Northeastern University
College of Professional Studies

June 06, 2023

ALY 6080: XN Project
CoverQuick Individual Draft

Submitted to:
Dr Chinthaka Pathum Dinesh, Prof
Herath Gedara, Faculty Lecturer

Presented by:
Abhilash Dikshit

CoverQuick

About Sponsor

- An AI-based programme called CoverQuick creates unique cover letters and resumes for job applications. By developing a customised cover letter and resume that are suited to the particular job application, CoverQuick hopes to assist job seekers stand out from the competition by utilising the most recent AI techniques.
- One of CoverQuick's distinctive characteristics is its capacity to produce a customised cover letter for each application, ensuring that applicants do not submit generic letters that fall flat with recruiters.
- Products : Prepare Resume , CoverLetter, Track Application, Resume Grader
- Number of Users : 5000 users (Ref till September 2022)

EXPLORATORY DATA ANALYSIS (EDA)

DATASET PROVIDED

A. WITH JOB DESCRIPTION
B. WITHOUT JOB DESCRIPTION

Research Questions:

1. What are the three industries that the majority of CoverQuick's users have applied (With Job Description Dataset)?
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 expected age and approximate experience level.
4. Determine trends in experience and skills for these target users.

Planning and Execution

1. EDA on job description dataset.
2. Dataset splitting for the respective columns which were in json and nested json format.
3. Identification and visualization for the top 3 industries that the majority of users have applied.
4. Identification and visualization for the approximate age range and experience level.
5. Identification and visualization for the trends in experience and skills for these target users.
6. Identification and visualization to discover the trends in demographics for the number of candidates registering to the website across globe for resume building.

(Note: For presentation, we have shown for with job description dataset analysis only)

EDA : WITH JOB DESCRIPTION DATASET

RAW DATASET

Total Rows: 11976
Total Columns: 3

FINAL DATASET AS OF NOW

Total Rows: 11976
Total Columns: 57

CoverQuick With Job Description Dataset:
==========

Display Raw Dataset:

	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 ...
...	...	...	...
11971	cleexyzag006ayg2vhr087als	{"awards": {"awards": []}, "header": {"role": ...	Assist with content ideation and creation, inc...
11972	cleec90b0005nyf2tlos9qc95	{"awards": {"awards": [{"name": "Honor Roll ",...	This person must excel in a fast-paced environ...
11973	cleey05qa000exd2up87uehkz	{"awards": {"awards": []}, "header": {"role": ...	In collaboration with the Senior Communication...
11974	cle0edrgo00a5wz2utru0nt5u	{"awards": {"awards": [{"name": "Honor Roll ",...	About the job\nYou've got 52 weeks a year to f...
11975	cleecwhm6006dyf2tsr12f761	{"awards": {"awards": [{"name": "Honor Roll ",...	About InsideTracker\n\nCreated by experts in a...

11976 rows × 3 columns

Display Type, Length, Shape about the dataset:

Type	Length	Shape
<class 'pandas.core.frame.DataFrame'>	11976	(11976, 3)

Display datatypes of respective columns in dataset:

	id	content	jobDescription
0	object	object	object

Showing max, min length and NA values:

Column	Max Length	Min Length	NA Count
id	25	25	0
content	52580	513	0
jobDescription	22567	1	4

