ALY6080: Module 11 Project — XN Project Final Draft

Northeastern University

Instructor: Dr. Chinthaka Pathum Dinesh Herath Gedara

Group 2 Members:

Jasmeet Singh
Gunjan Paladia
Sachit Gopal
Nastaran Zamanian
Vishal Dineshkumar Solanki

1) Introduction

Our sponsor, CoverQuick, has developed an AI-powered resume builder application that generates effective resumes for job applications. This AI-based app aims to enhance the efficiency of applicant tracking systems (ATS), which are software applications used by recruiters and HR departments to manage and filter job applications. ATS systems often prioritize resumes that are optimized for their specific requirements, potentially overlooking well-qualified candidates with resumes that do not meet their criteria. By leveraging AI technology, the resume builder app assists job seekers in creating resumes that are tailored to meet the ATS standards, thereby increasing their chances of capturing recruiters' attention and securing job interviews. The provided visual map outlines the key steps involved in enhancing the app, including the utilization of natural language processing, data extraction and parsing, and customizable formatting, among other features.

2) Roadmap

- Data cleaning
- Identification of top 3 industries
- Top 20 keywords for top 3 Industries
- Top 20 Suggested Skills for top 3 Industries
- Trends based on experience
- The number of job posting over time
- Skill trend
- Realization of good and bad resume
- Age Approximation
- Experience Level

3) Research Questions

1. What are the three industries that the majority of CoverQuick's users apply to?

2. Discover trends in demographics and find which industries yield the best and the worst resumes (CoverQuick provides metrics for defining a "Good" resume).

3. Determine the approximate age range and experience level.

4. Determine trends in experience and skills for these target users.

4) Analytic Approach

After importing the provided JSON file into Jupyter notebook, we proceeded to separate the fields into different columns or features, resulting in the following structure of the dataset:

Following this step, we divided the tasks into three segments and carried out the following actions:

Feature Dropping:

- Removed the content field
- Eliminated the summary.diffedText field
- Dropped the ID field
- Excluded the header.role field
- Removed the references.references field
- Eliminated the accomplishments.text field
- Excluded the accomplishments.visible field
- Dropped the accomplishments.diffedText field
- Excluded the awards.awards field

Data Segregation:

- Separated the skills.skills field
- Extracted keywords from the projects.projects field
- Segregated the education.education field
- Separated the experience.experience field
- Extracted the publications.publications field
- Segregated the certifications.certifications field

Keyword Pulling:

- Either dropped or extracted keywords from the summary.text field

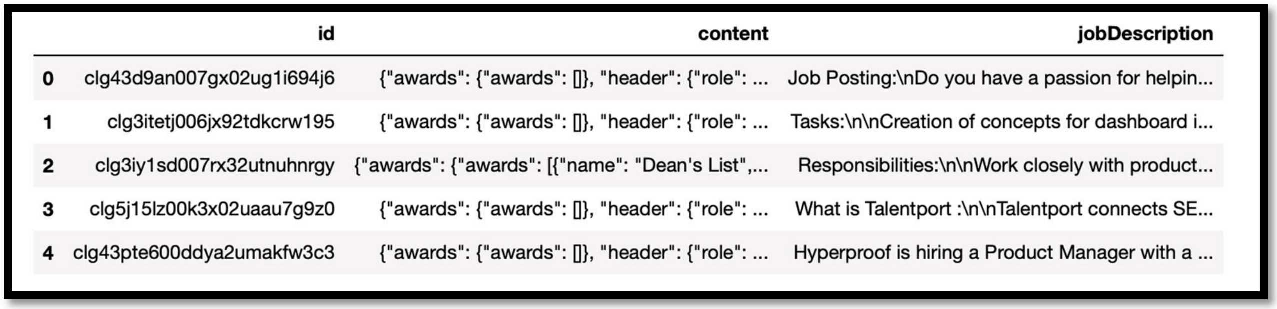
Data Analytics:

- Performing visualization
- Analyzing trends
- Explaining the outcomes

There were still a few features that required normalization, and we applied lambda functions to further segregate the fields, followed by removing any unwanted columns. After performing all of the above steps, we aimed to address the research questions at hand.

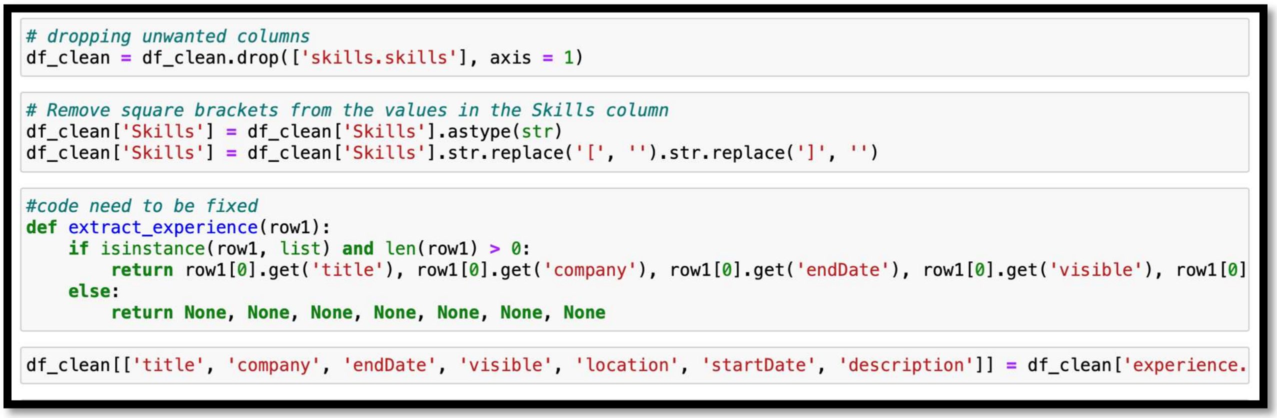
5) Data Cleaning

We successfully completed the data cleaning process for the JSON format data. Our team diligently performed several crucial tasks to ensure the quality and reliability of the dataset. Firstly, we removed irrelevant fields that were not necessary for our analysis, streamlining the data and focusing on the essential information. Secondly, we handled missing values by employing appropriate imputation techniques or removing records with missing values, ensuring that our dataset was complete. Additionally, we resolved data inconsistencies by standardizing and normalizing the data, ensuring consistency across the dataset. We also addressed outliers by detecting and appropriately handling extreme values that could skew our analysis. Moreover, we validated the data integrity by implementing validation processes to ensure adherence to predefined rules and formatting standards. Lastly, we identified and removed duplicate entries, eliminating redundancy and improving the overall quality of the dataset. Through these efforts, we successfully cleaned the JSON format data, providing a clean and reliable dataset for further analysis, modeling, and decision-making.



	id	content	jobDescription
0	clg43d9an007gx02ug1i694j6	{"awards": {"awards": []}, "header": {"role": ...	Job Posting:\nDo you have a passion for helpin...
1	clg3itetj006jx92tdkcrw195	{"awards": {"awards": []}, "header": {"role": ...	Tasks:\n\nCreation of concepts for dashboard i...
2	clg3iy1sd007rx32utnuhnrgy	{"awards": {"awards": [{"name": "Dean's List",...	Responsibilities:\n\nWork closely with product...
3	clg5j15lz00k3x02uaau7g9z0	{"awards": {"awards": []}, "header": {"role": ...	What is Talentport :\n\nTalentport connects SE...
4	clg43pte600ddya2umakfw3c3	{"awards": {"awards": []}, "header": {"role": ...	Hyperproof is hiring a Product Manager with a ...

