**Data Science: Knowledge Extraction from Employment Dataset**

**What Qualifications to Which Job?**

Saif Ali abdul-husseinabada

Department of Computer Science

Faculty of CS and Mathematics

University of Kufa

Dr. Kadhim AlJanabi

Department of Computer Science

Faculty of CS and Mathematics

University of Kufa

**Abstract**

As data science has become as important as the raw material in the new economic era, and since the knowledge nowadays represents a crucial factor in the success of different organizations, these two concepts came at the top priority of the most wanted jobs worldwide. The work in this paper presents a study and analysis of some important type of data which is employment data. The study consists of different mining algorithms including classification, association and clustering techniques.

The study shows that different jobs are highly related to the level of education. Different jobs are classified into groups according to the qualifications required for these jobs and the association between the jobs and the required level of education are studied and extracted from the logged data, and finally different clusters for the logged data were specified using modified K-Means algorithm.

Different tools and software including WEKA, XLSTAT and MATLAB were used to apply the required mining algorithms on the logged data.

***Keywords:*** Data Science, Data Mining, Association, Clustering, link analysis.

**Introduction**

Data Mining (DM), sometimes called knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful knowledge. This knowledge can be used in different levels of decision making to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to preprocess and analyze data from many different dimensions, classify, categorize cluster and identify the relationships [1,2,3].

Although DM is a relatively new term, the technology is not. Companies have used powerful computers to analyze huge amount of supermarket data and analyze market research reports for years. Continuous innovations in computer processing power, computer memories, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

In DM field, information represents an important term. For example, analysis of student achievements at the university like associations between achievements and educational environment data can provide *information*. On the other hand when the information can be converted into another higher level of representation, this in called knowledge. For example, summary information on student’s outcomes can be analyzed to get knowledge about the attributes highly required in getting jobs after graduation. Figure (1) shows the different levels of these concepts.[2,4,5]

****

**Figure (1). Different Data levels With Values and Volumes.**

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

DM techniques can be classified into five different fields; classification, clustering, association, prediction and link analysis. Each of these techniques has its own algorithms and applications in business context. Mining employment data set represents a crucial factor for different organizations in both public and private sectors. Many works have been carried out in this field

DM consists of five major elements: [1,2,6,7]

Extract, transform, and load (ETL) transaction data onto the data warehouse system.

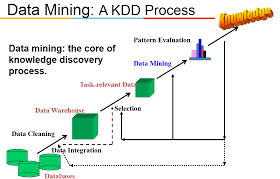
Store and manage the data in a multidimensional database system.

Provide data access to business analysts and information technology professionals.

Analyze the data by application software.

Present the data in a useful format, such as a graph or table.

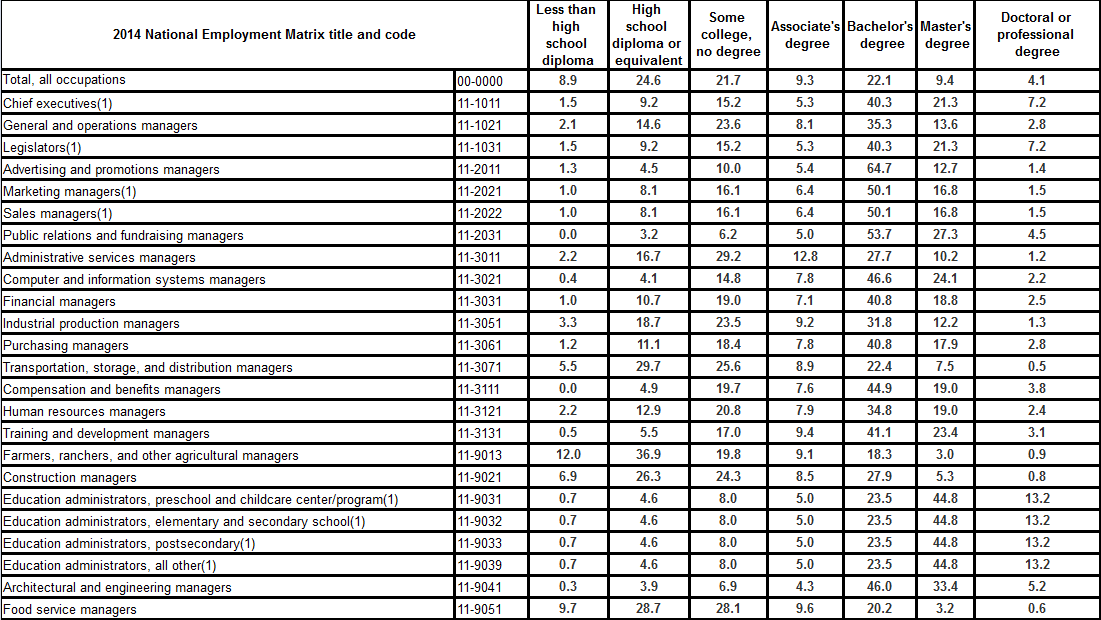
The whole KDD and DM process is shown in figure (2).



