**Analysis of jobs affected due to Automation**

Course BIA-660-C

Instructor: Rong Liu

**Submitted by : Group 5**

Team Members

Purva Khopkar

Parth Parab

Rohan Shan

Bhagyashree Shende

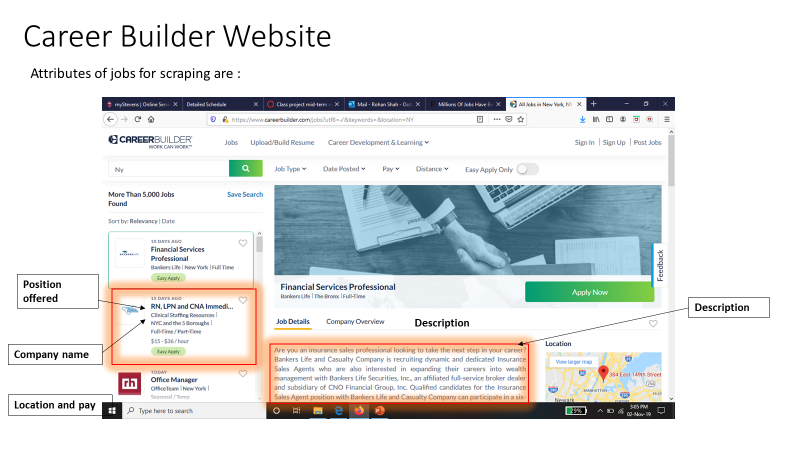
**Objective**

* Automation will displace many jobs over the next 10 to 15 years. Jobs of the future will use different skills and may have higher educational requirements. As AI progresses, some believe that it will steadily and inevitably take over large sectors of the workforce and will bring mass-scale unemployment and social unrest
* In this project, we are set to create an analysis which helps the workforce to keep up with the trends around them and develop the skills which cannot be easily replaced by automation.
* The analysis will recommend skills to laborers to level up with the changing industry standards.
* All skill sets are to be compared and checked for their relevance in the real world for now and for next few years too.

For this Analysis, we gathered information of various jobs on job posting website named CareerBuilder.