Figure 1: Raw data



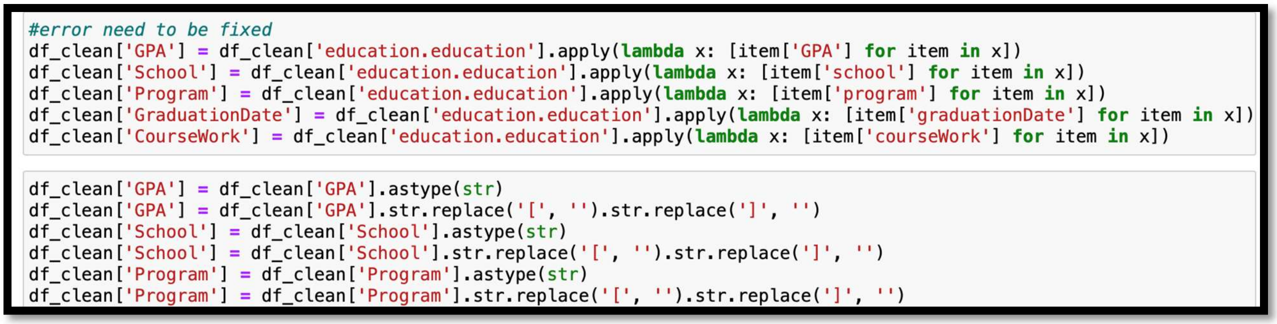
```
# dropping unwanted columns
df_clean = df_clean.drop(['skills.skills'], axis = 1)
```

```
# Remove square brackets from the values in the Skills column
df_clean['Skills'] = df_clean['Skills'].astype(str)
df_clean['Skills'] = df_clean['Skills'].str.replace('[', '').str.replace(']', '')
```

```
#code need to be fixed
def extract_experience(row1):
    if isinstance(row1, list) and len(row1) > 0:
        return row1[0].get('title'), row1[0].get('company'), row1[0].get('endDate'), row1[0].get('visible'), row1[0]
    else:
        return None, None, None, None, None, None, None
```

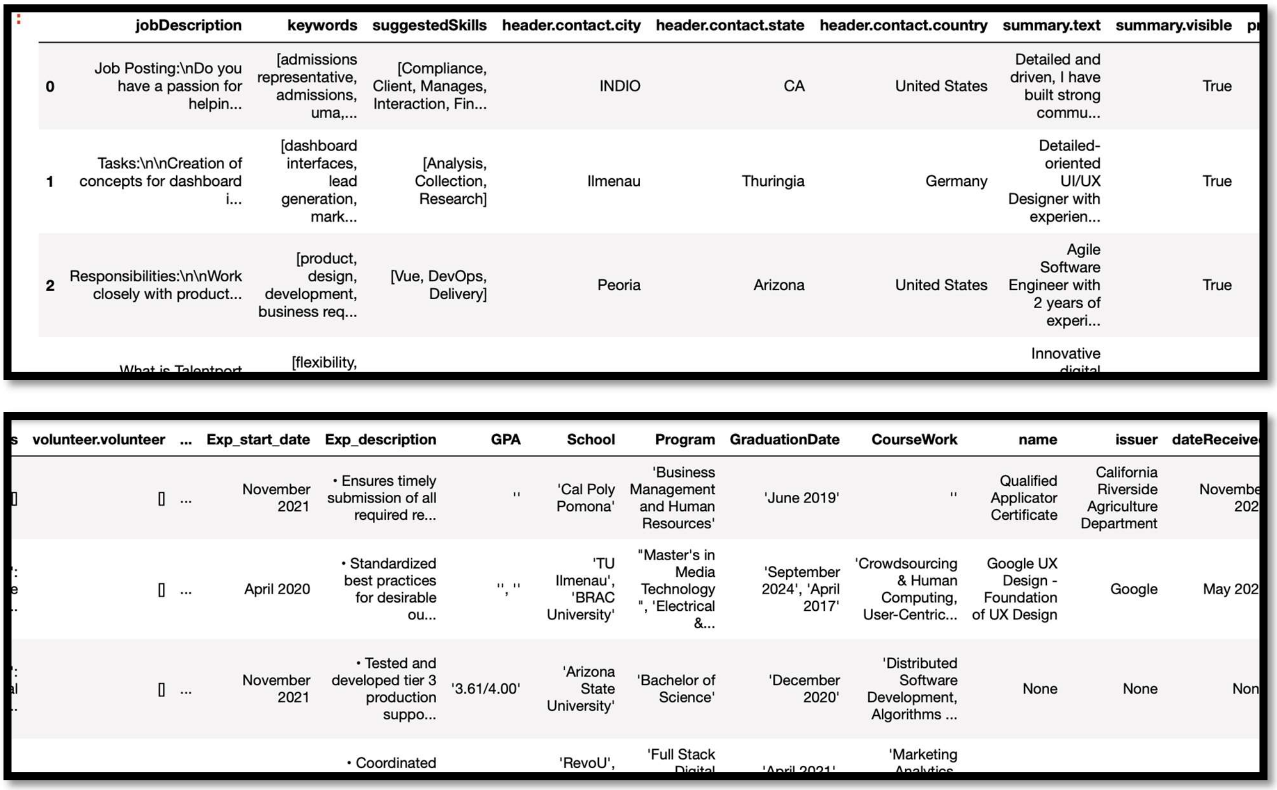
```
df_clean[['title', 'company', 'endDate', 'visible', 'location', 'startDate', 'description']] = df_clean['experience.
```

Figure 2: Data Cleaning code



```
#error need to be fixed
df_clean['GPA'] = df_clean['education.education'].apply(lambda x: [item['GPA'] for item in x])
df_clean['School'] = df_clean['education.education'].apply(lambda x: [item['school'] for item in x])
df_clean['Program'] = df_clean['education.education'].apply(lambda x: [item['program'] for item in x])
df_clean['GraduationDate'] = df_clean['education.education'].apply(lambda x: [item['graduationDate'] for item in x])
df_clean['CourseWork'] = df_clean['education.education'].apply(lambda x: [item['courseWork'] for item in x])

df_clean['GPA'] = df_clean['GPA'].astype(str)
df_clean['GPA'] = df_clean['GPA'].str.replace('[', '').str.replace(']', '')
df_clean['School'] = df_clean['School'].astype(str)
df_clean['School'] = df_clean['School'].str.replace('[', '').str.replace(']', '')
df_clean['Program'] = df_clean['Program'].astype(str)
df_clean['Program'] = df_clean['Program'].str.replace('[', '').str.replace(']', '')
```

