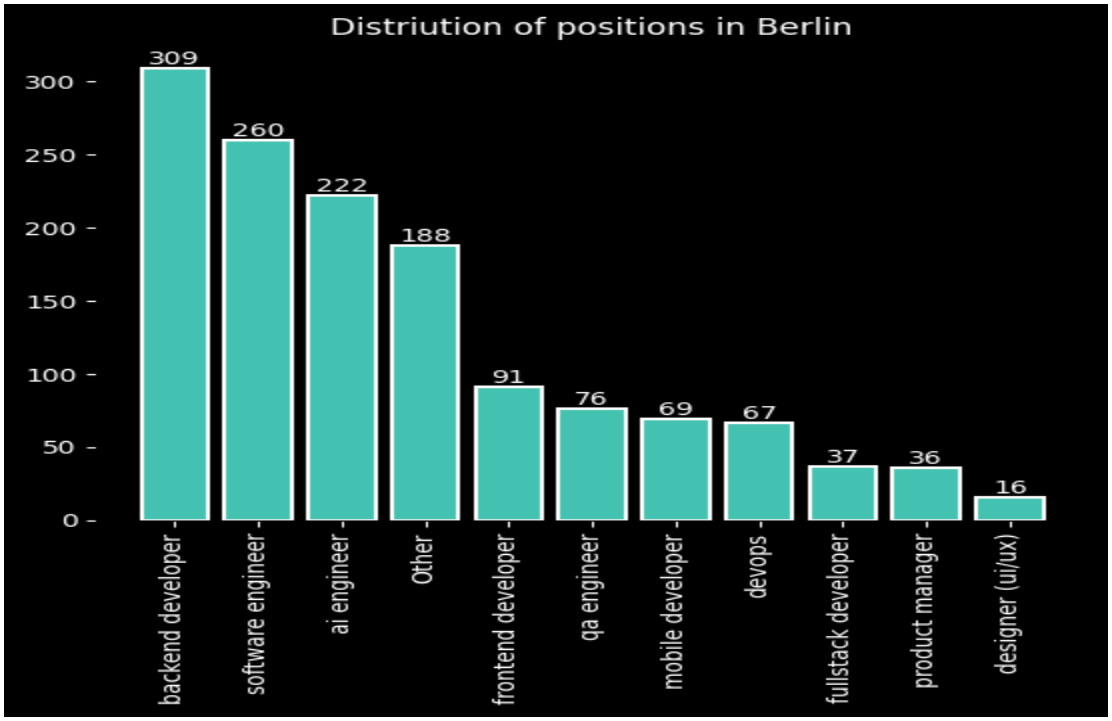
Distribution of the positions in the data: 📉



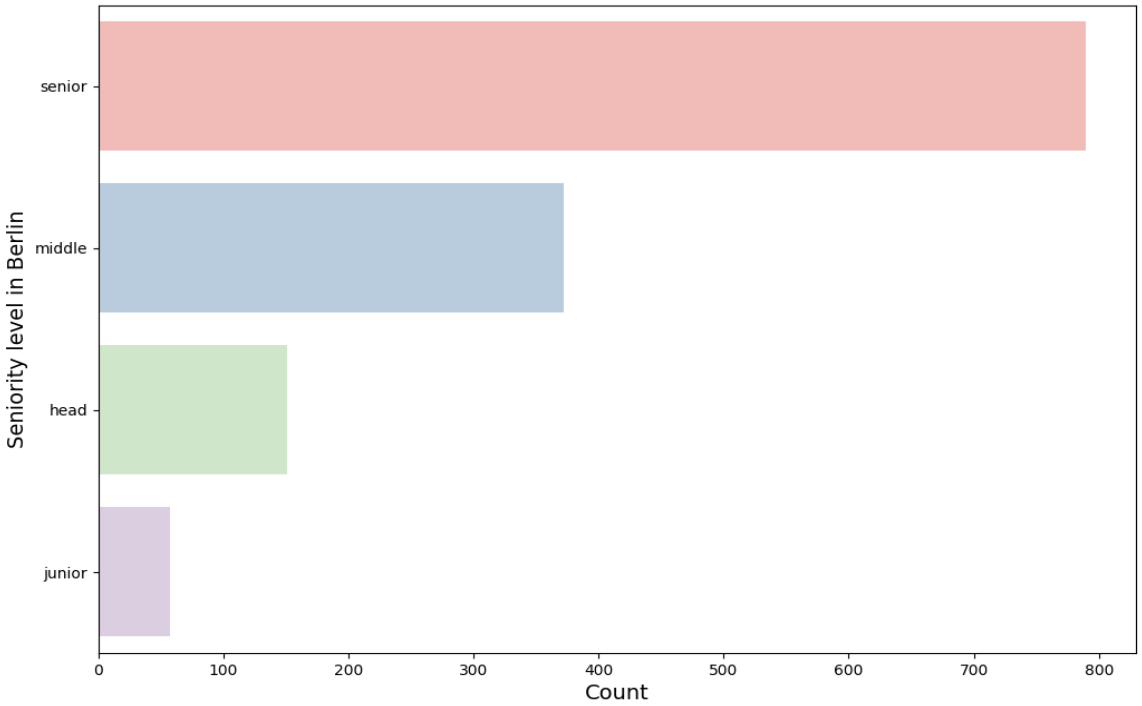
Interpretations ⚡

- It seems that the most frequent position is backend developer and software engineer and ai engineer

