Northeastern University
College of Professional Studies

ALY 6080: XN Project CoverQuick Individual Draft

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

Submitted by
Siddharth Alashi
002728528
Alashi.s@northeastern.edu

CoverQuick

About Sponsor

An AI-powered programme called CoverQuick creates original cover letters and resumes for job applicants. Its objective is to assist job searchers in standing out from the competition by producing customised cover letters and resumes that are targeted to certain job applications using the most recent AI algorithms. CoverQuick makes sure that candidates do not send generic cover letters that fall flat with employers by providing personalised cover letters. Additionally, the programme provides features like resume grading, cover letter generation, application tracking, and resume preparation. There were 5000 users of CoverQuick as of September 2022.

- Products : Prepare Resume , CoverLetter, Track Application, Resume Grader
- Number of Users : 5000 users (Ref till September 2022)

Dataset for Analysis

DATASET PROVIDED

A. WITH JOB DESCRIPTION
B. WITHOUT JOB DESCRIPTION

Research Question

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

- an EDA-based dataset of job descriptions.The nested and json-formatted columns make up the dataset.
- The top three industries for which the majority of users have submitted applications are identification and visualisation.
- A visual representation is used to identify and convey the general age range and level of experience.
- The experience and capability trends of these target consumers are described and visualised.
- There are identified and visualised demographic trends for the number of candidates from around the world registering for resume creation on the website.

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

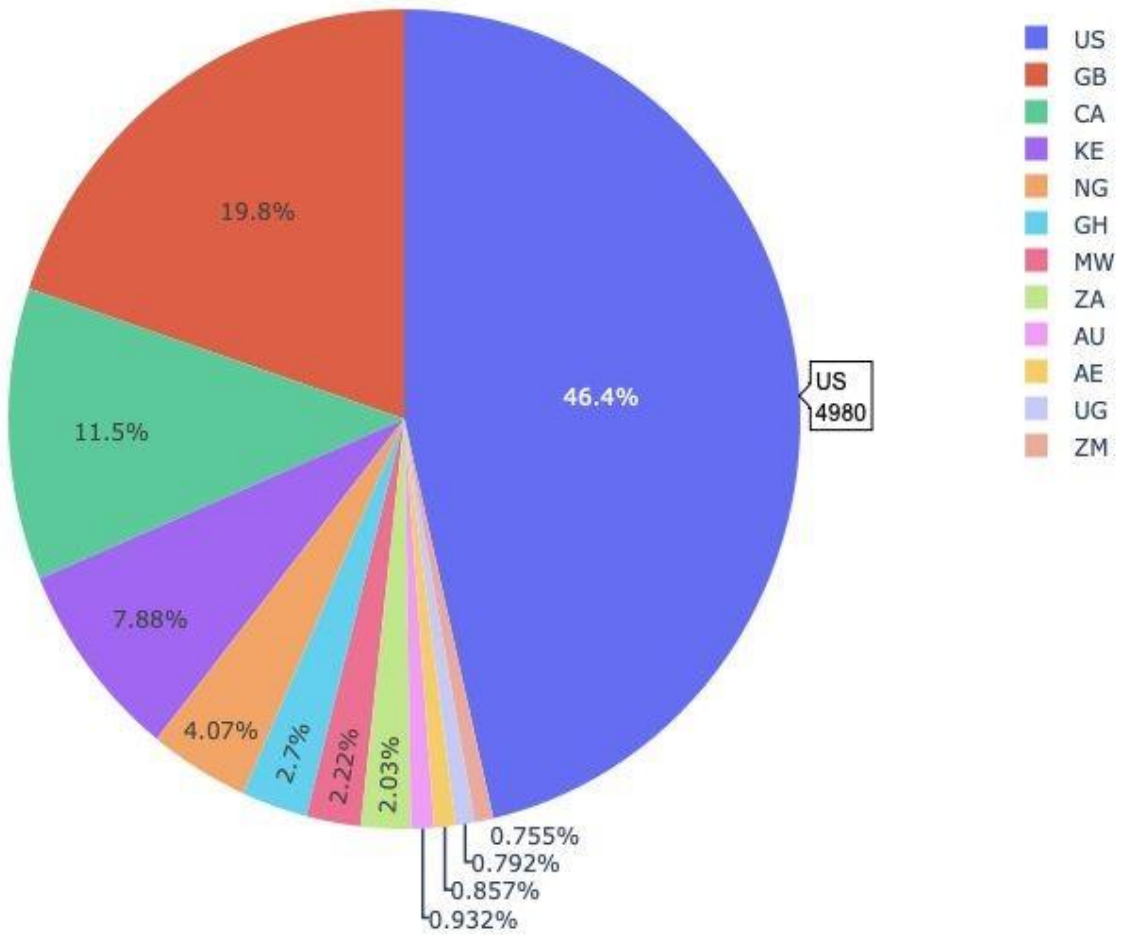
Showing max, min length and NA values:

