0. Scenario

I am a junior data analyst on the marketing analytics team at a hypothetical recruitment agency. Many of our clients highlight their strong language skills, which leads them to expect higher salaries, and they often complain when they do not receive the jobs they hoped for. The company's director wants to understand why clients with high language proficiency still struggle to secure suitable positions. He has asked me to analyze the data to identify the factors behind this issue.

1. Ask

The first step of the analysis is to define the problem we are trying to solve:

Why don't job seekers with strong language skills find the right job for them?

2. Prepare

The agency operates its own job advertising website, which connects employers with job seekers. This website will serve as the data source for our analysis. The first step is to perform a ROCCC assessment of the dataset:

- **R – Reliable:** The job advertisement data is provided directly by employers who submit their postings through our website forms.
- **O – Original:** The data is collected by the agency itself, making it an original source.
- **C – Comprehensive:** The dataset includes all key information needed for our analysis, such as required language level, expected proficiency, offered salary, and more.
- **C – Current:** The data is scraped immediately before the analysis, and it can be refreshed at any time because the scraping process is fully automated in Python.
- **C – Cited:** The dataset comes directly from the company's own data collection.

After scraping the data, extensive cleaning is required because employers often provide information in different formats, currencies, and structures.

3. Process

After loading the scraped raw data, several cleaning and transformation steps were performed:

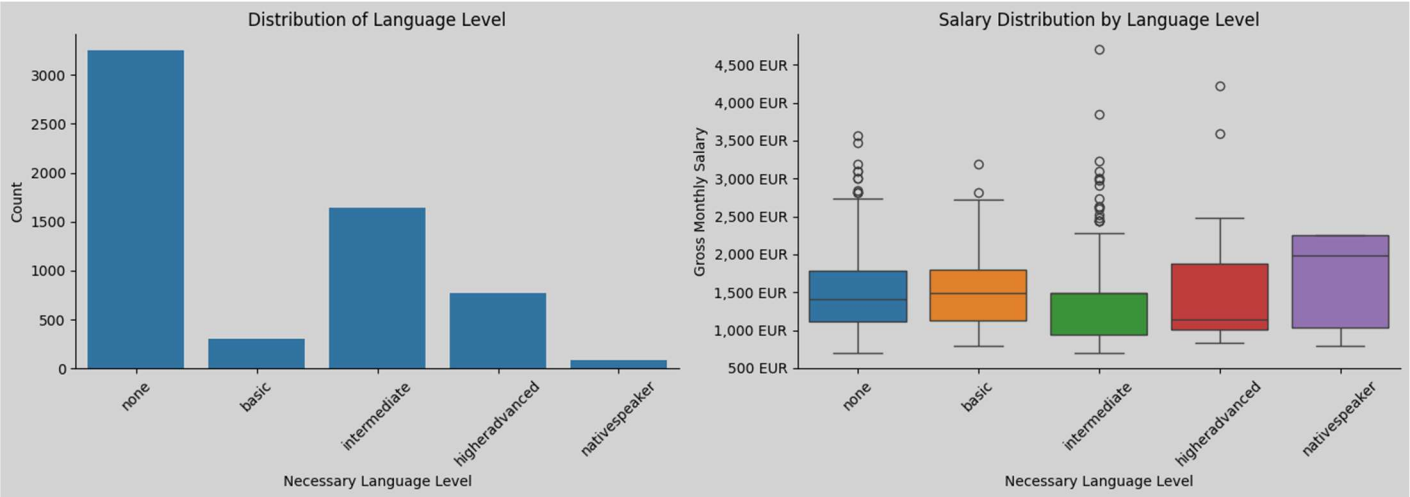
- Drop irrelevant, duplicate, constant columns.
- Rename columns for clarity, consistency.
- Parse salary text into numeric monthly salary (HUF), then convert to EUR
    - extract gross/net, currency (huf), and period (month, hour, year) using regex,

    - extract salary min/max (and prefixes if present),
    - compute an average salary,
    - normalizes into a gross monthly salary by:
        - yearly -> divide by 12
        - hourly -> multiply by ~168 hours/month
        - net -> convert to gross via dividing by 0.66
- Remove salary outliers by defining lower and upper bounds.
- Parse language skill into two separate columns: language nationality and language level.
- Drop redundant original text columns (salary, language skill).
- Remove rows missing key fields.
- Fill missing language fields with “none”.
- Filter to full-time jobs only
- Simplify experience labels (recoding), e.g.:
    - "Career starter/freshly graduated" → "starter"
    - "1-3 years experience" → "1-3 years"
    - etc.