Distribution of the positions in Berlin: 📉



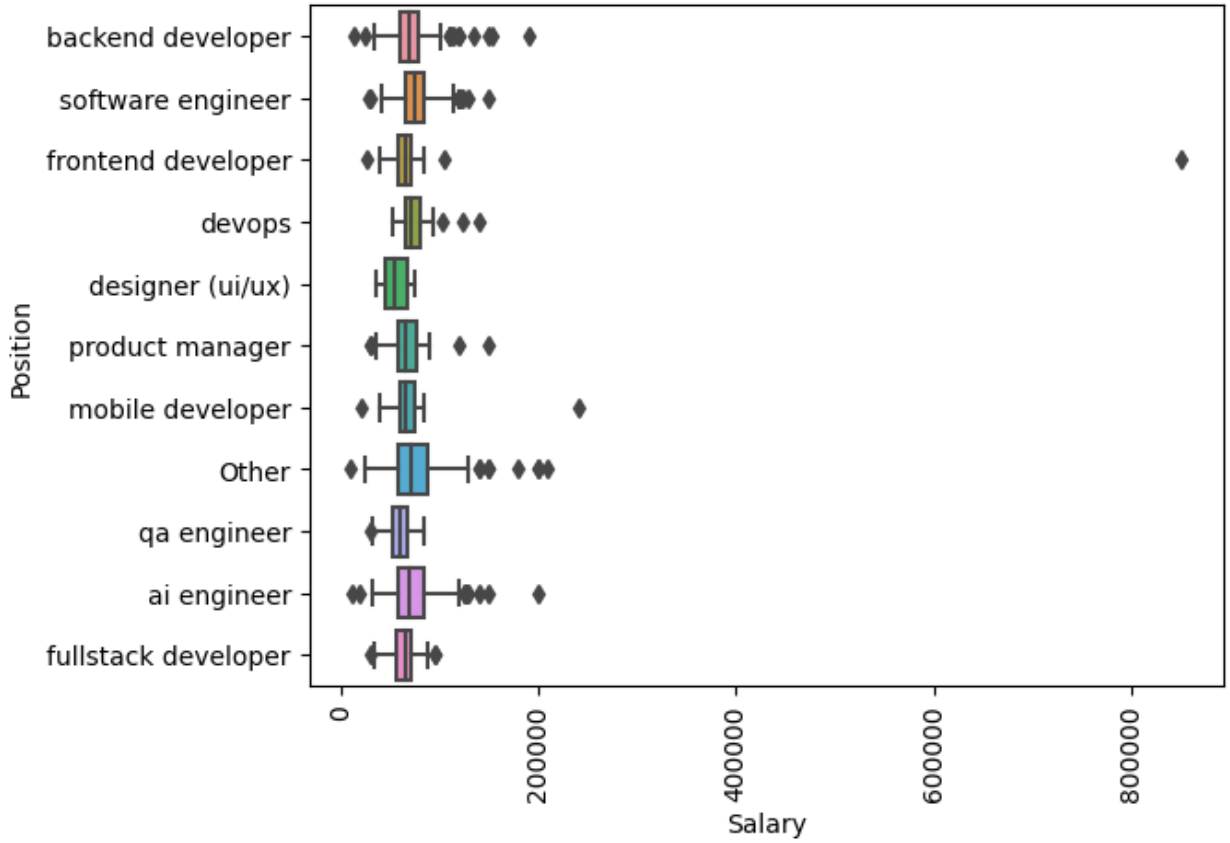
Distribution of the Seniority level: 📉



Interpretations ⚡

- From the graph it seems that the most frequent seniority level in the data is a senior level, so we can have a strong insights from this level to solve the question and then check other seniority levels so we can support our analysis
- we can also merge senior level with middle level as these two seniority levels are close to each other and the Salaries of then are also closed

