Introduction

Understanding the dynamics of the Egyptian job market is crucial for both economic planning and individual career development. This project aims to provide a comprehensive analysis of the current employment landscape in Egypt through the extraction and analysis of job listing data from Wuzzuf, a prominent online recruitment platform.

Data Collection

Data acquisition was achieved through web scraping, a technique employed to systematically collect information from websites. Initial investigation of Wuzzuf's robots.txt file revealed that job listing URLs were accessible, enabling the targeted retrieval of these links. The BeautifulSoup library was initially utilized for its efficiency in parsing HTML and extracting these URLs.

However, subsequent data extraction from individual job listings encountered challenges due to the dynamic loading of certain elements via JavaScript. Consequently, Selenium, a browser automation tool with robust JavaScript rendering capabilities, was employed to ensure complete data capture. While Selenium's execution speed is comparatively slower than HTML parsers, its ability to handle dynamic content proved essential.

To mitigate the performance bottleneck associated with Selenium, a parallel scraping approach was implemented. The URL dataset was partitioned, and individual team members concurrently scraped subsets of the data. The resulting data fragments were then aggregated into a unified dataset for subsequent processing. Following data collection, a rigorous data cleaning phase was undertaken to ensure data quality and consistency.

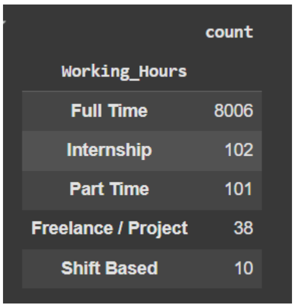
Data Cleaning

Title & Job_Category:

The dataset exhibited a significant discrepancy in the diversity of job categories and job titles. While the job category field was relatively clean, containing only 39 unique values, the job titles displayed substantial variation, with 4,848 unique values. This indicates that job categories were well-structured and standardized, whereas job titles lacked uniformity, potentially due to variations in job naming conventions, abbreviations, or inconsistent data entry. The high number of unique job titles suggests the need for further preprocessing, such as standardization or clustering, to enhance data consistency and usability for analysis.

Working_Hours:

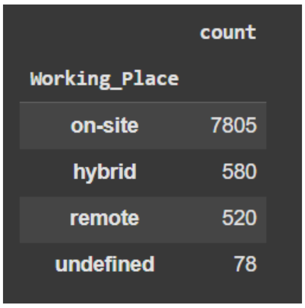
The Working Hours column had five distinct categories; no cleaning was needed. The distribution was highly imbalanced, with Full Time (8,006) being the most common, while Shift Based (10) was the least frequent.



	count
Working_Hours	
Full Time	8006
Internship	102
Part Time	101
Freelance / Project	38
Shift Based	10

Working_Place:

It contained a mix of valid categories (**On-site, Hybrid, Remote**) along with inconsistent and irrelevant entries such as job roles and industries (**e.g., Marketing, Sales, Engineering**). Additionally, **78 records lacked a specified working_place**. To clean this column, I retained only the three valid categories and assigned **"Unspecified"** to missing values while dropping all other inconsistent entries. This ensured a standardized and meaningful classification of working_place data.



	count
Working_Place	
on-site	7805
hybrid	580
remote	520
undefined	78

Company:

This column had values for only **2,000 records**, while the rest were missing (**NaN**). Since some companies prefer not to disclose their names, we filled the missing values with **"Not Specified"** to maintain consistency and completeness in the dataset.

Location:

The **Location** column had minimal occurrences of the **"\new company"** tag and **"\not verified"** entries, so we trimmed them. Additionally, unnecessary backslashes (**"\n"**) were removed. We also standardized governorate names by consolidating variations into a single record (e.g., **"Mandara-Alexandria"** → **"Alexandria"**) and removed any numerical values to ensure consistency.

Post_Date:

The **Post Date** column was recorded in a **"days ago"** format with no null values. We converted it to be relative to today's date. If the post date was given in hours, we defaulted it to **1 day** for consistency.

Number_of_Applicants:

For the **Number of Applicants**, entries labeled as **"Be the first to apply"** were replaced with **"0"** to ensure numerical consistency in the dataset.

Number_of_Positions:

For **Number of Positions**, missing values (**NaN**) were replaced with **"1"**, as it does not make sense for a job listing to have no available positions. This ensures completeness and accuracy in the data.

Experience:

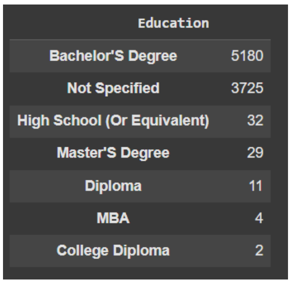
It was recorded in formats like **"from X to Y years"** or **"more than X years."** We extracted and kept only the **minimum experience value** (**min**). This is because applicants typically focus on the minimum required experience when applying for jobs, while those who meet the maximum experience limit are more likely to seek higher-level positions.

Career_Level:

It contained **six distinct categories**, all of which were well-defined and correctly formatted. Since there were no inconsistencies or missing values requiring adjustments, no cleaning was needed for this column.

Education:

It contained well-defined categories with no significant inconsistencies. Since all values were correctly formatted and meaningful, no cleaning was required for this column.



Education	
Bachelor'S Degree	5180
Not Specified	3725
High School (Or Equivalent)	32
Master'S Degree	29
Diploma	11
MBA	4
College Diploma	2

Salary:

It mostly contained "Confidential" or unspecified values. Additionally, there were errors where some records included **"Male," "Female," "Preferred Male," and "Preferred Female"** instead of salary values. These entries indicated gender preferences for certain positions rather than salary information.

To address this, we created a new **Gender** column to capture these details, categorizing entries as **Male, Female, Preferred Male, Preferred Female, and Both**. Meanwhile, in the **Salary** column, we replaced these incorrect values with **NaN** to ensure data consistency.

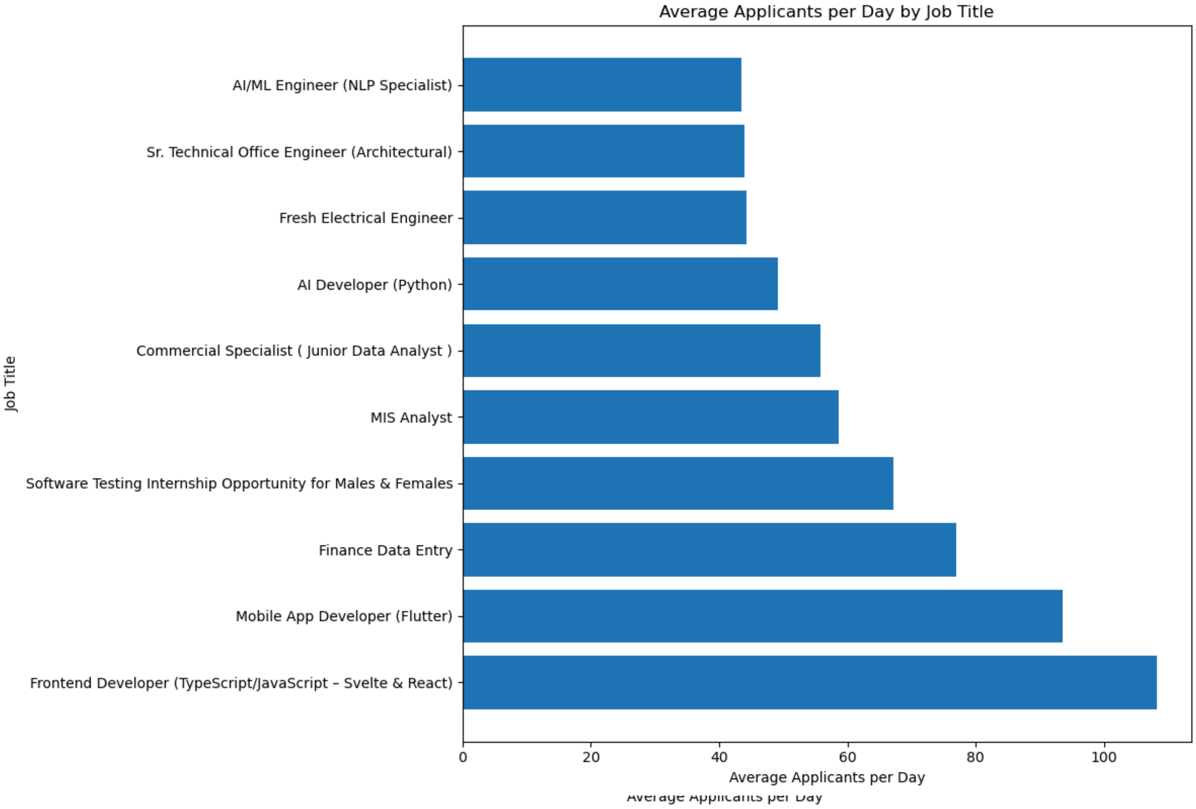
skills:

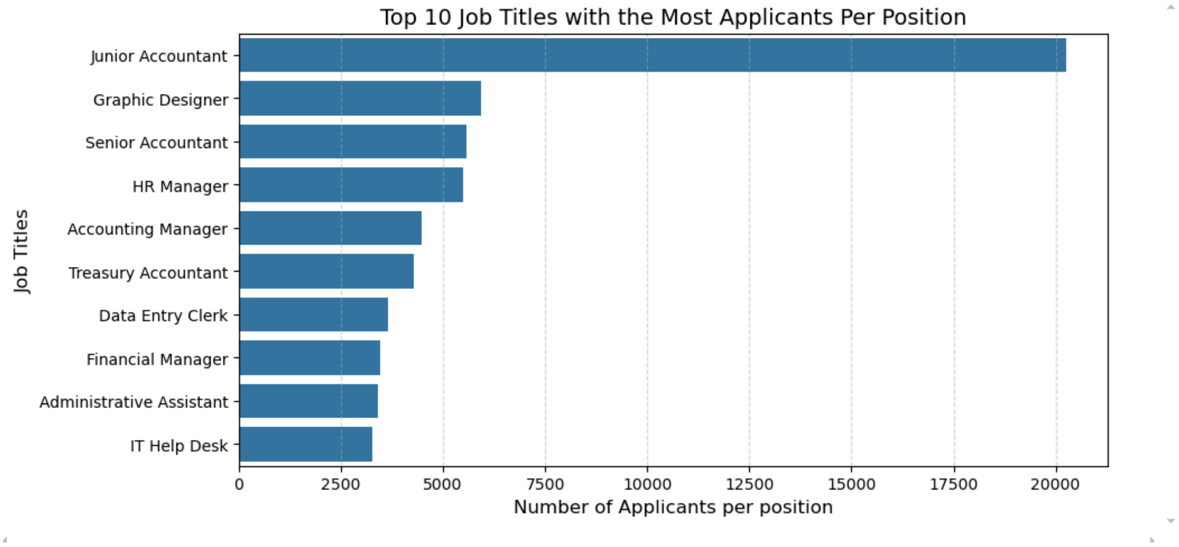
All skills were converted to lowercase and transformed into a list format for better processing. We focused on the top 200 skills, as the total number of unique skills was approximately 73,000, which is an extremely large value that cannot be easily visualized or processed efficiently.

Visualisation

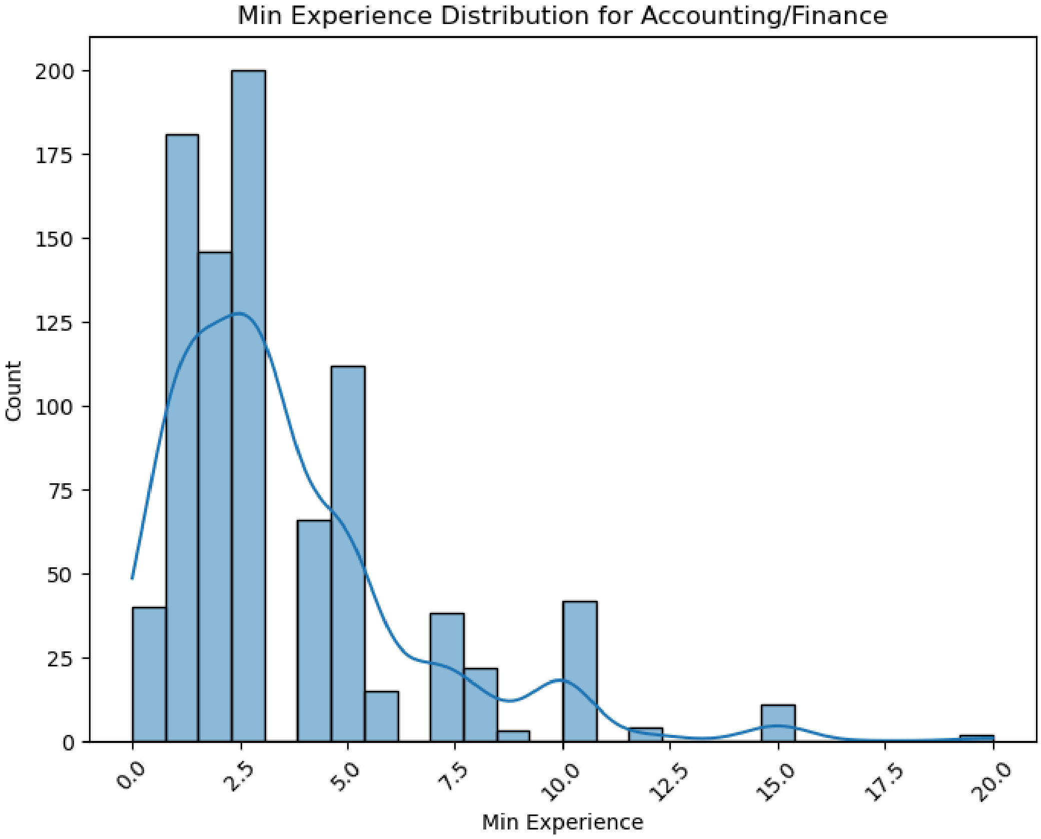
What are the positions/job categories most people apply for?

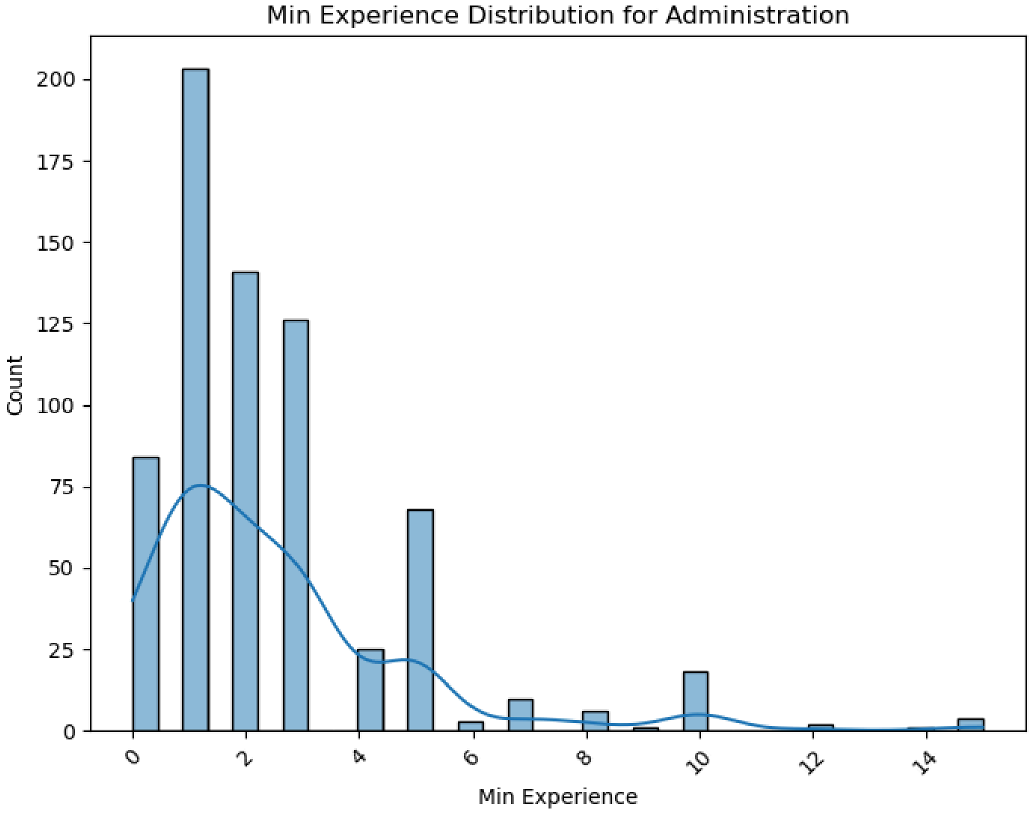
To account for the amount of time a job has been posted for, applicants per day was the chosen metric for this analysis instead of the total number of applicants.