**CareerBuilder Website**

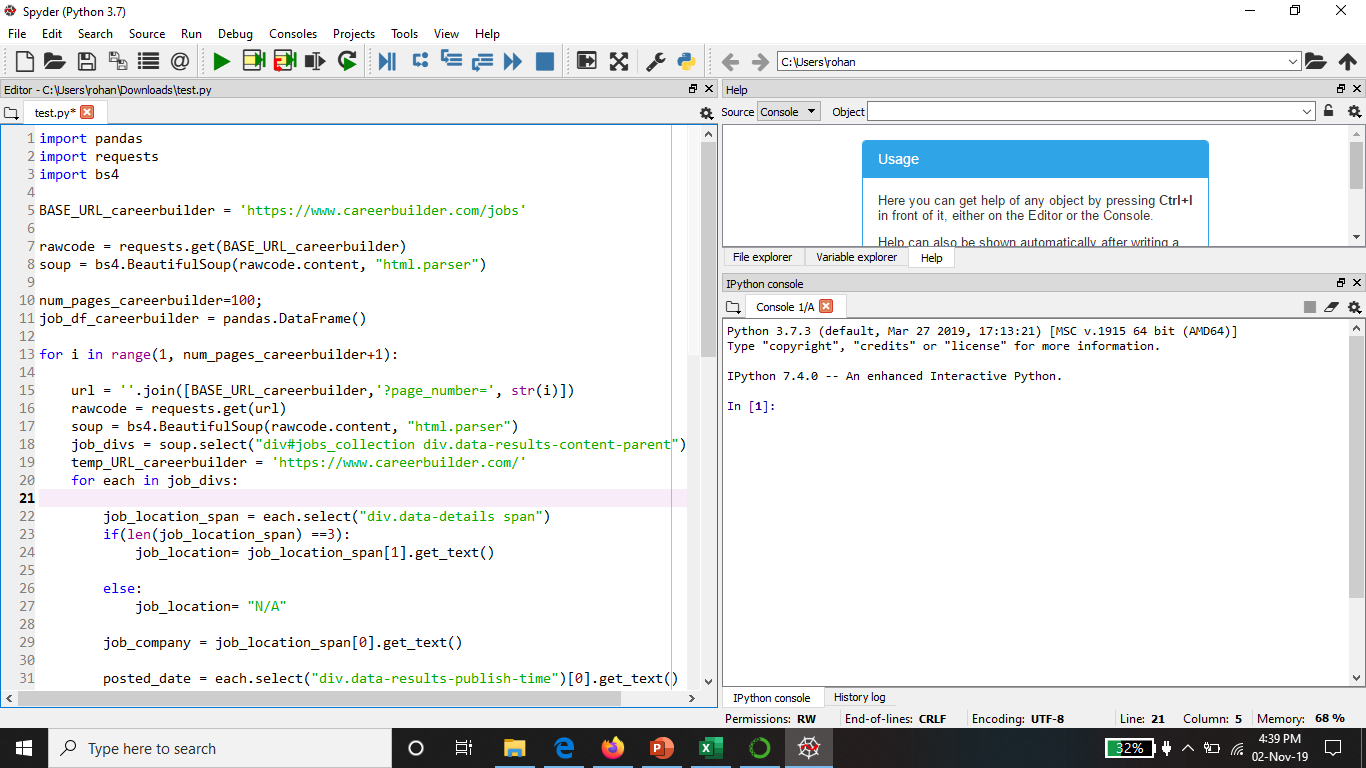
* CareerBuilder is an online employment website for job opportunities and advice. It is one of the leading online employment services in North America and Europe.
* It uses data science to evolve the human capital management space, with options that assist employers find, employ and manage fantastic people.