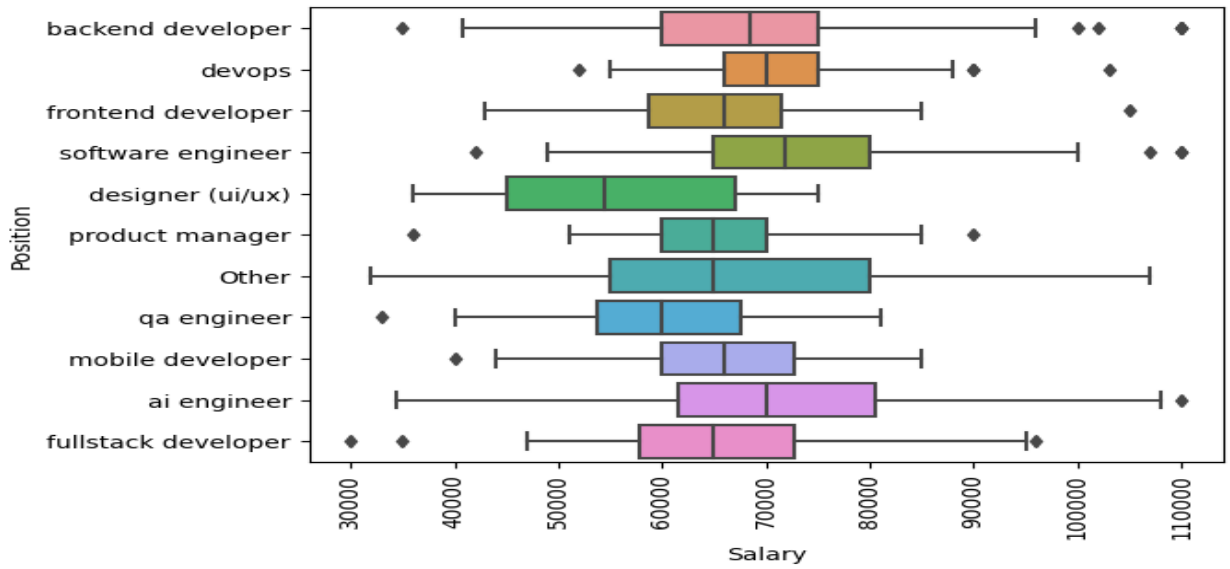
Distribution of the SALARY in Berlin: 📉



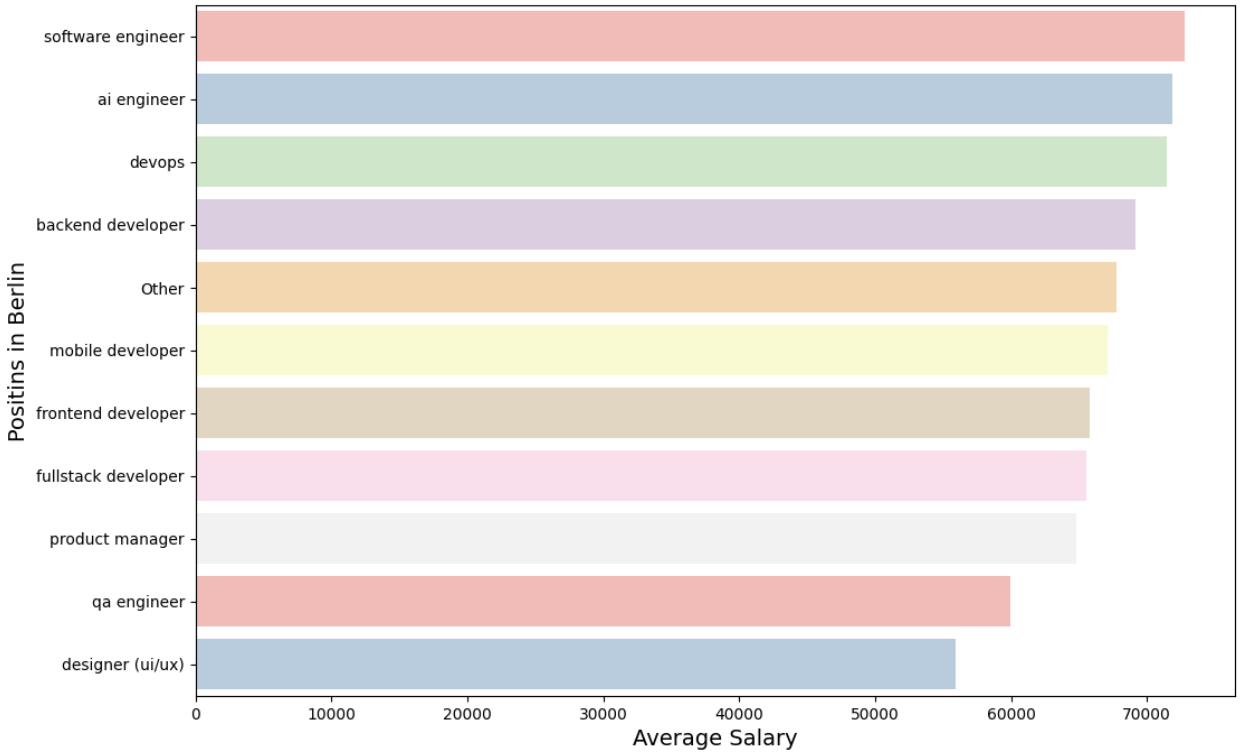
Interpretations ⚡

From this graph I notice a lot of outliers and start to remove them

Distribution of the SALARY in Berlin after removing outliers: 📉



Get average salary in Berlin for seniors & middle and start sorting them to find the highest: 📉



Interpretations ⚡

1. From the previous visualizations` I found that the top four position's salaries in Berlin are for four positions

2. Expectation
   high positions in ai engineering and software engineering (senior & middle) have higher salaries

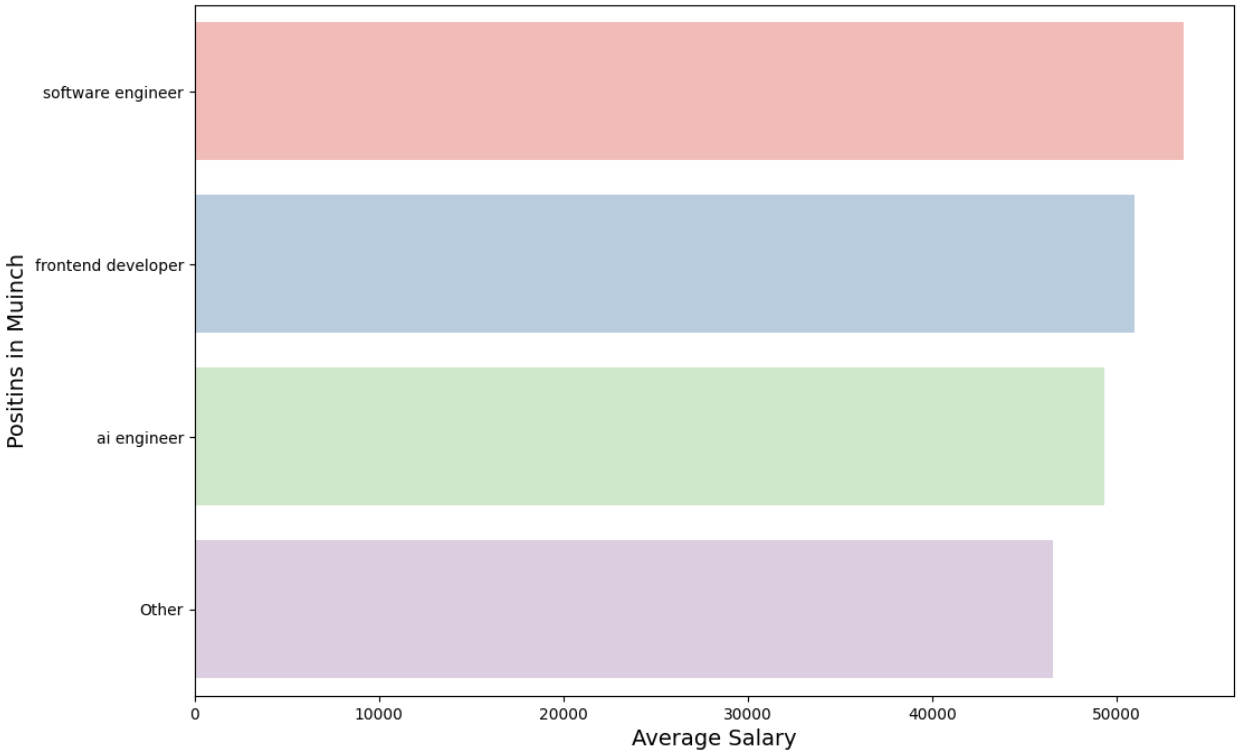
3. Results Interpretation

position	salary
software engineer	72811.167513
ai engineer	71848.548571
devops	71451.578947

4. From the analysis we find that the highest salary goes for software engineer and with a very small difference from the ai engineer so Expectation `Matched`

5. Additional insights
   - Back end and mobile developer are almost the same
   - Front end and full stack are almost the same
   - Previous Four positions is very close to each other in Salary
   - QA and Designer are the lowest salaries

Get average salary in Berlin for juniors and start Sorting them to find the highest: 📉



Interpretations ⚡

From the previous visualizations: I found that the highest average salary in Berlin FOR Junior POSITIONS is for three positions with this order:

position	avg_salary
software engineer	53666.666667
frontend developer	51000.000000

Expectation: software engineers and front-end developers need juniors (people that can work with a little experience) so I expect that they will have the highest salaries in Berlin for juniors

Data collected matches Expectation

Communicate Results 💡

Question's Answer: Software Engineer followed by ai Engineer but the highest is Software Engineer.

Q10: What is the relation between the number of people at some seniority level with the company Type?

