**HDSC Winter ‘22 Premier Project**

**Presentation: ANALYZING ONLINE JOB POSTS**

**A project by Team Gradboost**

**INTRODUCTION**

The demand for jobs in the labor market is continuously evolving with the use of technology to advertise job openings, hence, the need to understand the demand for certain job titles and professions. There is also a need to identify skills that are most frequently required by employers, how the distribution of necessary skills changes over time and making recommendations to job seekers and employers.

The online job market is a strong predictor of the local economy's overall labor need. Furthermore, data from online job posts is easier and faster to obtain, and they can be a more comprehensive source of information than traditional job postings such as those seen in newspapers.

The job market has made significant improvements in areas of advertising in recent times. The advent of social media and job/ career sites has made for easy access to a collection of available jobs in one’s environment. However, a challenge experienced by potential employees and employers is how to get appropriate job postings matched with candidates that are qualified / suitable for them.

As a result, this project sets out to perform analysis on an online job posting dataset and make inferences regarding the most in demand job titles, varying skills in demand and setting up a recommendation system to help potential employers link up with appropriate employees.

**PROBLEM STATEMENT**

This project is aimed at answering the following questions:

1. Finding out the jobs that are in highest demand
2. Identifying skills mostly required by employers
3. Discovering the distribution of how necessary skill changes over time
4. Making recommendations to job seekers and employers

**DATA UNDERSTANDING**

The dataset used for this project was obtained from Kaggle competition. Each row represents a job post. The dataset representation is tabular, but many of the columns are textual/unstructured in nature. Most notably, the columns JobDescription, JobRequirement, RequiredQual, ApplicationP and AboutC are textual. The column jobpost is an amalgamation of these various textual columns. The dataset contains about 19,000 job postings from 2004 to 2015 posted on Career Center, an Armenian human resource portal. Analysis done on this data will be able to understand the nature of the ever-changing job market, as well as the overall demand for labour in the economy. The data was originally scraped from a Yahoo! Mailing group.

**DATA EXPLORATION**

The dataset can be downloaded from [here](https://www.kaggle.com/madhab/jobposts). There are 24 columns to describe various job posts.

**Key Observations:**

1. Data entries in the jobpost column was scraped from the web and was processed to create other columns.

2.There are repeated entries of similar jobposts in the ‘jobpost’ column, consequently, jobposts from the same company posted in the same year and month are duplicated.

3. The dataset has a lot of null values (30%).

4. The only significant columns to answer the objectives are columns with the required information ‘Title’, ‘IT’, ‘Company’, ‘JobDescription, ‘JobRequirment’, ‘RequiredQual’, Month, and ‘Year.

5. The dataset used in this project like every real-life data needs to be cleaned as null values, duplicated values and redundant columns/rows are present.

**DATA-CLEANING ACTIVITIES**

1. Columns in the dataset were renamed for easy identification

2. All duplicated rows in the jobpost column were removed using the drop\_duplicates() function.

3. Null values in Title column was dropped as it would have no significant effect on the data set.

4. The next action carried out was the creation of a new dataframe (job\_demand) containing a subset of the data with only the columns required to answer each of the objectives.

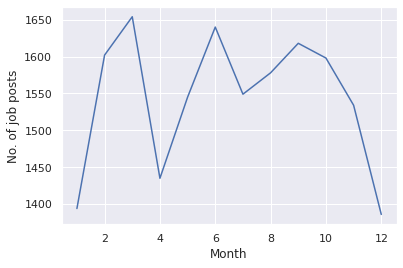
**ANALYZING THE CLEANED DATASET**

This is performed to give proper insight about the dataset in terms of its summary statistics and visualization. To solve the questions posed by our objectives, the following analysis were carried out on the dataset.

An initial count to determine the year with the highest number of job postings was done which revealed 2012 as the highest. Similarly, we were able to show that March had the highest number of job postings and a steady decline of job posting started from September till December.