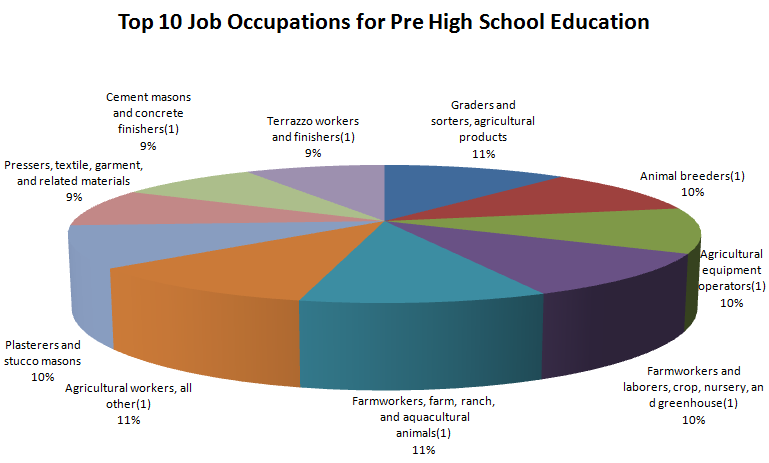
**Figure (2). Knowledge Discovery in Databases (KDD) Process.**

**Data Collection**

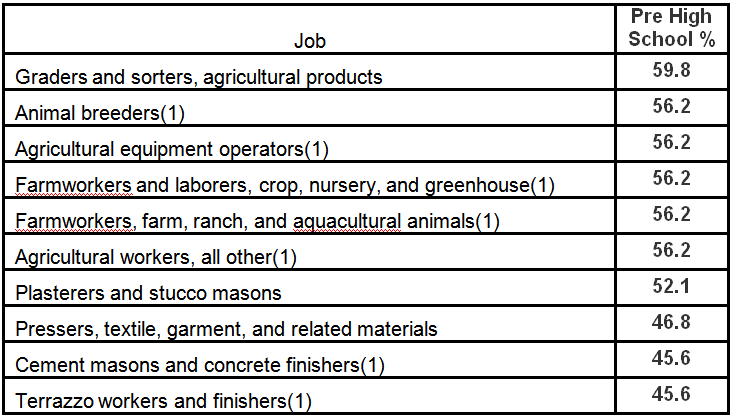
More than 800 records of free dataset available on the internet related to different job titles with the required degree of education (starting from less than high school to doctoral and professional degree).



**Data Analysis-Association and Link Analysis**

In Data Mining techniques, association and link analysis represent an issue of great importance because both gives good indication about relationships between a specific job and the required qualification. The data given in Table(II) and the diagram given in figure(3) represent the association between the job occupation related to a pre high school qualification.