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

NULL count that was useful for future visualisation.

COUNTRY_CODE Non-Null and Null Count:

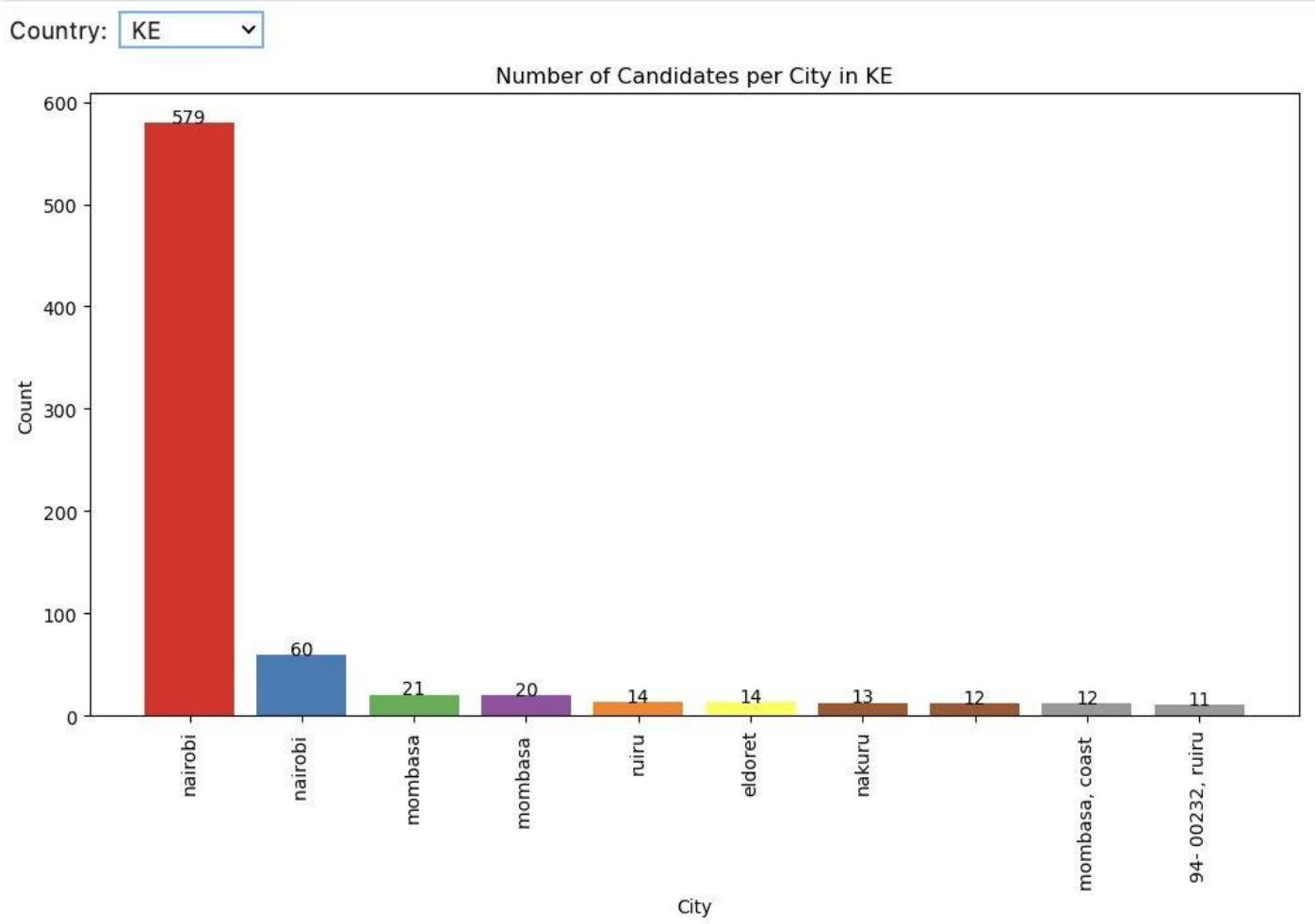
	Total Count	Non-Null Count	Null Count
0	11976	11976	0