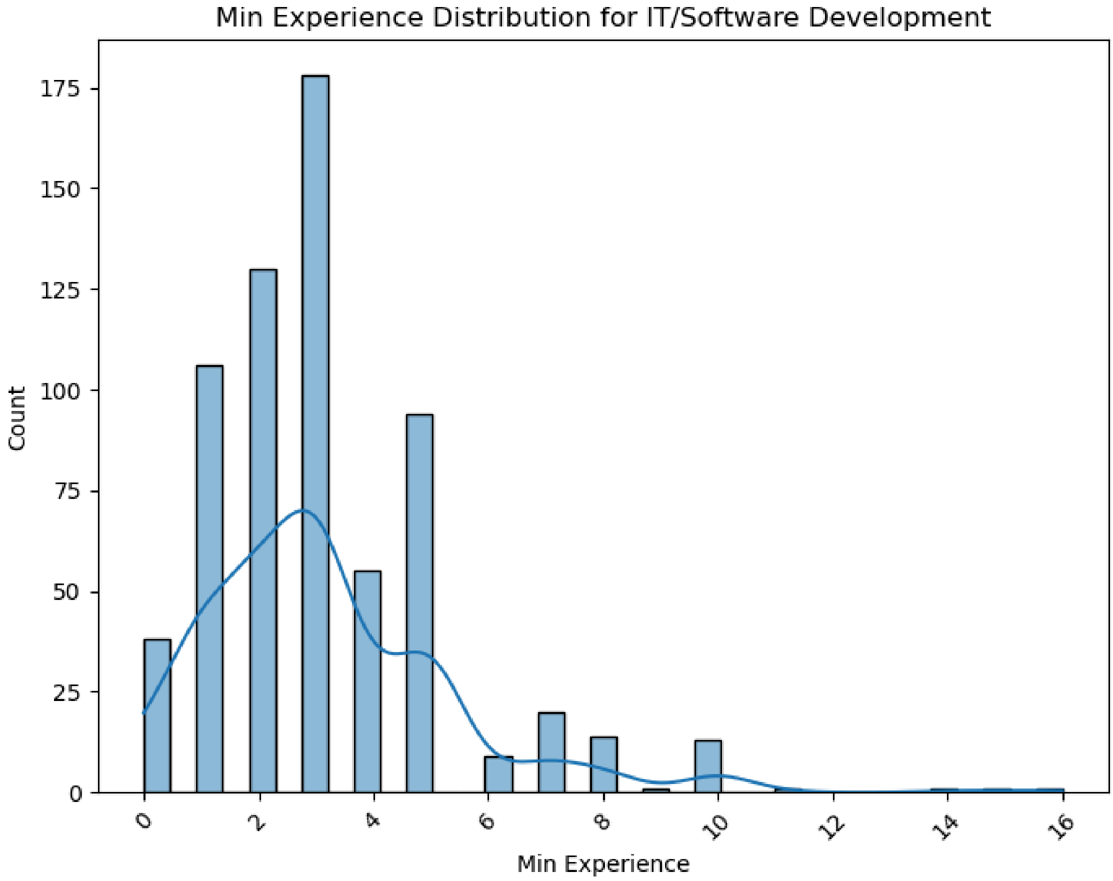


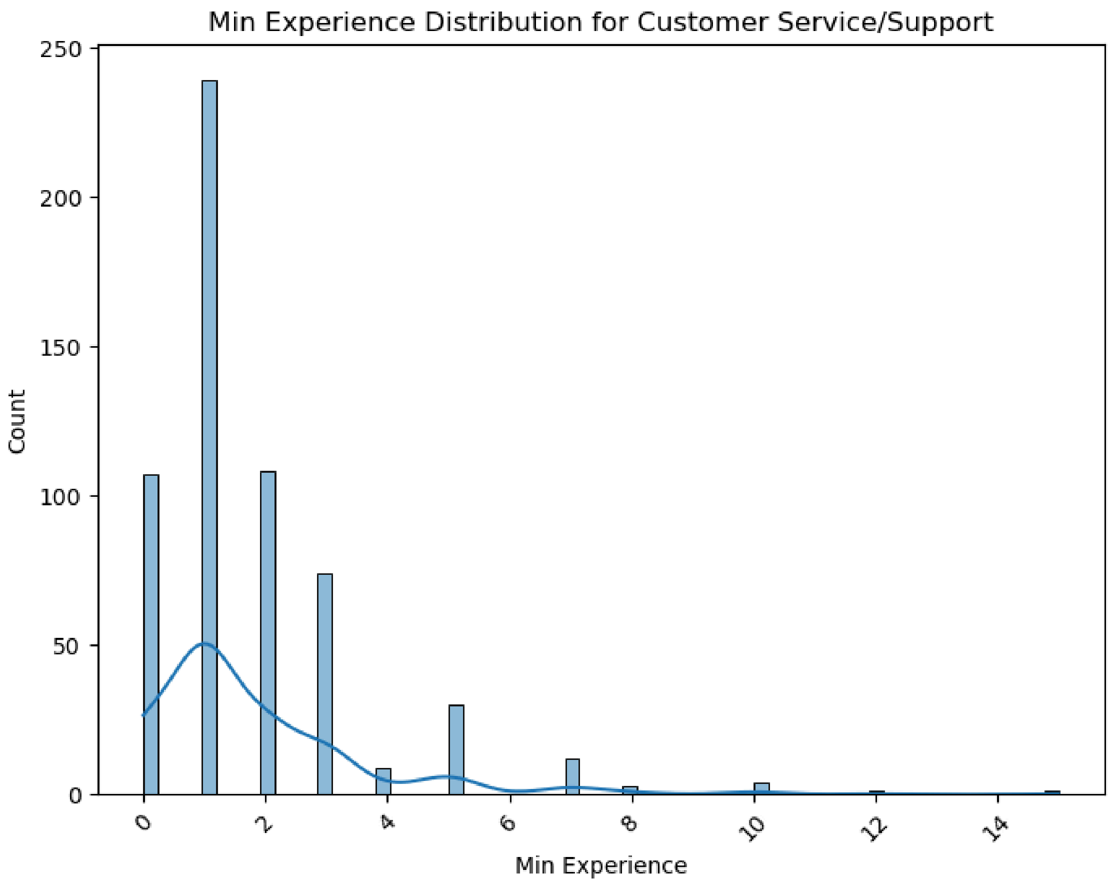


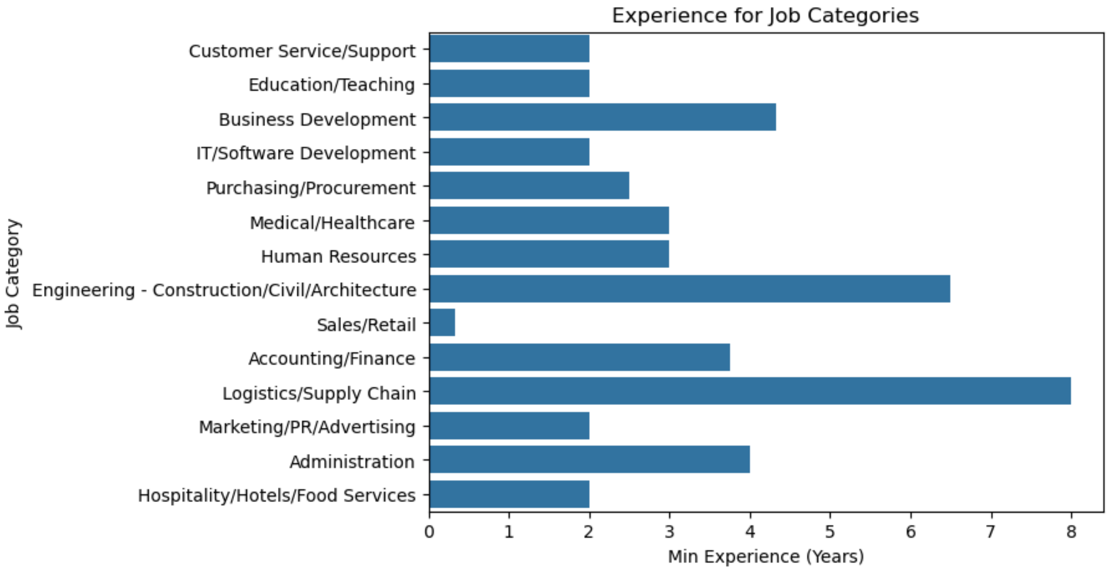
What is the distribution of the minimum required experience for each job category?




Min Experience Distribution for Administration
Count
0
25
50
75
100
125
150
175
200
0
2
4
6
8
10
12
14
Min Experience

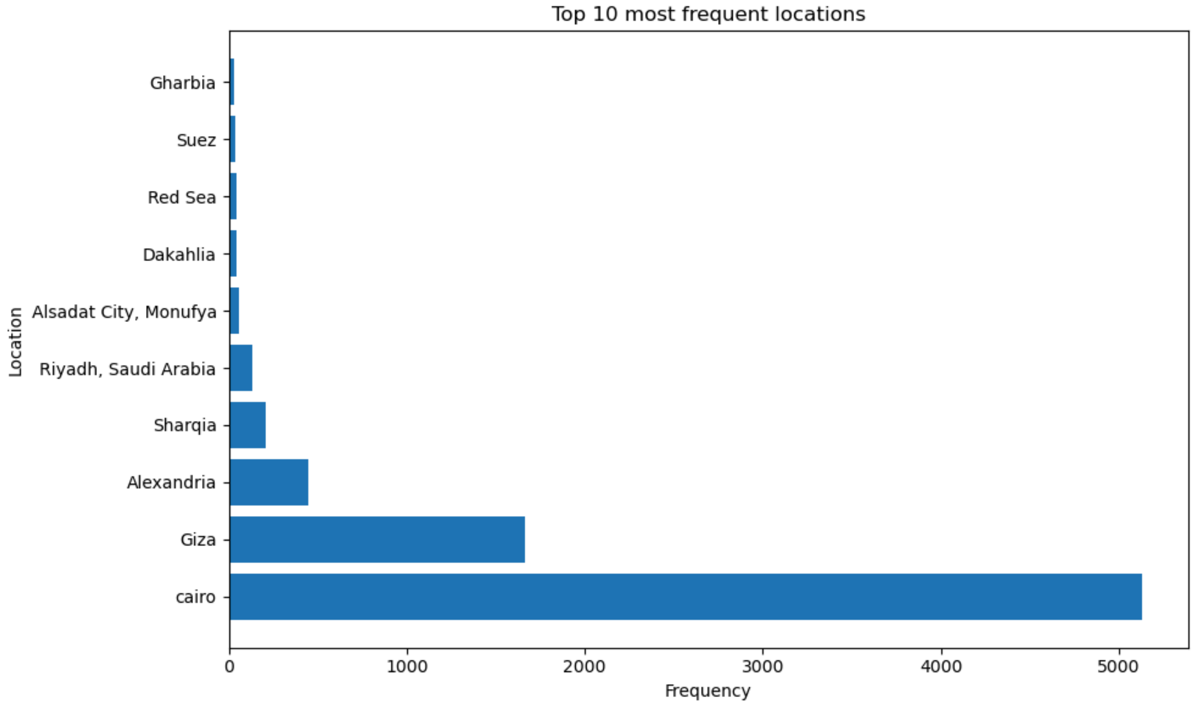

Min Experience Distribution for IT/Software Development
Count
0
25
50
75
100
125
150
175
0
2
4
6
8
10
12
14
16
Min Experience