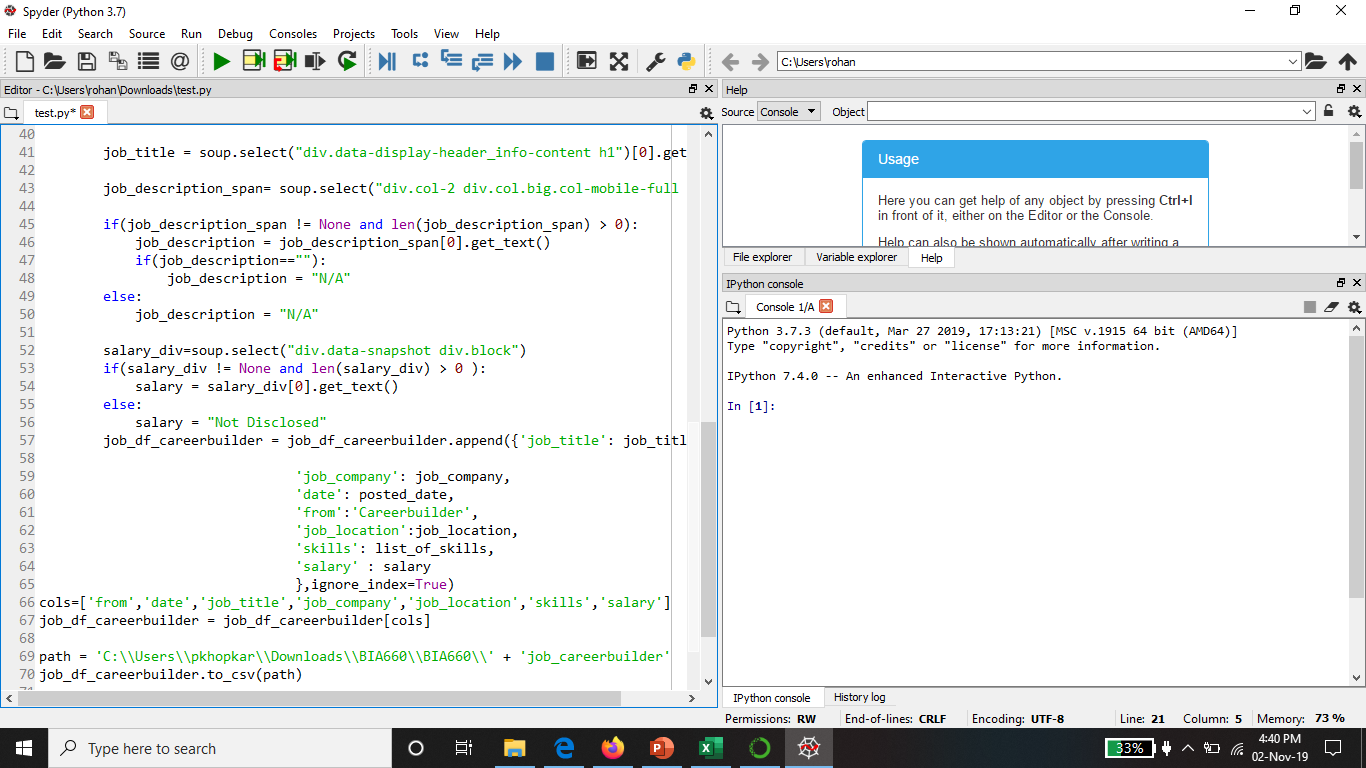


**Web Scraping**

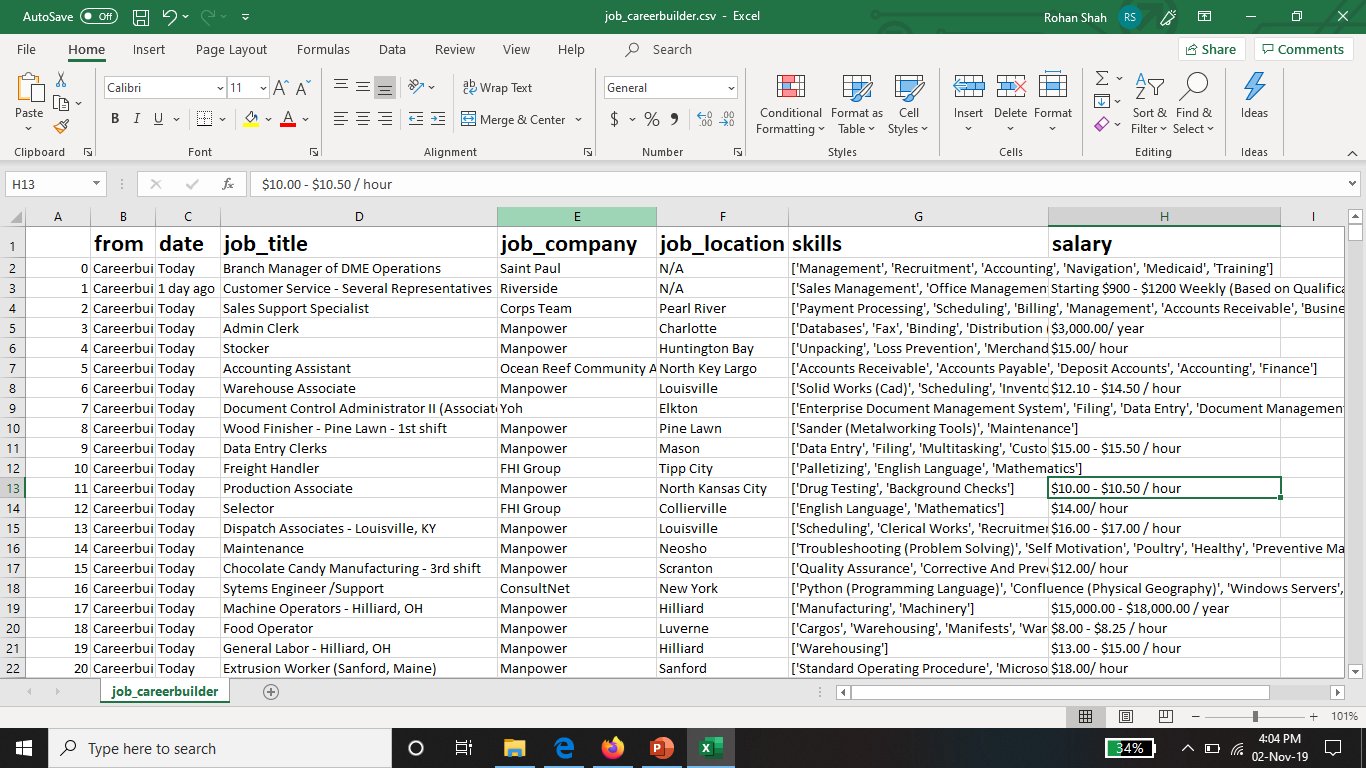
* We Scrapped first 100 pages of Careerbuilder website having 25 jobs each page to get details of 2500 Jobs for analysis and stored the data into file
* We used BeautifulSoup library and scrapped the following information about each Job position
* Job title
* Company name
* Location
* Skills required
* Date of posting,
* Description of the Job
* Pay per hour