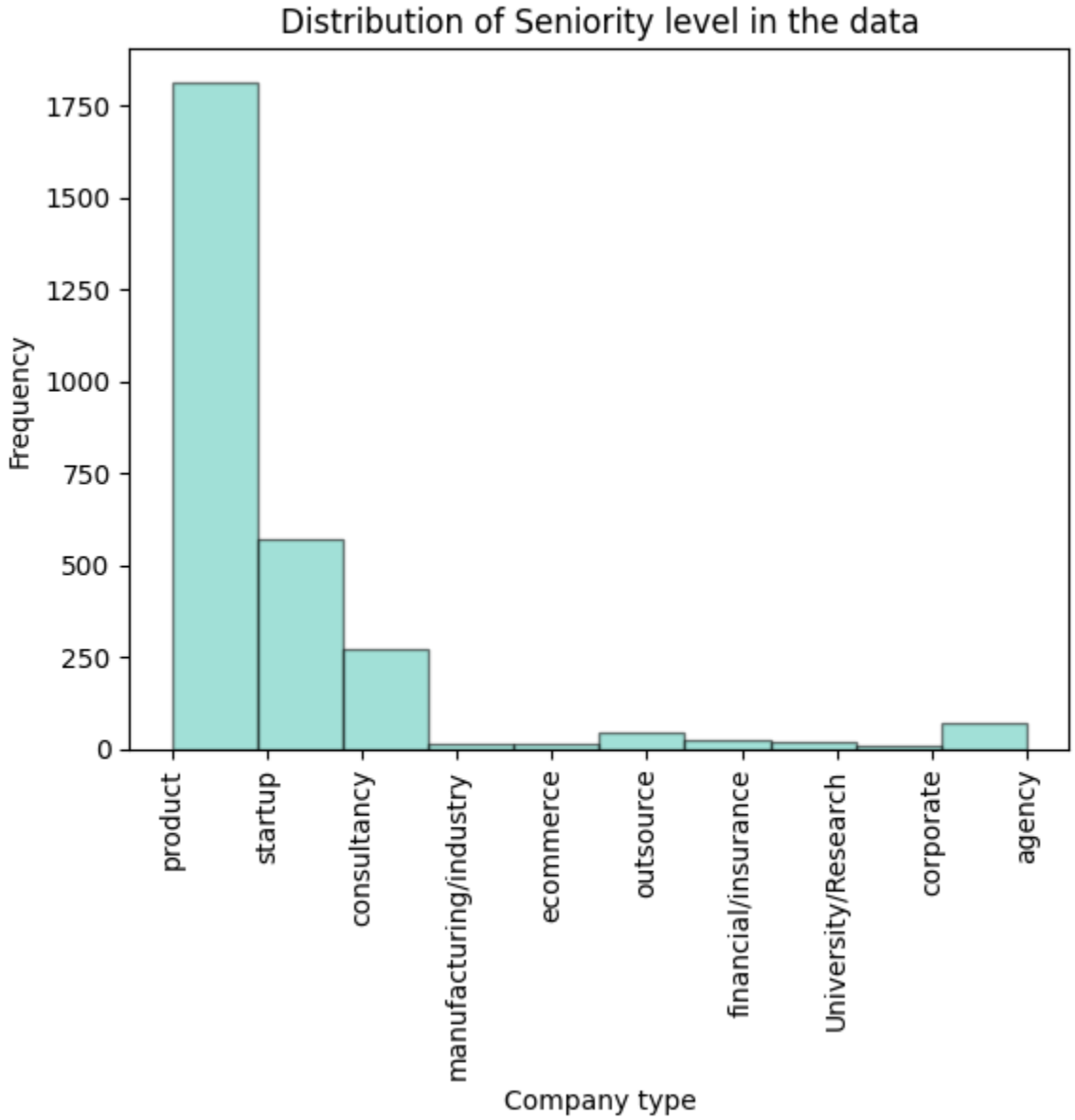
Set Expectation ❓

- Seniors will be the most frequent seniority level in all companies
- Head and lead will be the least frequent seniority level in all companies
- The number of juniors compared to seniors in the company will depend on company type (startups will contain more juniors)

Data collected matches Expectation

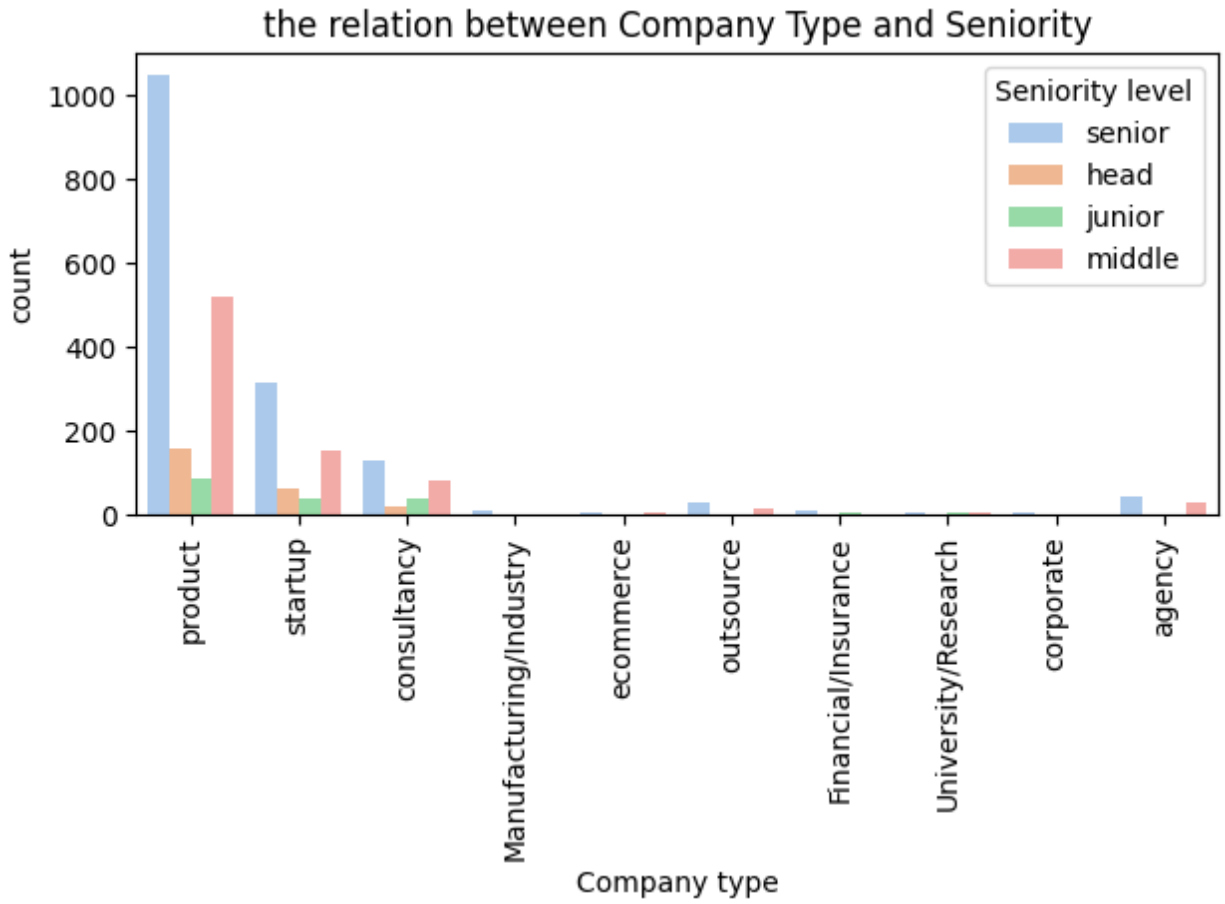
Explore Data

Distribution of the Company type 📉



Interpretations ⚡

From the graph, we will find that most common Company type is the product.