Fig: Display of Raw Dataset

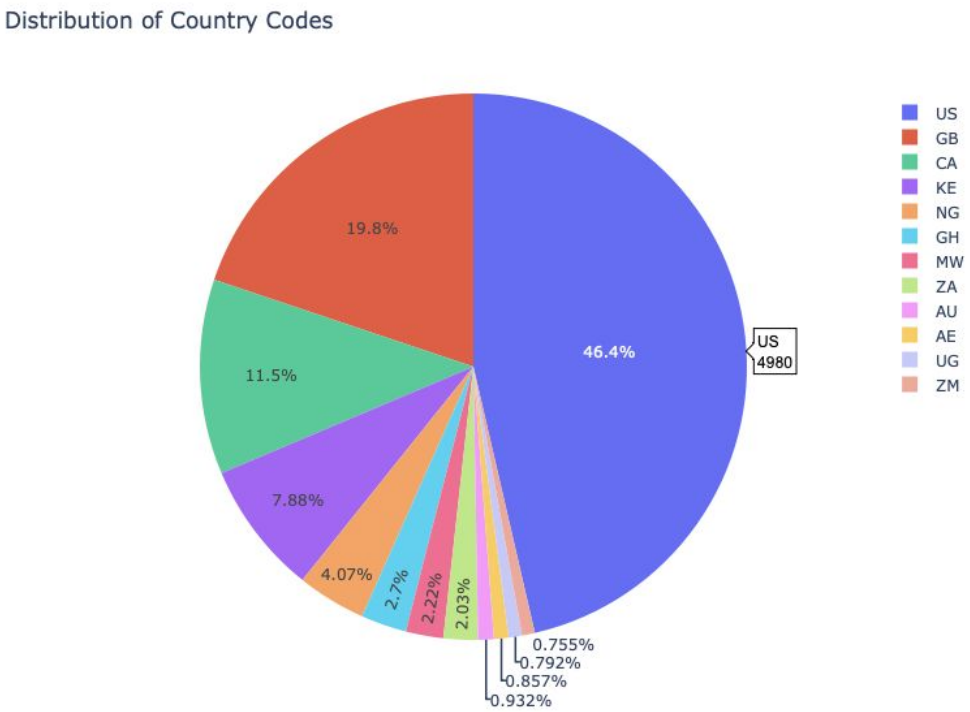
Distribution of Country Codes w.r.t ID

Country Code was defined using Country Column and after successful analysis we have 0 NULL count which helped us for further visualisation.

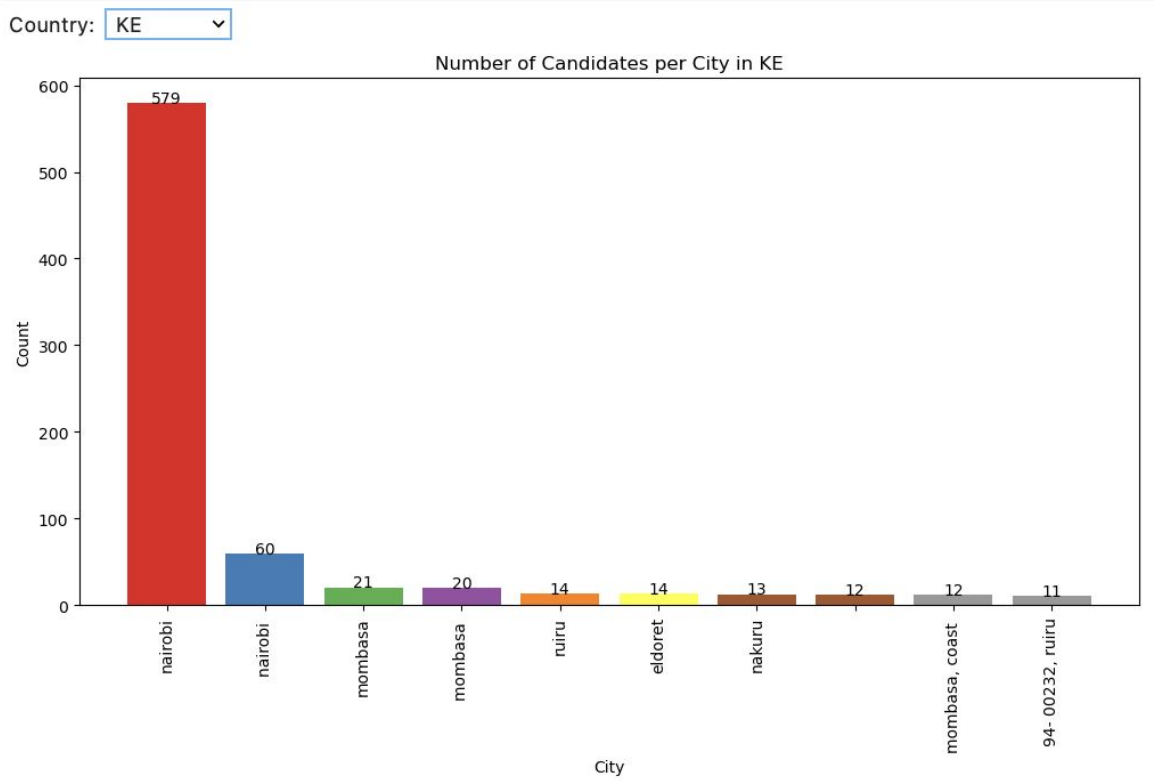
```
COUNTRY_CODE Non-Null and Null Count:
=====================================
   Total Count  Non-Null Count  Null Count
0        11976           11976           0
```

Demographics Explanation:
Top 3 countries as per number of users.