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



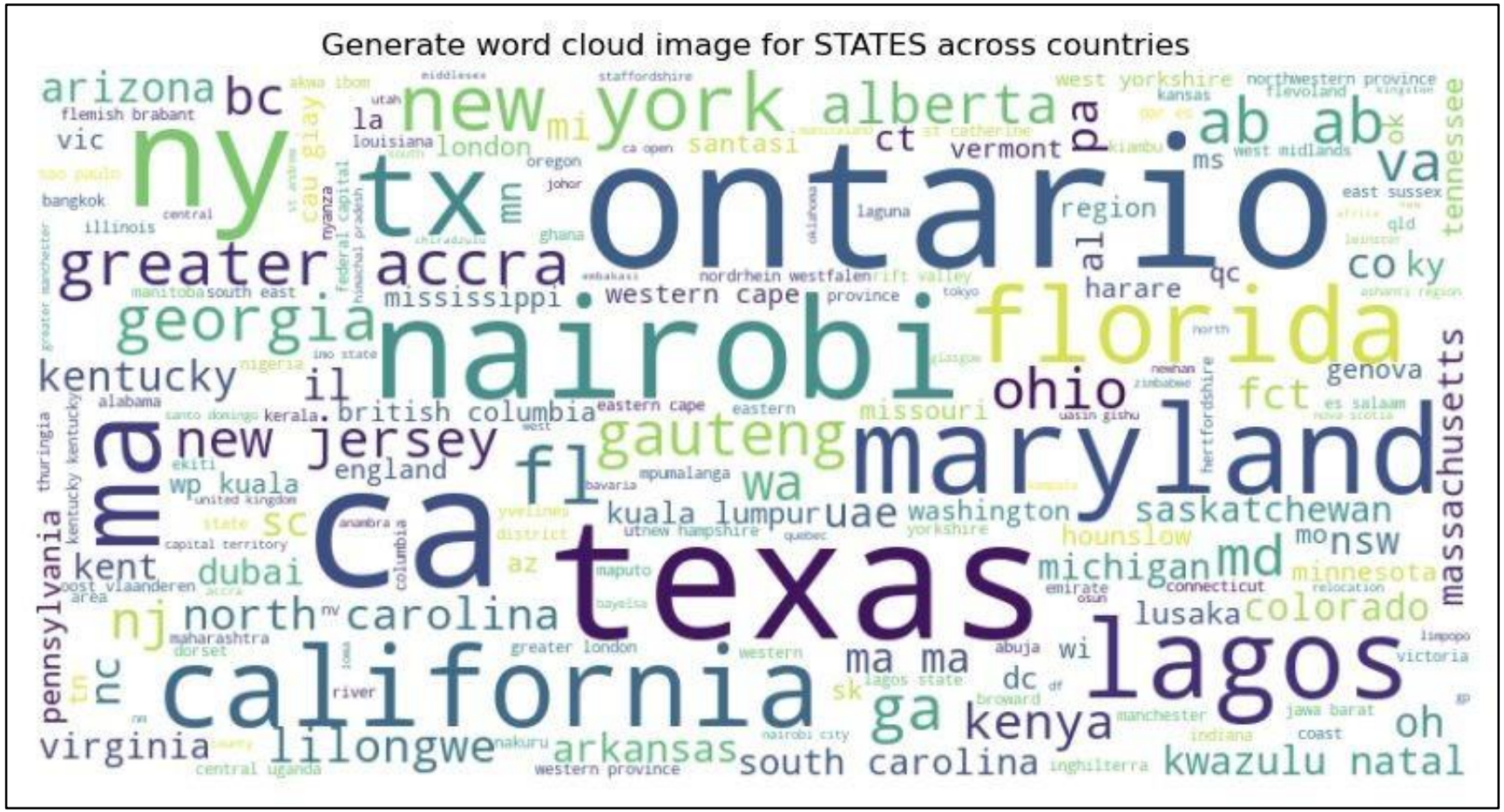
By selecting the countries from the drop-down option, we may see how many people have registered on CoverQuick's website for resume creation.