Hypothesis test 👍

Using: chi2_contingency
P-value: 5.069630716309754e-12
Result: P-value <0.05 there is a relation

Interpretations ⚡

1. It seems that seniority level is the most frequent level regardless of company type, However, the difference in the

number of senior and junior employees is larger in product companies compared to startups, and for the head and lead seniority level, they appeared little in all types as we expect

2. This result may suggest that product companies may have a greater emphasis on hiring experienced and senior-level employees, potentially due to the need for specialized skills and expertise in product development. On the other hand, startups may be more willing to hire junior-level employees and provide them with opportunities for growth and advancement.

3. The lead and head level is less frequent in all types of companies. This may suggest that these higher-level positions are relatively rare and require significant experience and expertise to attain.

Q11: Can we infer that the most paying position in Berlin is the highest paid in all other cities?

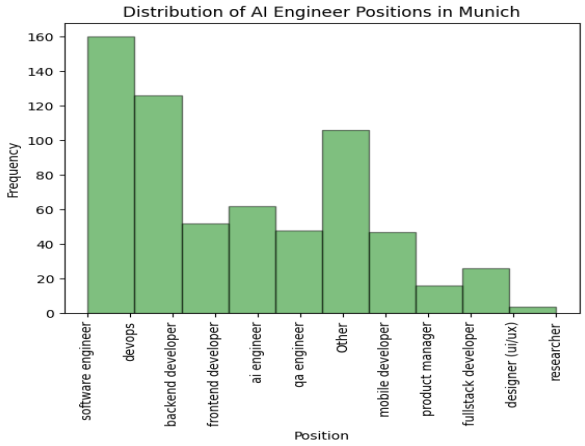
Set Expectation ❓

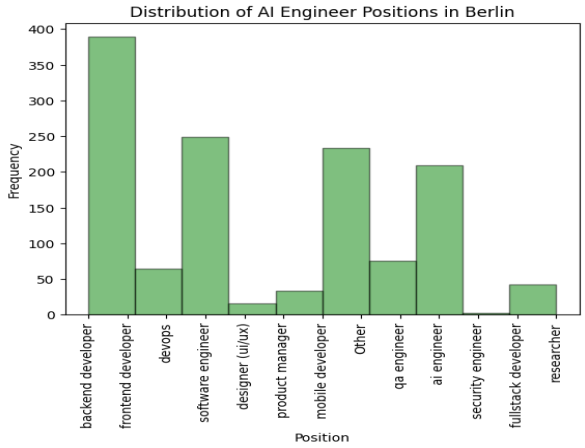
I think can't infer as Berlin is the City of Germany and it is make sense that it will have varieties of positions including the highest one