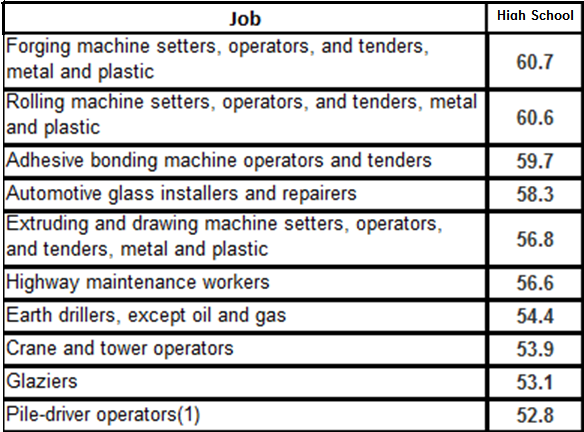
**Table (II) Job distribution related to pre high school**



**Figure (3) Job Distribution VS Pre High School**

Data given in Table(III) and the diagram given in figure(4) represent the association between the job occupation related to high school qualification.

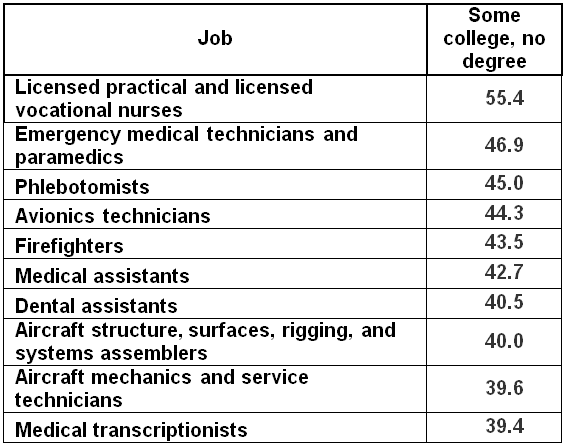
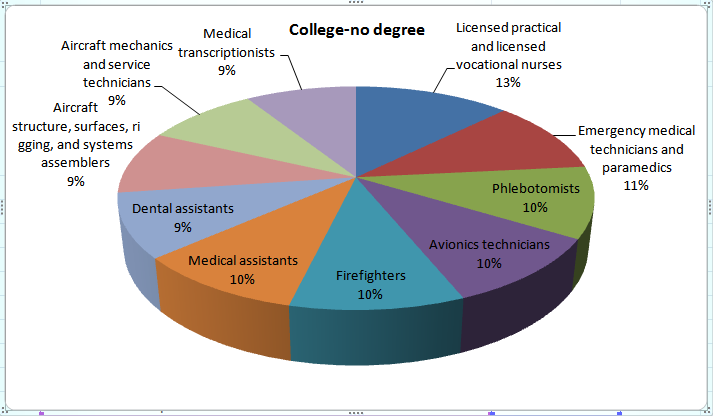
Table (III) Job distribution related to high school



**Figure (4). Job Distribution Related to High School**

The jobs related to college and above qualifications are highly different from the jobs for high school and pre high school degrees. Table (IV) and Figure(5) give the association and link between lobs associated to colleges without degree qualification.