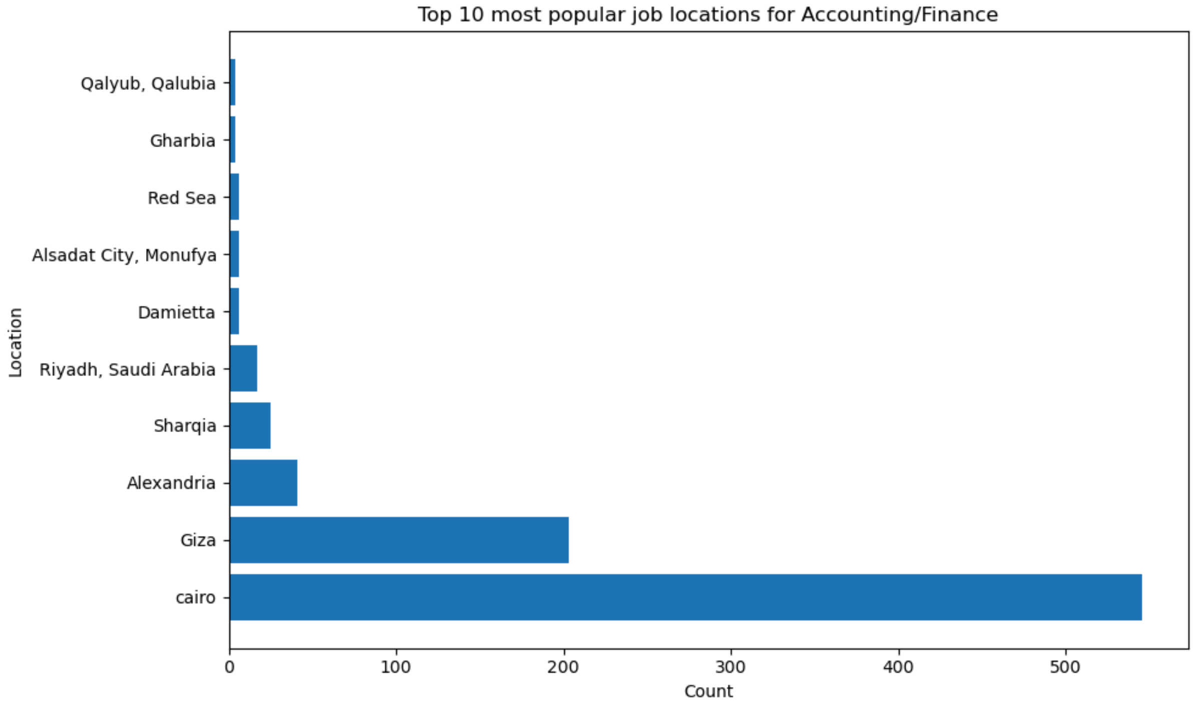


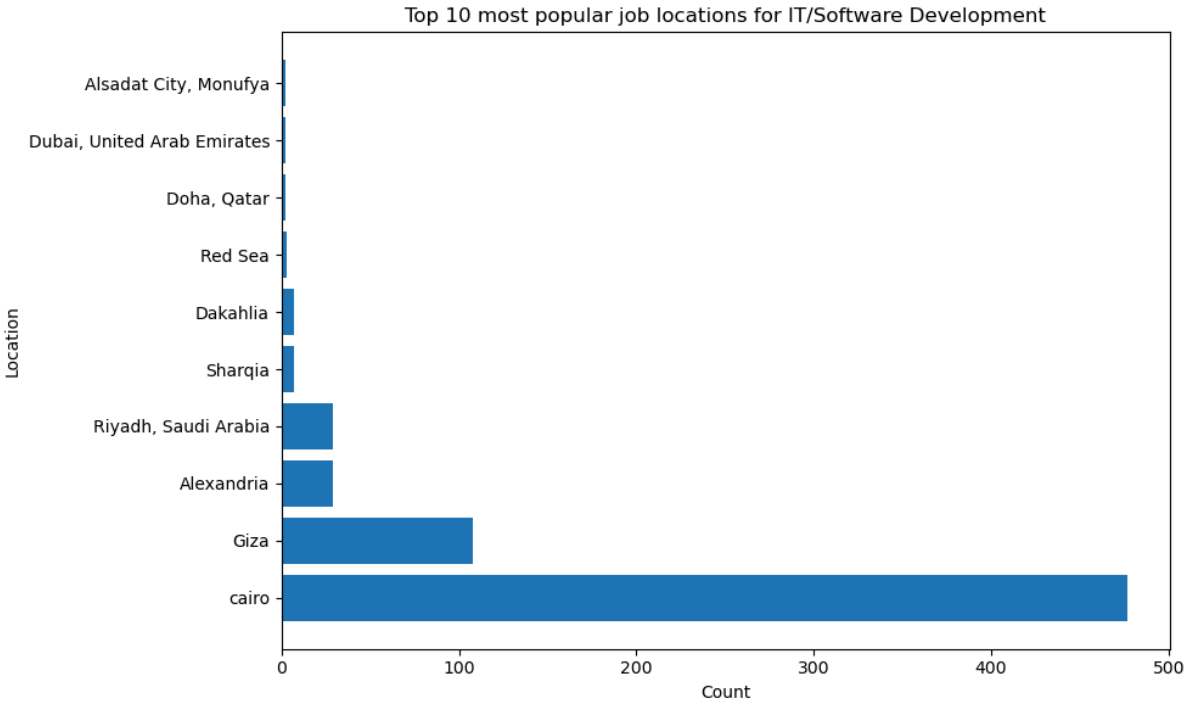
It appears that the required minimum experience ranges from 1 to 3 according to job category.

Where are most jobs concentrated?

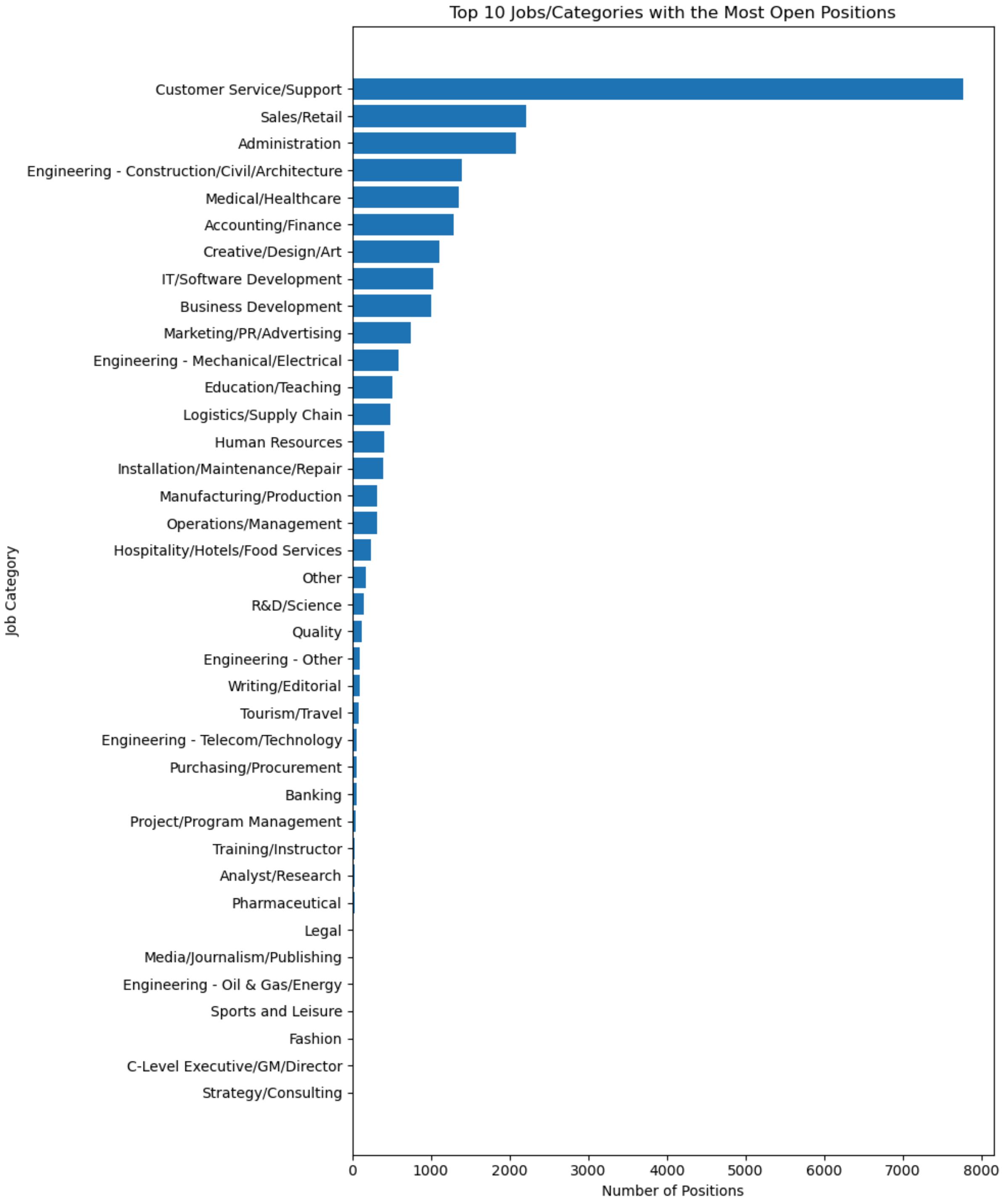


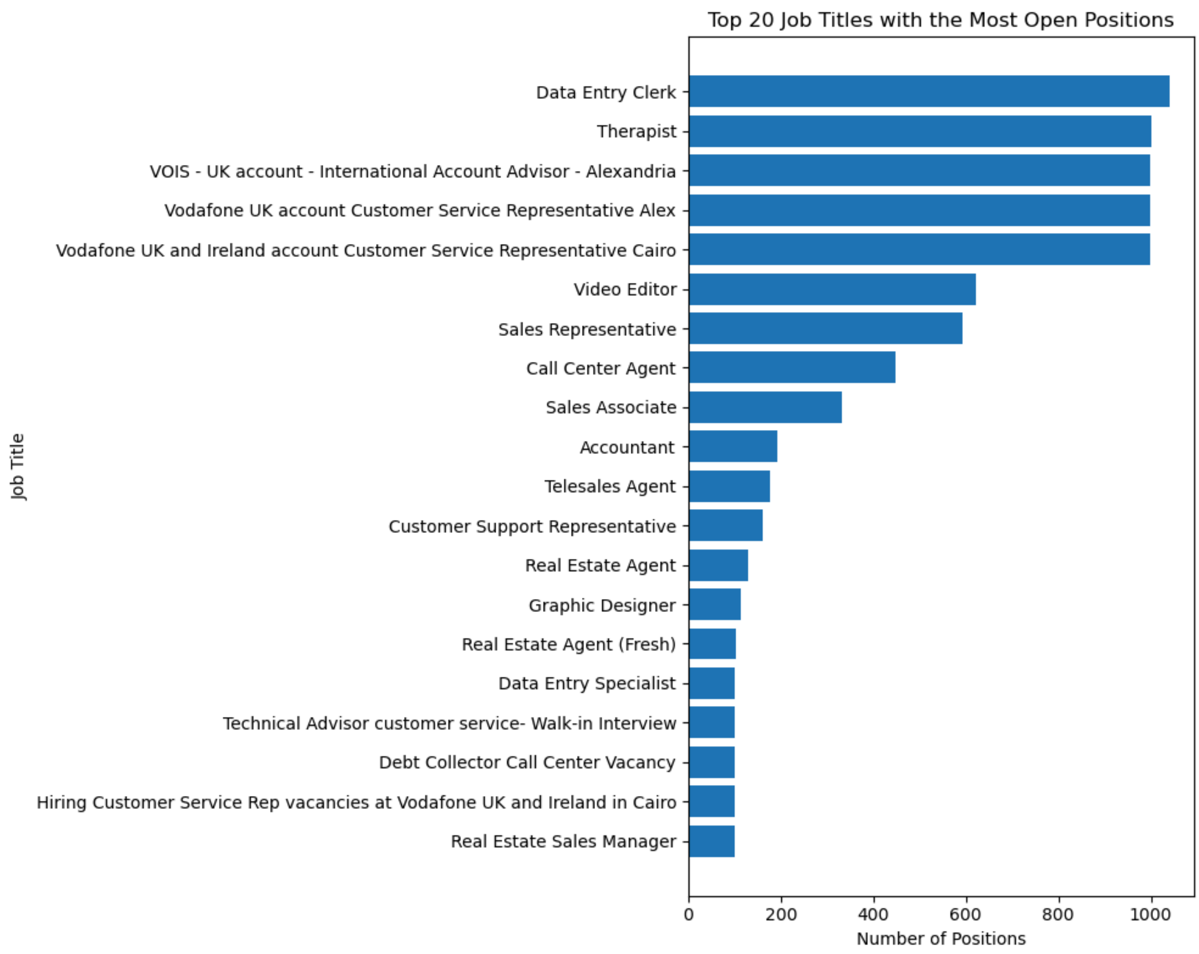
If we instead look for specific job categories, they show the same distribution with minor differences.




Top 10 most popular job locations for IT/Software Development
Location
Alsadat City, Monufya
Dubai, United Arab Emirates
Doha, Qatar
Red Sea
Dakahlia
Sharqia
Riyadh, Saudi Arabia
Alexandria
Giza
cairo
0
100
200
300
400
500
Count

What jobs / categories have the most open positions?



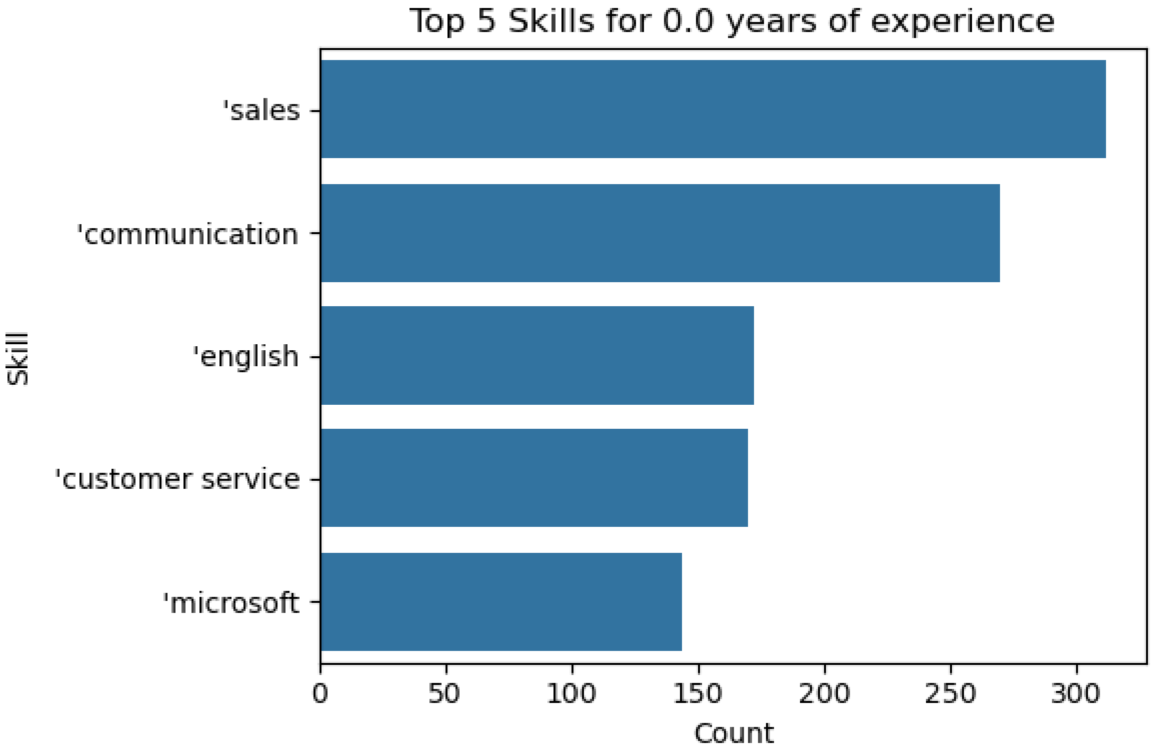


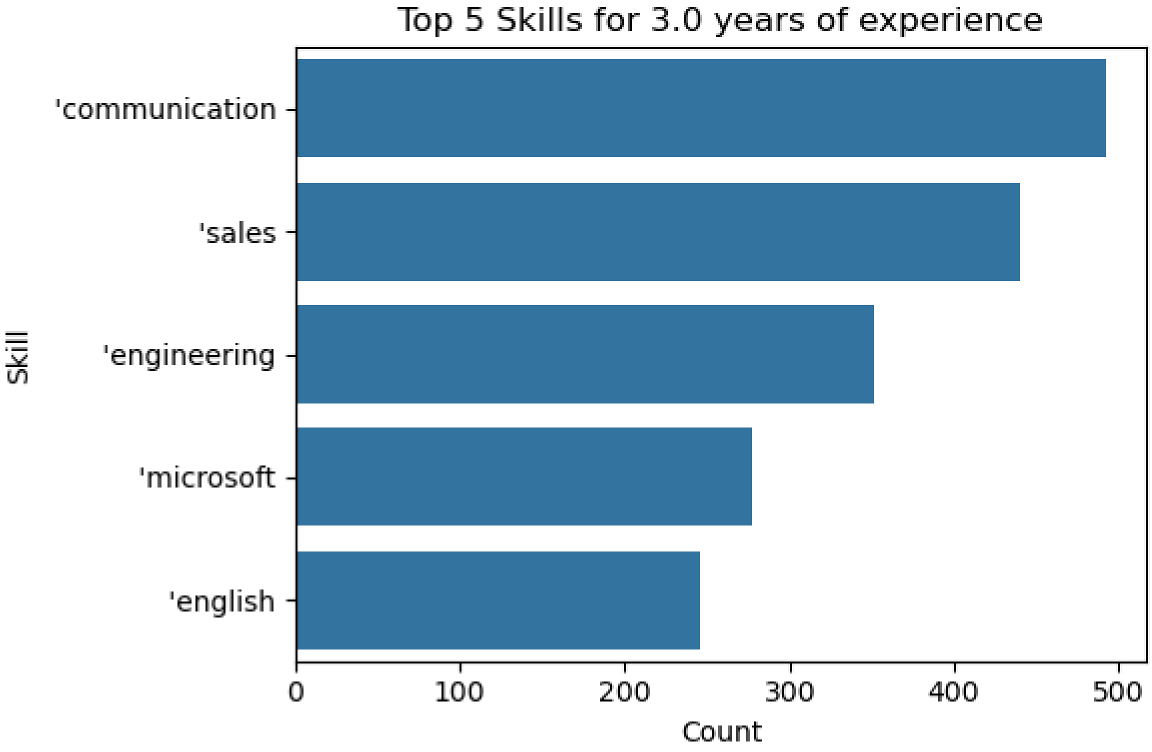
We do need a lot of therapists.

