I J C T A, 9(20), 2016, pp. 221-229
© International Science Press

Analyzing State Board's Examination Results Using Data Mining Techniques and R Programming

Manpreet Kaur*, Ashutosh Pandey and Ashish Garg*****

ABSTRACT

Data mining techniques are being widely used to extract useful information from large amounts of raw data available in the form of big data. Big data refers to large volumes of data that may be structured or unstructured. This data is generated at an expeditious rate and in different forms. Through this work, we try to present some pertinent results and areas of improvement that need to be considered in state board's high school examinations. To calculate the required results R programming was used as it provides various data manipulating and processing techniques. The results presented here have been derived using R-programming's SQLDF library that makes querying large datasets easy through its SQL-friendly commands and the graphs were plotted using Microsoft excel. The prime motive was to provide recommendations so that the overall high school results of the state may be improved.

Keywords: Data Mining, Data Cleaning, R programming, SQL

1. INTRODUCTION

With the advent of 21st century, the most prominent progress has been made in the amount of data that is being generated every day. From satellite data to social networks, from Research and Development to education, data is being generated everywhere at a tenacious rate. Around 2.5 Quintillion bytes of data is being generated every day [1]. Dough Laney of Gartner Research, articulated the most widely accepted definition of Big data as the three Vs. of Big Data: Volume, Velocity and Variety. [2]

Volume: Massive amounts of data are generated through geo satellites revolving around the Earth, sensors used to determine climatic changes, unstructured data streaming in from social media etc.

Velocity: The unprecedented speed with which data is being generated calls for near-real time analysis so that relevant and useful information can be derived through it.

Variety: The data flows around in a number of different formats – It may be structured, as in traditional databases or in the form of unstructured text documents, email, video, audio, stock ticker data and financial data.

This data is generated in three forms, namely, – Structured, Semi Structured and Unstructured. This data needs to be handled, processed and analyzed as it may or may not contain vital information. Not only this data mining [3] helps in retrieving information but also, it helps in removing redundant data that may lead to wastage of storage space and inconsistency. Numerous attempts are being made to efficiently and quickly analyze these large quantities of data.

* Department of Computer Science and Engineering Graphic Era University, Dehradun, Uttarakhand, India, *Email: manpreet317@gmail.com*

** Department of Computer Science and Engineering Graphic Era University, Dehradun, Uttarakhand, India, *Email: ashutoshpandey2710@gmail.com*

*** Department of Computer Science and Engineering Graphic Era University, Dehradun, Uttarakhand, India, *Email: ashish.garg@iiitb.org*

Here we have presented our work that was based on the data describing the examination results of state board's high school examination. The data was in the form of CSV (Comma Separated Values) file which is a form of Semi-structured data. It contained the marks of 6 different subjects along with the students' registration numbers, school name, district code, taluq code and attendance in each examination. The goal was to analyze and retrieve all the information that could be extracted in order to help the schools and students build a better education environment for the students. For accomplishing the purpose, R programming language was used and the graphs were plotted. Different parameters such as District Code, Taluq Code were used to group the results and suggest which subjects need improvement and other similar suggestions which are presented in the Results section later.[4]

2. DATA CLEANING AND PROCESSING

As mentioned earlier, the goal of the data analysis was to provide a deep insight into the results of students appearing for state board's high school examination so as to build a better education system for the students. The data was in CSV format (Column Separated Values, semi-structured). The columns depicted various parameters related to a student. Studying the data manually would be a tedious as well as a slow task without giving much details and a deep insight into the root cause of the loopholes in the education system as the database itself was humongous. Thus, data mining techniques were used to query and retrieve useful information from the same database, thereby proficiently gathering pertinent results in a span of lesser time. As it was an extensive database (approximately 825 thousand rows), it required data cleaning also. Data Cleaning [5] refers to the process of filling in missing values, smoothening of noisy data, identifying or removing outliers, and resolving inconsistencies. For accomplishing this and the tasks mentioned ahead the SQLDF library of R programming language [6] was utilised. R makes it easier to process CSV files as it treats them similar to the way Relational Database Management Systems treat Relational Databases.