1. **Nairobi** has 579 users
2. **Mombasa** : 21 users.
3. Kenya's Nakuru and Nanyuki cities had the fewest users, with 13, 12, and 11 users, respectively.

WordCloud for Maximum States of Applicants

- 1. The Provinces and States of Nairobi, Lagos, Texas, Ontario, and Gauteng participated in the global application process.

- 2. The Delta States, Mefu South, and Dar Es Salaam are the provinces with the fewest candidates.



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

Skillsas per Description

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

Competitive Skills required : SQL, Cr
Data Analysis, Analysis

WordCloud of Most Common Skills

WordCloud of Most Common Skills description

Final Dataset : After DataCleaning

- We now have 57 columns and their corresponding data types, which will enable us to continue answering our study questions.

	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
```

Age Range and Experience Prediction

- 1. Based on predetermined time criteria, experience levels are classed. The three levels are labelled "BEGINNER" for experience durations up to one year (365 days), "INTERMEDIATE" for experience periods of one to two years (365 to 730 days), and "ADVANCED" for experience periods of two to three years (730 to 1095 days).

- 2. Next, a new column named "EXP_DURATION" is added to the DataFrame to reflect the length of experience in days. Additionally, a brand-new column named "EXP_LEVEL" is included, classifying experience levels in accordance with pre-established standards.

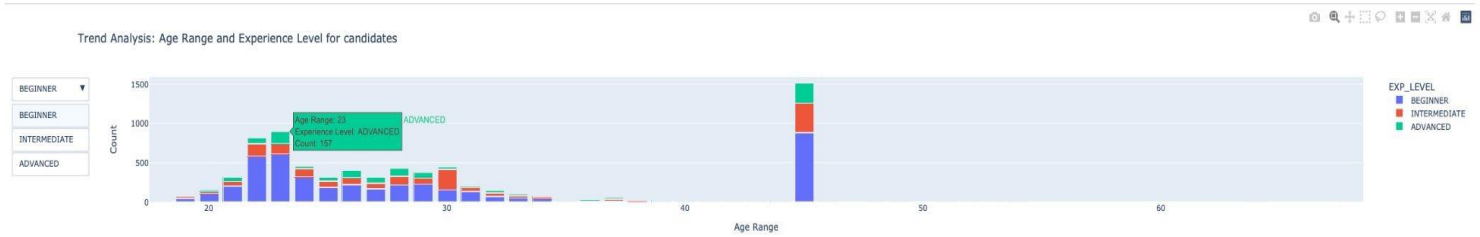
Determine the approximate age range and experience level:

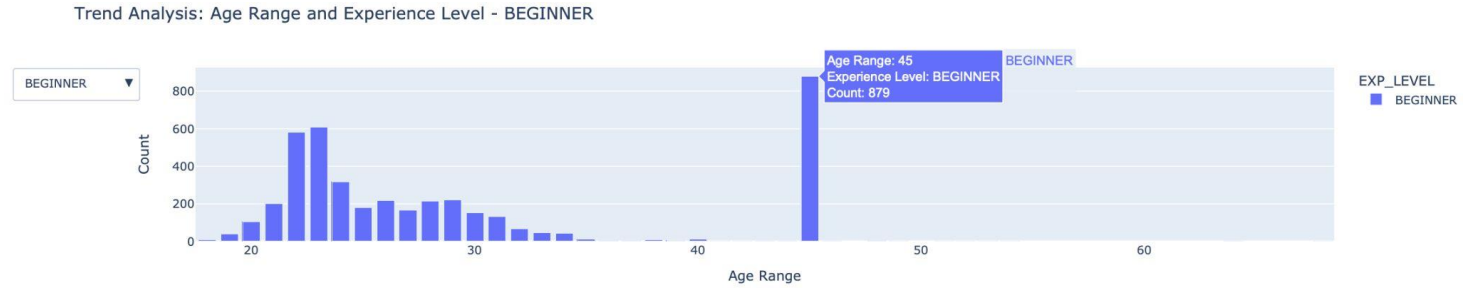
	ID	COUNTRY_CODE	BIRTH_YEAR	AGE_RANGE	EXP_DURATION	EXP_LEVEL
0	CLG43D9AN007GX02UG1I694J6	US	1997	26	NaN	NaN
1	CLG3ITETJ006JX92TDKCRW195	DE	1995	28	699.0	INTERMEDIATE
2	CLG3IY1SD007RX32UTNUHNRGY	US	1998	25	672.0	INTERMEDIATE
3	CLG5J15LZ00K3X02UAAU7G9Z0	GB	1998	25	122.0	BEGINNER
4	CLG43PTE600DDYA2UMAKFW3C3	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

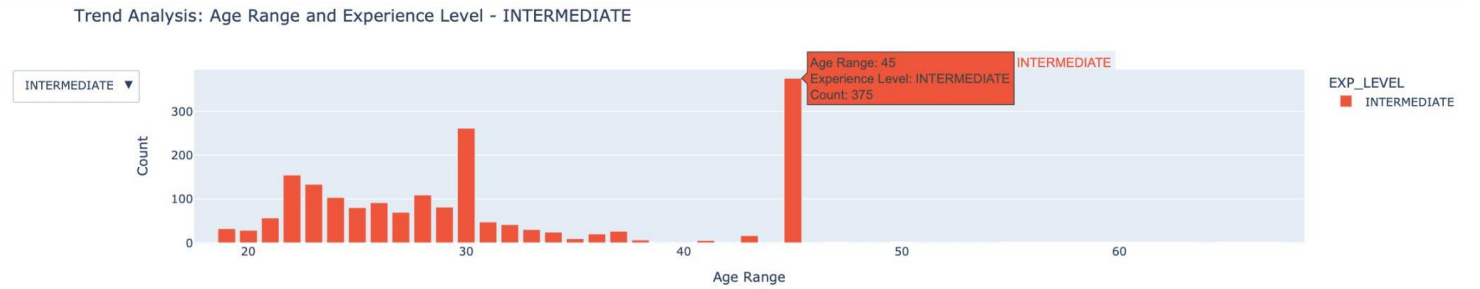
11976 rows × 6 columns

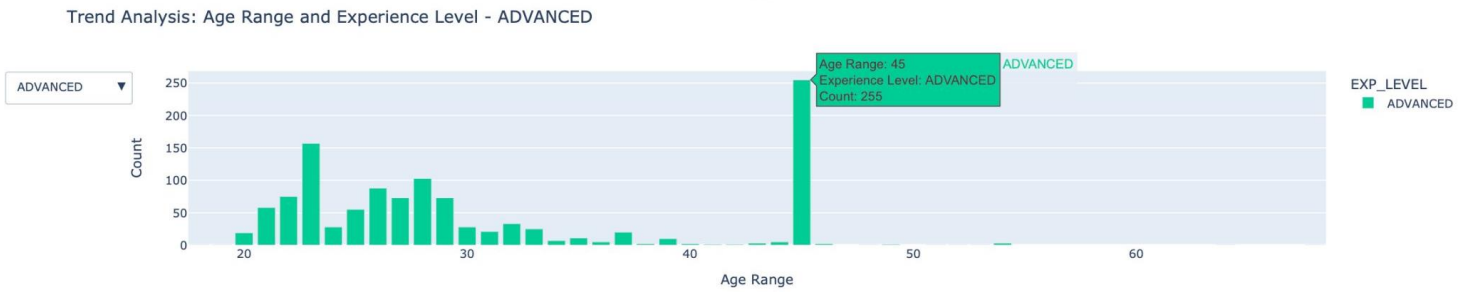
Determine the expected age and experience level

- The chart showcases age ranges (18-24, 25-34, 35-44, 45-54, and 55+) on the horizontal axis, while the vertical axis indicates the number of candidates. Each bar on the chart is color-coded to represent different experience levels.