1. United States has maximum user: 46.4% of total user : 4980 applicants
2. Kenya (KE) has 19.8%
3. Great Britain (GB) : 11.5%



Number of Candidates per City in Kenya

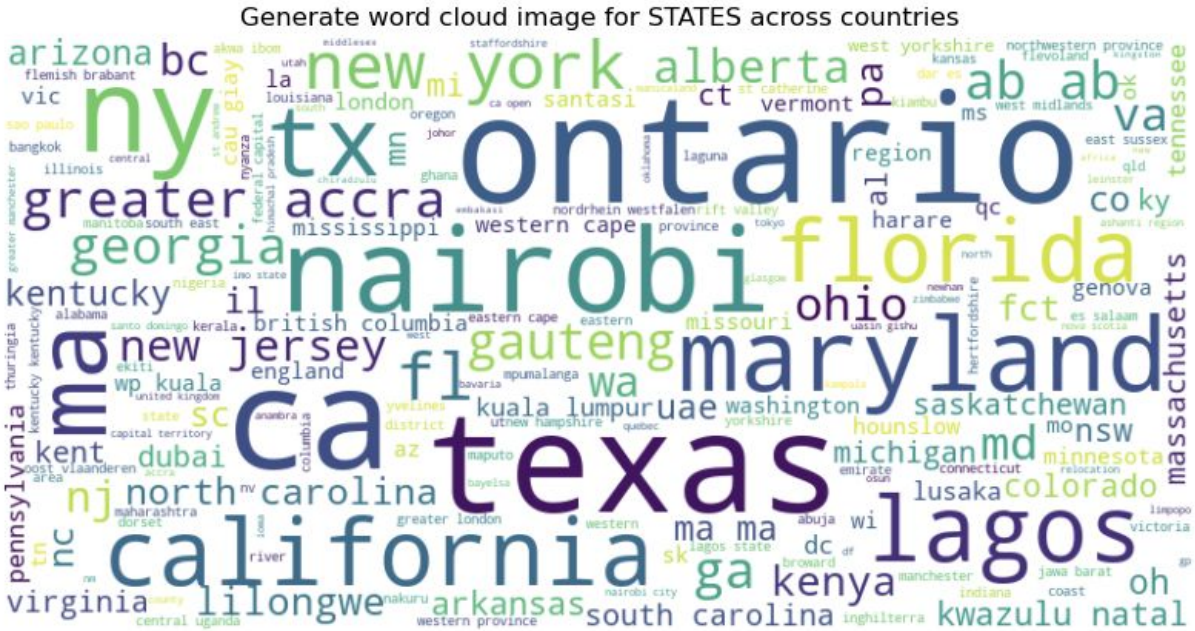


We can determine how many users have registered to CoverQuick's website for resume building by specifying the countries from the drop down.

1. Nairobi with 579 users
2. Mombasa : 21 users
3. Least number of users in Kenya Nakuru, nanyuki cities.: 13, and 12 and 11 respectively

Generate word cloud image for STATES across countries

- Nairobi, Lagos, Texas, Ontario, Gauteng are the Provinces, States applicants in the worldwide
- Delta states, mefu south, dar es, are the provinces which has less number of applicants



WordCloud of Most Common Skills and Skills description

Common Skills Among the users

- Management
- Technical skills
- Soft Skill (good communication, leadership skills, interpersonal skills)
- Language
- DevOps
- Software

Skills as per Description

1. Customer Service
2. Problem Solving
3. Time Management
4. Microsoft Office

Competitive Skills required : SQL, Critical thinking, Data Analysis, Analysis

WordCloud of Most Common Skills

WordCloud of Most Common Skills description

Final Dataset

- After the cleanup, we have 57 columns and their respective data type which will help us to answer our research questions going forward.