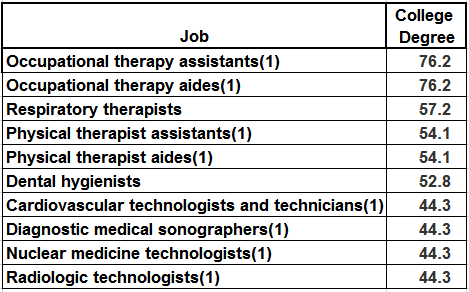
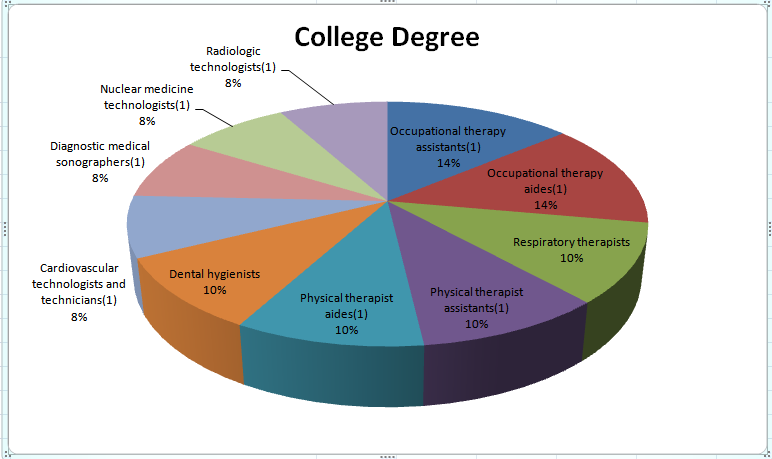
**Table (IV) Job distribution related to college, no degree**



**Figure (5). Job Distribution Related to College, no Degree**

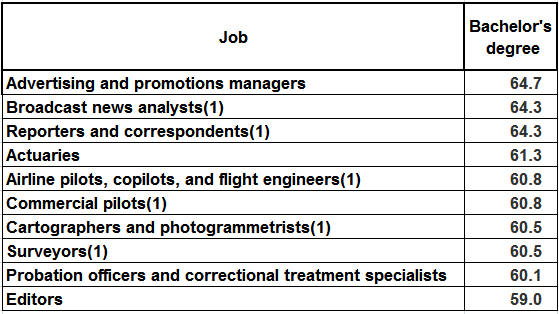
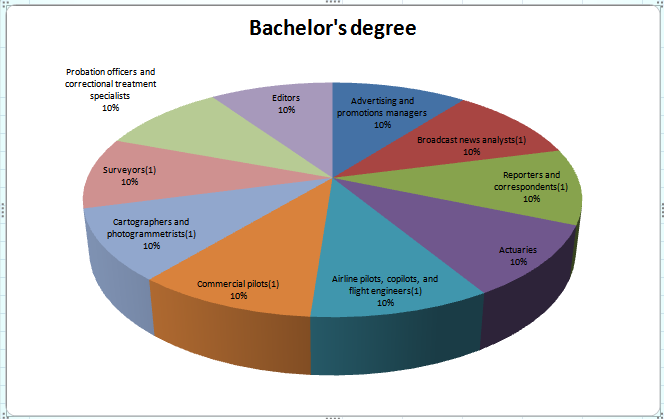
The remaining figures and tables show the associations between different fob occupations and the required qualifications (college, Bachelor, Master, Doctoral and post doctoral degrees).

**Table (V) Job Distribution Related to College Degree**



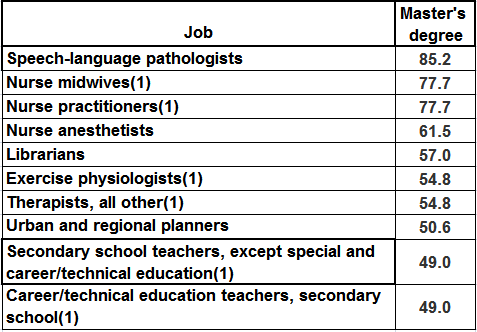
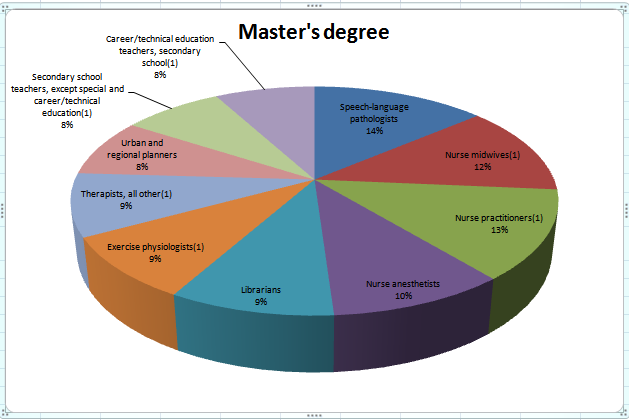
**Figure (6). Job Distribution VS College Degree**