3. WORK DONE

The next step involved the processing of data [7]. For this SQL queries were written using the SQLDF library and run against our database (a CSV file). The resultant queries and conclusions drawn from the results are listed here along with the graphs that depict them.

1) The first task was to find the percentage of students failed in any particular subject district-wise. Here, we have shown the results of 3 queries for the subjects L1 (code for Kannada language based upon frequency), L2 (code for English language based upon frequency) and S2 (code for Science subject based upon frequency). It was found that the subject S2 has the poorest performance among all 6 subjects, hence it was chosen for comparison purposes. Also, subjects L3 (code for Hindi language based upon frequency), S1 (code for subject Mathematics based upon frequency) and S3 (code for subject Social Science based upon frequency) had results similar to those for L1 and L2 and hence, they are not shown for the purpose of brevity. It is noteworthy that the district with code "SS" (SS refers to BIDAR district) has the highest percentage of failed students in all 3 subjects.

Thus, immediate attention is needed to curb this growing menace of failing students in the district Bidar.

Also districts "QQ", "CA", "QA" and "OO" (codes for districts Gulbarga, Chikkaballapur, Yadgir, Bijapur respectively) are present in the top-10 largest percentage of failing students in each district, subject wise. Hence these districts also need to bolster and improve the overall performance of their students in each subject.

Figure 1.1 shows the 10 districts with the highest number of failed students in subject L1. District Bidar (code SS) clearly has the highest percentage of failed students. Other districts that follow are QQ (code for Gulbarga) and QA (code for Yadgir). These districts should be taken care of seperately while the reformative measures are being taken for improving the performance in the subject L1.

(Note – Codes for districts refer to SS- Bidar, QQ- Gulbarga, QA- Yadgir, OO- Bijapur, RA- Koppal, CA- Chikkaballapur, RR- Raichur, DA- Madhugiri, AN- Bangalore North)

(Note – Codes for districts refer to SS- Bidar, CA- Chikkaballapur, QA- Yadgir, EA- Chamarajanagar, DA- Madhugiri, IA- Davanagere,, RA- Koppal, QQ- Gulbarga, KK- Shimoga, OO- Bijapur)

Similar to Figure 1.1, Figure 1.2 shows the top 10 districts who have the highest percentage of failed students in subject L2. Again, district Bidar has the highest percentage of failed students.

(Note – Codes for districts refer to SS- Bidar, QA- Yadgir, CA- Chikkaballapur, QQ- Gulbarga, OO- Bijapur, AS- Bangalore South, AN- Bangalore North, CC- Kolar, EA- Chamarajanagar)