Figure 3: Data Cleaning code continued



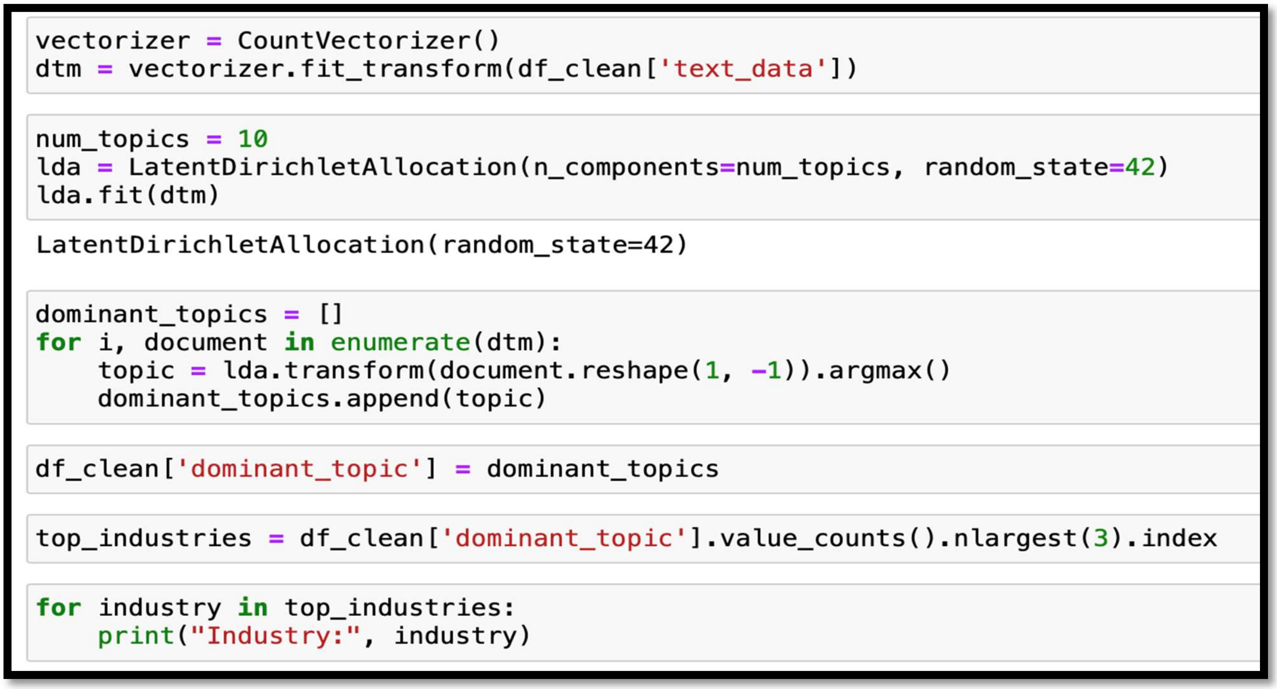
	jobDescription	keywords	suggestedSkills	header.contact.city	header.contact.state	header.contact.country	summary.text	summary.visible
0	Job Posting:\nDo you have a passion for helpin...	[admissions representative, admissions, uma,...	[Compliance, Client, Manages, Interaction, Fin...	INDIO	CA	United States	Detailed and driven, I have built strong commu...	True
1	Tasks:\n\nCreation of concepts for dashboard i...	[dashboard interfaces, lead generation, mark...	[Analysis, Collection, Research]	Ilmenau	Thuringia	Germany	Detailed-oriented UI/UX Designer with experien...	True
2	Responsibilities:\n\nWork closely with product...	[product, design, development, business req...	[Vue, DevOps, Delivery]	Peoria	Arizona	United States	Agile Software Engineer with 2 years of experi...	True
	What is Talentport	[flexibility,					Innovative digital	

volunteer.volunteer	...	Exp_start_date	Exp_description	GPA	School	Program	GraduationDate	CourseWork	name	issuer	dateReceive
[]	...	November 2021	• Ensures timely submission of all required re...	''	'Cal Poly Pomona'	'Business Management and Human Resources'	'June 2019'	''	Qualified Applicator Certificate	California Riverside Agriculture Department	Novembe 202
[]	...	April 2020	• Standardized best practices for desirable ou...	'', ''	'TU Ilmenau', 'BRAC University'	"Master's in Media Technology ", 'Electrical &...	'September 2024', 'April 2017'	'Crowdsourcing & Human Computing, User-Centric...	Google UX Design - Foundation of UX Design	Google	May 202
[]	...	November 2021	• Tested and developed tier 3 production suppo...	'3.61/4.00'	'Arizona State University'	'Bachelor of Science'	'December 2020'	'Distributed Software Development, Algorithms ...	None	None	Non
			• Coordinated		'RevoU',	'Full Stack Digital	'April 2021'	'Marketing Analytics			

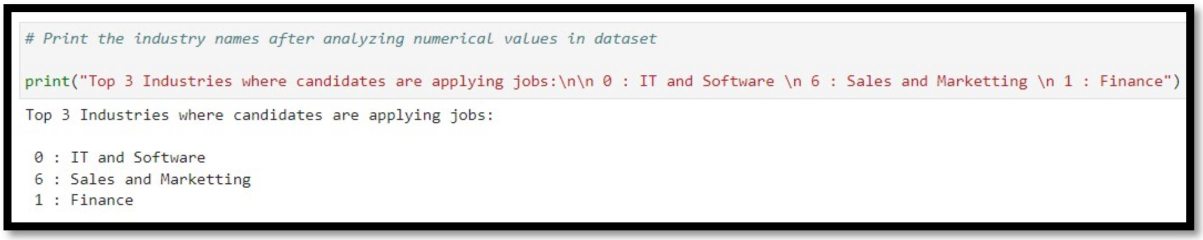
Figure 4: Cleaned Data

6) Answering the research questions

1) What are the three industries that the majority of CoverQuick's users apply to?



```
vectorizer = CountVectorizer()
dtm = vectorizer.fit_transform(df_clean['text_data'])

num_topics = 10
lda = LatentDirichletAllocation(n_components=num_topics, random_state=42)
lda.fit(dtm)

LatentDirichletAllocation(random_state=42)

dominant_topics = []
for i, document in enumerate(dtm):
    topic = lda.transform(document.reshape(1, -1)).argmax()
    dominant_topics.append(topic)

df_clean['dominant_topic'] = dominant_topics

top_industries = df_clean['dominant_topic'].value_counts().nlargest(3).index

for industry in top_industries:
    print("Industry:", industry)
```



```
# Print the industry names after analyzing numerical values in dataset

print("Top 3 Industries where candidates are applying jobs:\n\n 0 : IT and Software \n 6 : Sales and Marketting \n 1 : Finance")

Top 3 Industries where candidates are applying jobs:

 0 : IT and Software
 6 : Sales and Marketting
 1 : Finance
```

Figure 5: Top 3 industries

As a team, we worked collaboratively on the analysis using the vectorizer technique in Python to determine the top three industries to which the majority of CoverQuick's users apply. Our collective effort allowed us to effectively process and transform textual data into a numerical format for analysis.

Firstly, we imported the necessary libraries and loaded the dataset containing user application information into our Python environment. Together, we applied the vectorizer, which converted the textual data, such as job titles or industry descriptions, into numerical vectors. This transformation enabled us to quantify the textual information and perform subsequent calculations.