After completing all cleaning steps, the processed dataset was saved for use in the analysis phase.

4. Analysis and Share

Firstly, we analyzed the relationship between expected language skill level and salary.

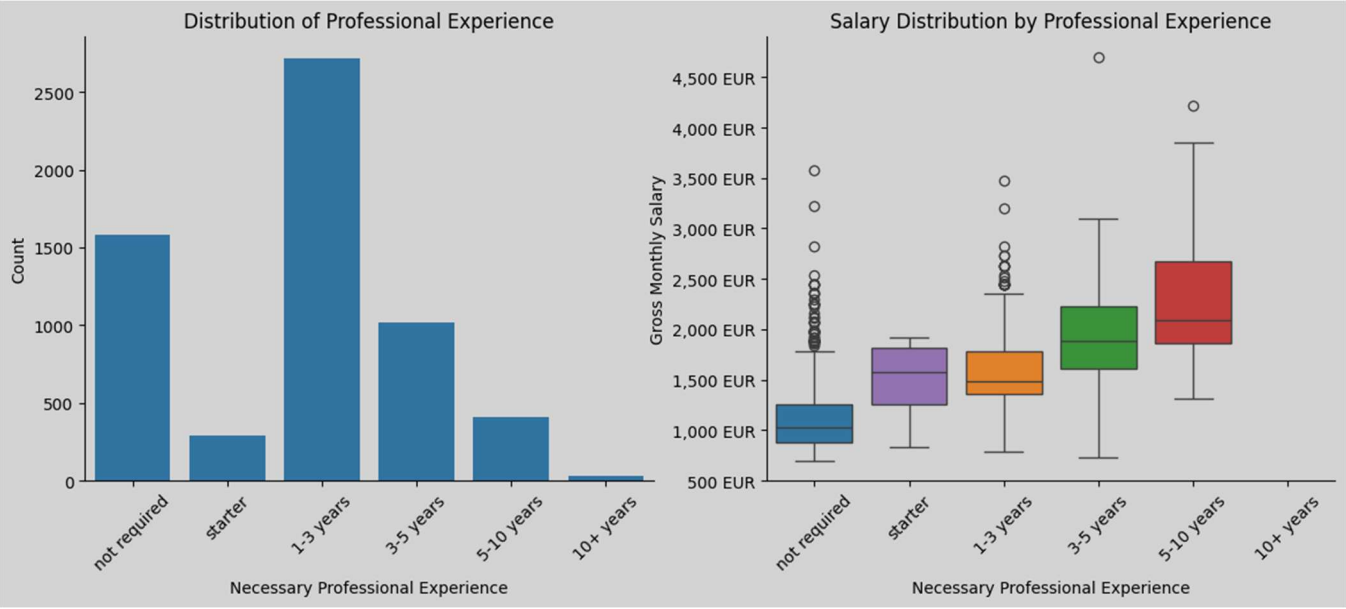


As we can see, most jobs that require language skills demand an intermediate level of proficiency. The distribution of language levels is noticeably uneven, with a significant difference in the number of listings across levels.

While many of our job seekers expected clear differences in salary based on language proficiency, the data shows a different picture. The none, basic and intermediate levels follow almost identical salary patterns. Although the more advanced levels show slightly higher salaries, the gap is still much smaller than expected. This also explains why many of the complaints come from clients at basic, intermediate or higher-advanced levels, where the salary difference is not as significant as they assume.

This suggests that language proficiency may not be a primary factor influencing salary. Instead, profession type and years of experience likely play a more significant role in determining compensation.

Next, we examined the relationship between professional experience level and salary.



As we previously assumed, salary appears to be more strongly influenced by work experience than by language proficiency. One interesting insight is that individuals at the "Starter" level and those with "1–3 years" of experience tend to earn roughly the same salary. This suggests that having 1–3 years of experience may not offer a significant salary advantage over having no experience at all.

A possible conclusion is that early-career professionals may need to accumulate at least 3 years of experience before seeing a noticeable increase in salary.

5. Act

The analysis shows that, in general, language skill level does not influence salary as much as many job seekers believe. Based on these findings, our agency can explain to clients that professional experience has a much stronger impact on salary. This means that

candidates with several years of experience are likely to achieve better results by applying for roles that match their expertise, rather than focusing mainly on positions with high language requirements.