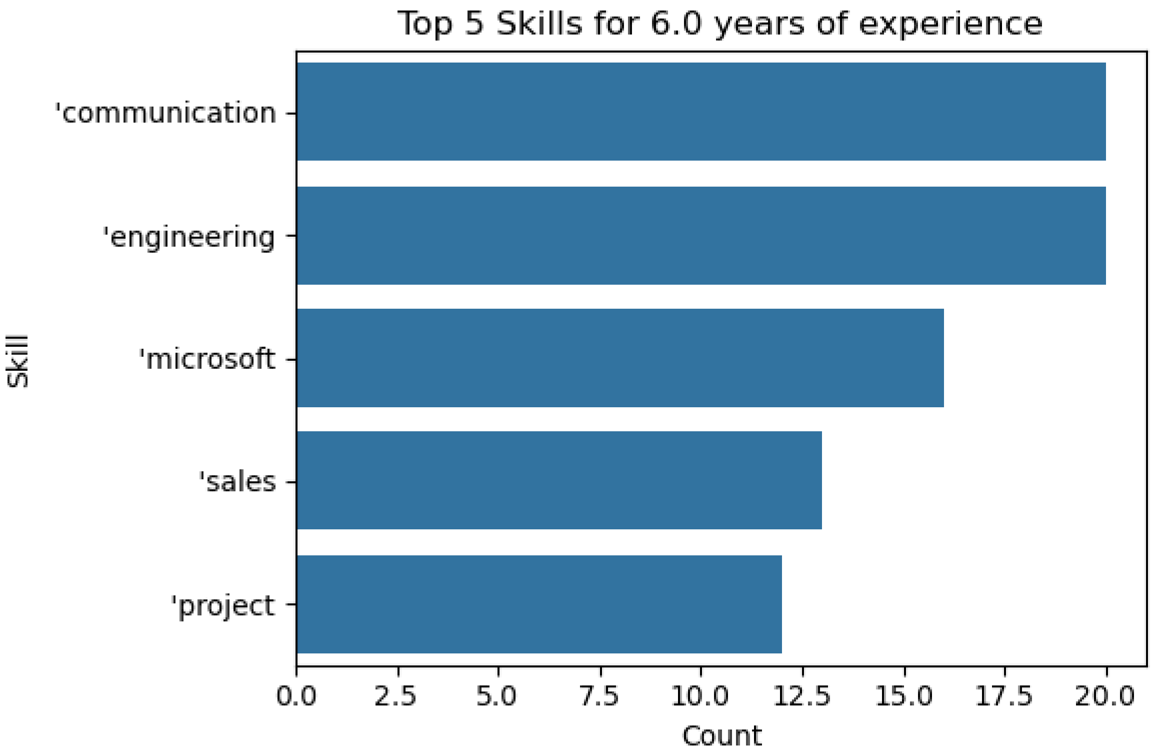
Confirm the education distribution (according to category)

There was a better graph somewhere.

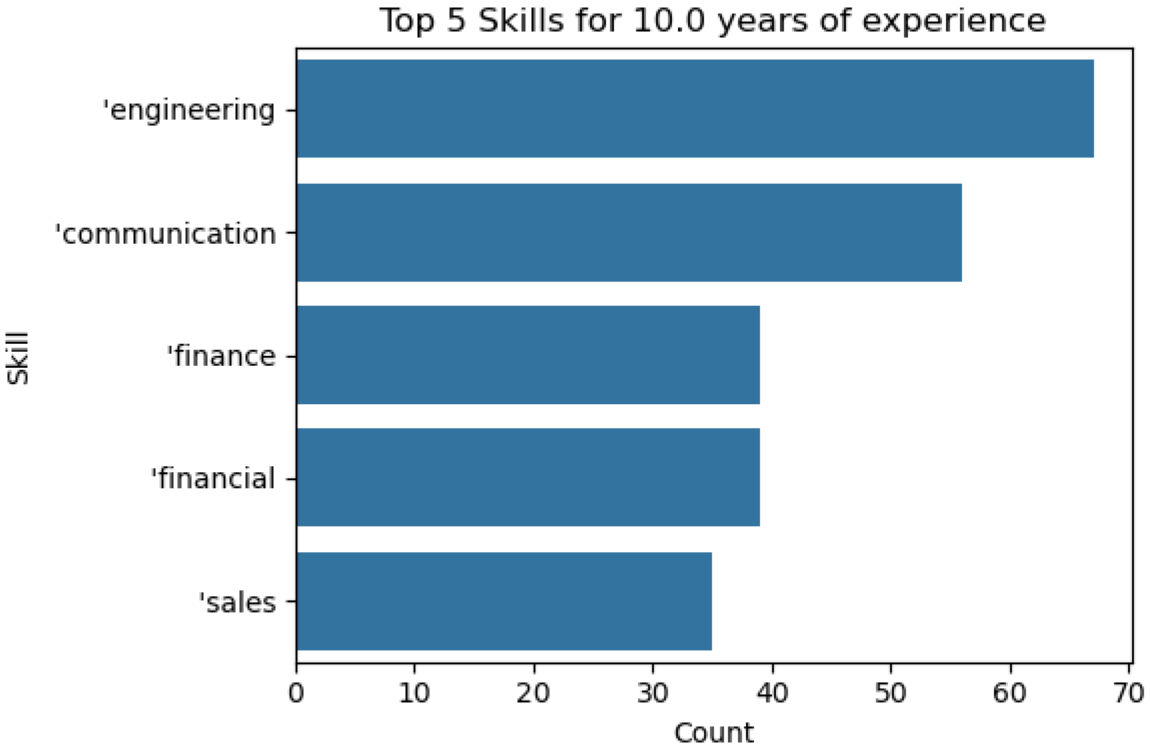
What are the most wanted skills according to experience / career level?



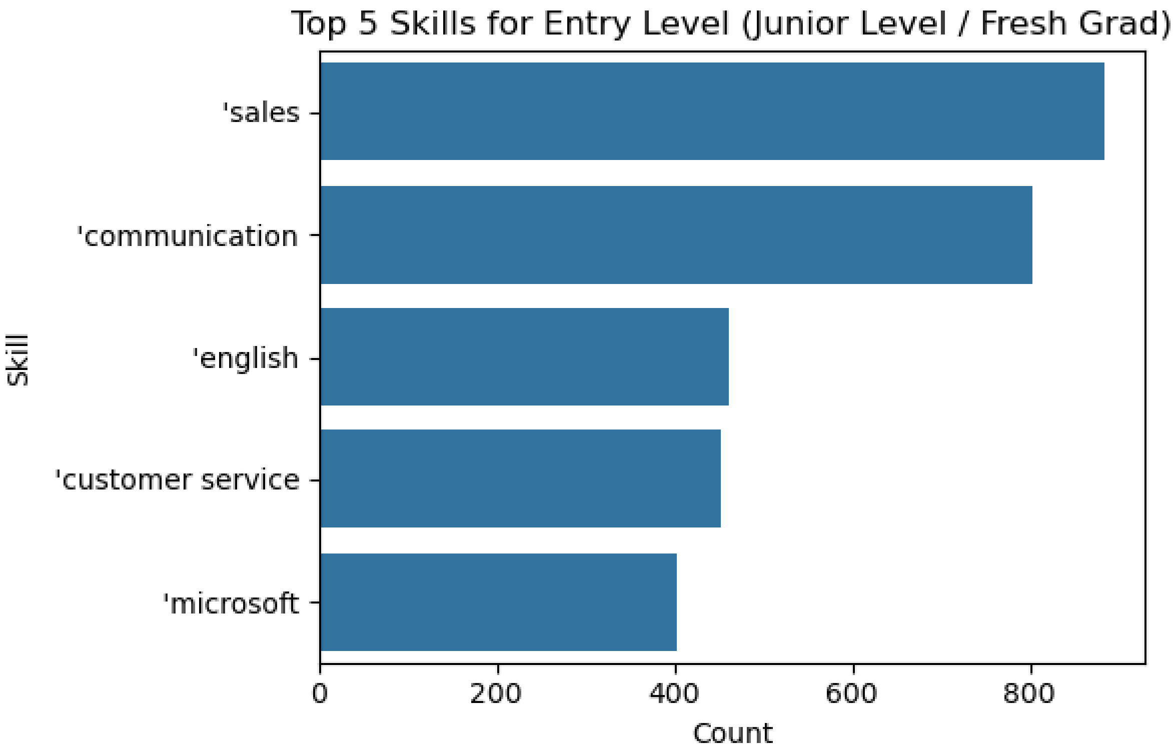




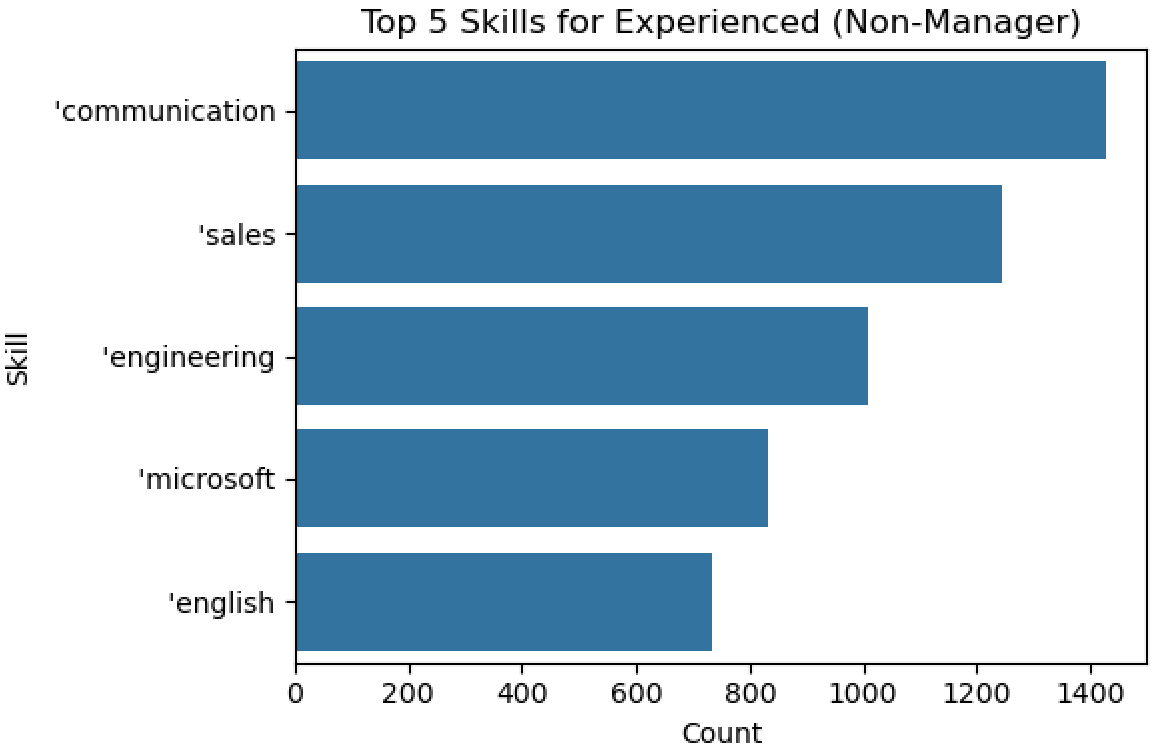
Top 5 Skills for 6.0 years of experience



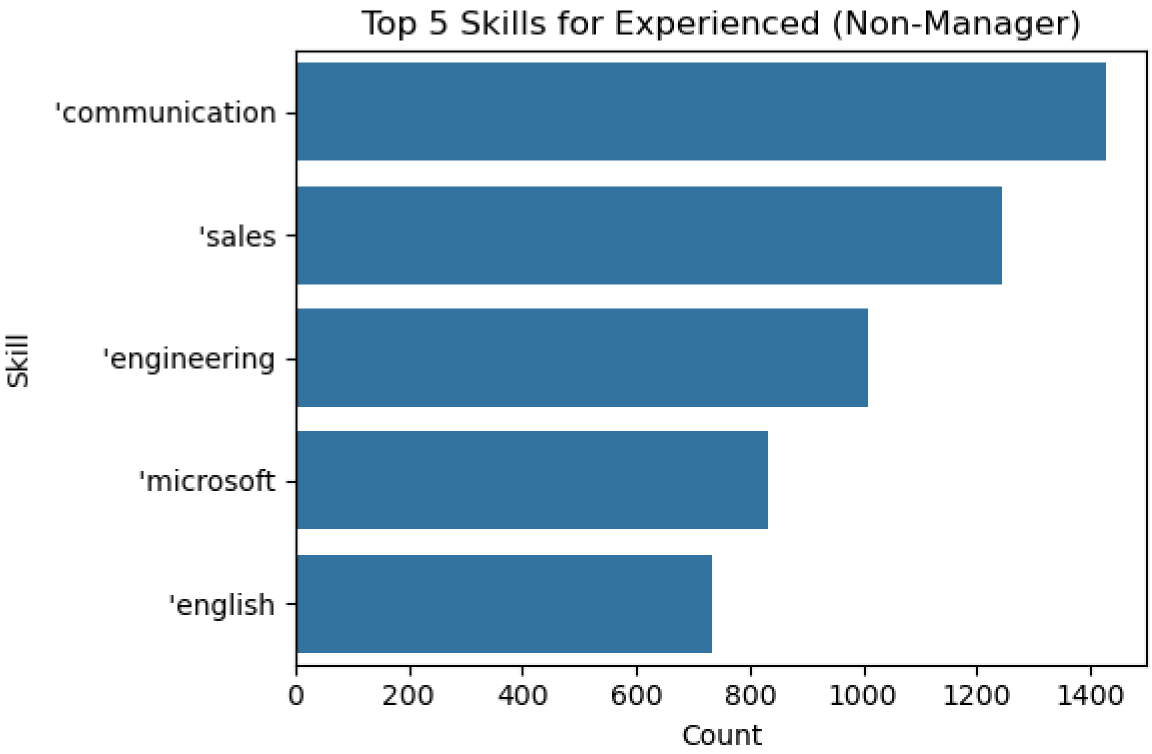
Top 5 Skills for 10.0 years of experience

