Chapter 4: Data Mining using MS-Excel: Distributions & Graphs

Information Visualization (IV) is use of computers to generate interactive visual representations to explain and understand specific features of data. The basic principle of IV is to present data in a visual form and use human perceptual abilities for their interpretation. Visualizations improve comprehension. It allows recognition of essential features and helps making important inferences. Colour, form, movement and spatial position in IV generate good user interfaces too. In this chapter, a study has been conducted and presented on the enrolment data using graphs and distributions obtained using MS-Excel. Sections after basic introduction are titled as justification of data source, strategy for data cleaning followed by results in the form of informative visuals to study various divides in the society and then and discussions about them.

4.1 Introduction

Computers are used widely in the field of education by students, teachers & administrators alike. Administrators typically work with computers in education if their educational institution is storing data of various kinds, supporting an LMS (Learning Management System) or an e-Learning portal, audio-video channels and repositories etc.

Data may be generated at various relevant levels like enrolment, leave/attendance, assessment, migration, drop-out, scholarships, fee, placements etc. Similar data are generated at the staff or employee end. In LMSs’ log files, assignments (with date/time), assessments, support links, help links, usernames, e-mail ids, images, power point slide shows, other documents generate data for analysis. Repositories are multimedia data bases containing video and audio files along with usernames, e-mail ids, comments, images and slide shows as well. Studying educational data sets can aid in suggesting pedagogies, site modification, intelligence services & page recommendations in the long run. Personality and performance of every pupil have to be judged throughout the academics and the employment/career to
clearly identify any patterns and correlations in the data, in turn to predict & project the right path thereby guiding the pupil accordingly.

The focus of this report (and this research) is enrolment data. Mining (analysis) of educational data to visualize the educational settings, solve the problem of students, find any hidden patterns if present, is called EDM. Broadly, there are two major parts of techniques in DM: primitive techniques and advanced techniques. Here we are concentrating only on the primitives.

4.2 Justification of Data Source

Enrolment data of disabled students was obtained from SRD, IGNOU in December 2009. There are two tables: data set – PH091 for cycle 1, 2009 and the data set – PH092 for 2nd cycle, 2009 where PH standing for Physically Handicapped. A total of 3020 student records were provided. The sample size is large enough to provide good confidence for the presented data analysis. As sample size increases, we can estimate more interactions, which typically are smaller and have relatively larger standard errors than main effects [95]. In order to ensure that the enough data is available to generate the conclusions with a certain degree of confidence, statistical power is checked. Statistical power is the likelihood or probability of detecting a shift or difference for a given sample size in the presence of variation. Statistical power is usually dependent on the scale of measurement or type of data, the sample size, the difference to be detected, and the noise level of the experimental environment. In this research study, different types of analysis have been performed; we used the general guideline for ensuring the number of records is suitable as per the graph below. The graph below is reproduced from reference [96] which shows that having more than 200 to 300 records ensure 0.90 statistical power. Typically for nominal data more than 2000 records ensure more than 0.90 to 0.95 statistical power. As indicated before, the dataset used here has
more than 3000 records so the statistical power of the current data set is enough to draw conclusions presented in this study.

![Figure 21: Plot of Power versus Sample Size, reproduced from [96]](image.png)

<table>
<thead>
<tr>
<th>File Name</th>
<th>No. of Columns</th>
<th>No. of Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH091</td>
<td>120</td>
<td>1840</td>
</tr>
<tr>
<td>PH092</td>
<td>128</td>
<td>1180</td>
</tr>
</tbody>
</table>

Table 2: Description of data files

The data files are in dbf format (Figure 22) and can be easily imported in MS-Excel (Figure 23) using FoxPro.
Figure 22: Data file in MS-Visual Studio’s Visual FoxPro displaying first column

Figure 23: Data file in MS-Excel

After data cleaning (using pivot table) some interesting patterns are obtained. The graphs obtained from this analysis are shown below and discussed in the next chapter. The research methodology has been followed from reference [89]. A wide variety of DM methods are available such as prediction, clustering, relationship mining, discovery with models, and distillation of data to obtain and present knowledge.

4.3 Strategy for Data Cleaning

Data files are imported from FoxPro to MS-Excel because every column can be viewed separately now. Most values are standardized already and discrete, as can be verified
from the university website & prospectus. Records are matched to check that there is no repetition of a student’s enrolment number (primary key). Necessary transformation can be done, e.g., to get age from date of birth. So, overall MS-Excel turns out to be a good tool for data cleaning. The given dataset is called dirty data because it is full of blanks and errors.