- By selecting various experience levels from the dropdown menu, the chart dynamically updates to display trend analysis specific to the chosen experience level. The chart's title also adapts to provide focused insights based on the selected experience level.
- Hovering over each bar reveals additional details, such as the age range, experience level, and the corresponding count of candidates.
- The visualization aims to present candidate distribution across age ranges and experience levels in a visually appealing and easily understandable manner. It facilitates swift and comprehensive comprehension of the data.



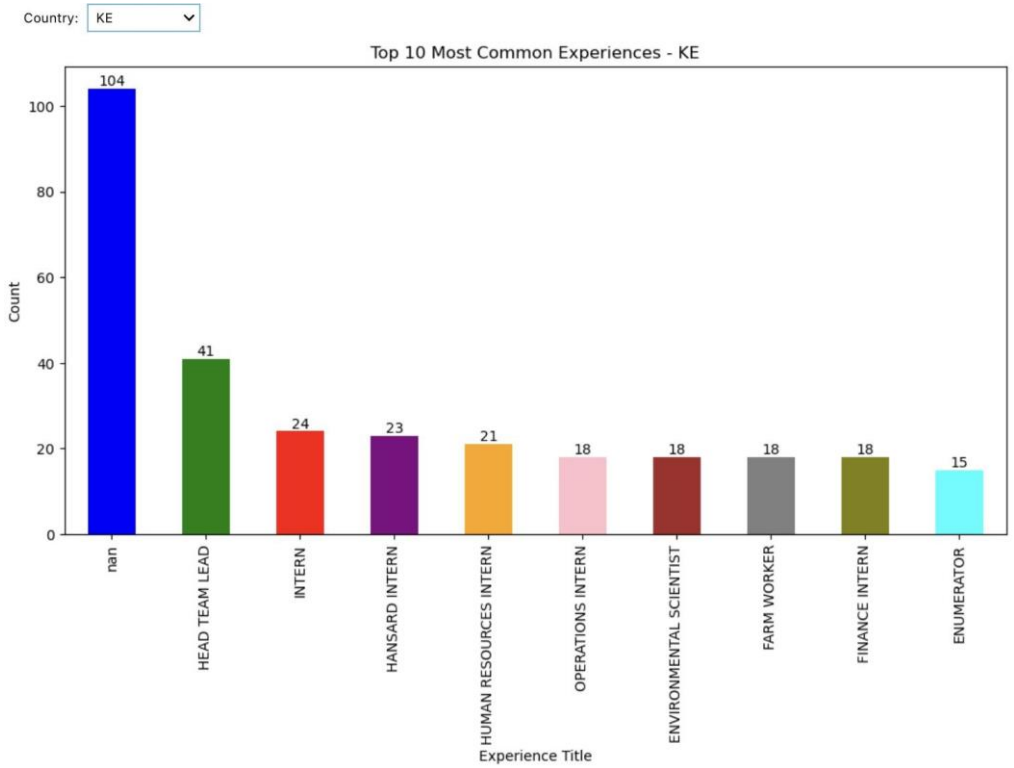


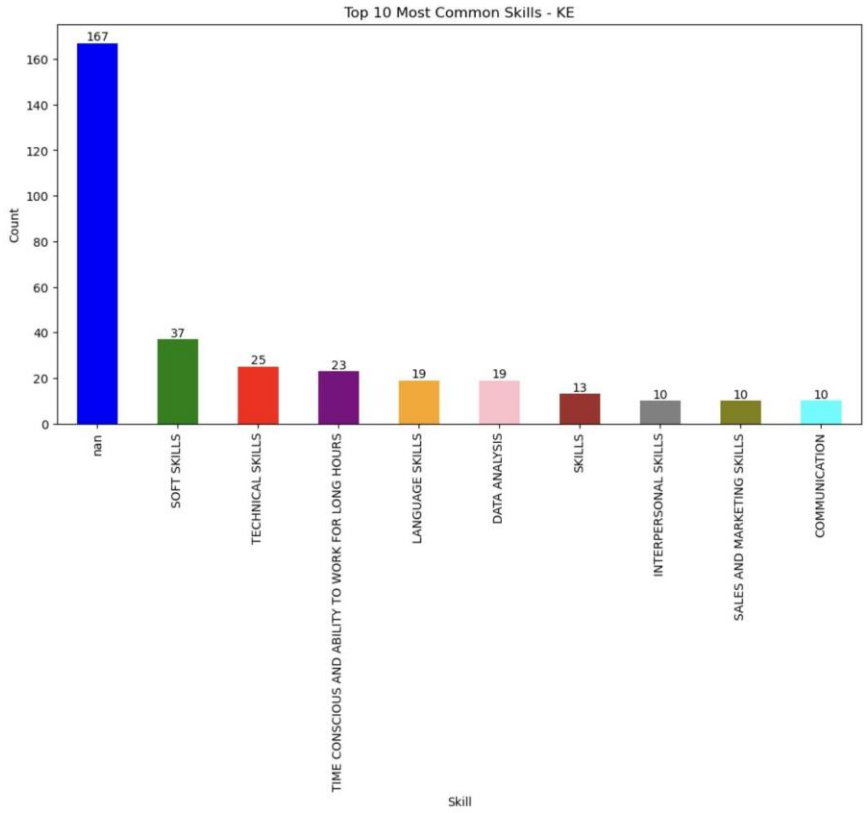




Determine trends in experience and skills for the target users.

A bar chart may be used to visually depict the top 10 experiences and skills. Using a dropdown menu, users may choose specific countries to display the top 10 skills and experiences associated with particular country. Thanks to this graphical representation, it is easy to easily understand the most common experiences and skills across different cultures.