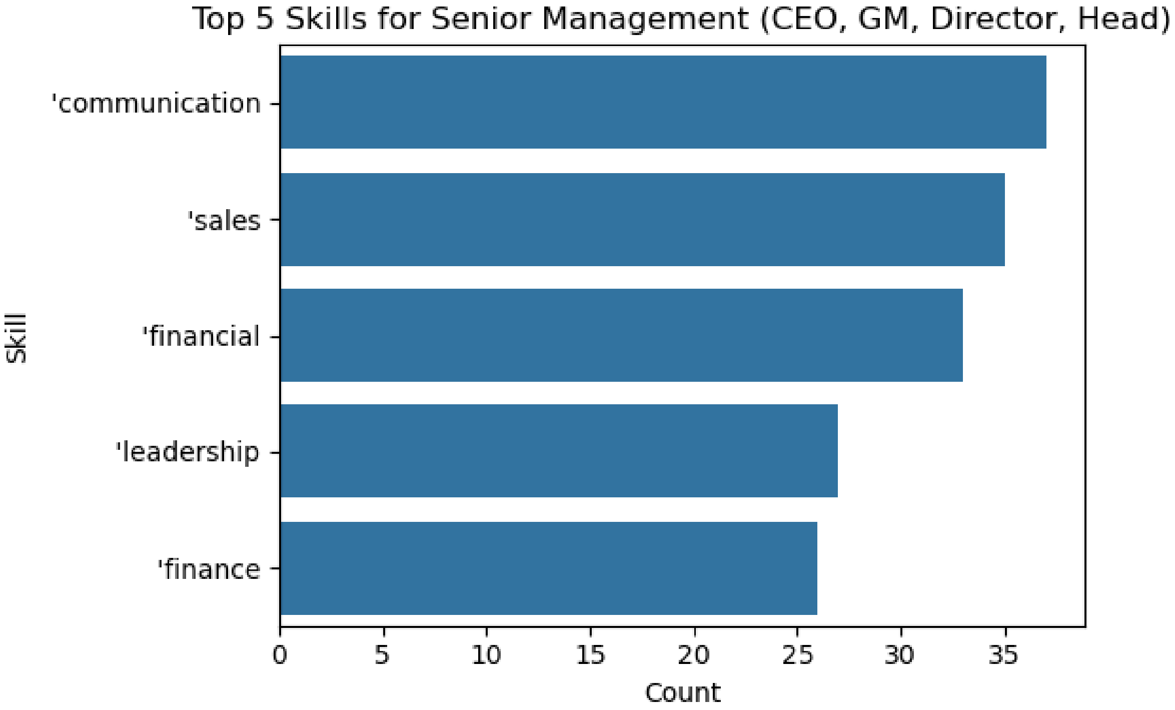


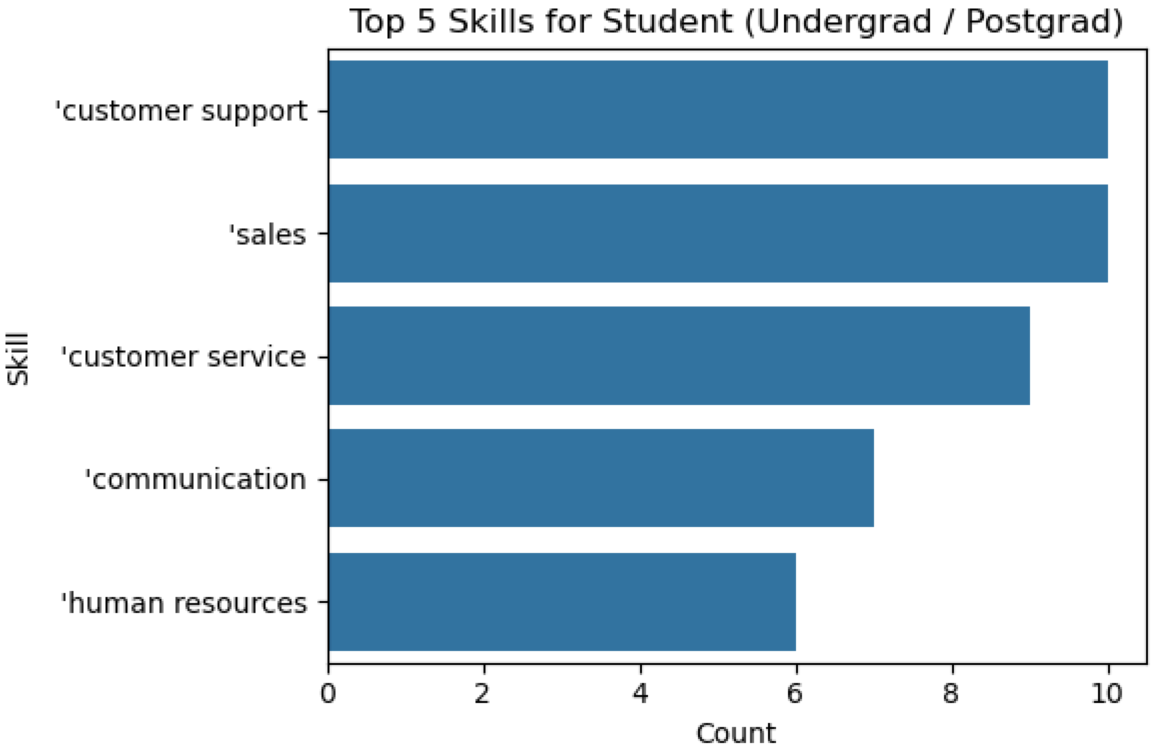
Top 5 Skills for Entry Level (Junior Level / Fresh Grad)
Skill
'sales
'communication
'english
'customer service
'microsoft
0
200
400
600
800
Count



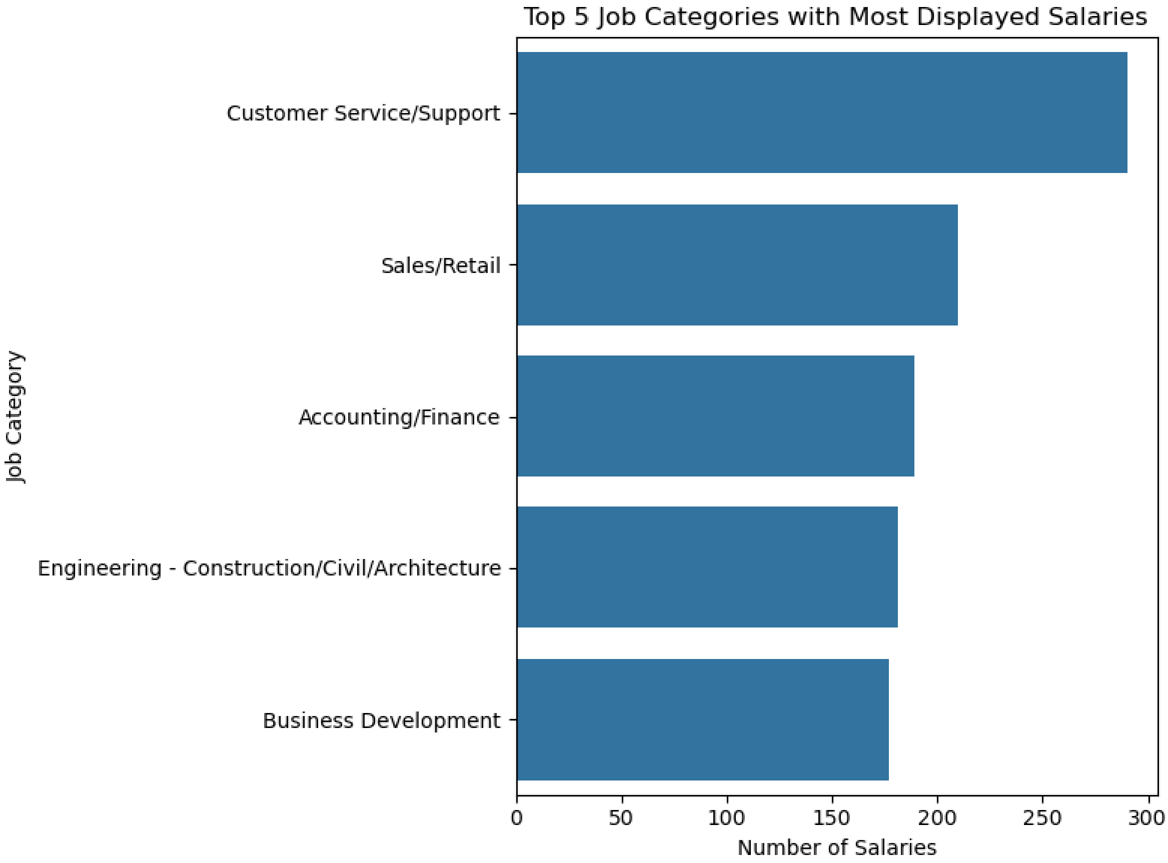
Top 5 Skills for Experienced (Non-Manager)
Skill
'communication
'sales
'engineering
'microsoft
'english
0
200
400
600
800
1000
1200
1400
Count

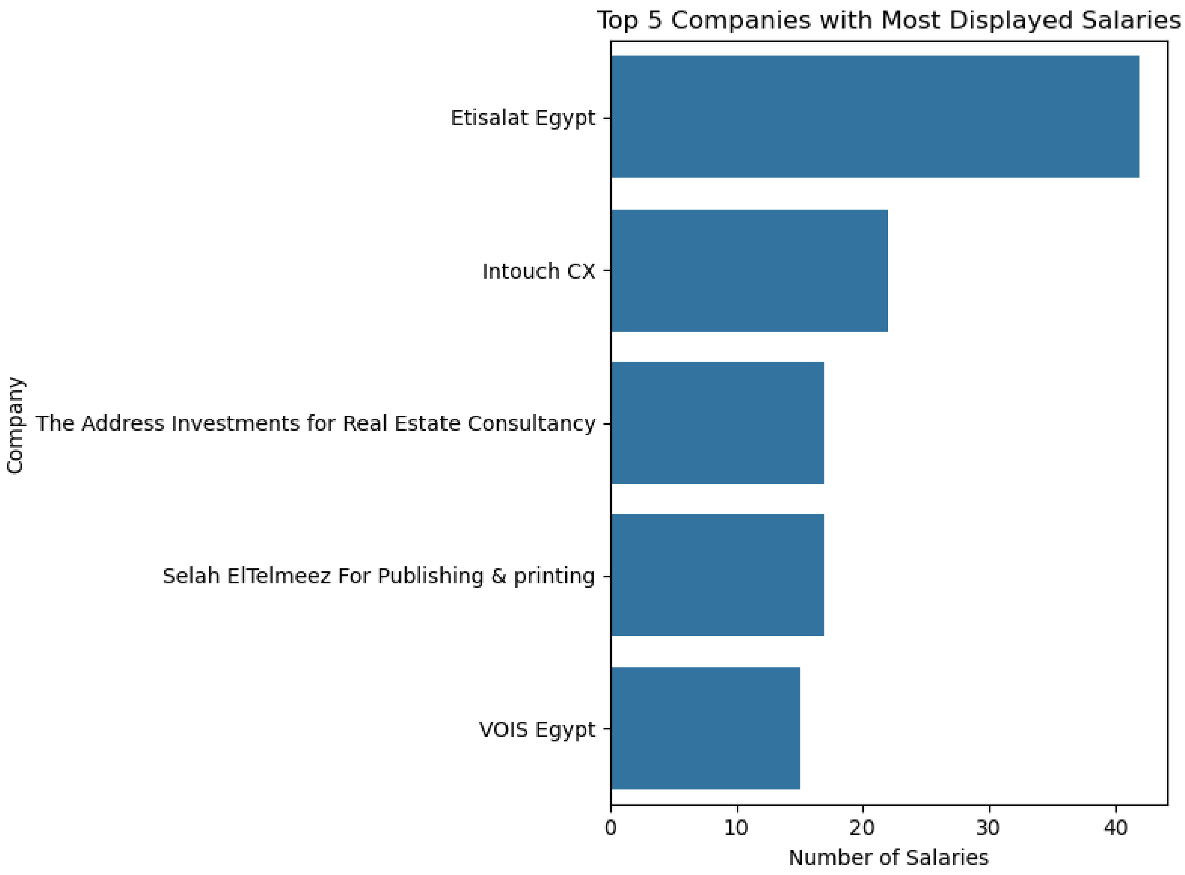

Top 5 Skills for Experienced (Non-Manager)
Skill
'communication
'sales
'engineering
'microsoft
'english
0
200
400
600
800
1000
1200
1400
Count


Top 5 Skills for Senior Management (CEO, GM, Director, Head)
Skill
'communication
'sales
'financial
'leadership
'finance
0
5
10
15
20
25
30
35
Count



What job categories / companies actually have salaries?