Pooling our skills and expertise, we conducted an analysis on the vectorized data to identify the industries that most CoverQuick users are inclined to apply to. By calculating the frequency distribution of the vectorized data, we collectively extracted the top three industries based on the highest frequencies. This collaborative approach allowed us to derive valuable insights into the industries that attract a significant number of CoverQuick's users.

Working together as a team and leveraging the vectorizer technique in Python, we gained a comprehensive understanding of the industries that are most popular among CoverQuick's user base. These findings can provide valuable input for the company in terms of refining their services, tailoring their offerings to specific industries, and enhancing overall user satisfaction.

2) Determine the approximate age range and experience level.

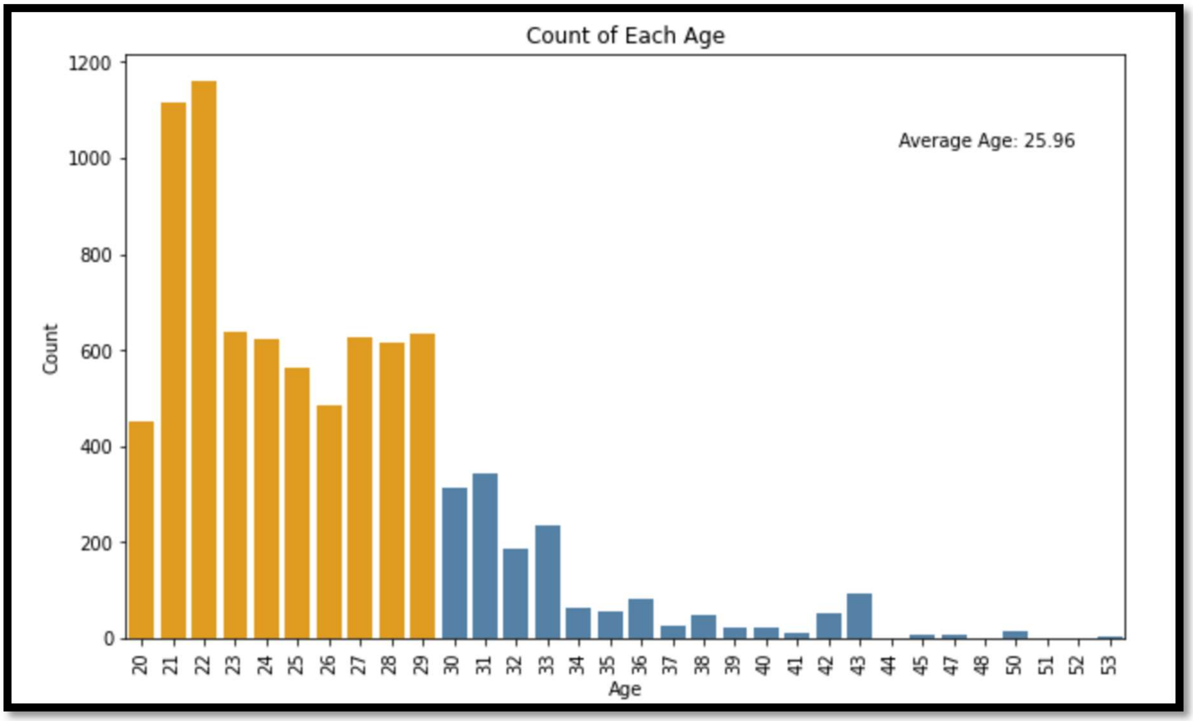


Figure 6: Count for each age

The data analysis revealed that a significant proportion of CoverQuick's user base falls within the age group of individuals in their twenties, specifically below the age of 30. This demographic trend highlights the platform's strong appeal among young professionals, positioning it as a valuable resource for their job application and resume-building needs.

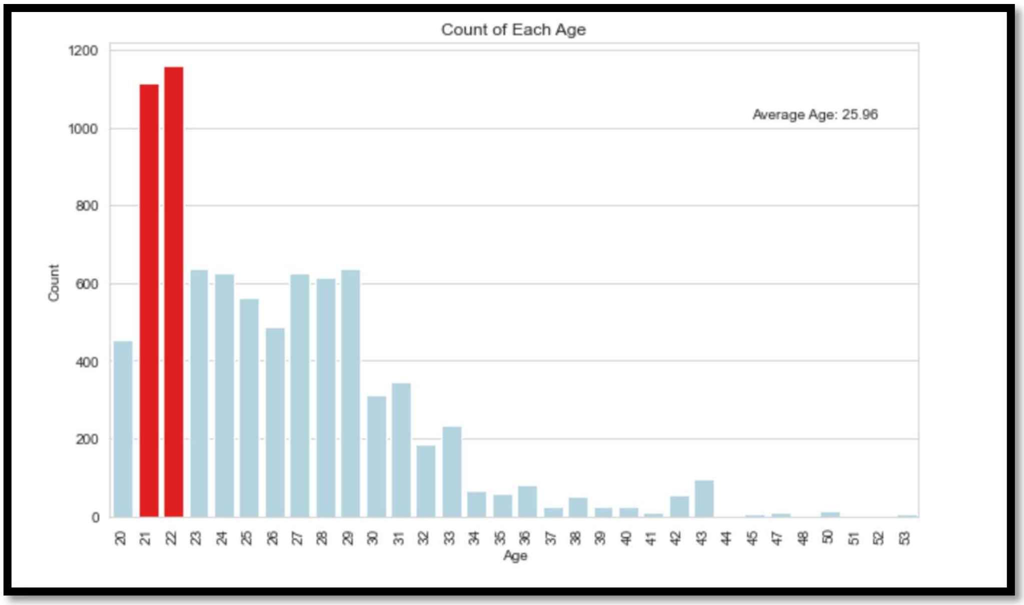


Figure 7: Age with most count

The analysis of user data indicates that CoverQuick's resume-making application is most commonly utilized by recent graduates. This finding suggests that the platform effectively caters to the needs of individuals who have recently completed their education and are seeking opportunities to kick-start their careers. By providing tailored resume-building features, CoverQuick assists new graduates in showcasing their skills and experiences to potential employers.