‘Cold start’ is a stage in Data Mining, when a data miner has to start from scratch or ‘zero’ as in this study. Typical real world data sets which are unformatted (raw) need to go through data cleaning steps to be successfully used in a study [75]. After formatting the data appropriately, pivot table feature in excel is used for data cleaning. Suppose the variable under consideration is ‘State’. Various possible occurrences of State ‘Delhi’ can be counted as in Figure 24 blanks and wrong fields also got marked (‘Del’, ‘Dilli’). This also explains that why sometimes local understanding of the database can be crucial.

\[\text{Figure 24: Pivot table used to count variables & detect blanks and wrong code}\]

In the following sub section use of pivot tables for data cleaning is explained in detail: how to check uniqueness, definition, measure frequency, assure accuracy, find and correct errors, mark invalid data, and show outliers.
4.3.1 Using Pivot Tables for Data Cleaning

Data Cleaning is the process of detecting and (if possible) removing errors. It is a part of Data Mining process. Importance of data cleaning increases manifolds when data files from various similar sources (context) are picked up in raw format. Then these files are checked for errors, missing values and correct them as much as possible using forms or algorithms. There are tools available to clean datasets e.g. Dataflux [97]. Considering the cost factor and to obtain good understanding, making such tool or using spread sheet to start with the process MS-Excel program is a good choice. In this section use of ‘Pivot Table’ feature has been explained with the help of a dummy file and MS-Excel 2010.

Data cleaning stage involves various other tasks. A given data cleaning problem can be solved using existing software or tools. Dataflux by SAS (Statistical Analysis System) is a comprehensive collection of programs (procedures) for data cleaning. It has an interactive front end. Such tools allow writing queries also. A spread sheet, on the other hand allows user to browse to fresh & raw data sets. VBA (Visual Basic Applications) & macro are features useful while cleaning and analysing data sets.

Data cleaning step consists of various sub steps like including accuracy & validity of entered data (ranges, dates), identifying primary key & missing values, and eliminating redundancy or duplication. In the following sub sections each such step is discussed, one by one, separately.
4.3.2 Uniqueness

Consider primary key ‘enrolment number’ in any university admission/enrolment data set. To check its uniqueness, first a column is made in front of it called ‘Count’. The column is then filled up with 1s as shown in Figure 26, till the row where the first column ends. Then the Pivot Table option from the Insert menu is selected. The source table is selected as shown in Figure 27. The pivot table is inserted in the spread sheet or workbook at the location of choice. On dragging the ‘Enrolment Number’ field to ‘Row labels’ and ‘Count’ field to ‘∑ values’ in Figure 28, it is clearly evident from ‘Count of Count’, that all values are unique i.e. there is no repetition.
Figure 26: Using Pivot table-1

Figure 27: Using Pivot table-2
4.3.3 Verify Variable Definition

Length and certain characters or numbers can be unique to a field.

Figure 28: Using Pivot table feature to check uniqueness of values of primary key

Figure 29: Checking length of a field
4.3.4 Finding frequency

Consider variable ‘Program’ in Figure 30. Make a pivot table as shown above. The output shows that 5 people opted for BCA and 2 for MP program.

Figure 30: Counting frequency

4.3.5 Accuracy of the data

Consider variables ‘Fee’ and ‘Program’. In Figure 31 below, an error is indicated upon cross checking, for BCA program there are two different values of fees in the same year. Such errors can be corrected because the correction is available already.

Figure 31: Accuracy of the data
4.3.6 Errors and corrections

Some errors are spotted easily and can be corrected if an obvious correction is known.

For example, values M, Male, Female as below:

![Image of a spreadsheet showing corrected data]

Figure 32: Error Correction

4.3.7 Invalid data

Since there is no ‘Medium’ of Instruction as D4 (invalid data) is detected below.

More such invalid values can be detected in dates, age etc.

![Image of a spreadsheet showing invalid data]

Figure 33: Validity check
4.3.8 Outlier

A sudden leap or drop in graph can be detected as outlier. Maximum and minimum values can also be detected using Pivot tables and described using appropriate statistical functions. Percentiles, percentage, standard deviation, mean, mode, median, etc. functions can be used on certain columns like marks obtained in TEE (Term End Examination), assignments, project & viva-voce for the purpose of Report Writing. An example of such learning analytics is [68] where data is recorded at each and every step of calculation using educational software and analysed to improve students’ learning.