	ID	KEYWORDS	SUGGESTEDSKILLS	ROLE	CITY	STATE	SUMMARY	ACCOMPLISHMENTS	COUNTRY_CODE	SKILL_DESCRIPTION	...	PBL_DETAILS	PBL_PUBLISHER	CRT_NAME	CRT_ISSUER
0	CLG43D9AN007GX02UG1I694J6	[admissions representative, admissions, uma,...	[Compliance, Client, Manages, Interaction, Fin...		INDIO	CA	DETAILED AND DRIVEN, I HAVE BUILT STRONG COMMU...		US	VERBAL, WRITTEN, AND VISUAL COMMUNICATION, GOA...	...	NaN	NaN	QUALIFIED APPLICATOR CERTIFICATE	CALIFORNIA RIVERSIDE AGRICULTURE DEPARTMENT
1	CLG3ITETJ006JX92TDKCRW195	[dashboard interfaces, lead generation, mark...	[Analysis, Collection, Research]		ILMENAU	THURINGIA	DETAILED-ORIENTED UI/UX DESIGNER WITH EXPERIEN...		DE	FIGMA, SKETCH, ADOBE XD, FRAMER, MIRO, UXPIN, ...	...	NaN	NaN	VISUAL ELEMENTS OF USER INTERFACE DESIGN	CALIFORNIA INSTITUTE OF THE ARTS
2	CLG3IY1SD007RX32UTNUHNRGY	[product, design, development, business req...	[Vue, DevOps, Delivery]		PEORIA	ARIZONA	AGILE SOFTWARE ENGINEER WITH 2 YEARS OF EXPERI...		US	JIRA, VSC, MYSQL, GIT, BITBUCKET, GITHUB, POS...	...	NaN	NaN	NaN	NaN
3	CLG5J15LZ00K3X02UAAU7G9Z0	[flexibility, international exposure, dream ...	[]		MALANG		INNOVATIVE DIGITAL MARKETING PROFESSIONAL WITH...		GB	MARKETING ANALYTICS, WEBSITE ANALYTICS, PRODUC...	...	NaN	NaN	NaN	NaN
4	CLG43PTE600DDYA2UMAKFW3C3	[product roadmaps, new features, product enh...	[Curiosity]		CALGARY	AB	PASSIONATE JOB SEEKER WITH STRONG ORGANIZATION...		CA	CRITICAL AND ANALYTICAL THINKING, TIME MANAGEM...	...	NaN	NaN	SCRUM MASTER CERTIFICATION	LEARN QUEST
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11971	CLEEXYZAG006AYG2VHR087ALS	[content, ideation, creation, camera, cont...	[Instagram, Calendar, TikTok]		BROOKLYN	NY	DEPENDABLE VIDEOGRAPHER AND VIDEO EDITOR WITH ...		US	PROJECT MANAGEMENT, SELF-DRIVEN, PRODUCTION PL...	...	NaN	NaN	NaN	NaN
11972	CLEEC90B0005NYF2TLOS9QC95	[adobe premiere, adobe after effects, adobe ...	[Broadcast, Promotional, Broadcast & Promotion...		PEABODY	MA			US	FACEBOOK LIVE, TWITCH, OBS, XSPLIT	...	NaN	NaN	LEARN HTML COURSE	CODECADEMY
11973	CLEEY05QA000EXD2UP87UEHKZ	[write, content, graphics, imagery, social...	[]		CRIVITZ	WI	PERSONABLE AND HARDWORKING PROFESSIONAL WITH E...		GB	NaN	...	NaN	NaN	INTERNATIONAL ORGANIZATION MANAGEMENT	UNIVERSITY OF GENEVA
11974	CLE0EDRGO00A5WZ2UTRU0NT5U	[accountable, challenges, social media, bes...	[]		PEABODY	MA	PROFESSIONAL WITH OVER A DECADE OF EXPERIENCE ...		US	FACEBOOK LIVE, TWITCH, OBS, XSPLIT	...	NaN	NaN	LEARN HTML COURSE	CODECADEMY
11975	CLEECWHM6006DYF2TSR12F761	[product, product team, product managers, s...	[Communicate]		PEABODY	MA			US	FACEBOOK LIVE, TWITCH, OBS, XSPLIT	...	NaN	NaN	LEARN HTML COURSE	CODECADEMY

11976 rows × 57 columns

```
df.dtypes
```

ID	object
KEYWORDS	object
SUGGESTEDSKILLS	object
ROLE	object
CITY	object
STATE	object
SUMMARY	object
ACCOMPLISHMENTS	object
COUNTRY_CODE	object
SKILL_DESCRIPTION	object
SKILL	object
EDU_GPA	object
EDU_MINOR	object
EDU_AWARDS	object
EDU_SCHOOL	object
EDU_PROGRAM	object
EDU_LOCATION	object
EDU_COURSEWORK	object
EDU_GRADUATIONDATE	datetime64[ns]
EDU_GRAD_YEAR	int64
BIRTH_YEAR	int64
AGE_RANGE	int64
VLNTR_TITLE	object
VLNTR_ENDDATE	datetime64[ns]
VLNTR_LOCATION	object
VLNTR_STARTDATE	datetime64[ns]
VLNTR_DESCRIPTION	object
VLNTR_ORGANIZATION	object
EXP_TITLE	object
EXP_COMPANY	object
EXP_ENDDATE	datetime64[ns]
EXP_LOCATION	object
EXP_STARTDATE	datetime64[ns]
EXP_DESCRIPTION	object
PRJ_LINK	object
PRJ_TITLE	object
PRJ_SKILLS	object
PRJ_ENDDATE	datetime64[ns]
PRJ_STARTDATE	datetime64[ns]
PRJ_DESCRIPTION	object
REF_NAME	object
REF_EMAIL	object
REF_PHONENUMBER	object
REF_RELATIONSHIP	object
PBL_DATE	datetime64[ns]
PBL_LINK	object
PBL_NAME	object
PBL_DETAILS	object
PBL_PUBLISHER	object
CRT_NAME	object
CRT_ISSUER	object
CRT_DATERECEIVED	datetime64[ns]
AWD_NAME	object
AWD_ISSUER	object
AWD_DETAILS	object
AWD_DATERECEIVED	datetime64[ns]
AWD_DESCRIPTION	object

dtype: object

Determine the approximate age range and experience level

- Predefined duration thresholds are established in order to classify experience levels. These levels are classified as 'BEGINNER' for experience durations up to 1 year (365 days), 'INTERMEDIATE' for durations ranging from 1 to 2 years (365 to 730 days), and 'ADVANCED' for durations between 2 to 3 years (730 to 1095 days).
- Subsequently, the DataFrame is modified to incorporate an additional column called 'EXP_DURATION', which denotes the duration of experience calculated in days. Additionally, an 'EXP_LEVEL' column is introduced, which classifies the experience level based on predefined thresholds.