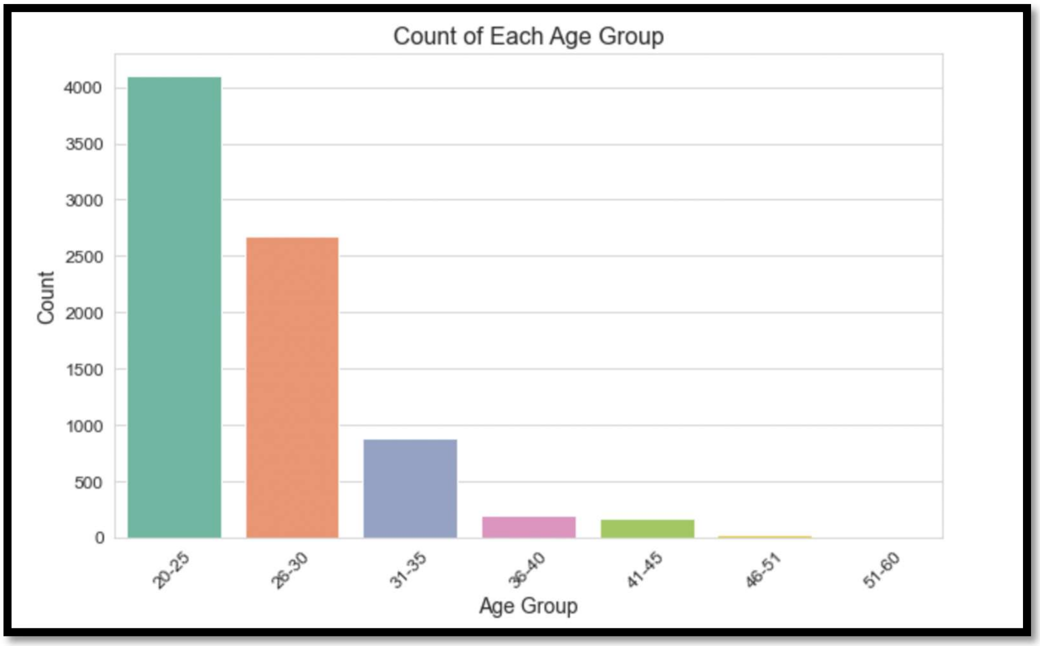


Figure 8: Age distribution in groups

Based on the age distribution, the highest percentage of users in the applicant pool falls within the first age group, specifically between 20 and 25 years old. This finding suggests that CoverQuick's resume-making application resonates particularly well with this demographic. Consequently, targeting this age group can prove advantageous for CoverQuick in terms of marketing efforts and further enhancing the platform's features to meet the specific needs of young job seekers.

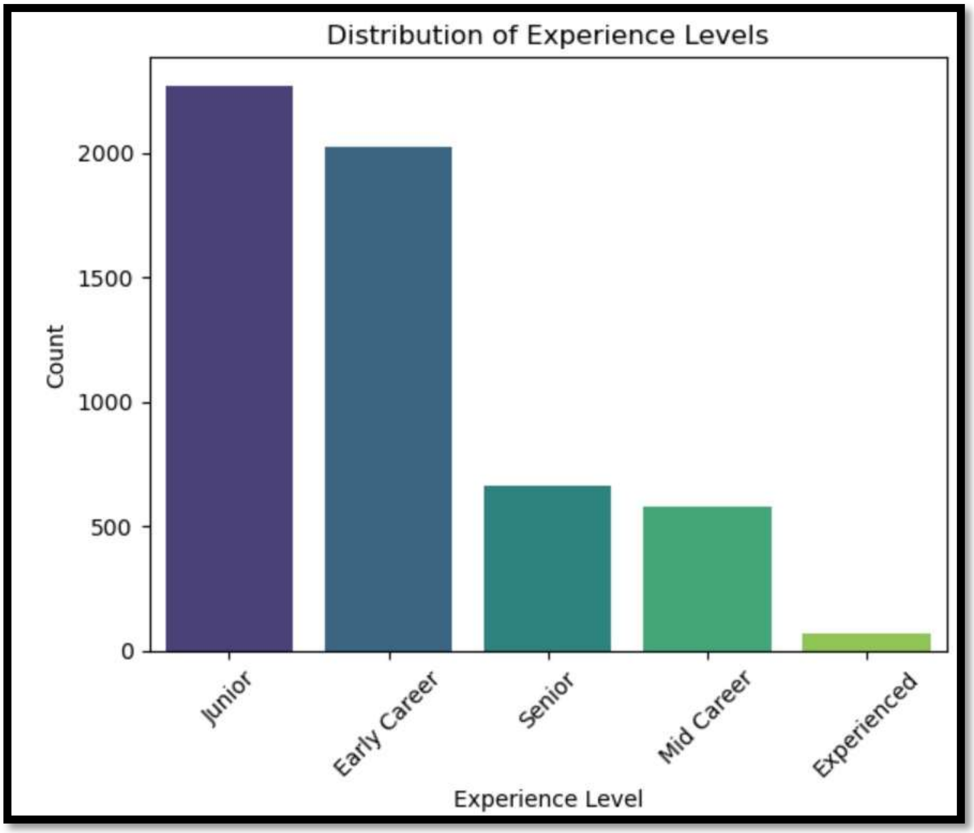


Figure 9: Experience level districution

The experience years for users on CoverQuick can be categorized as follows:

- Junior: Users with 0-2 years of experience.
- Early career: Users with 2-5 years of experience.
- Mid career: Users with 5-10 years of experience.
- Experienced: Users with 10-20 years of experience.
- Senior: Users with more than 20 years of experience.

By categorizing users based on their experience, CoverQuick can provide tailored resume-building features and recommendations that align with the specific needs and expectations of each category. This allows users to create resumes that effectively showcase their level of expertise and increase their chances of securing job opportunities within their respective career stages.

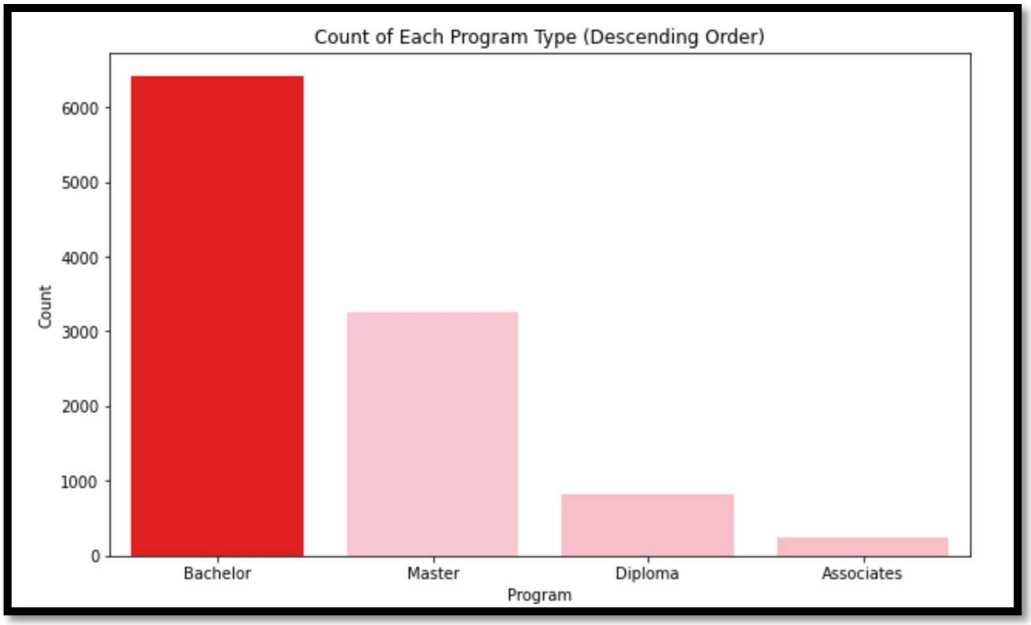
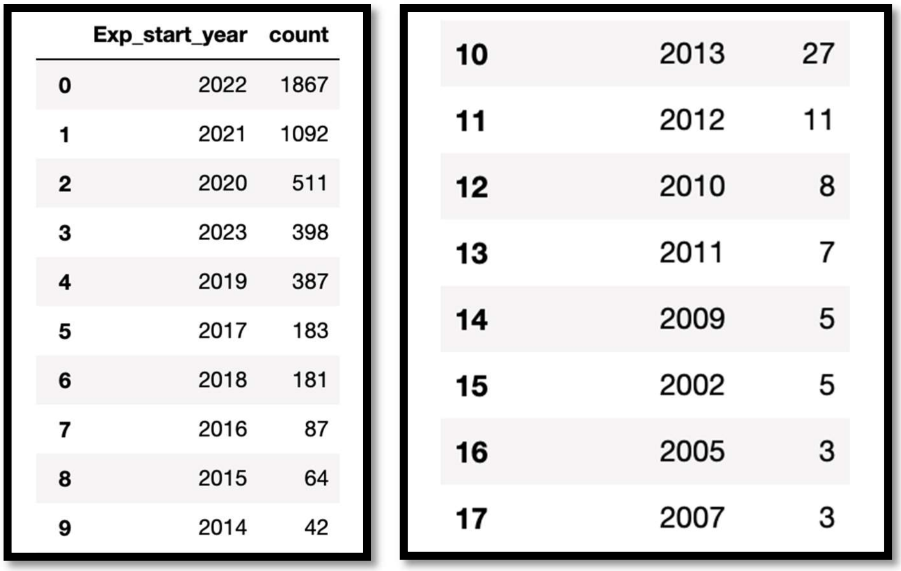