Code snippet for scrapping :





**Scrapped data from Careerbuilder**



**Data Cleaning and Processing-**

After scrapping the data, we observed few inconstancies in it. So, we cleaned and prepossessed the data in order to make it consistent

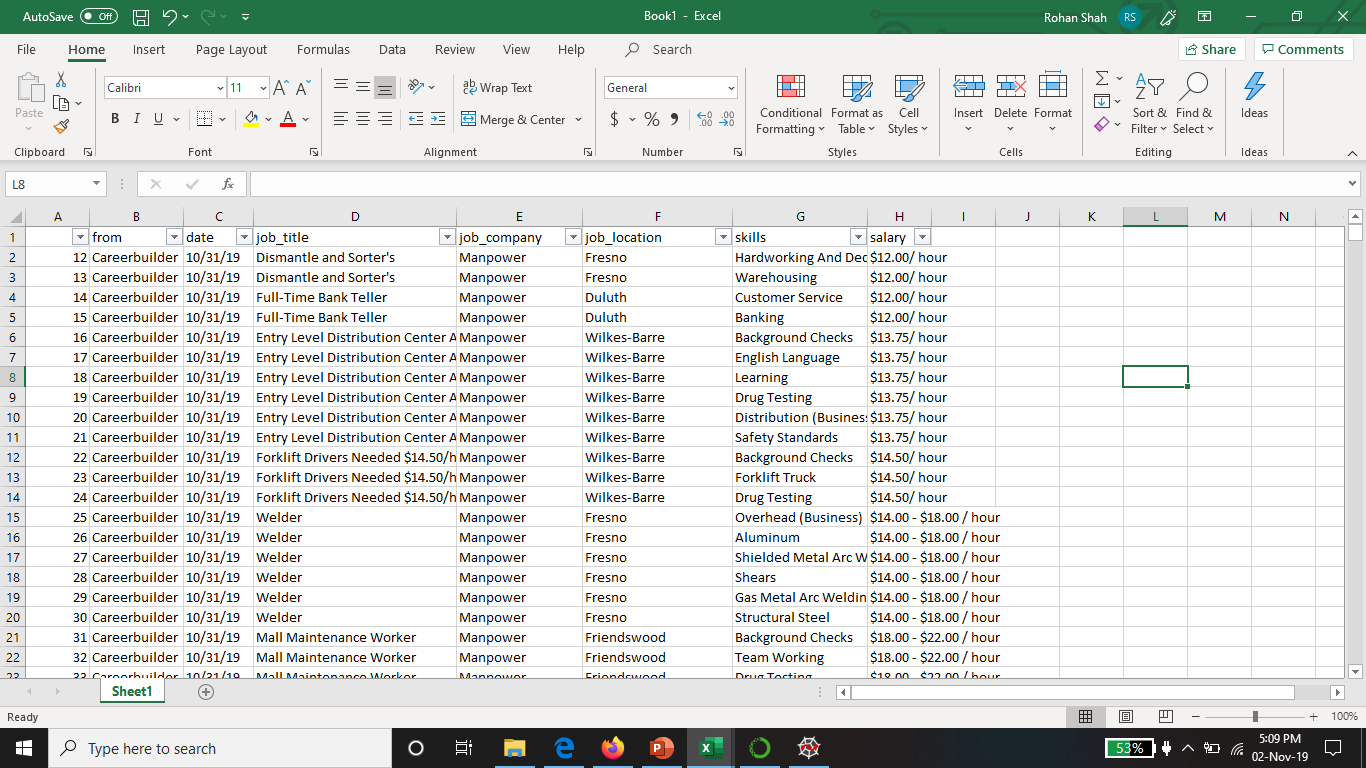
Following steps were performed in this step :