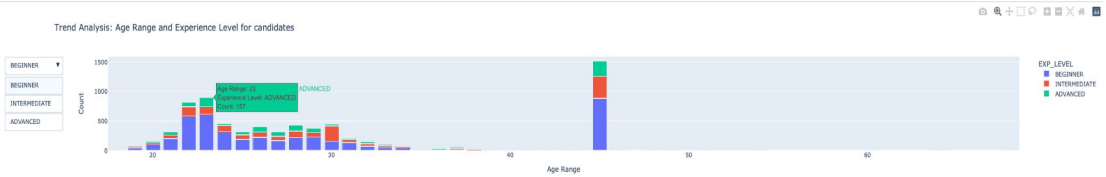
```
Determine the approximate age range and experience level:
=========================================================
```

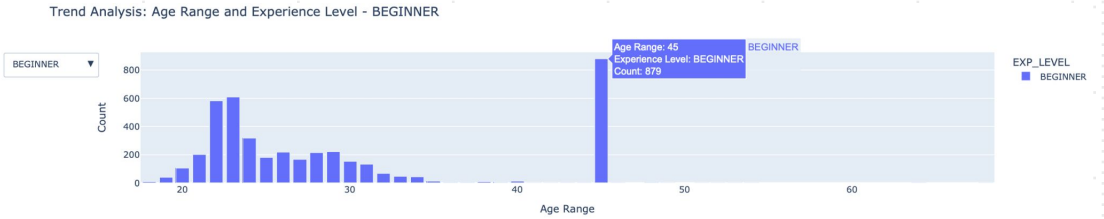
	ID	COUNTRY_CODE	BIRTH_YEAR	AGE_RANGE	EXP_DURATION	EXP_LEVEL
0	CLG43D9AN007GX02UG1I694J6	US	1997	26	NaN	NaN
1	CLG3ITETJ006JX92TDKCRW195	DE	1995	28	699.0	INTERMEDIATE
2	CLG3IY1SD007RX32UTNUHNRGY	US	1998	25	672.0	INTERMEDIATE
3	CLG5J15LZ00K3X02UAAU7G9Z0	GB	1998	25	122.0	BEGINNER
4	CLG43PTE600DDYA2UMAKFW3C3	CA	1999	24	92.0	BEGINNER
...	...	...	...	...	...	...
11971	CLEEXYZAG006AYG2VHR087ALS	US	1987	36	519.0	INTERMEDIATE
11972	CLEEC90B0005NYF2TLOS9QC95	US	1998	25	NaN	NaN
11973	CLEEY05QA000EXD2UP87UEHKZ	GB	1993	30	730.0	INTERMEDIATE
11974	CLE0EDRGO00A5WZ2UTRU0NT5U	US	1998	25	NaN	NaN
11975	CLEECWHM6006DYF2TSR12F761	US	1998	25	NaN	NaN

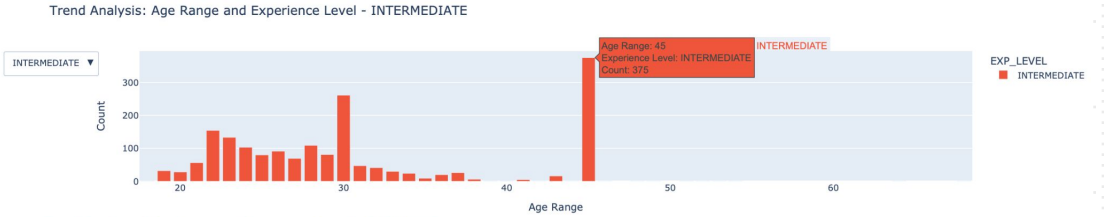
11976 rows × 6 columns

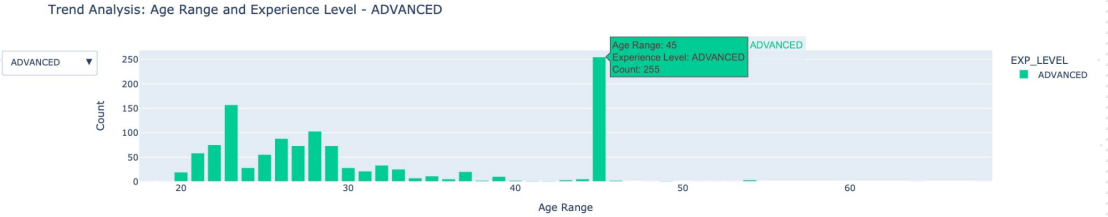
Determine the expected age and experience level