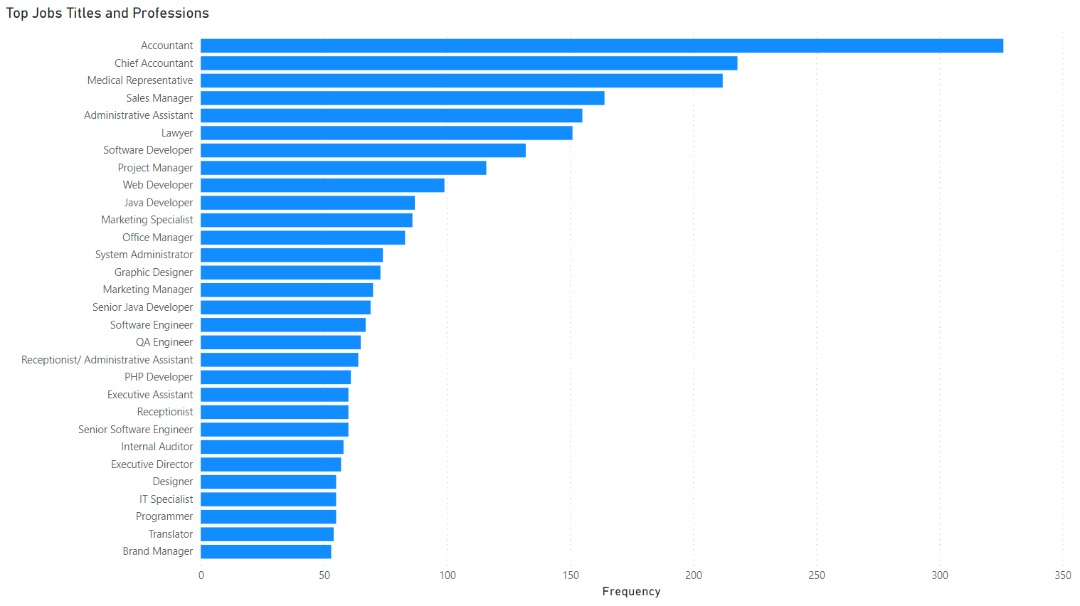


**Plot of Job postings per Year**

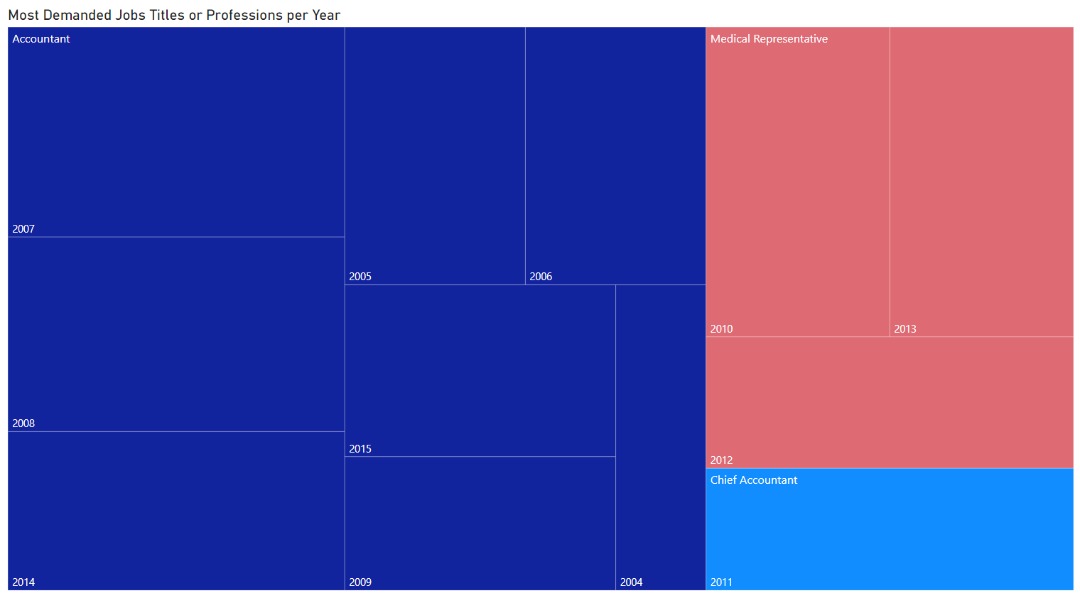


**Plot of Job posting per month**

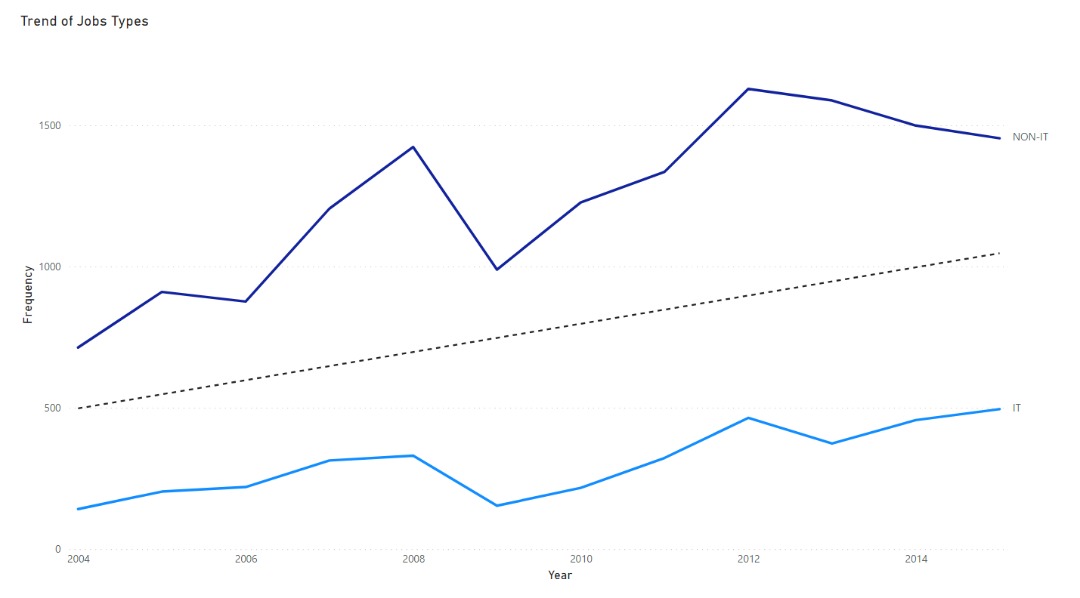
Using the title column of the job\_demand dataset, we were able to discern the top 30 jobs between 2004 and 2015. It was revealed that accountant positions were in high demand over the years closely followed by chief accountants with medical representatives making up the top 3 most in demand jobs.



In terms of yearly job demand, accountant positions were highest in most years with only medical representatives beating it to the top spot in years 2010, 2012, and 2013.



The demand for IT jobs was also seen to have been continuously lower than that of NON-IT jobs over the years indicating a larger majority of the populace in Armenia work in NON-IT industries.