* **Changed date format to MM/DD/YYYY**

Job posted currently does not show date of posting the job on website but rather shows posting time in the format as “Today” or “1 day ago” etc. We replaced the current date posting format to actual date like mm/dd/yyyy.

* Removed jobs having missing values of Job Location.
* Analyzed for probable outliers and removed them from existing data
* Skills were collected in the form of list for each Job, so created a separate entry of each skill so that further analysis on each skill can be perform

**Cleaned and processed data**



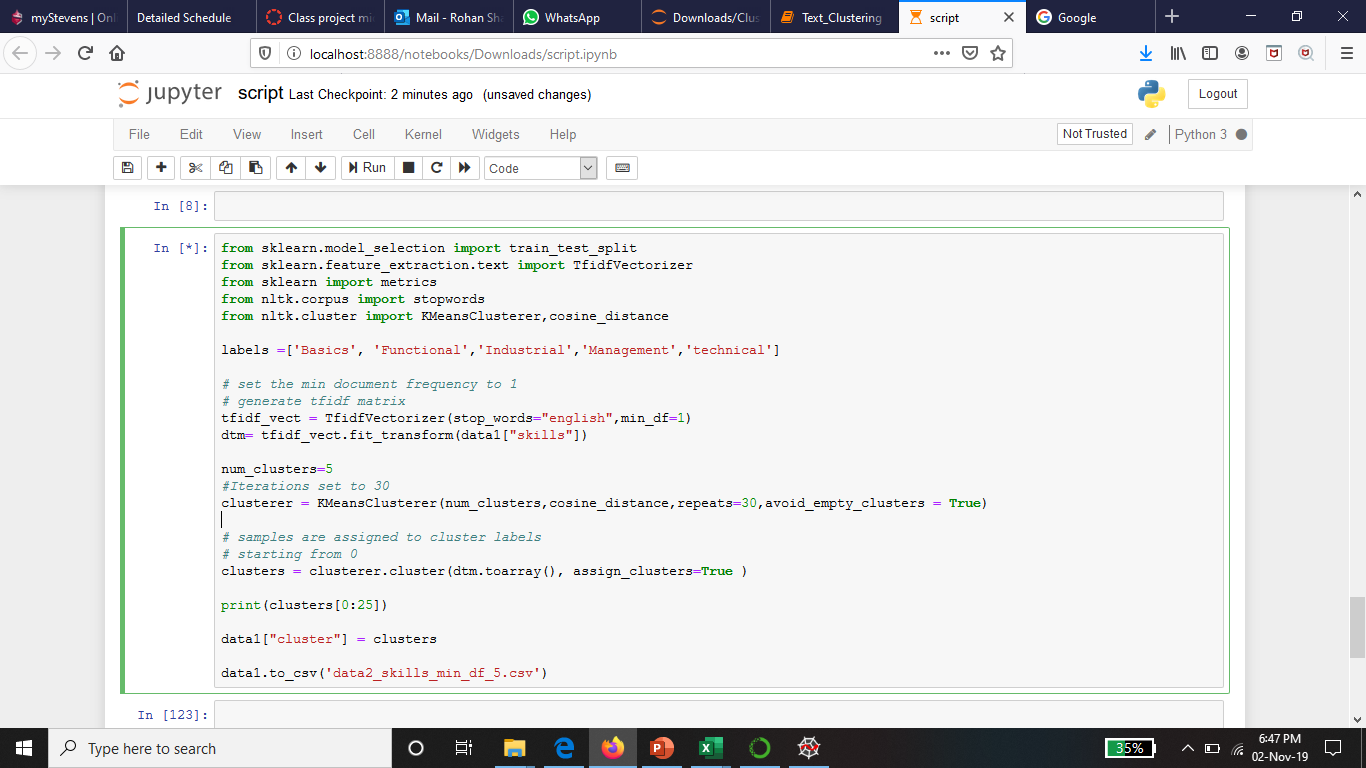
**Exploratory data analysis**

* From the above data, it can be observed that there are no labels available. Hence, we performed clustering on the existing dataset which is part of Unsupervised learning