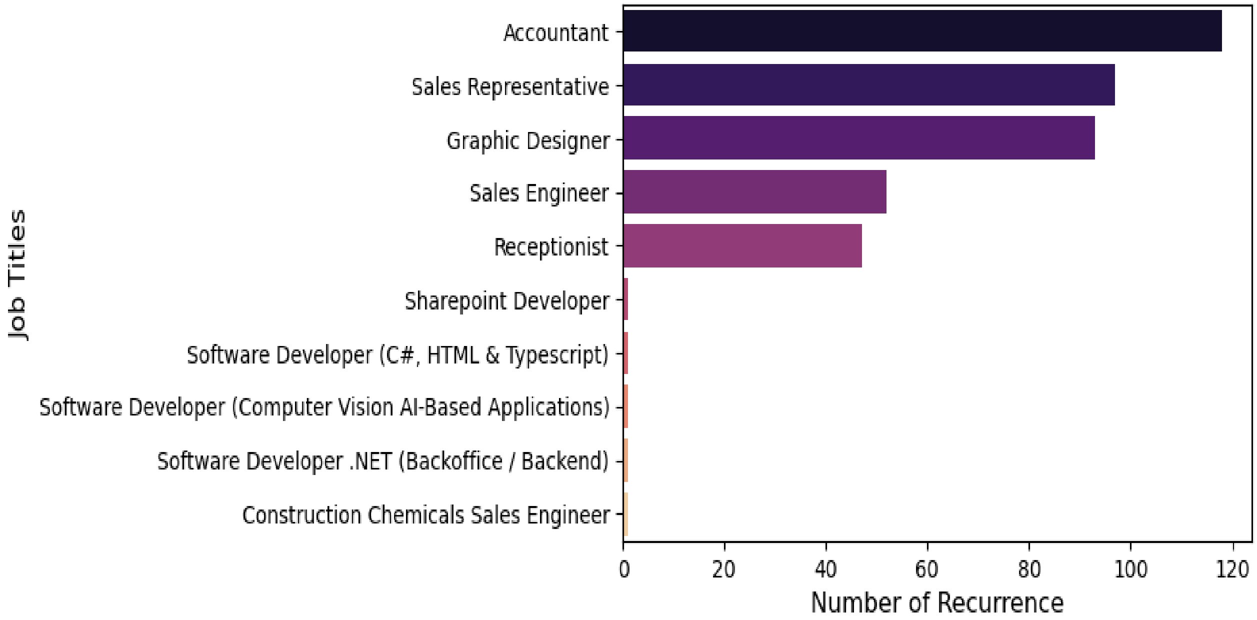


Career Level vs experience

There was a better graph somewhere

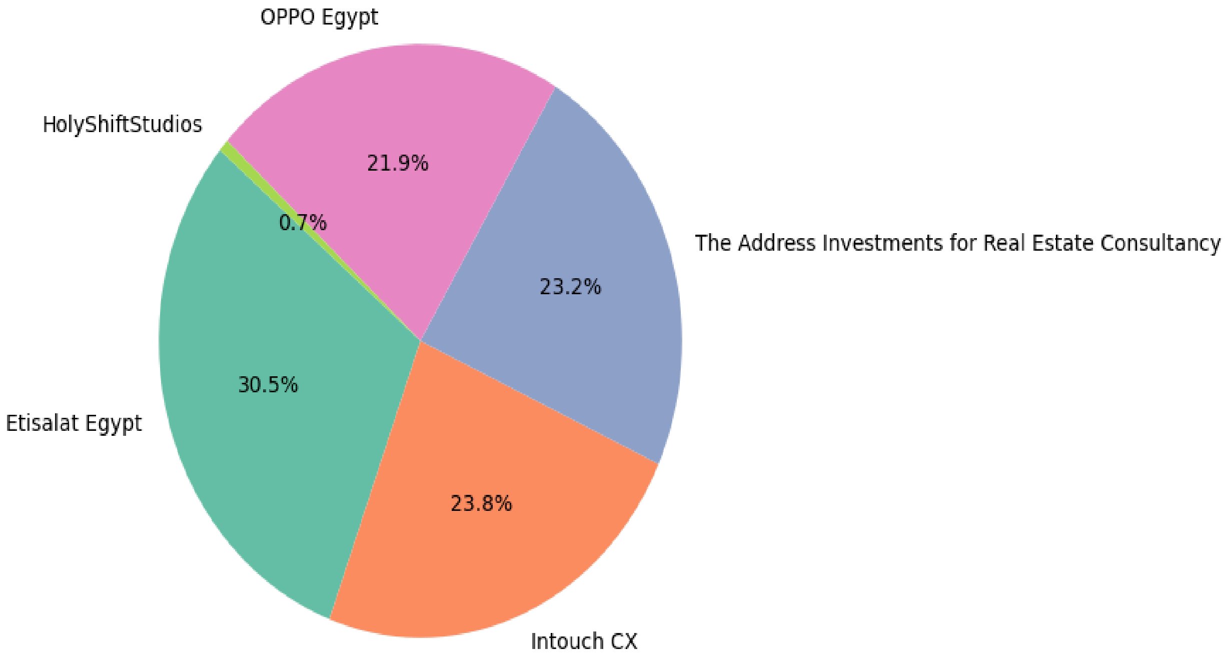
Khalid——

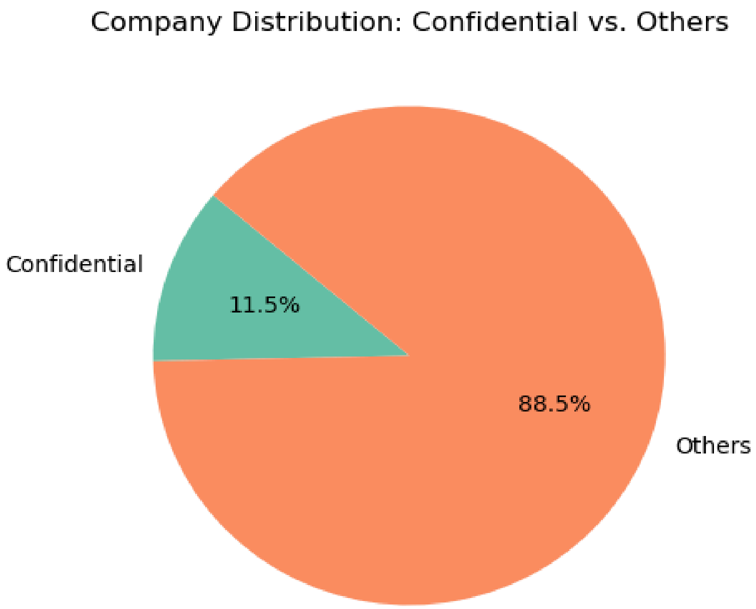
1- Most and Least Frequent Job Titles



The data indicates that the **Accountant** position is the most frequently required on the website, followed by **Sales** roles. In contrast, **Graphic Designers** and **Developers** positions appear much less frequently.

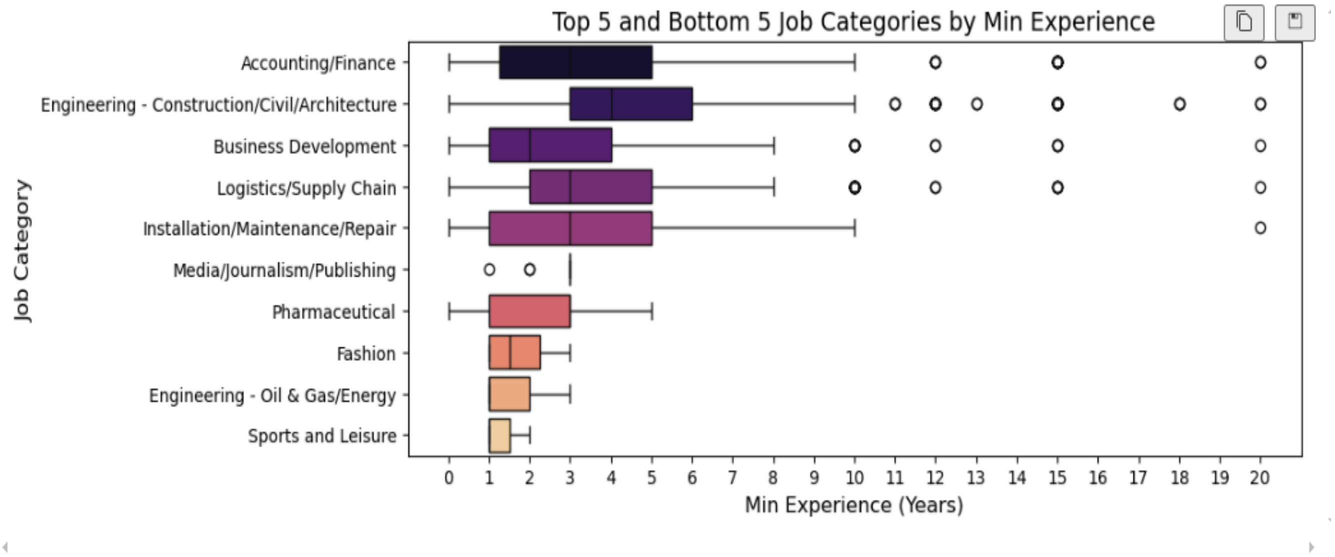
2- Most and Least Frequent Companies





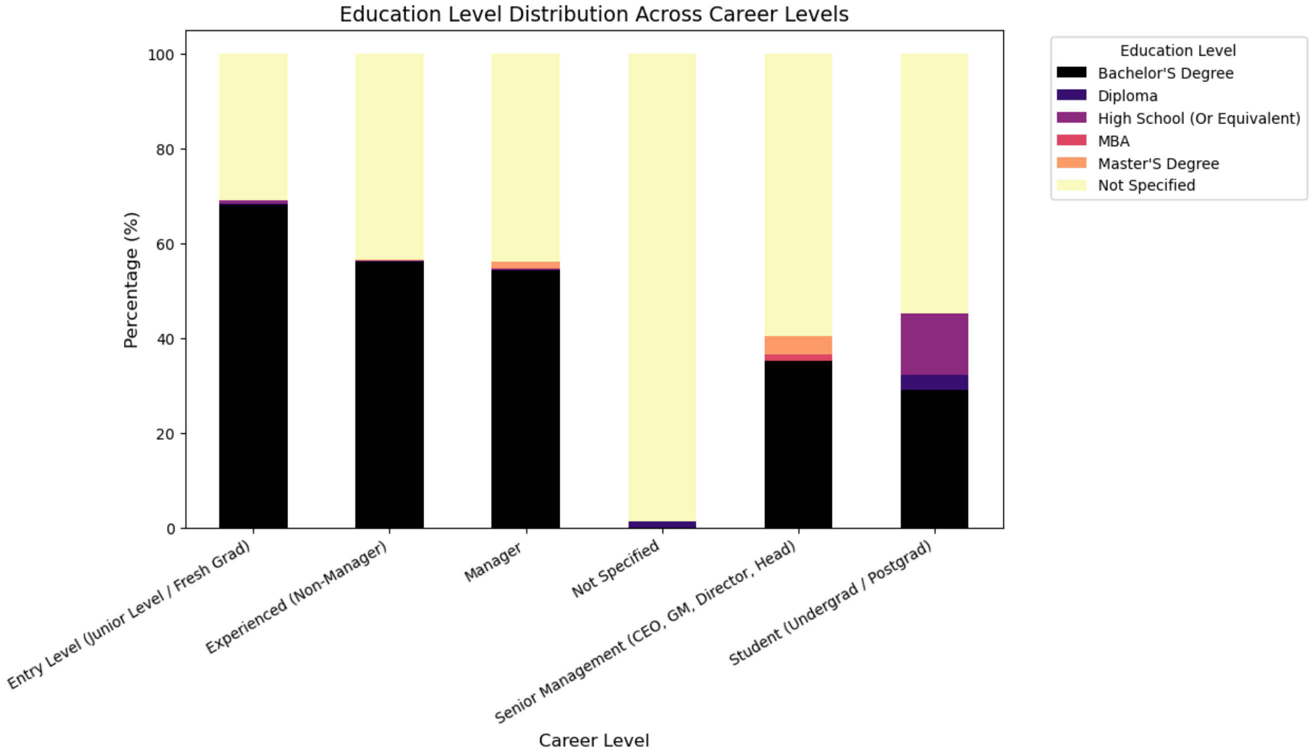
The plots indicate that **Etisalat** has the highest number of job postings on the website, while **HolyShiftStudios** has the least. Additionally, a significant portion of companies choose not to disclose their names, as reflected by the high percentage of 'Confidential' entries in the company column.

3- Job Categories and Their Minimum Experience Requirements.



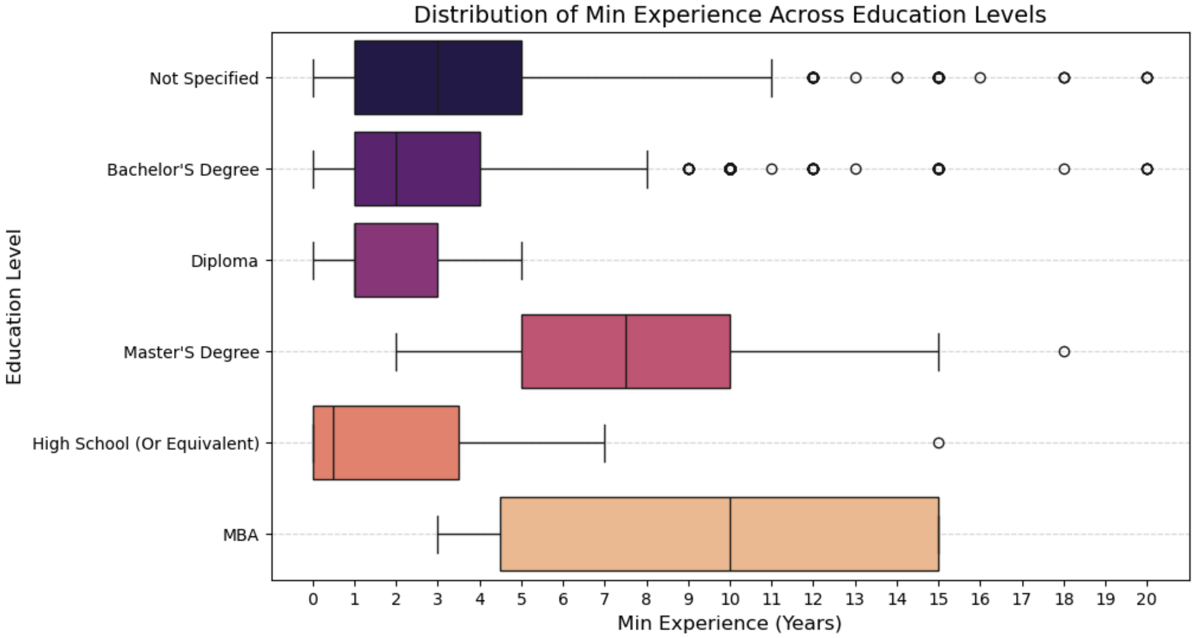
This plot displays the distribution of minimum experience required across the top 5 and bottom 5 job categories. **Accounting/Finance** and **Engineering - Construction/Civil/Architecture** require a broader range of experience, whereas categories like **Fashion** and **Sports and Leisure** tend to have lower experience requirements. Additionally, some job categories show significant outliers, indicating occasional positions that demand exceptionally high experience levels.

4- Relationship Between Career Level and Education Requirements.



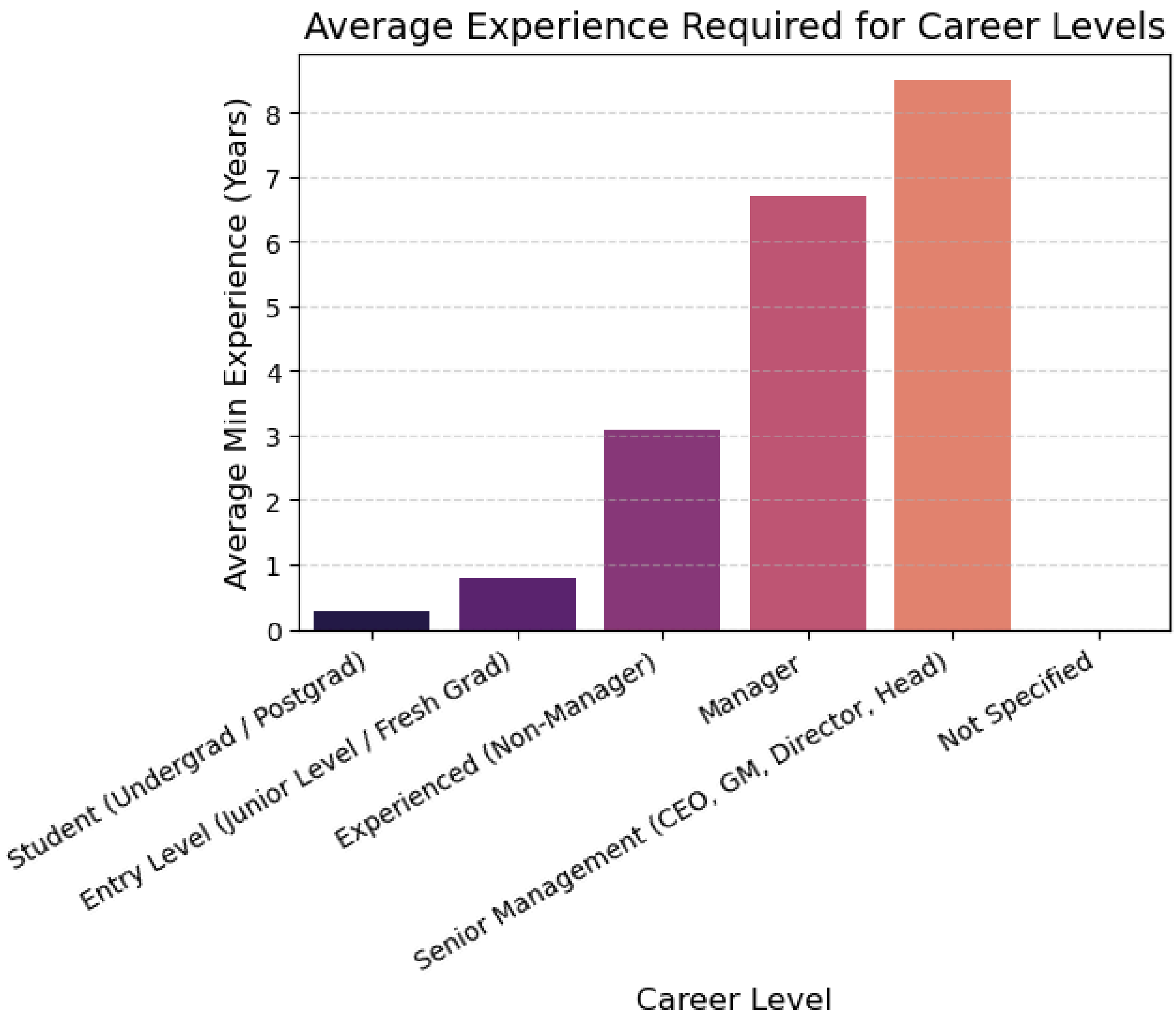
This chart illustrates the distribution of education levels across different career levels. **Bachelor’s degrees** dominate most career levels, especially at the **entry-level and experienced positions**. However, as career levels advance to **managerial and senior management roles**, there is a slight increase in **MBA and Master's degree holders**. Interestingly, a significant percentage of job postings have their education requirements listed as **'Not Specified'**, especially at the **Senior Management and Not Specified** career levels, indicating flexibility or lack of strict education criteria for these roles.