4.4 Results

In this section, some of the results are presented as obtained using pivot table feature of MS-Excel, while doing the data analysis conducted on disabled students of IGNOU, who enrolled for various courses in the year 2009. Relevant codes have been put as appendices A to D at the end of this thesis.

Result 1: 54.11% students are of young age group (Figure 34).

Result 2: 37.11% students enrolled for Master of Political Sciences (MP) and paid an average fee (Figure 35).
Result 3: 68.01% opted for English Medium (Figure 36).

Result 4: More than 70% students are male (Figure 37).
Result 5: 65% students are pass outs of past 10 years only (Figure 38).

Result 6: 54.4% students them are unemployed (Figure 39).
Result 7: 60% students are from urban areas (Figure 40).

Result 8: More than 71% students don’t have email ids (Figure 41).

Result 9: Around 12% of the students have indicated that they have locomotor impairment. 72.5% students did not fill this column or entered wrong value (Figure 42).
Result 10: Students again could not fill the column for previous qualifications properly. 20.8% have done graduation or its equivalent. This confirms with the result that most of the students have enrolled in MP program of School of Social Science (SOSS), Figure 43. Reference for the codes used in the distribution graph is provided below.
Result 11: In this improperly filled column representing those students who are already registered with IGNOU or not, following statistics are observed, 5% of students are already registered with IGNOU and are enrolled in two courses simultaneously (Figure 44).

![Figure 44: Distribution showing percentage of students already registered with IGNOU](image)

Result 12: In another poorly filled column of division, 30% students have obtained 2nd division (Figure 45).
4.5 Discussion with possible actions

Result 1 is self-explanatory. Figure 34 has an approximate shape of a normal distribution[98] as often exhibited in biological data[99]. This also resonates well with the model of our current education system where most people like to study or focus on their career value addition in their twenties or early thirties preferably. This kurtosis curve is skewed to the left and a bit slant on the right. It is also an indicator of generational divide, because all age groups are not represented. Distance learning is by definition for all age groups. Many of IGNOU’s programs have no age limit. Considering the category or type of students such pattern is still understandable.

Result 2 is due to the fact that these students find it easy to do humanities or social sciences courses because there is no help in the form of artificial limbs & training to use them in laboratories (sciences). Science laboratories have no accessibility equipment or area.

English medium books are comparatively easily available in India at higher education level. IGNOU however plans to launch courses in regional languages. More steps that can be taken are – to encourage translation of books/texts in all subjects and to make them accessible.
– brail translation, audio books (record and release), video field tours and online repositories of all these educational media.

Result 4 shows that disability is more common in males in this data sets. Or it could mean gender divide. So, this data raises more questions about infant & child mortality rate, gender biases.

Result 5 is self-evident. Students are in their 20s, so they have mostly passed out in recent past. Figure 38 is shape of a chi-square distribution[98, 99], with one outlier, ‘current’ year pass outs.

Result 6 requires action from Governments to create accessible jobs to increase employment. This static indicates economic gap which can rob a person off his/her freedom and lead to loss of confidence.

Result 7 indicates the possibility of urban area students having better accessibility to these courses. This needs to be verified by designing a focused future study for the same and improving awareness and accessibility in other areas as well. Figure 40 exhibits a knowledge divide.

Result 8, no e-mail ids indicate lack of access to internet and technology for a substantial number of disabled students.

Result 9, to help and promote the enrolment of disabled students in IGNOU, a study of assistive technology and some tools are suggested in Chapter 7.

Result 10 & 11 again suggest that enrolment forms must contain accessibility features. Or, some assistance should be provided as the students fill forms.

Result 12 shows that teachers will have to work hard with the students. This static again shows knowledge gap. To improve the learning of these students Adaptive Educational Hypermedia tools and Assistive Technology techniques have been suggested in Chapter 7.
Overall, options being filled need to be simplified and coded differently. Every field has options like A1, B2…, A1, B2,…, A1, B2 which can be very confusing. These options should be replaced by short forms like for Physical Handicap: PHN1, PHN2…, for Gender: M, F…., for financial status: l, lm, m, um, h (lower, lower middle, middle, upper middle, higher).

4.6 Summary

This chapter achieves a part of the objective of visualization through generating IVs and writing results, conclusion & reports based on them. This study also meets another objective of studying various divides in society: digital, economical, and generational and gender. These divides in the society might be limiting the interactions/education for learners in general and disabled students.