**Table (VI) Job Distribution Related to Bachelor’s Degree**



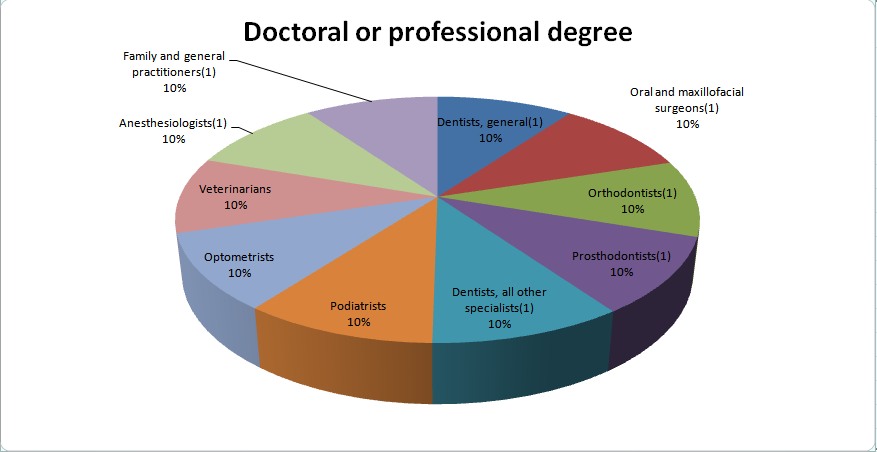
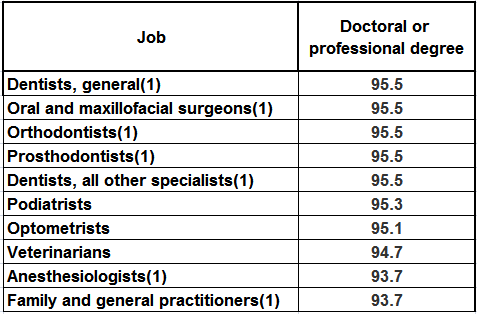
**Figure (7) Job Distribution VS Bachelor’s Degree**