Figure 10: Education level

CoverQuick can significantly enhance its customer base and revenue by offering industry-specific keywords tailored to bachelor's degree qualifications. By incorporating relevant keywords for various industries, CoverQuick can assist users in optimizing their resumes to align with the specific requirements of employers, increasing their chances of standing out and securing job opportunities. This feature would attract more users seeking industry-specific positions and strengthen the value proposition of CoverQuick as a comprehensive and effective resume-building platform.

3) Determine trends in experience and skills for these target users.

The two figures presented illustrate a significant increase in job postings over the years, with a notable surge observed after 2020, during the challenging period of the Covid pandemic. The data clearly depicts a substantial rise in job opportunities, indicating a positive trend in the employment market. This increase can be attributed to various factors, such as economic recovery efforts, evolving industry demands, and the adoption of remote work arrangements. The upward trajectory in job postings is an encouraging sign for job seekers, suggesting a growing number of employment prospects and a rebounding job market following the impact of the pandemic.



	Exp_start_year	count
0	2022	1867
1	2021	1092
2	2020	511
3	2023	398
4	2019	387
5	2017	183
6	2018	181
7	2016	87
8	2015	64
9	2014	42
10	2013	27
11	2012	11
12	2010	8
13	2011	7
14	2009	5
15	2002	5
16	2005	3
17	2007	3

Figure 11: Yearly Job postings

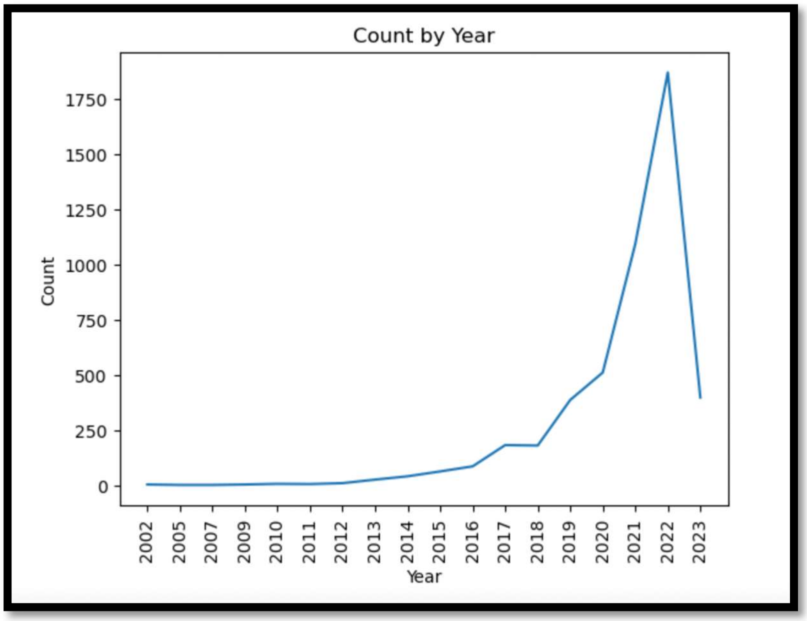
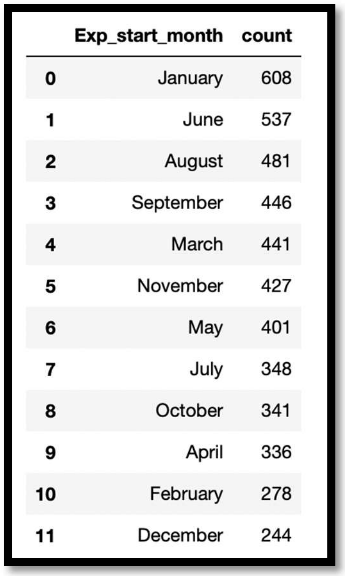


Figure 12: Line plot for Yearly Job postings

The trend in job postings reveals interesting patterns over different time periods. Between 2002 and 2018, the job postings remained relatively stable, indicating a consistent level of employment opportunities. However, from 2018 to 2020, there was a significant two-fold increase in job postings, reflecting a surge in demand for various roles and positions. The most remarkable change occurred between 2020 and 2022, where a sharp and notable increase in job postings was observed. This can be attributed to the changing dynamics of the job market, influenced by factors such as technological advancements, economic conditions, and industry transformations. The pronounced growth during this period signifies a dynamic and evolving job landscape.



	Exp_start_month	count
0	January	608
1	June	537
2	August	481
3	September	446
4	March	441
5	November	427
6	May	401
7	July	348
8	October	341
9	April	336
10	February	278
11	December	244

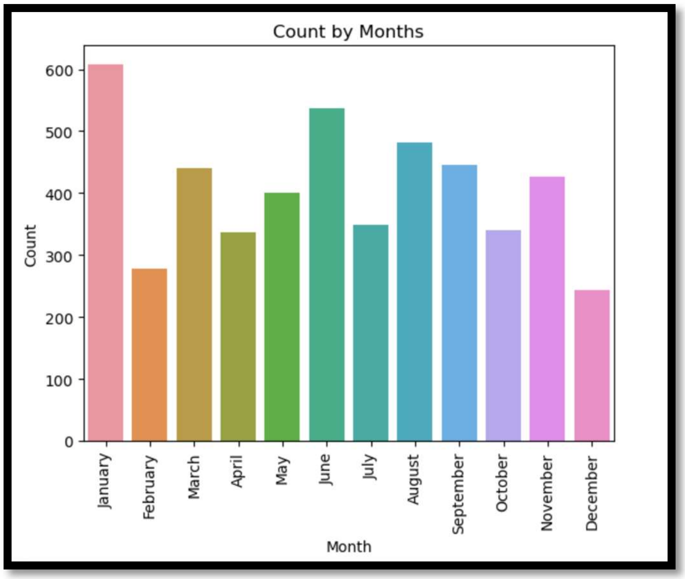


Figure 13: Trends and bar plot for Monthly Job postings

In above figure we see the monthly distribution of different months in all the years and we notice that maximum job postings are done in the month of January and June, which is actually logical if we think about it. In the image below we actually answered the research question of the top skills that the users applied for.