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:

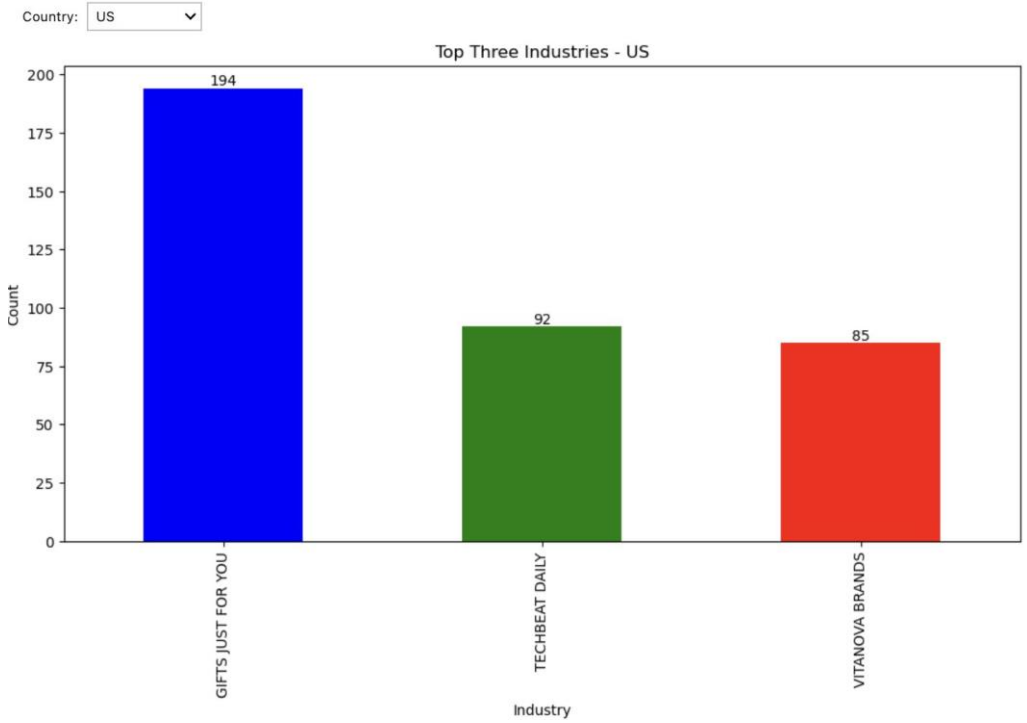
Experience	Count
nan	104
HEAD TEAM LEAD	41
INTERN	24
HANSARD INTERN	23
HUMAN RESOURCES INTERN	21
OPERATIONS INTERN	18
ENVIRONMENTAL SCIENTIST	18
FARM WORKER	18
FINANCE INTERN	18
ENUMERATOR	15

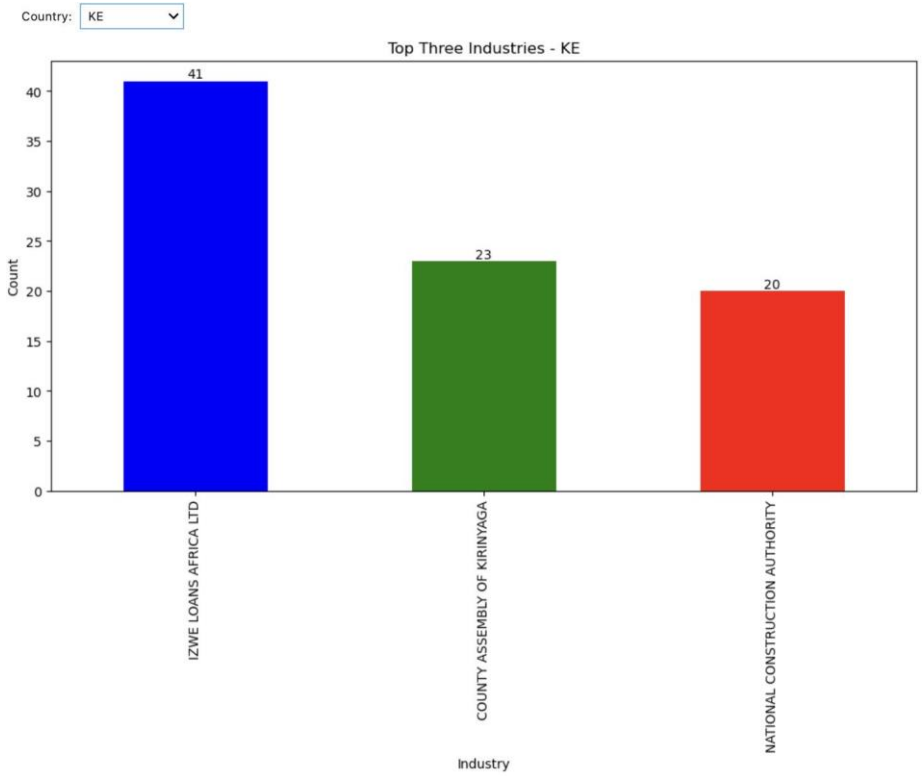
Filter Top three industries

Based on the dataset with job descriptions, the three sectors and companies that receive the highest number of job applications are 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
```

To determine the top three industries for each nation, you can select them from the menu provided.





Demographics trends and Resume Quality

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 count the total number of words in each column to get the word count in certain columns. This word count is then added to the DataFrame as a new column with the name "RES_LEN."The word count is then graded using a score-card, with 'POOR' being assigned if it is **less than 300 and 'GOOD' being assigned if it is 300 or above.** By using this mapping, we may classify the word count according to predetermined standards.

RES_LEN Non-Null and Null Count:

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

Action Words

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 : In Resume

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 evaluate the following criteria:
1. Checking for the absence of pronouns.
2. Identifying excessive bullet length and penalizing it using our scorecard generator.

Using our scorecard, we will consider all the values in the respective columns we have created. Based on the final column called "SCORECARD," we will determine whether the resume is classified as good or bad.

In terms of future plans, we aim to utilize the scorecard analysis to suggest and modify resumes by considering relevant columns for suggested key skills and market analysis specific to the jobs the candidates are applying for. This will provide a better understanding to the users through showcasing the analysis.

However, there are certain limitations to consider, such as not having control over the database, lack of overview on the current code base, the need for higher computing power, and the possibility of additional paid services.

Thank You!