**Table (VII) Job Distribution Related to Master degree**



**Figure (8) Job Distribution VS Master Degree**

**Table (VIII) Job Distribution Related to Doctoral or Professional Degree**

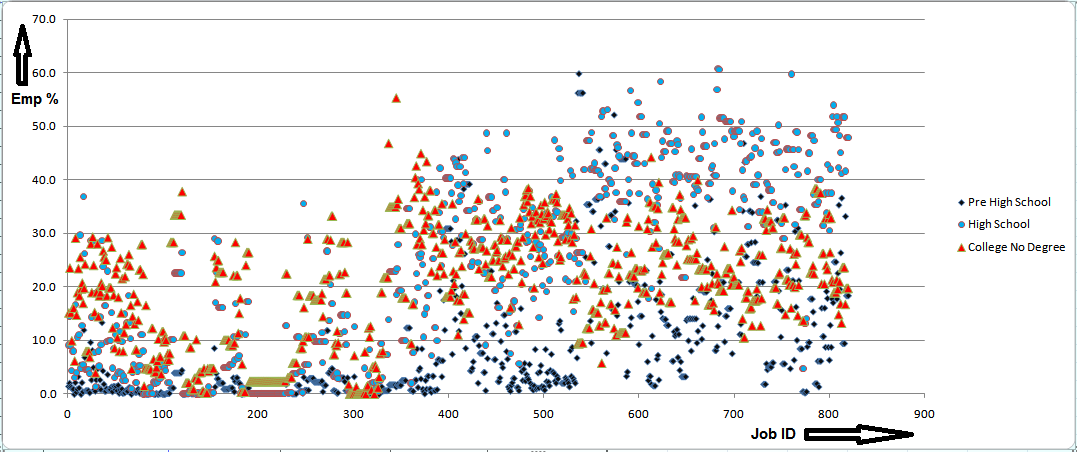


**Figure (10) Job Distribution VS to Doctoral or**

**Professional Degree**

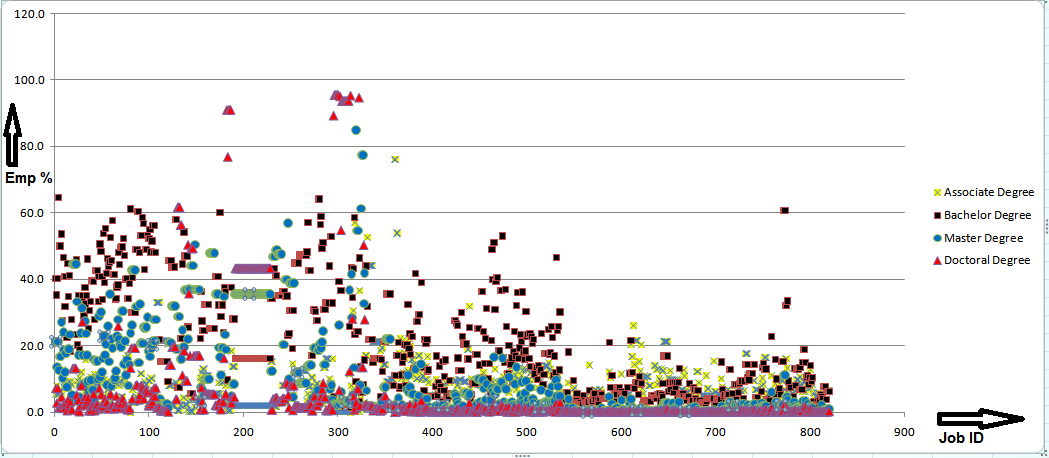
**Data Analysis – Clustering Techniques**

Clustering represents one of the most DM techniques since it groups different objects together according some similarities and dissimilarities between these objects. That is why this technique was used in this paper. Different jobs and qualifications where grouped together according to the similarities between them. Figure(11) shows the groups of the jobs for pre high school, high school and college no degree qualification, whereas Figure(12) shows the groups of the jobs for pre college degree, Bachelor, Master and Doctoral degree qualification.



**Figure (11). Clusters of Different Jobs Related to Pre High School, High**

**School and College no Degree Qualification.**



**Figure (12). Clusters of Different Jobs Related to College through**

**Doctoral Qualification.**

**Results and Conclusions**

According to the data collected and the analysis of the data in both distributions and clustering show in figures (2) through figure(12) and tables (I) through table(VIII), it is obvious that different jobs in both computer science and other fields are highly related and linked by the qualifications that vary from pre secondary school through diploma and higher degrees. From the study in this paper, jobs available to pre secondary schools are categorized as technicians, operators, unskilled workers and technology assistants, whereas top management jobs, analysts and problem solvers require university degree as minimum.

From the same figures and tables, and according to the three groups of qualifications

Pre high school, high school and college no degree

College, bachelor and master degree

Doctoral and post doctoral degree

Different types of jobs are highly related to each qualification.

**References**

[1] M. Steinbach, P.-N.Tan and V. Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN: 0-321- 32136-7.

[2] Jiawei Han and MichelineKamber “Data Mining: Concepts and Techniques” 2nd ed., Morgan Kaufmann, 2006.

[3] Deborah Osborne, MA, Susan Wernicke, MS, “Introduction to Crime Analysis: Basic Resources for Criminal Justice Practice, The Haworth Press, New York, London, Oxford, 2003.

[4] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

[5] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed., 2005, ISBN 0-12-088407-0

[6] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.

[7] Tianyi Wu, Yuguo Chen and Jiawei Han, “Association Mining in Large Databases: A Re- Examination of Its Measures”, in Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, Sept. 2007.