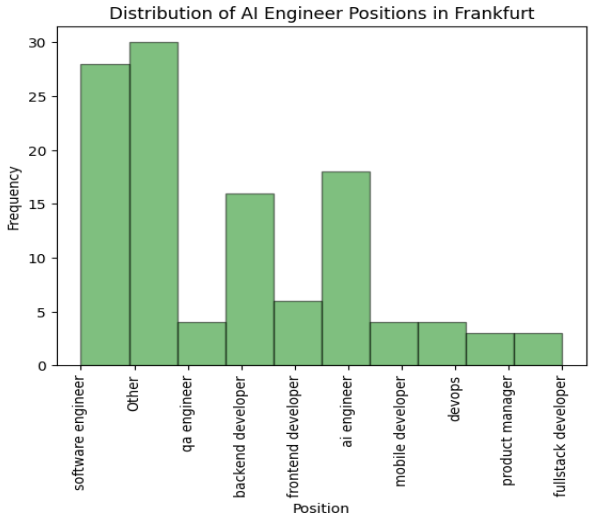
Data collected matches Expectation

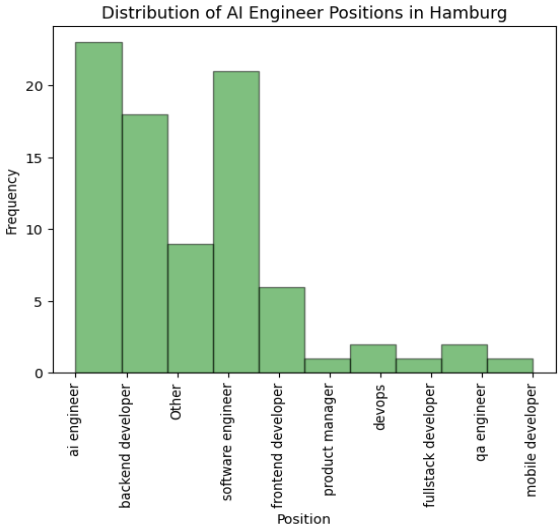
Explore Data

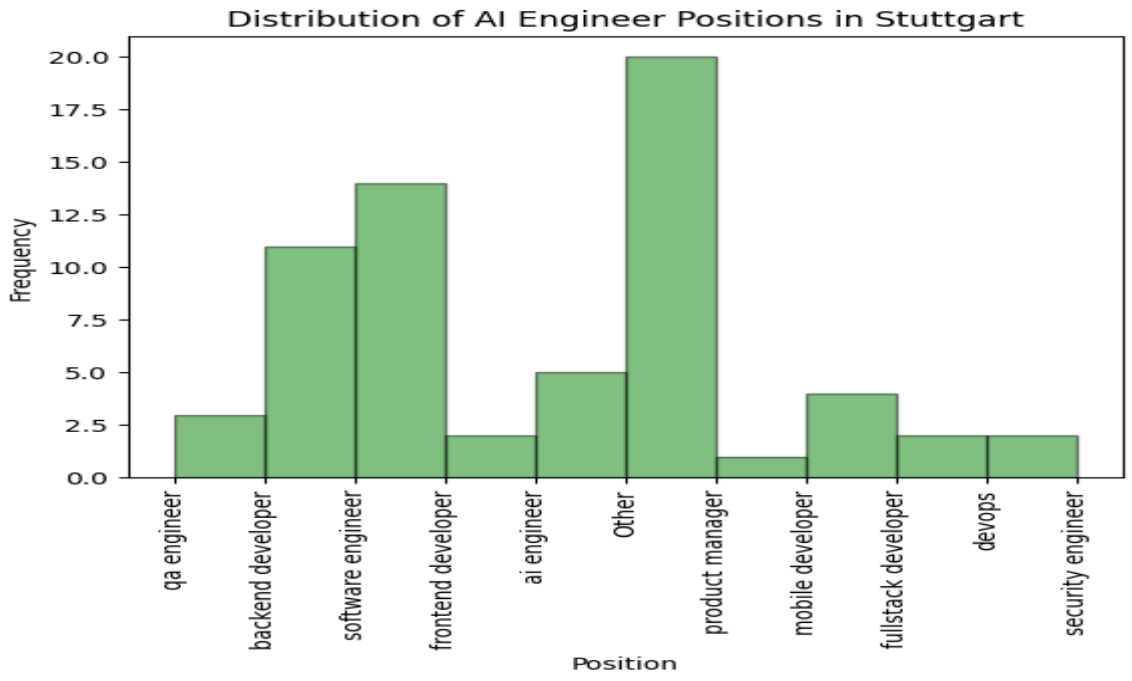
Choose only the Cities that have software engineer positions: Distribution of the Position in each city 📉