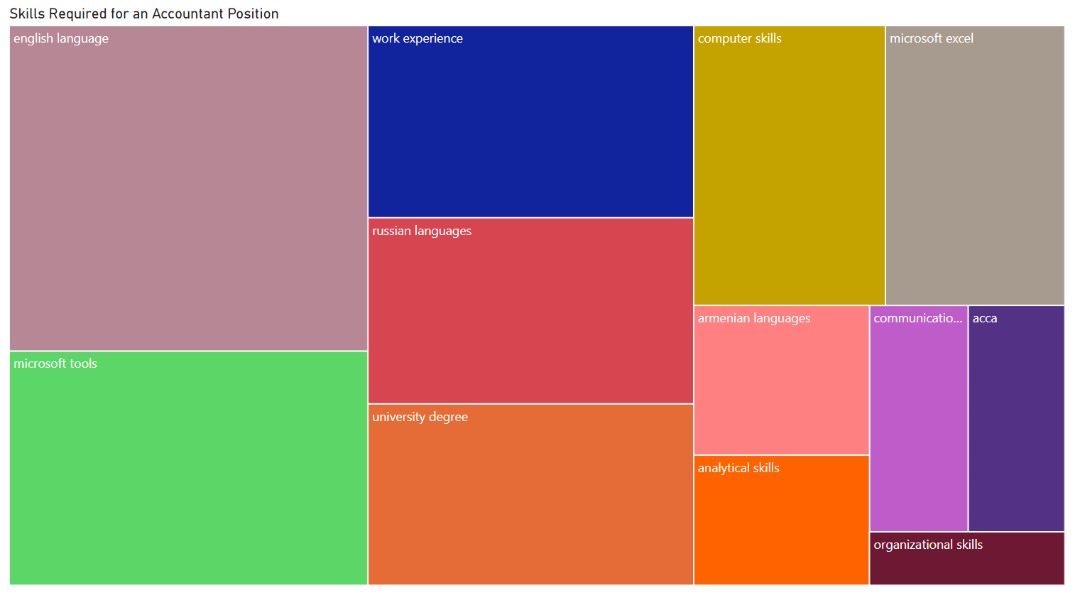


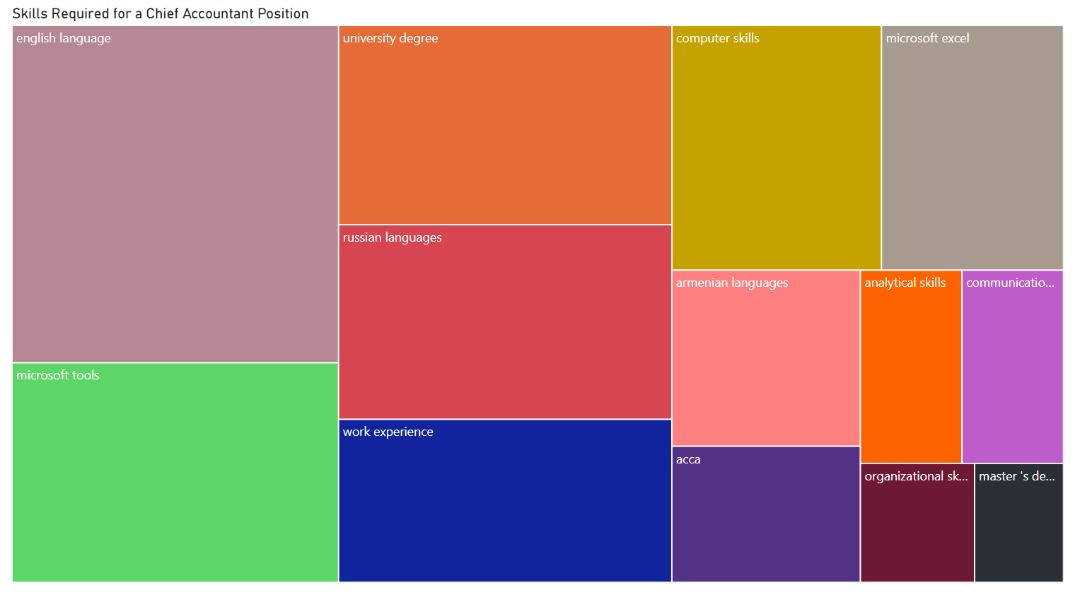
**TEXT ANALYSIS USIN NATURAL LANGUAGE PROCESSING KITS**

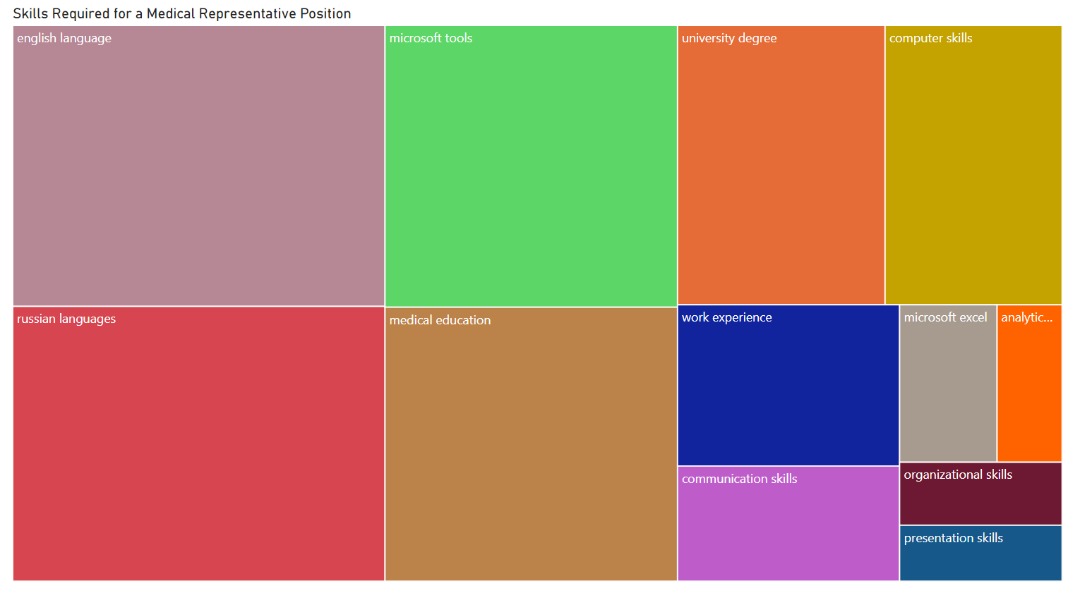
The dataset being mostly non-structural or textual in nature had to analyzed using NLP toolkits such as lemmatization for proper vocabulary, textblob for processing the data to retrieve noun phrases, and removing stopwords.