* Clustering is done to group similar skills into same clusters.
* K-means algorithm is used to perform clustering of skills. It helped summarization, compression and efficiently finding nearest neighbors.

**Steps for K-means are as follows:**

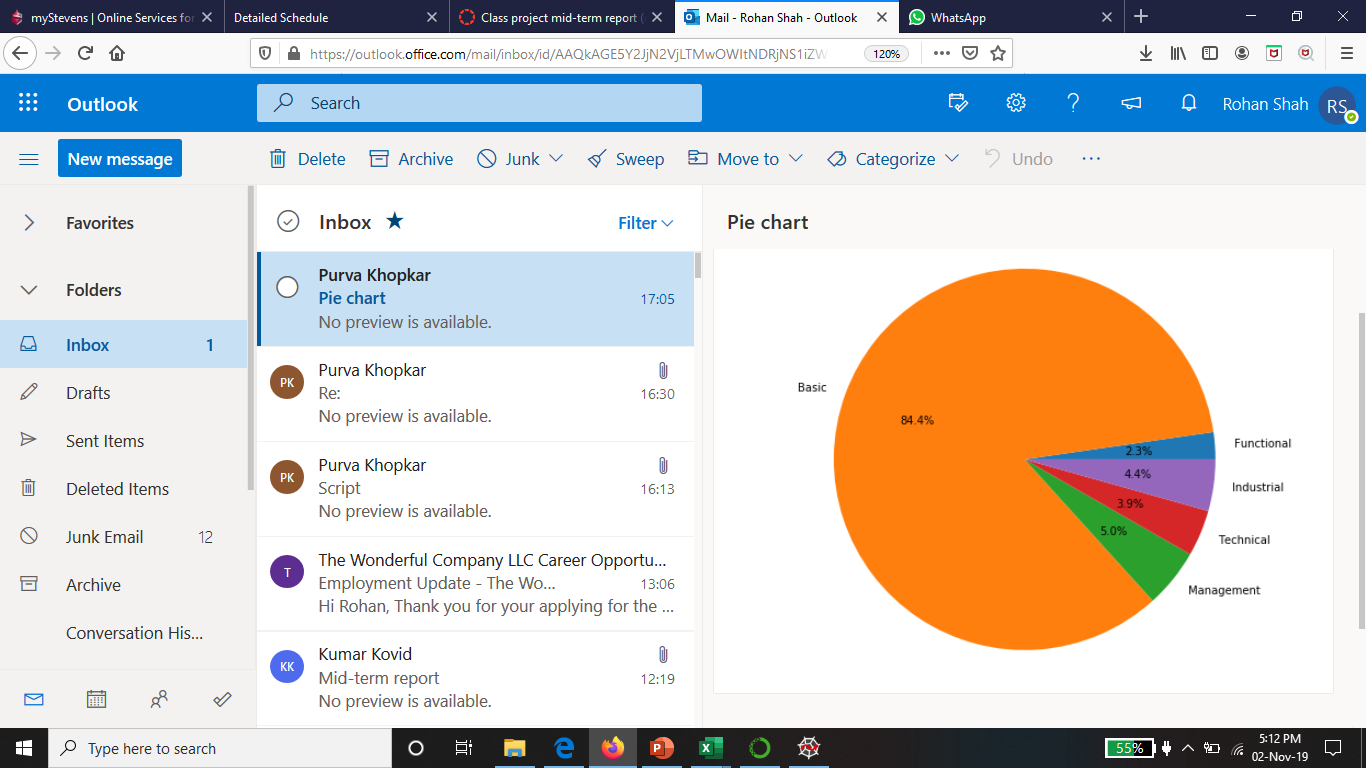
* Cluster skillsets into K clusters
* Select a point as initial centroid
* Create a cluster by assigning each point to its nearest centroid by distance.
* Most commonly occurring centroids are chosen.
* The method used for calculating distance is Cosine similarity.
* Cosine Similarity is highly used for calculating text similarity among documents