- The x-axis represents the age ranges, ranging from 18-24, 25-34, 35-44, 45-54, and 55+. The y-axis represents the count of candidates. Each bar in the chart is segmented into different colors representing different experience levels.
- By selecting different experience levels from the dropdown menu, the chart dynamically updates to show the trend analysis specifically for that experience level. The title of the chart also changes accordingly to provide focused insights.
- Hovering over each bar provides additional information, including the specific age range, experience level, and the corresponding count of candidates.
- The visualization aims to provide a clear and visually appealing representation of the distribution of candidates across different age ranges and experience levels, allowing for a quick and comprehensive understanding of the trends in your project data.



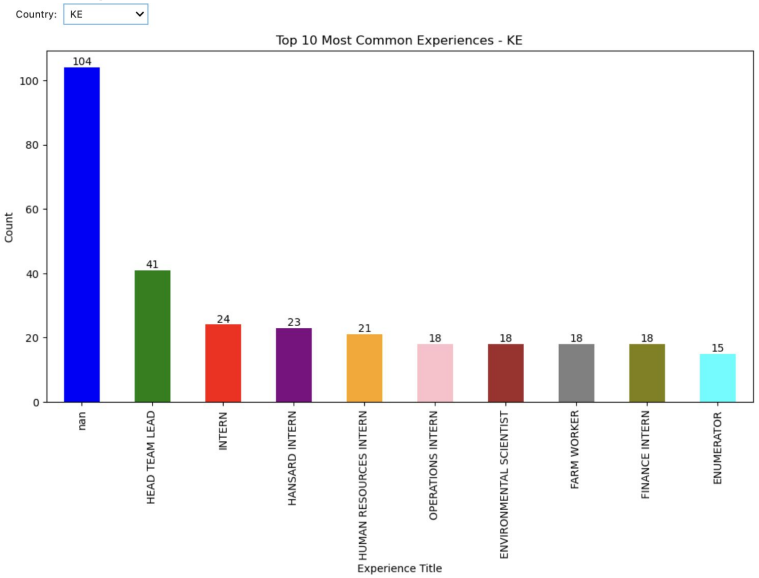


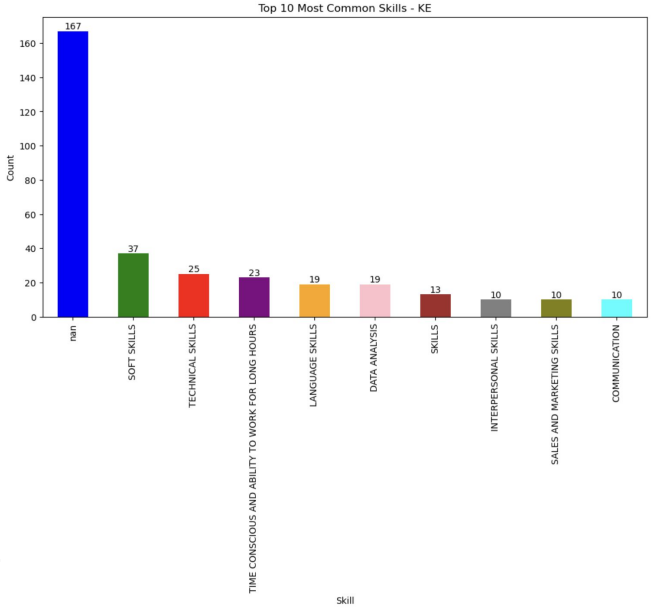




Determine trends in experience and skills for the target users.

A bar chart can be utilized to visually represent the top 10 prevalent experiences and skills. By utilizing a dropdown menu, users have the option to choose specific countries and observe the corresponding top 10 experiences and skills associated with those countries. This graphical representation allows for a clear understanding of the most common experiences and skills across different countries.