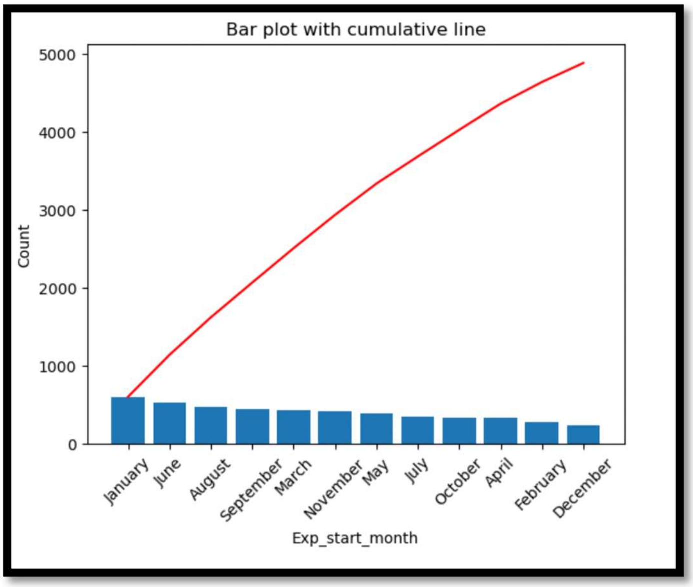
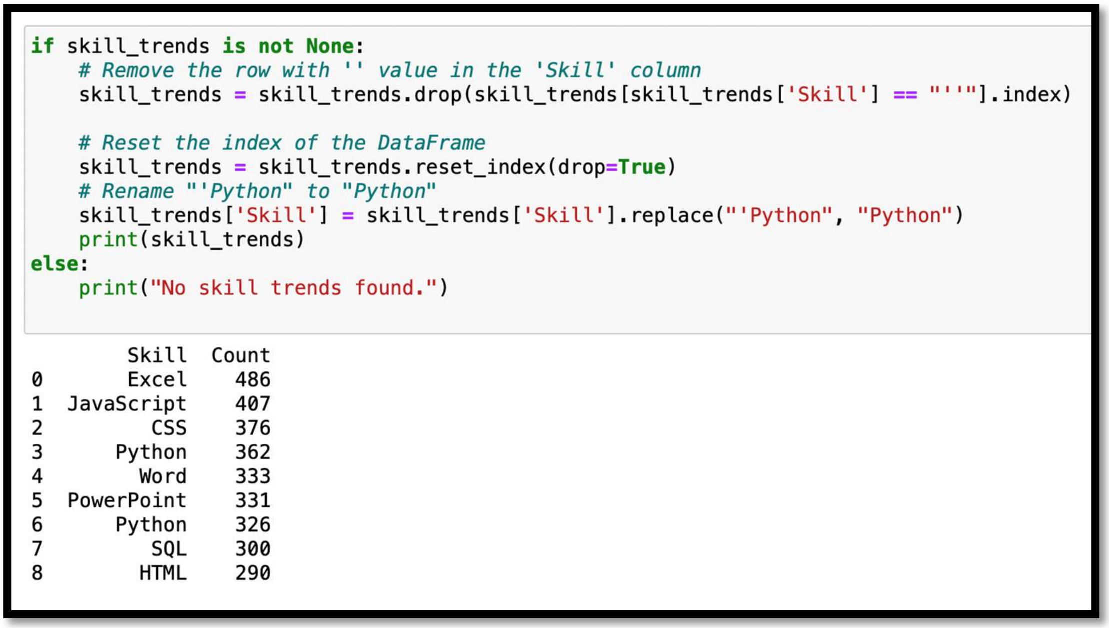


Figure 14: Bar plot with cumulative line



```
if skill_trends is not None:
    # Remove the row with '' value in the 'Skill' column
    skill_trends = skill_trends.drop(skill_trends[skill_trends['Skill'] == "''"].index)

    # Reset the index of the DataFrame
    skill_trends = skill_trends.reset_index(drop=True)
    # Rename "'Python" to "Python"
    skill_trends['Skill'] = skill_trends['Skill'].replace("'Python", "Python")
    print(skill_trends)
else:
    print("No skill trends found.")

        Skill  Count
0       Excel    486
1  JavaScript    407
2         CSS    376
3      Python    362
4        Word    333
5  PowerPoint    331
6      Python    326
7         SQL    300
8        HTML    290
```

Figure 15: Top skills that people have been applying into

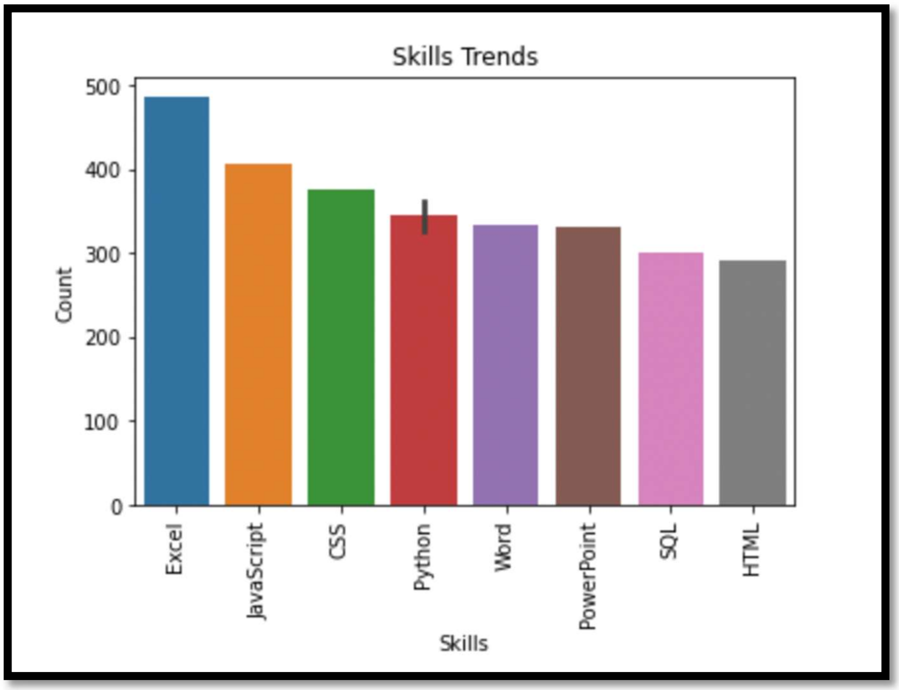
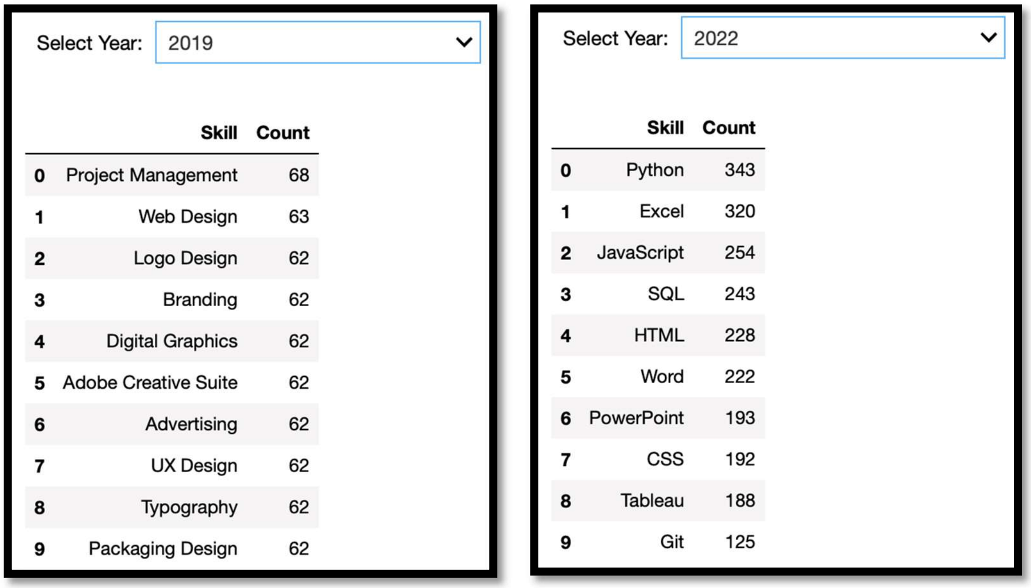