**Code snippet for K-means clustering of Skills**

****

**Analysis of Clusters**

* Using K-means algorithm 5 clusters were created.
* Text clustering was performed on skills and there were 5 highly common categories
* Basic
* Functional
* Industrial
* Technical
* Management skills

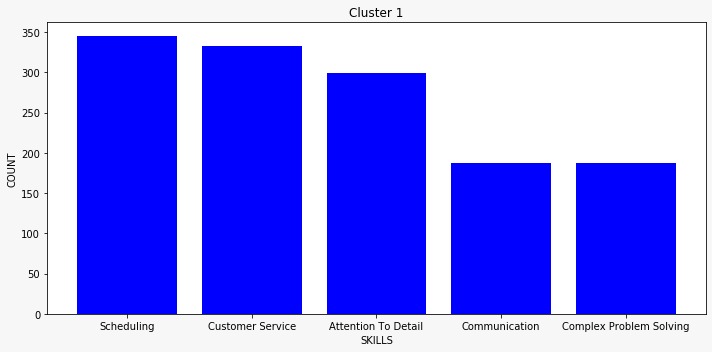


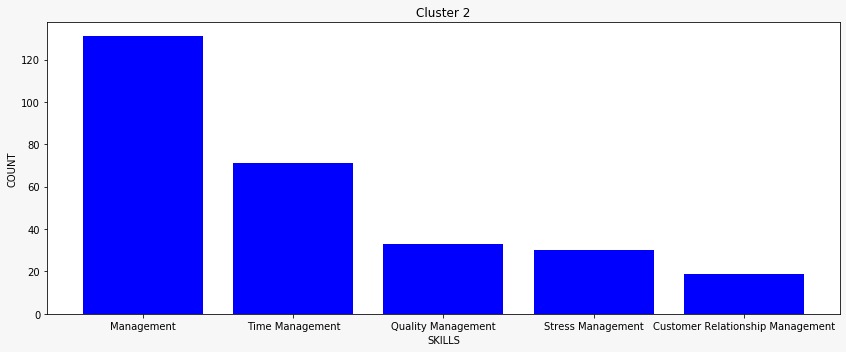
*Fig1 : Pie chart showing all the categories of skills*