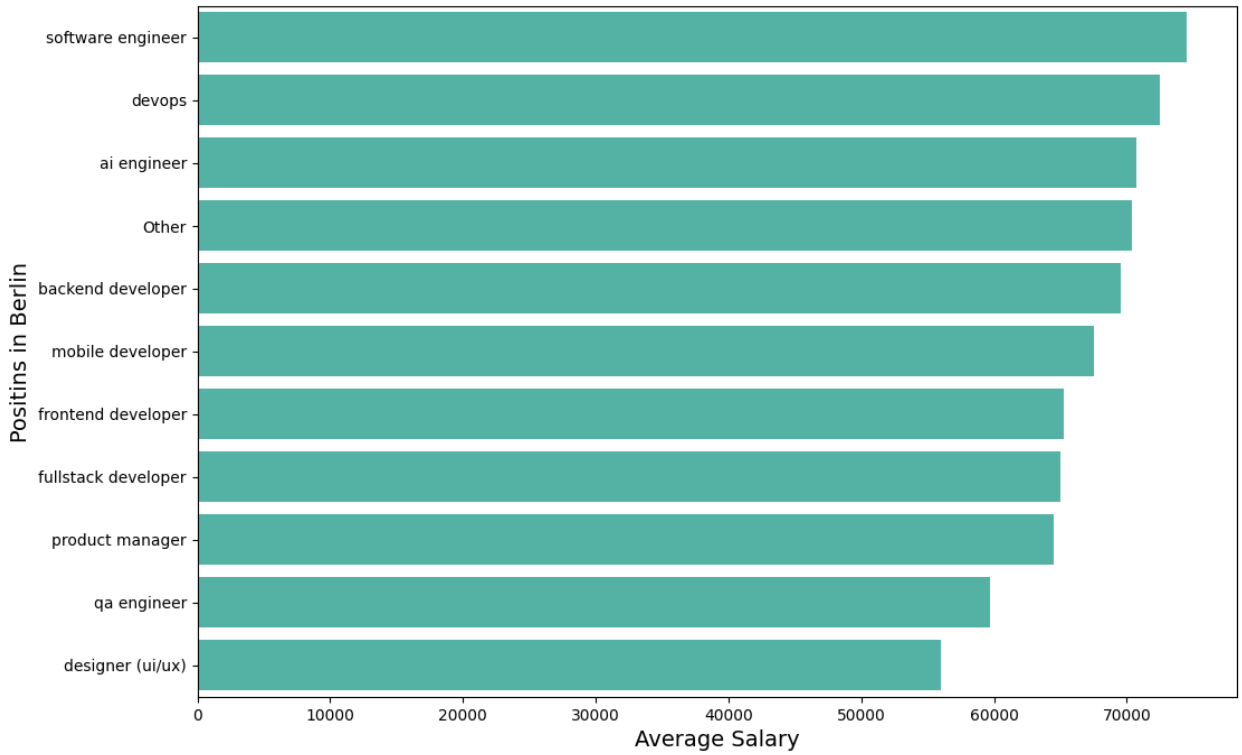


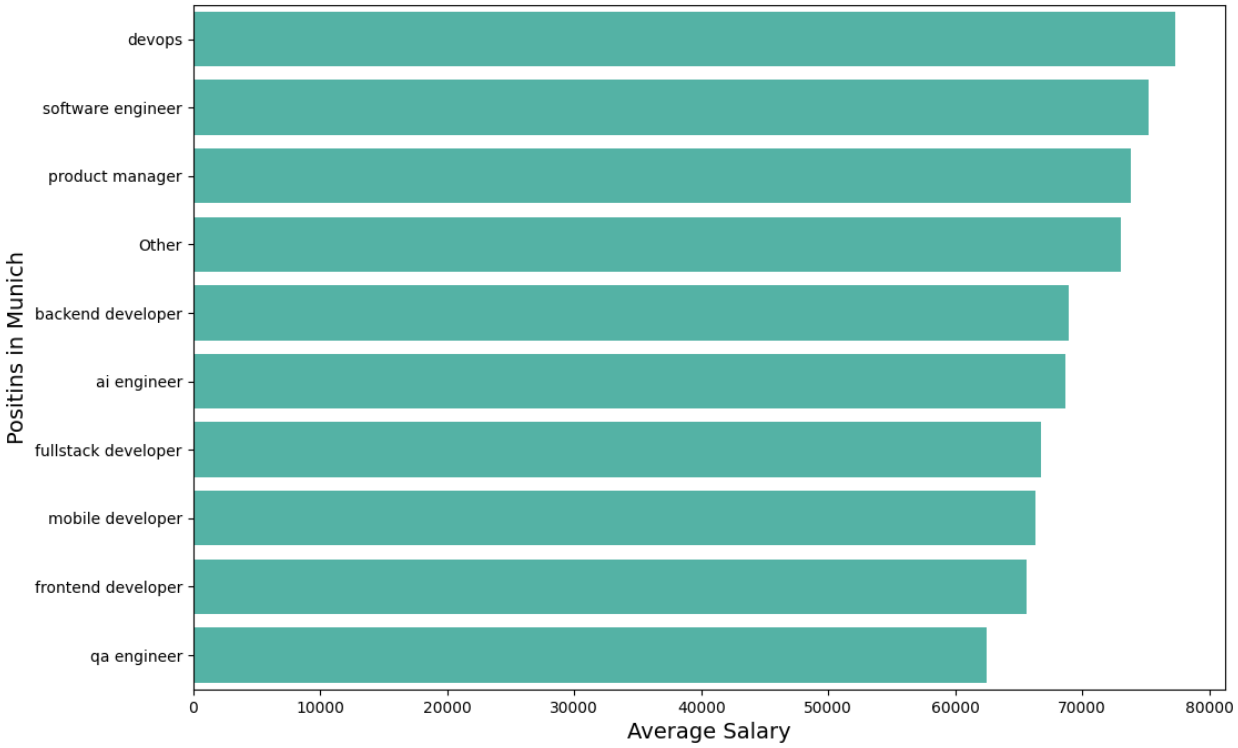


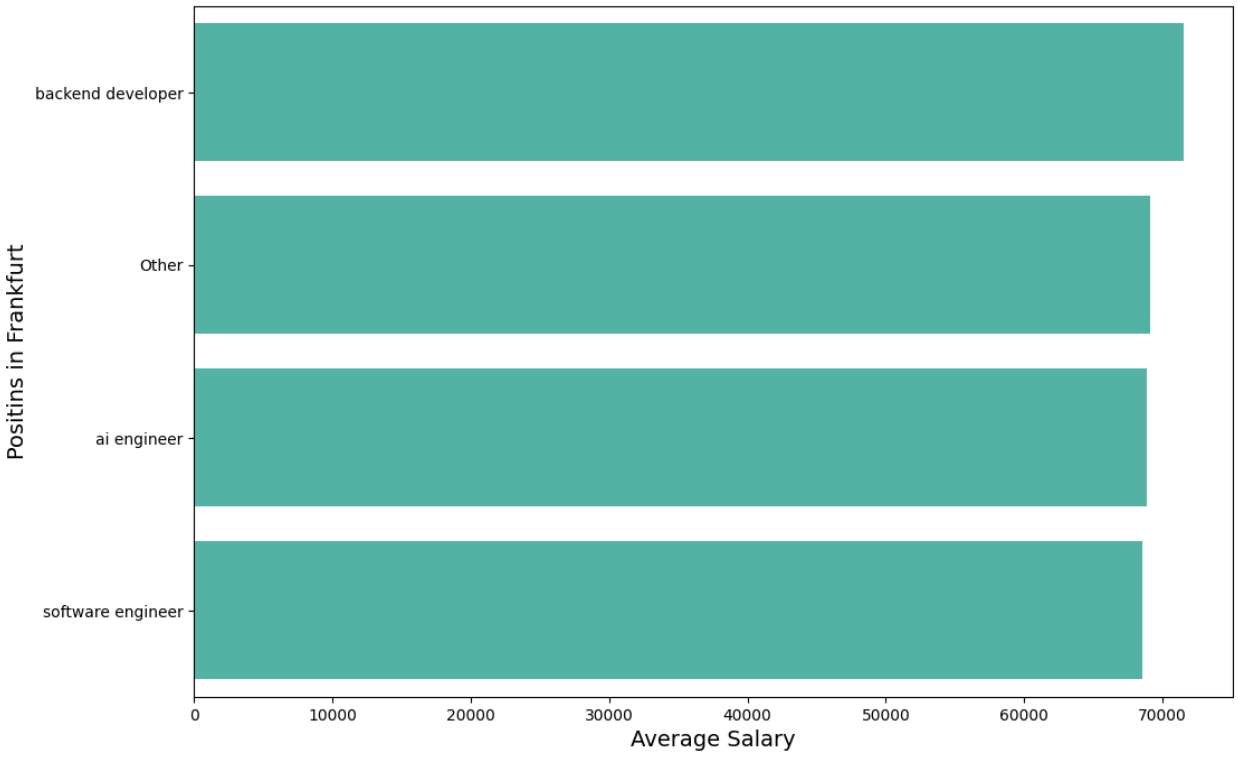


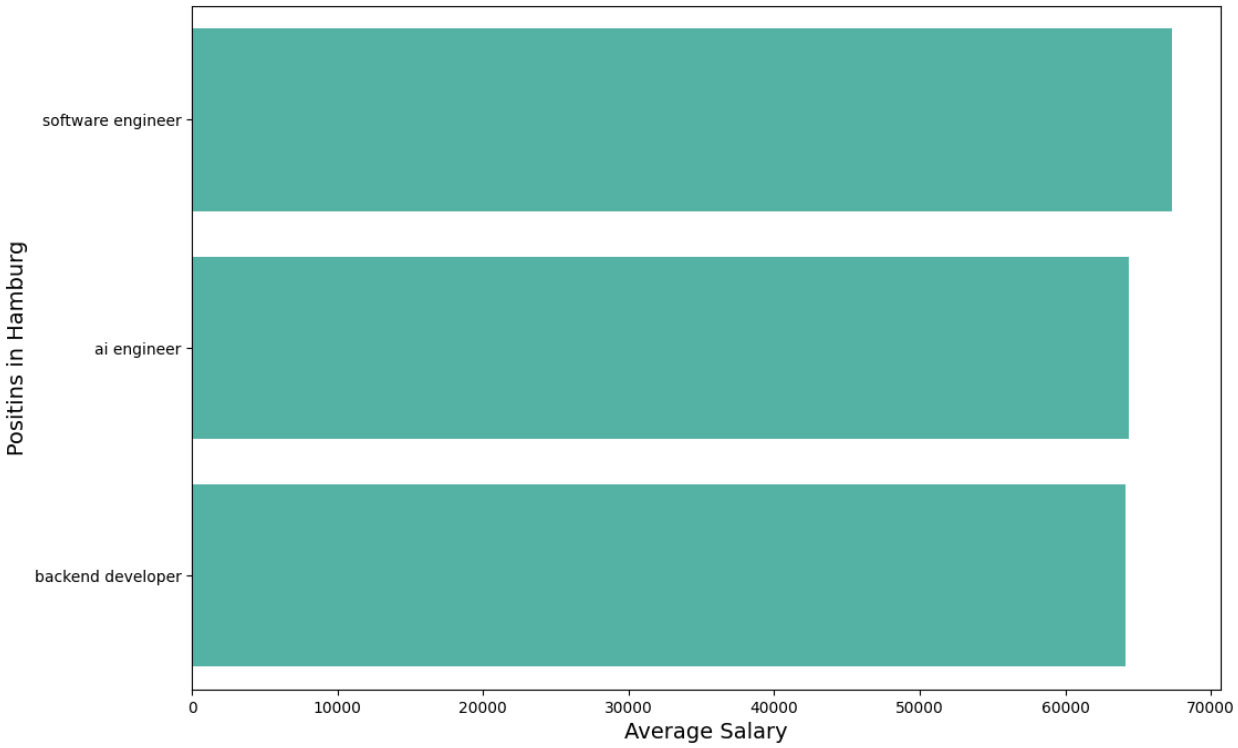


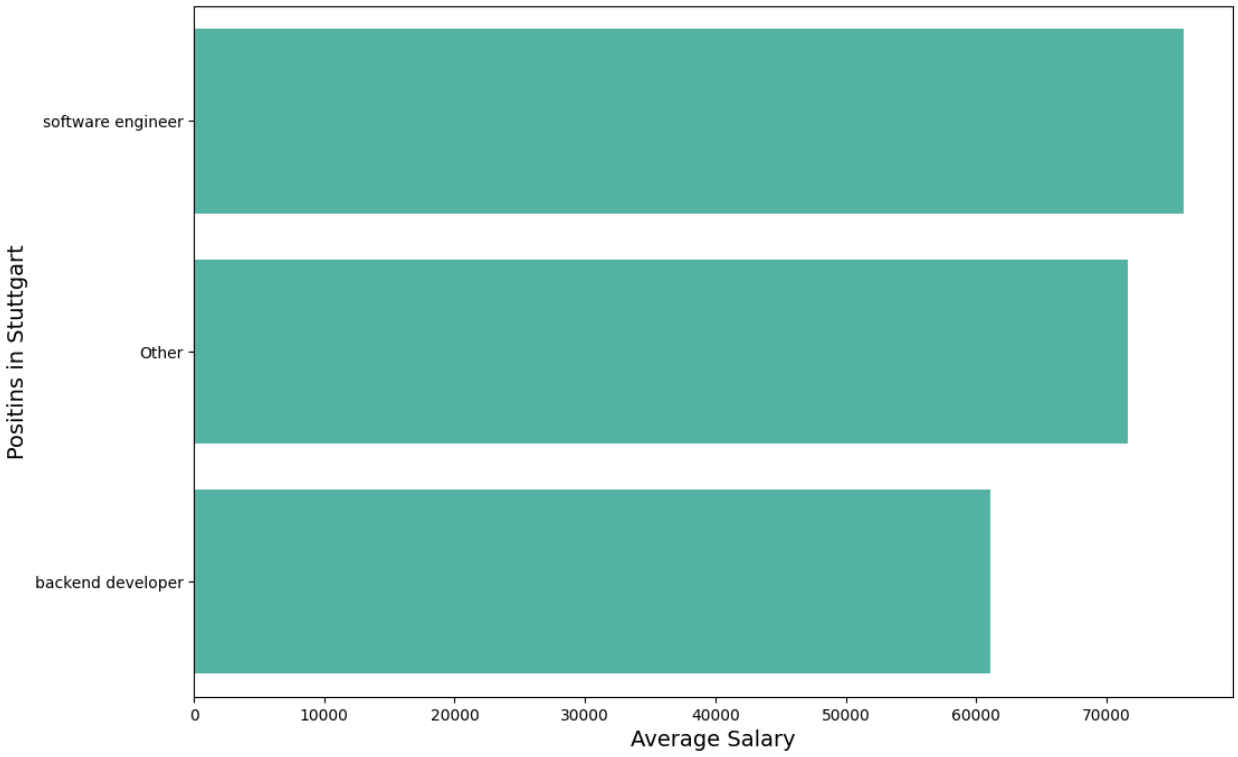
Average salary for each position in each city 📉




devops
software engineer
product manager
Other
backend developer
ai engineer
fullstack developer
mobile developer
frontend developer
qa engineer
Positins in Munich
0
10000
20000
30000
40000
50000
60000
70000
80000
Average Salary


backend developer
Other
ai engineer
software engineer
Positins in Frankfurt
0
10000
20000
30000
40000
50000
60000
70000
Average Salary


Positins in Hamburg
software engineer
ai engineer
backend developer
0
10000
20000
30000
40000
50000
60000
70000
Average Salary


Positins in Stuttgart
software engineer
Other
backend developer
0
10000
20000
30000
40000
50000
60000
70000
Average Salary

Interpretations ⚡

From the distributions, we can notice that Stuttgart has software engineer as the highest, Hamburg position has software engineer as the highest but there is no significant difference between avg software engineer salary and other positions, Munich has software engineer in the top two but not the highest one, Frankfurt has low average for software engineer salary but there is no a big difference with it and other positions .

From the graphs it's clear that we can't infer anything except for Stuttgart.

Hypothesis test 👍

- software engineer salary in Munich is not highest (p-value=0.409)(position=devops)