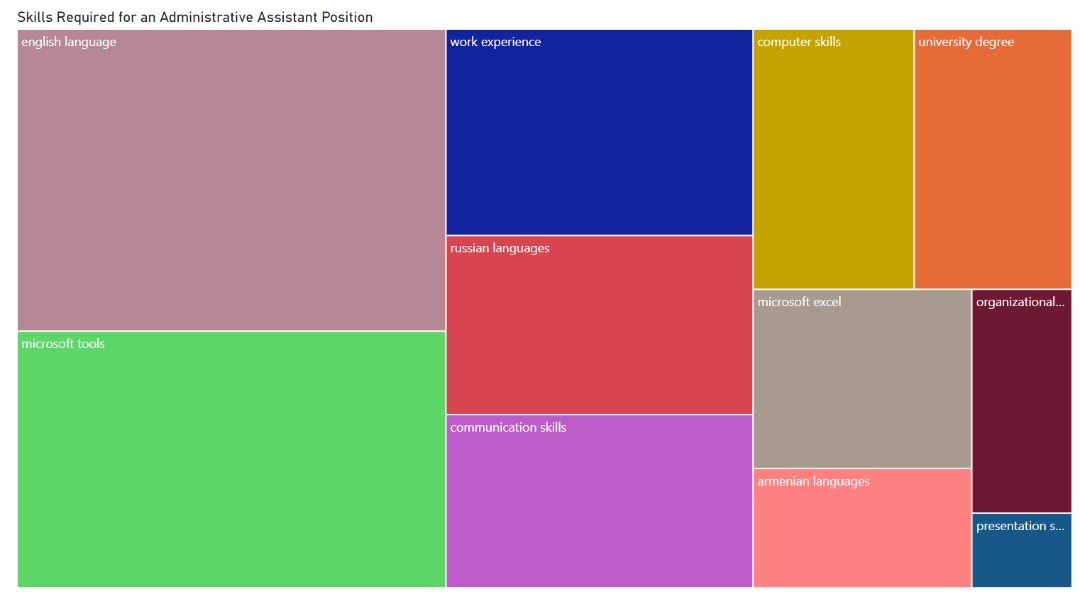
The ‘Required Qualification’ columns were used for this analysis as it gives us an idea of the skills required for each of the job postings. The steps involved in this approach include:

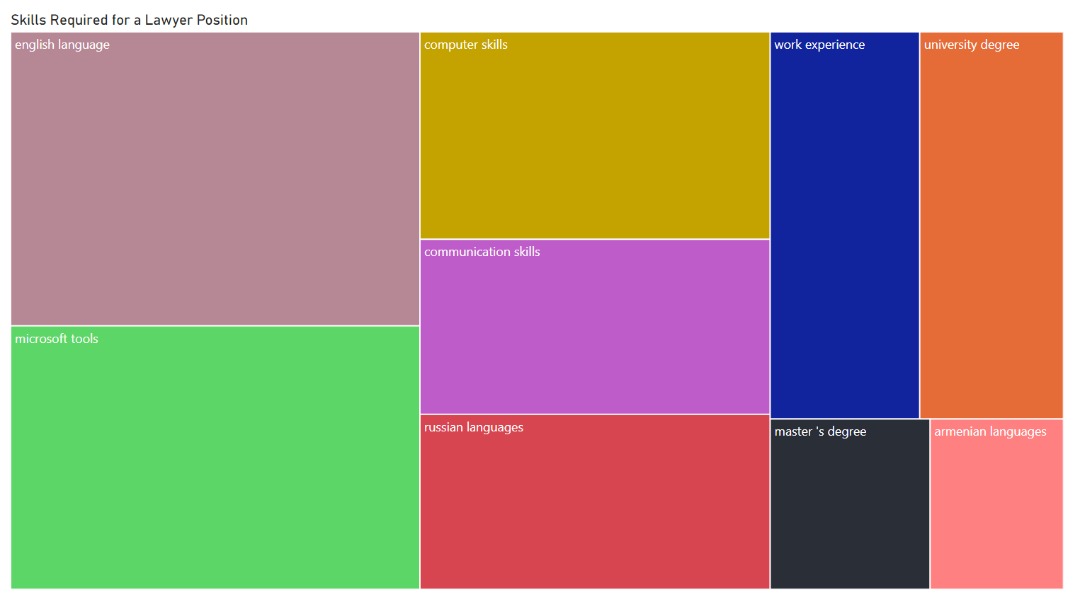
1. Initial though was to remove all stopwords and punctuations in the text
2. Using the textblob tool kit, noun phrases were removed for each title, it was assumed that skills present in the required qualification column would most likely fall under the noun phrase category.
3. For each entry in the title column, the corresponding value in the required qualification column was pulled and saved in a dictionary.
4. The results were then counted and plotted in a tree map to show the most in demand skills from the dataset
5. For describing the changes in skills required over the years, a similar function is performed. However, this time columns used for the analysis were the ‘Year’ and ‘Required Qualification’ column.
6. The results of the most in demand skills were then plotted over time from 2004 to 2015.