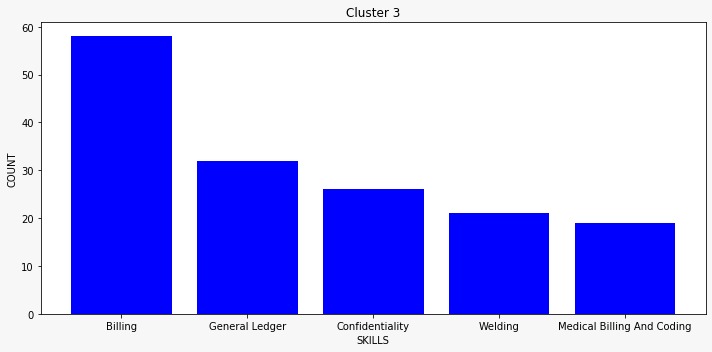
* From the pie chart, it can be observed that category **Basic skill** is highest among all the categories

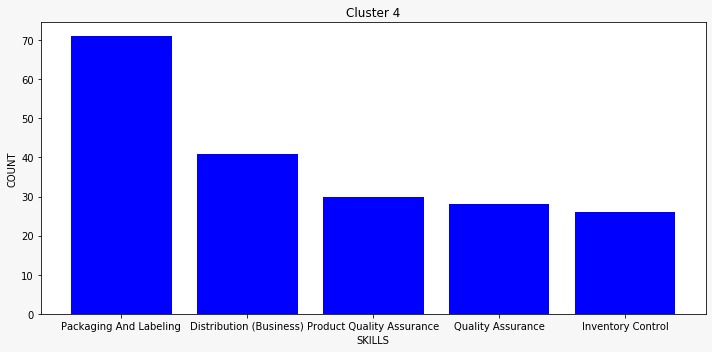
**Top 5 skills in each cluster**

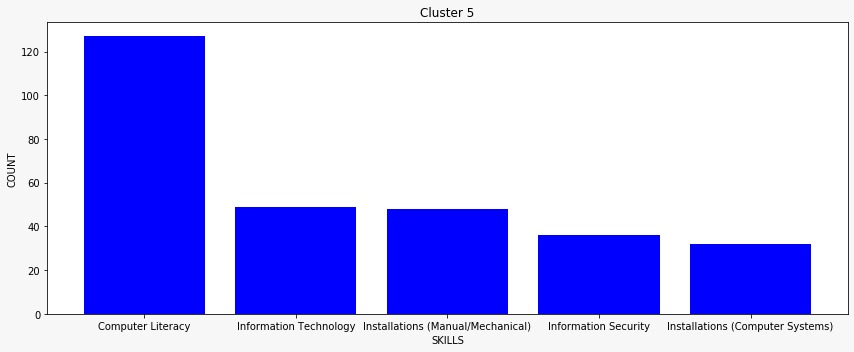
We also did the analysis of Top 5 skills in each cluster











**Clusterwise Wordcloud:**

Cluster 1 -

**A picture containing sitting, red, track

Description automatically generated**

**Cluster 2 – Functional Skills**

**A close up of a newspaper

Description automatically generated**

**Cluster 3- Industrial Skills**

**A close up of a newspaper

Description automatically generated**

**Cluster 4 – Technical Skills**

**A picture containing text, newspaper, sitting, table

Description automatically generated**

**Cluster 5 – Management Skills**

**A picture containing sitting, table, green

Description automatically generated**

**Steps completed till now in the project along with the team member is given below :**

|  |  |
| --- | --- |
| Steps Completed | Team Members |
| Scrapping of Careerbuilder website to scrape data of 2500 Jobs | Purva and Bhagyashree |
| Data cleaning and preprocessing | Parth |
| Exploratory Data Analysis | All members |
| Clustering ( K-means) | Purva and Bhagyashree |
| Report preparation | Rohan |

**Action Plan**

* Testing of existing K-mean clustering using Test data to evaluate performance of the model
* As per evaluation, fine tune the model for entire data set
* We are planning to perform the clustering on Historical data of Jobs to get the good mix of skills
* We are also planning to perform Topic modelling using Latent Dirichlet allocation to summarize the skills
* We will perform analysis of the finalized clusters of skills for Highly and less likely automatable skills

References:

<https://www.onetonline.org/find/descriptor/browse/Skills/>