Top 10 Most Common Skills:

Skill	Count
nan	167
SOFT SKILLS	37
TECHNICAL SKILLS	25
TIME CONSCIOUS AND ABILITY TO WORK FOR LONG HOURS	23
LANGUAGE SKILLS	19
DATA ANALYSIS	19
SKILLS	13
INTERPERSONAL SKILLS	10
SALES AND MARKETING SKILLS	10
COMMUNICATION	10

Top 10 Most Common Experiences:

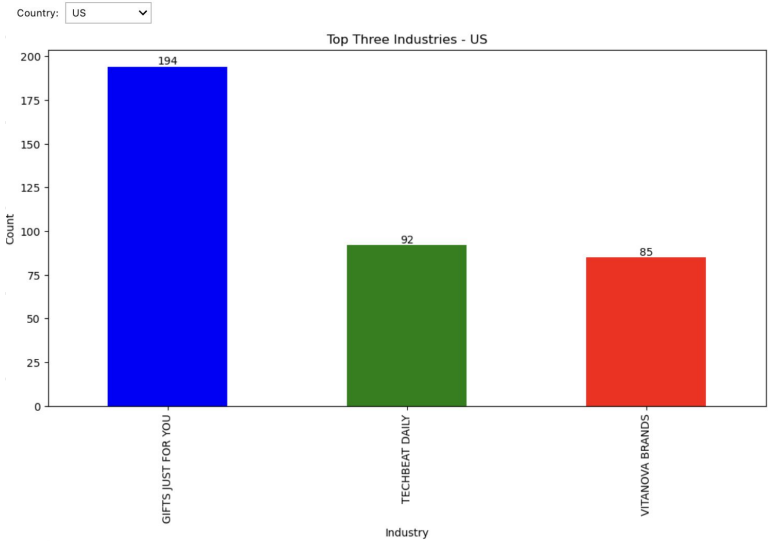
Experience	Count
nan	104
HEAD TEAM LEAD	41
INTERN	24
HANSARD INTERN	23
HUMAN RESOURCES INTERN	21
OPERATIONS INTERN	18
ENVIRONMENTAL SCIENTIST	18
FARM WORKER	18
FINANCE INTERN	18
ENUMERATOR	15

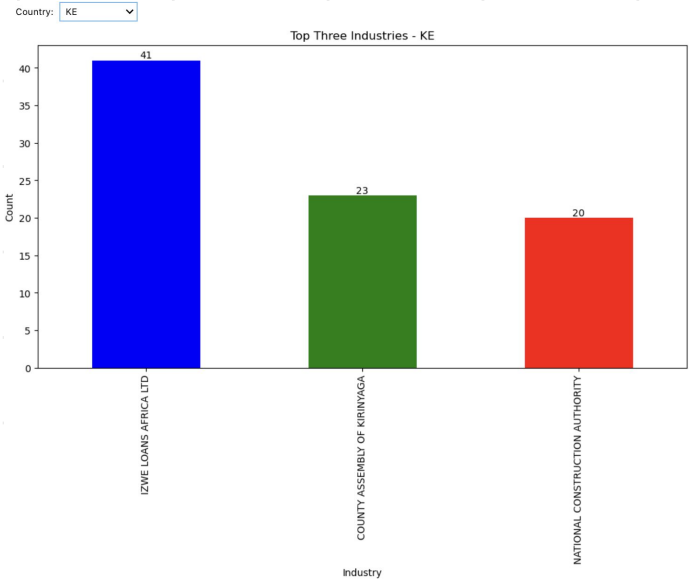
Top three industries that the majority of CoverQuick's users have applied:

As shown below, from the total dataset with job description, we can see that the top three industries/company candidates apply to are mostly GIFTS JUST FOR YOU (194), LISAP (163), and EDMONTON FIRE RESCUE (159).

```
Top Three Industries:
GIFTS JUST FOR YOU                          194
LIVINGSTONIA SYNOD AIDS PROGRAMME (LISAP)   163
EDMONTON FIRE RESCUE                        159
```

We can also select the countries from the dropdown and get the top three industries per country wise.





Discover trends in demographics and find which industries yield the best and the worst resumes.

Resume Optimality Criteria

1. Important Sections: This may include and not be limited to: work experience, education, projects, as the most important and relevant sections.
2. Resume Length: The solid resume length may be between 300-500 words, however; if the length is outside this range, it may not mean a resume is poor.
3. Use of action verbs: Direct use of action verbs in the bullets of a resume will ensure a resume will perform better.
4. No use of pronouns: Resumes should not contain pronouns such as I, we or me written in the document.
5. Excessive bullet points: A resume experience or section should not have an excessive number of bullet points. If a section has over 10 bullet points, it is looked upon unfavourably.
6. Spelling Mistakes: A resume with spelling errors is immediately penalized against.
7. Excessive sentence or bullet length

RESUME LENGTH

EXP_DURATION	EXP_LEVEL	RES_LEN
nan	NaN	188
699.0	INTERMEDIATE	259
672.0	INTERMEDIATE	188
122.0	BEGINNER	161
92.0	BEGINNER	136
...	...	...
519.0	INTERMEDIATE	145
nan	NaN	484
730.0	INTERMEDIATE	155
nan	NaN	349
nan	NaN	451

We iterate over the specified columns and count the total number of words. The word count is then added as a new column 'RES_LEN' to the DataFrame.