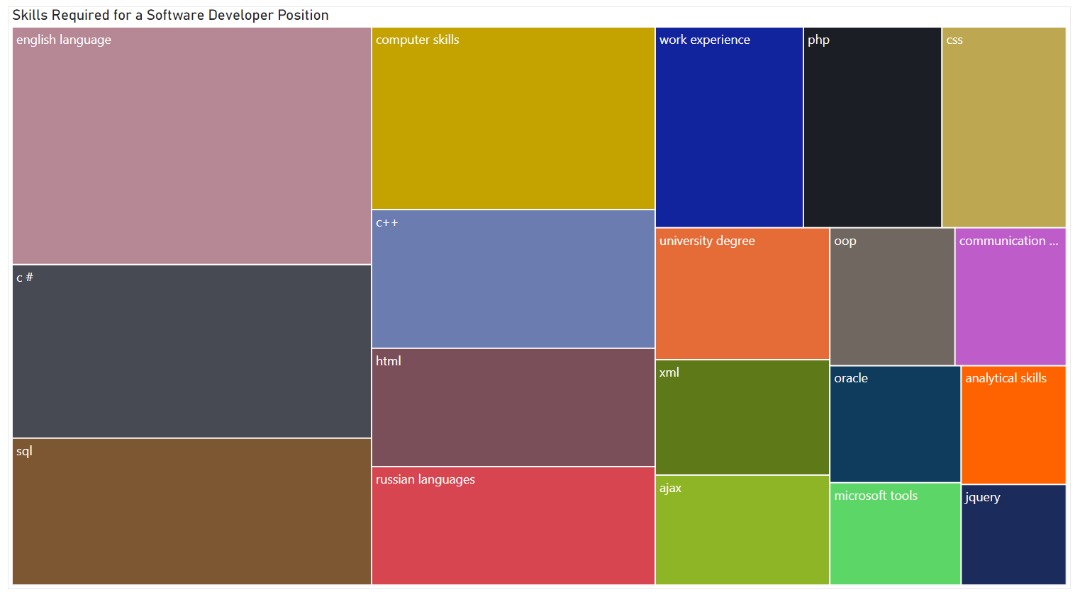


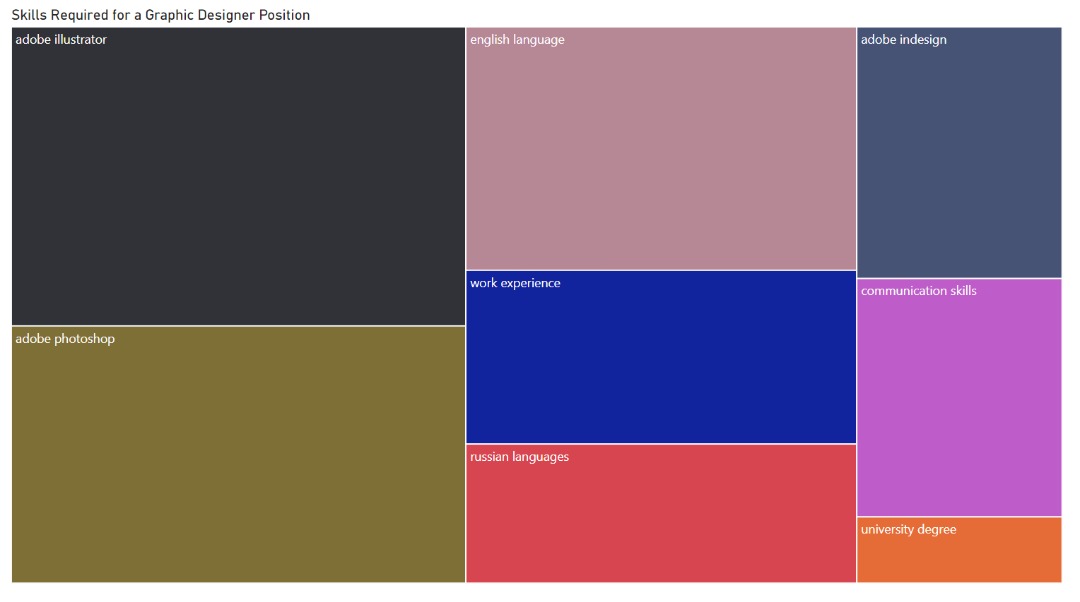


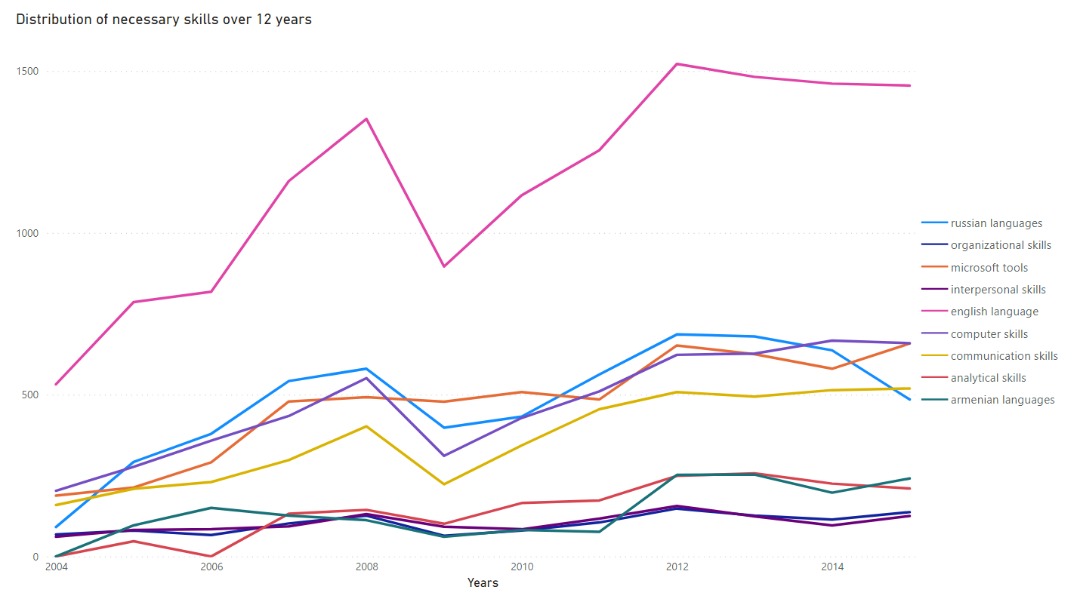












**CONCLUSION**

Based on the data analysis, years with most job posting were analyzed, and investigations were made on the job title with the greatest number of listings, which revealed accountant positions were in high demand between 2004 to 2015. Missing values in the dataset and errored data were the challenges faced during analysis of this dataset. Meanwhile, a suggested improvement is to make an analysis on the about column to get the company with the highest number of job postings and how job postings has varied for each company over the years.

Inferences were made from the analysis performed and the following recommendations were proposed:

**RECOMMENDATIONS**

1. There are more or greater opportunities in IT related jobs as the current data shows more people are likely employed in NON-IT jobs.
2. Job seekers looking to change careers should be more inclined towards the IT sector of the economy.
3. The decrease in job postings during the ‘EMBER’ months should be taken into consideration for job seekers as stifler/tougher competition during this period could lead to difficulty in securing jobs.