5- Does Higher Experience Correlate with Higher Education?



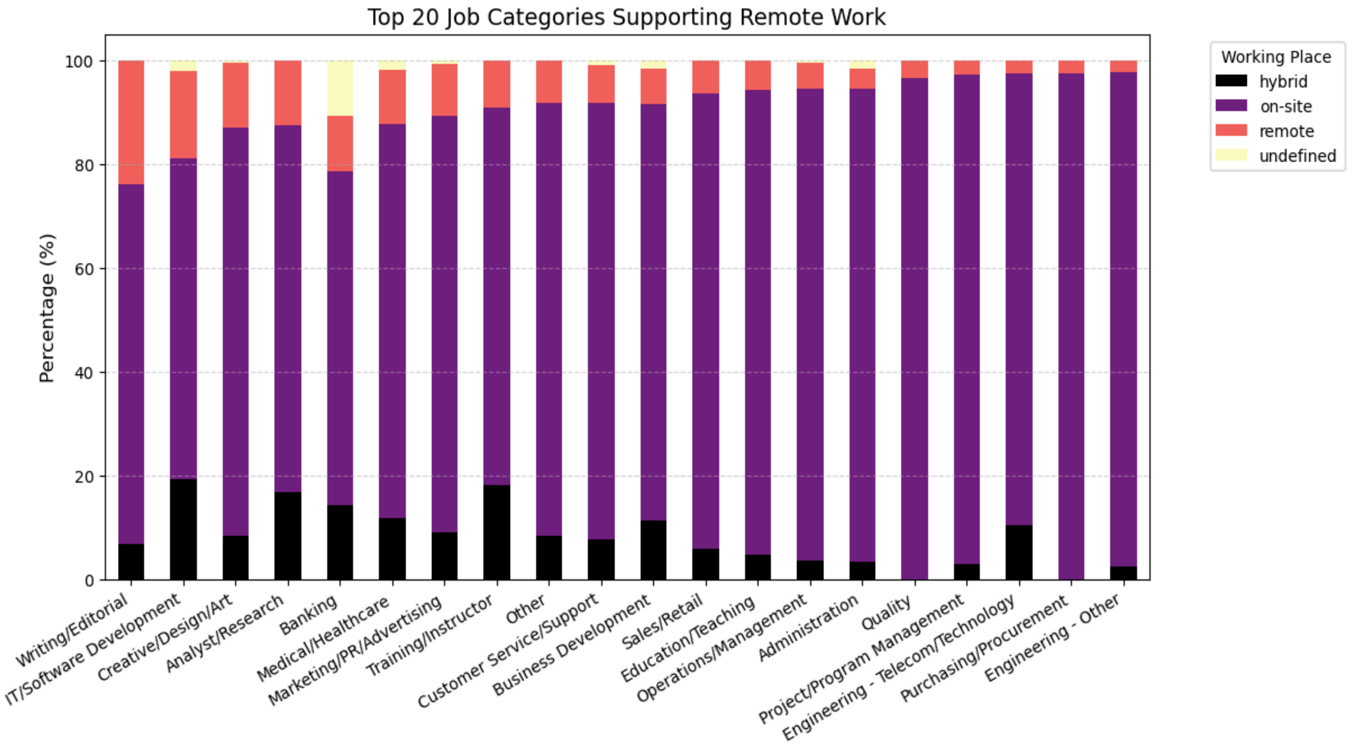
The plot indicates that **higher education levels, such as MBA and Master's degrees, generally require more years of experience**, with their median values being higher than those of Bachelor's and Diploma holders. However, there are also cases where **high school graduates and bachelor's degree holders have significant experience requirements**, likely due to industry-specific demands. The **"Not Specified"** category spans a wide range, suggesting that many job postings do not explicitly mention education requirements.

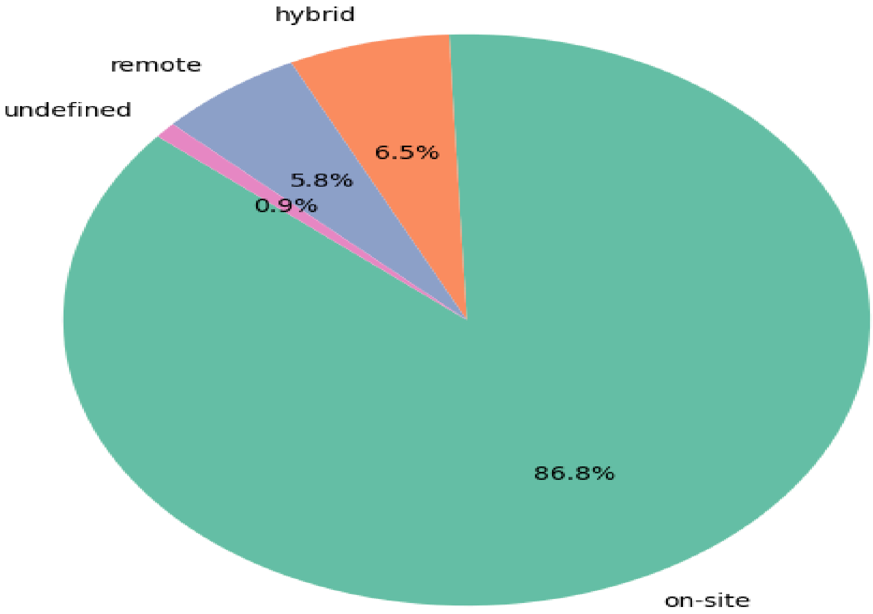
6- How Career Levels Correlate with Work Experience.



This bar chart illustrates the average minimum experience required for different career levels. As expected, higher career levels such as 'Manager' and 'Senior Management' demand significantly more years of experience compared to entry-level or student positions. The 'Not Specified' category has no specified required experience.

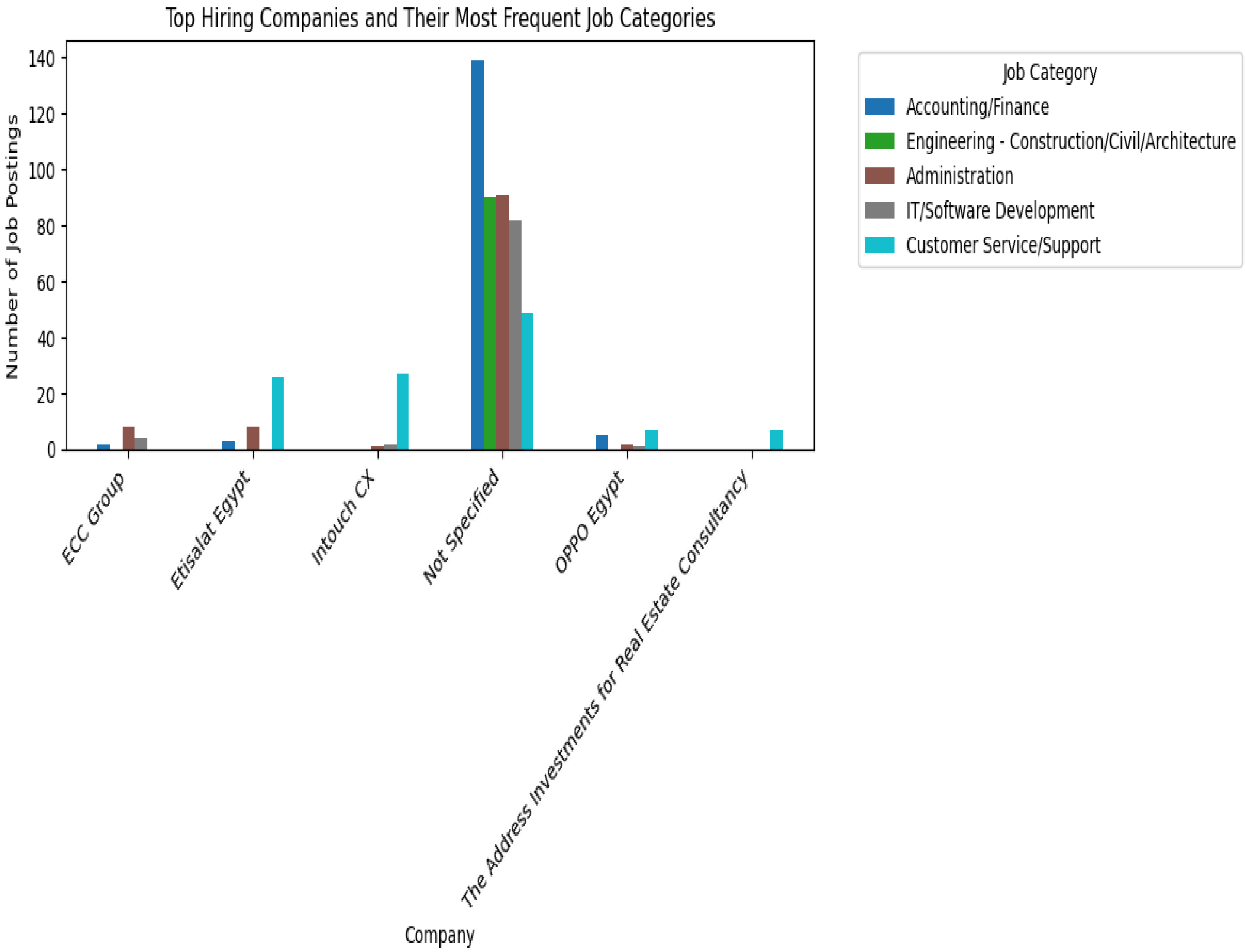
7- How Job Categories Influence Work Location (Remote, Hybrid, On-Site).





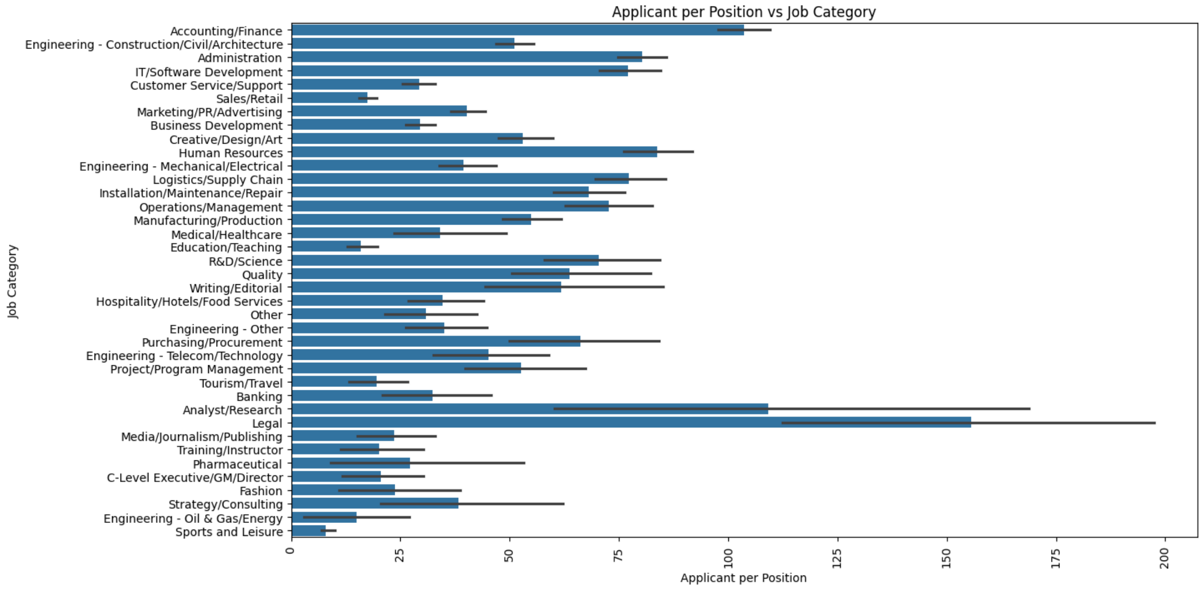
The bar chart displays the distribution of remote, hybrid, on-site, and undefined work arrangements across the top 20 job categories. **IT/Software Development** and **Writing/Editorial** have the highest remote/hybrid flexibility, while **Engineering, Administration, and Customer Service** remain largely on-site. Fields like **Creative/Design and Research** show a mix of work modes. The chart highlights industries best suited for remote work, overall the most jobs required on-site workers and a little required remote, and hybrid.

8- Top Hiring Companies and Their Most In-Demand Job Categories.



The "Not Specified" category has the highest number of job postings across multiple fields, followed by other companies with varying demands in fields such as Accounting/Finance, Engineering, Administration, IT/Software Development, and Customer Service/Support. This visualization highlights the most in-demand job categories within leading hiring companies.

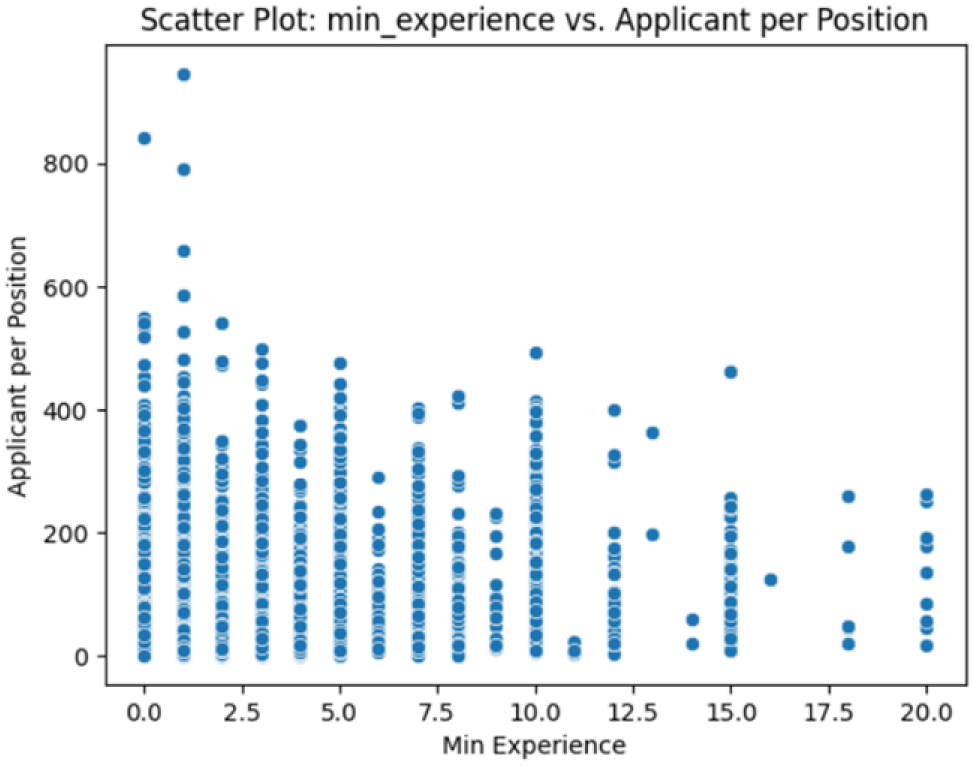
1. Which job categories have the highest competition in terms of applicants per position?