Finally, we will map it to 'POOR' if it's below 300 and 'GOOD' if it's greater than or equal to 300 using our score-card.

```
RES_LEN Non-Null and Null Count:
==================================
   Total Count  Non-Null Count  Null Count
0        11976           11976           0
```

ACTION VERBS

EXP_DURATION	EXP_LEVEL	RES_LEN	ACTN_VERB
nan	NaN	188	47
699.0	INTERMEDIATE	259	58
672.0	INTERMEDIATE	188	56
122.0	BEGINNER	161	49
92.0	BEGINNER	136	41
...	...	...	...
519.0	INTERMEDIATE	145	40
nan	NaN	484	116
730.0	INTERMEDIATE	155	36
nan	NaN	349	89
nan	NaN	451	109

```python
import nltk
from nltk import word_tokenize
from nltk.corpus import stopwords
from nltk.corpus import wordnet

# Download necessary NLTK resources if not already downloaded
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
ACTN_VERB Non-Null and Null Count:
==================================
   Total Count  Non-Null Count  Null Count
0        11976           11976           0
```

SPELLING MISTAKES

EXP_DURATION	EXP_LEVEL	RES_LEN	ACTN_VERB	SPLNG_MSTK
nan	NaN	188	47	1
699.0	INTERMEDIATE	259	58	1
672.0	INTERMEDIATE	188	56	1
122.0	BEGINNER	161	49	1
92.0	BEGINNER	136	41	1
...	...	...	...	...
519.0	INTERMEDIATE	145	40	1
nan	NaN	484	116	1
730.0	INTERMEDIATE	155	36	1
nan	NaN	349	89	1
nan	NaN	451	109	1

```
'a'
'aardvark'
'ab'
'aback'
'abacus'
'abalone'
'abandon'
'abandoned'
'abandoning'
'abandonment'
'abandons'
'abase'
'abased'
'abate'
'abated'
'abatement'
'abates'
'abattoir'
'abba'
'abbas'
'abbess'
'abbey'
"abbey's"
'abbeys'
'abbie'
'abbies'
'abbot'
"abbot's"
'abbots'
'abbott'
"abbott's"
'abbreviate'
'abbreviated'
'abbreviation'
'abbreviations'
'abby'
"abby's"
'abc'
'abdal'
'abdicate'
'abdicated'
'abdicating'
'abdication'
'abdomen'
"abdomen's"
'abdomens'
'abdominal'
'abdominals'
'abduct'
'abducted'
```

Output of this cell has been trimmed on the initial display.
Displaying the first 50 top outputs.
Click on this message to get the complete output.

```
!pip install spellchecker
```

```
import pandas as pd
from spellchecker import SpellChecker
```

```
splng_mstk.SPLNG_MSTK.unique()
```

```
array([ 1,  0,  2, 24,  9,  3,  5,  4])
```

SPLNG_MSTK Non-Null and Null Count:

==========

	Total Count	Non-Null Count	Null Count
0	11976	11976	0

IMPORTANT SECTIONS

EXP_DURATION	EXP_LEVEL	RES_LEN	ACTN_VERB	SPLNG_MSTK	IMP_SEC
nan	NaN	188	47	1	1
699.0	INTERMEDIATE	259	58	1	1
672.0	INTERMEDIATE	188	56	1	1
122.0	BEGINNER	161	49	1	1
92.0	BEGINNER	136	41	1	1
...	...	...	...	...	...
519.0	INTERMEDIATE	145	40	1	1
nan	NaN	484	116	1	1
730.0	INTERMEDIATE	155	36	1	1
nan	NaN	349	89	1	1
nan	NaN	451	109	1	1

We included and not be limited to: work experience, education

We are checking both the conditions and mapping value to 1 in IMP_SEC

```
imp_sec.loc[~imp_sec['EXP_DURATION'].isna() &
(imp_sec['EDU_GRAD_YEAR'] != 1900), 'IMP_SEC'] = 1
```

```
1    8486
0    3490
```

NEXT ACTION, FUTURE ACTION AND LIMITATION

We will be checking

- No use of Pronouns
- Excessive Bullet length -> Penalise it using our score card generator

SCORE CARD:

We will be considering all the values in the respective columns that we have created and based on the the final column "SCORECARD" we will be defining if the resume is good or bad resume.

FUTURE SCOPE:

Based on our scorecard analysis and considering relevant columns for Suggested Key Skills and Market analysis for respective Jobs which the candidate will be applying for, we will be suggesting and modifying the resume and showcase the analysis for better understanding to the user.

LIMITATIONS

- No control over the database.
- No overview on the current code base.
- Higher computing power.
- Additional services which can be paid.

Thank You!

```
def gratitude():
    print("Thank you.")

```