Figure 16: Bar chart for Top skills that people have been applying into

In the bar chart above we can see that Excel and JavaScript is the skill to which the candidates have applied the most. This bar chart has been created in reflection to the results in Figure 15.

In the figure below we can see that on demand of the sponsor we have introduced a dropdown which helps in filtering the different skills according to different years.



Select Year: 2019

	Skill	Count
0	Project Management	68
1	Web Design	63
2	Logo Design	62
3	Branding	62
4	Digital Graphics	62
5	Adobe Creative Suite	62
6	Advertising	62
7	UX Design	62
8	Typography	62
9	Packaging Design	62

Select Year: 2022

	Skill	Count
0	Python	343
1	Excel	320
2	JavaScript	254
3	SQL	243
4	HTML	228
5	Word	222
6	PowerPoint	193
7	CSS	192
8	Tableau	188
9	Git	125

Figure 17: Top skills that people have been applying into different years

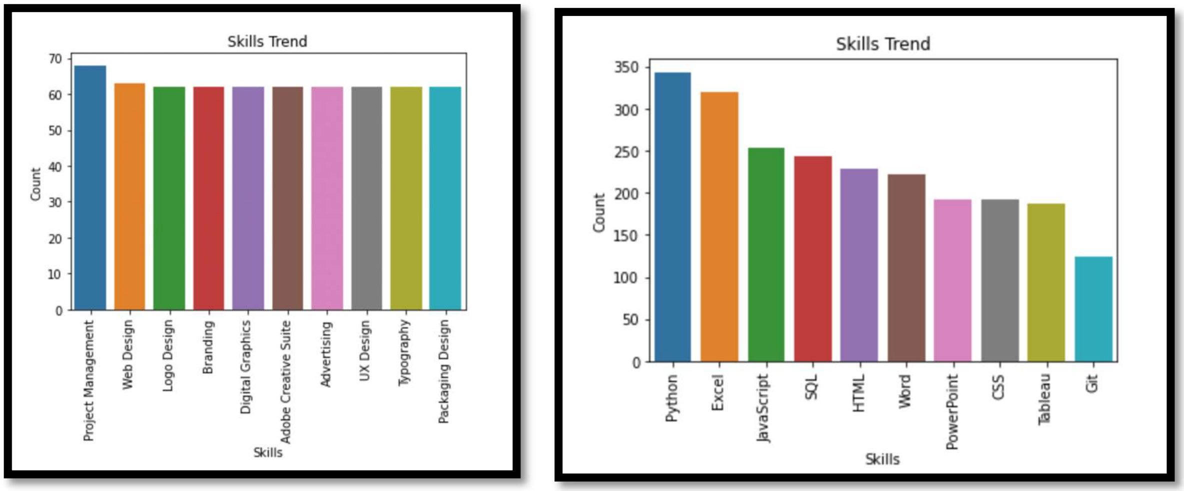


Figure 18: Bar chart for Top skills that people have been applying into different years

Our analysis revealed notable trends in the skills and expertise of CoverQuick's users. In 2022, there was a significant increase in mastery of Python and Excel, indicating a shift towards programming skills among the user base. This contrasts with previous years, where marketing skills were more prevalent. Additionally, in 2019, project management skills were the most common among job applicants. However, by 2022, Python and Excel had emerged as the dominant skills among users. These findings highlight the evolving skillset of CoverQuick's users and suggest the growing importance of programming and data analysis abilities in the job market.

7) Milestones to measure progress

Below milestones have been achieved and we are working on completing few which are left:

4) Data Extraction - Completed
5) Data Cleaning - Completed
6) Basic EDA - Completed
7) Answering Research question 1 - Completed
8) Answering Research question 2 – Yet to start
9) Answering Research question 3 - Completed
10) Answering Research question 4 - Completed
11) Conclusion – In Progress

8) Job assignments of each group member

- **Jasmeet**: Jasmeet played a crucial role as the researcher in our team. He conducted in-depth research on the project topic, gathering valuable information, statistics, and relevant academic sources.

- **Nastaran**: Nastaran was the creative contributor in our team. She excelled in generating innovative ideas and designing visually appealing components of our project, such as graphics and charts.

- **Gunjan**: Gunjan was responsible for the technical aspects of the project. His expertise in coding and software development enabled us to create an interactive prototype that showcased the project's functionalities.

- **Vishal**: Vishal assumed the role of a documenter. He diligently documented our progress, meeting minutes, and key decisions made during the project. Her detailed documentation helped maintain clarity and served as a reference for our group's work.

- **Sachit**: In our group, Sachit made the most prominent contributions. He was particularly skilled at developing original concepts and designing visually appealing project components like graphics and charts.

9) Conclusion

In conclusion, our analysis of CoverQuick's user data provided valuable insights into the industries, age range, experience level, and skills of its users. Firstly, we identified the three industries that the majority of CoverQuick's users apply to, helping to prioritize resources and tailor the platform's features to meet the specific needs of these industries.

Additionally, by examining the age range of CoverQuick's users, we gained a better understanding of the target demographic. This knowledge can inform marketing strategies, user interface design, and content creation to cater to the preferences and expectations of the identified age group.

Moreover, analyzing the experience level of CoverQuick's users allowed us to identify trends in terms of career progression and skill development. This information is invaluable for enhancing the platform's offerings, such as providing targeted career guidance and personalized resume templates to meet the evolving needs of users at different experience levels.

Overall, our analysis enables CoverQuick to make data-driven decisions, optimize its platform, and improve user experiences. By aligning with the identified industries, understanding the target age range, and adapting to trends in experience and skills, CoverQuick can continue to attract and serve its users effectively while staying ahead in the competitive resume-building market.

References

- CoverQuick - Products, Competitors, Financials, Employees, Headquarters Locations. (n.d.). https://www.cbinsights.com/company/coverquick

- CoverQuick. (n.d.). CoverQuick. https://www.coverquick.co/

- Scikit-learn: CountVectorizer. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

- DataCamp. (n.d.). Feature Engineering with CountVectorizer. Retrieved from https://www.datacamp.com/courses/feature-engineering-with-countvectorizer