Insights:

This graph shows that the majority of applicants on Wuzzuf apply for the Legal category, and the least thing people apply for is Sports and Leisure.

2. How does the minimum required years of experience for a job affect the number of applicants per position?



Insights:

1. **Negative Correlation:**

   - As **minimum experience** increases, the **number of applicants per position** decreases.
   - This suggests that **entry-level jobs (0-2 years experience) receive the highest number of applicants**, whereas positions requiring more experience attract fewer applicants.

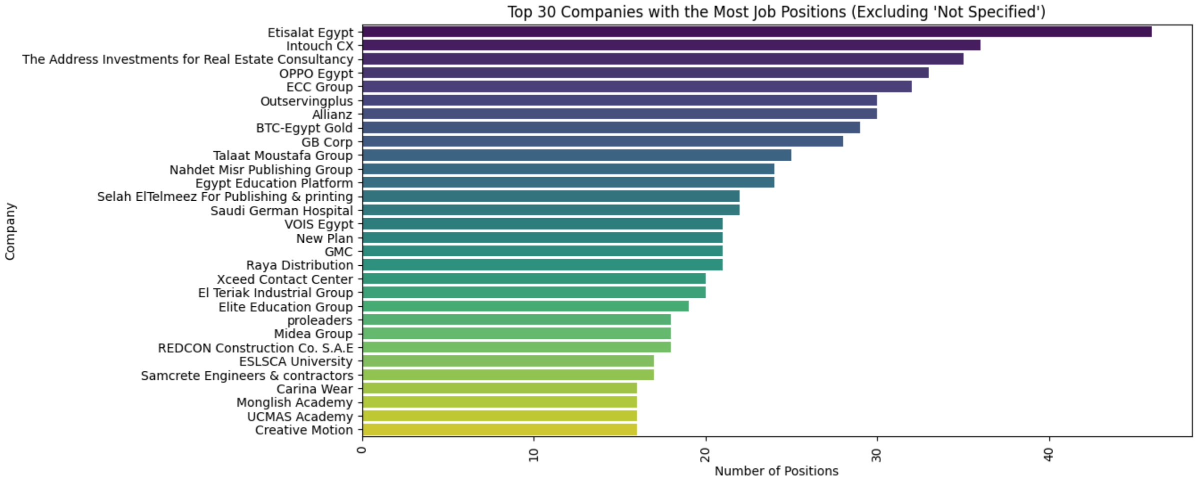
2. **High Competition for Low-Experience Jobs:**

   - Jobs requiring **0 to 5 years of experience** have a **high applicant count**, with some positions exceeding **800 applicants per position**.
   - This indicates that many job seekers fall into the **early-career category**, leading to higher competition.

3. **Less Competition for Senior Roles:**

   - Positions demanding **10+ years of experience** see a **lower number of applicants per position**.
   - This could be due to fewer qualified candidates in the job market for senior roles.

Which companies in Egypt are offering the most job positions, and how do they compare in terms of hiring activity?

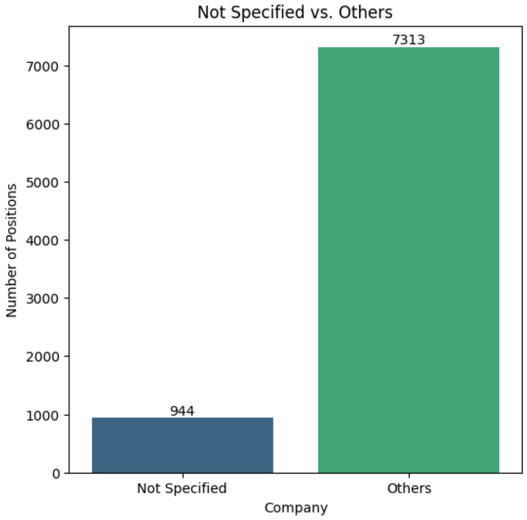


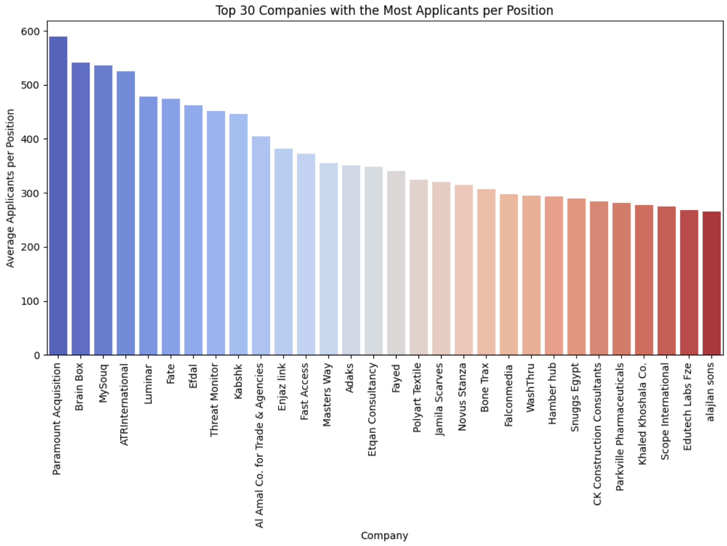
Insights:

1. **Top Hiring Company**: *Etisalat Egypt* leads as the company with the highest number of job positions.
2. **Industry Distribution**: The top companies span across various industries, including **telecommunications (Etisalat Egypt, VOIS Egypt, OPPO Egypt), real estate (Talaat Moustafa Group, The Address Investments), insurance (Allianz), and education (Elite Education Group, ESLCA University, UCMAS Academy)**.

3. **Hiring Trends**: Larger corporations, particularly in telecom, real estate, and education, appear to be among the most active recruiters.
4. **Job Market Demand**: The high number of positions in these companies suggests growing sectors in Egypt, with potential job opportunities concentrated in these industries.

What is the proportion of job positions that fall under the "Not Specified" category compared to those categorized under "Others"?



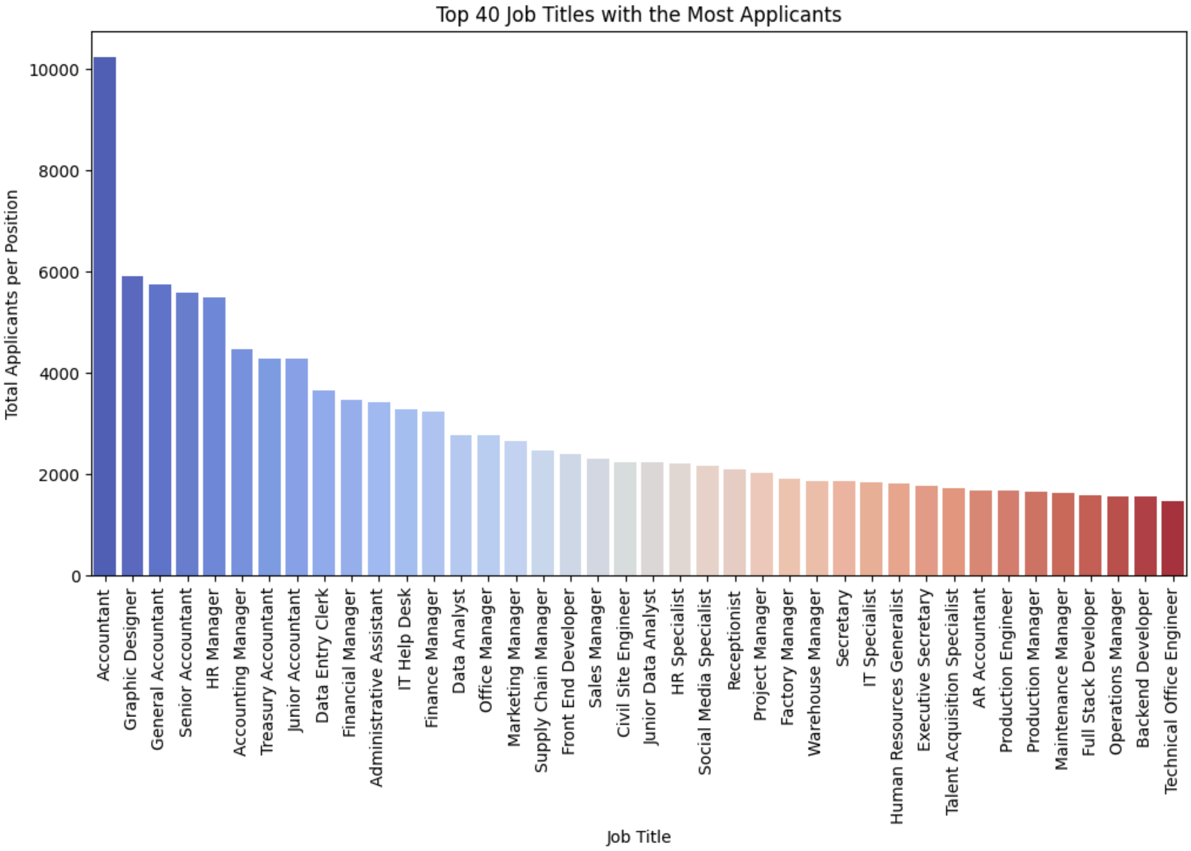


Insights:

1. The "Not Specified" category represents approximately 11.4% of the total positions people apply for (944 out of 8,257).

2. **Paramount Acquisition** has the highest competition, with nearly **600 applicants per position**, making it the most sought-after company in this dataset.
3. The top companies, including **Brain Box, MySouq, ATR International, and Luminar**, all have over **500 applicants per position**, indicating high demand.
4. The competition gradually decreases across the list, with the 30th-ranked company (**Alajan Sons**) receiving around **270 applicants per position**.
5. The large disparity in applicant numbers suggests that some companies are significantly more attractive to job seekers, potentially due to factors like reputation, salary, benefits, or job stability.

Which job titles have the highest number of applicants per position?

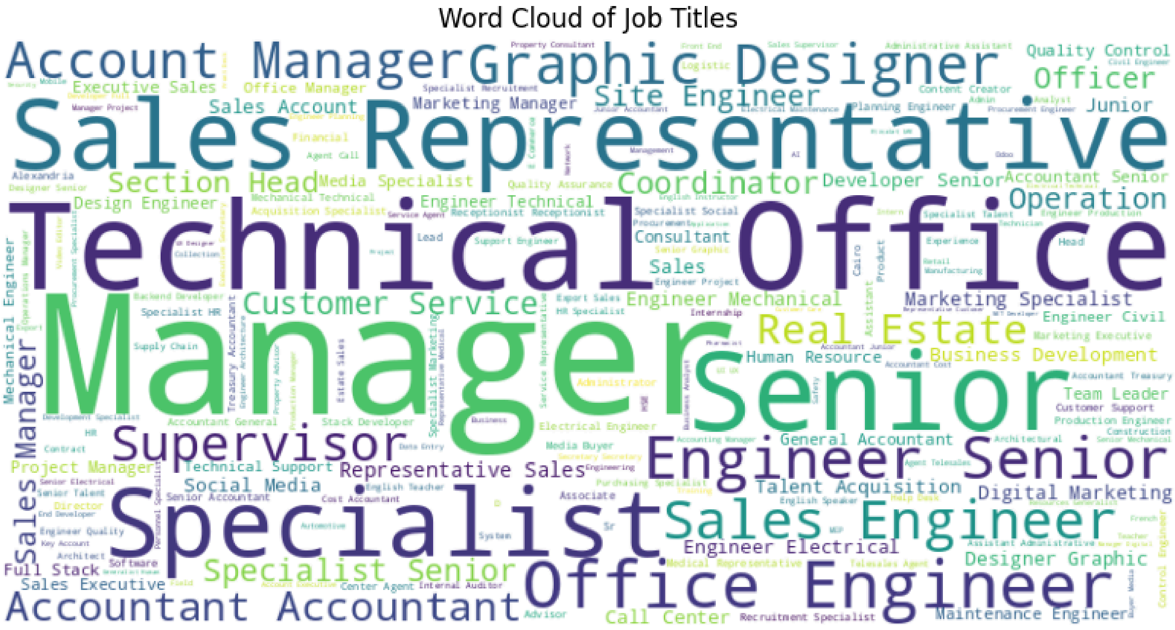


Insights:

1. **Accountant** roles receive the highest competition, with over **10,000 applicants per position**, making it the most sought-after job in this dataset.
2. **Graphic Designer** and **General Accountant** roles follow, with approximately **6,000 applicants per position**, indicating strong demand in both finance and creative fields.
3. Other finance-related positions such as **Senior Accountant, Treasury Accountant, and Financial Assistant** also have high competition, showing the popularity of accounting-related careers.
4. HR and data-related roles, including **HR Manager, Data Analyst, and Finance Manager**, also attract thousands of applicants, suggesting a significant job market focus in these areas.
5. Engineering and tech roles, such as **Front End Developer, Full Stack Developer, Backend Developer, and Technical Office Engineer**, appear toward the lower end of the top 40, with **under 3,000 applicants per position**.

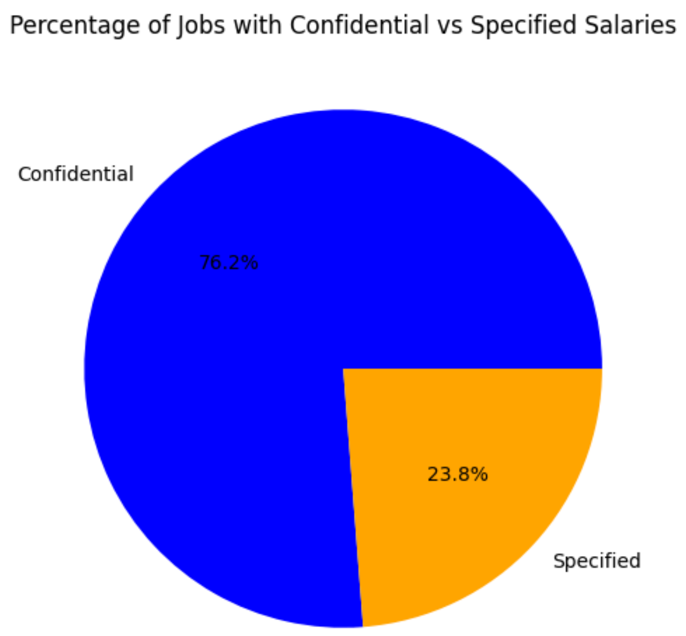
The overall trend suggests that **finance, administrative, and HR jobs** are in much higher demand compared to **tech and engineering roles**, which may indicate either fewer openings in these fields or a stronger supply of candidates.

What are the dominant career fields based on the word cloud?



Insights:

1. **Engineering & Technical Roles:** Words like **"Engineer," "Technical Office," "Site Engineer," and "Office Engineer"** suggest a strong demand for engineering professionals.
2. **Management & Senior Roles:** The frequent occurrence of **"Manager," "Senior," "Supervisor," and "Executive"** implies a high number of leadership-related job postings.
3. **Sales & Marketing:** Titles such as **"Sales Representative," "Marketing Specialist," and "Customer Service"** indicate that sales roles are widely available.
4. **Accounting & Finance:** The presence of **"Accountant," "Treasury," and "Financial"** shows a significant focus on finance-related positions.



Insights:

Most of the companies don’t mention the expected salary range for positions. High-demand roles and senior-level positions tend to have more confidential salaries, whereas entry-level and standardized roles often disclose salary details.