- software engineer salary in Munich is highest Munich (p-value=0.000) (position=backend developer)
- software engineer salary in Munich is highest Munich (p-value=0.000) (position=frontend developer)
- software engineer salary in Munich is highest Munich (p-value=0.003) (position=ai engineer)
- software engineer salary in Munich is highest Munich (p-value=0.000) (position=qa engineer)
- software engineer salary in Munich is highest Munich (p-value=0.000) (position=mobile developer)
- software engineer salary in Munich is not highest (p-value=0.699)(position=product manager)
- software engineer salary in Munich is highest Munich (p-value=0.004) (position=fullstack developer)
- software engineer salary in Hamburg is not highest (p-value=0.483)(position=ai engineer)
- software engineer salary in Hamburg is not highest (p-value=0.394)(position=backend developer)

- software engineer salary in Stuttgart is highest Stuttgart (p-value=0.011) (position=backend developer)
- software engineer salary in Frankfurt is not highest (p-value=0.568)(position=backend developer)
- software engineer salary in Frankfurt is not highest (p-value=0.943)(position=ai engineer)

Interpretations ⚡

- Munich → has one failed with devops as it is higher than software engineer (we can infer if all values p-value with all position in this city after filtration is <0.05)
- Stuttgart → valid
- Hamburg → failed (there is not a big difference between avg software salary and other positions .
- Frankfurt → failed

Finally we and all these conditions so we can't infer.