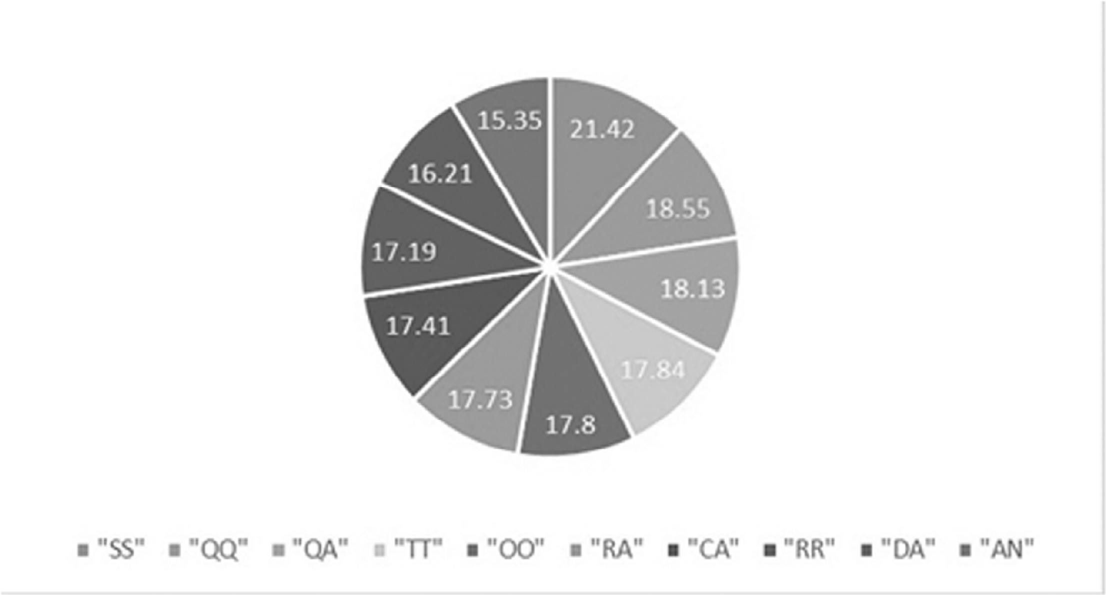

Figure 1.1: Percentage of Students failed in subject L1

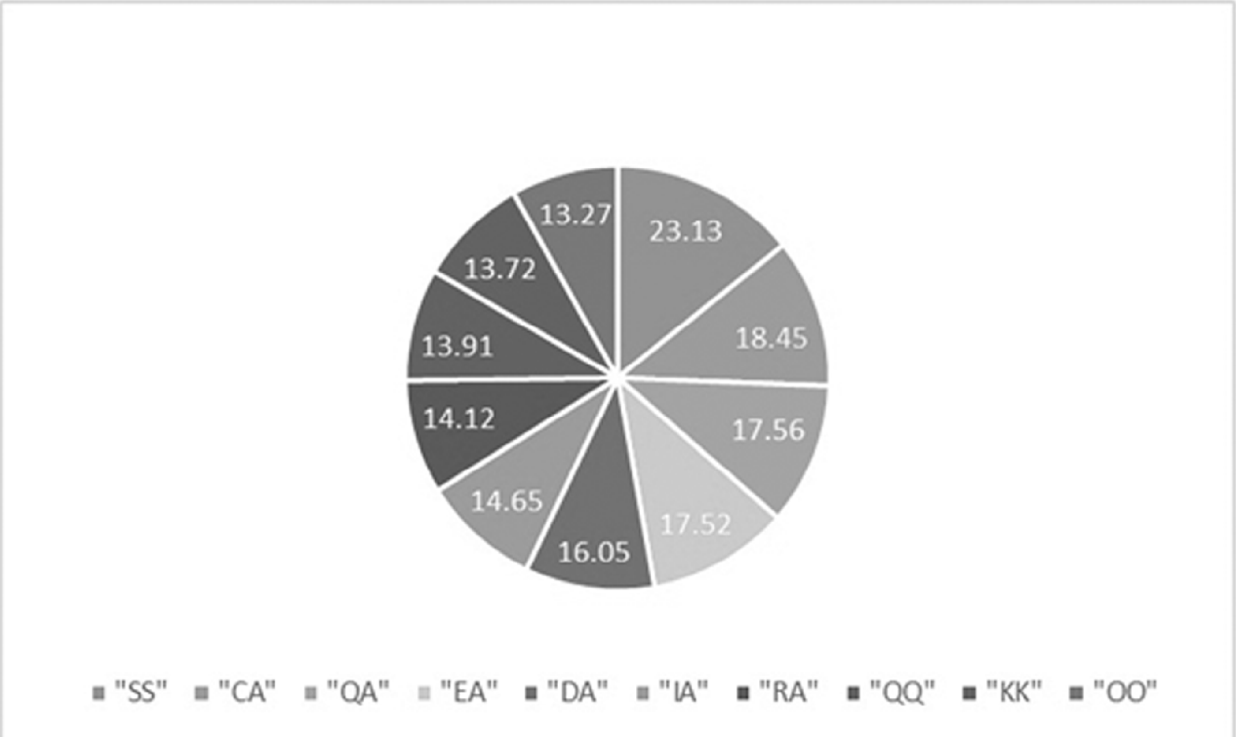

Figure 1.2: Percentage of students failed in subject L2

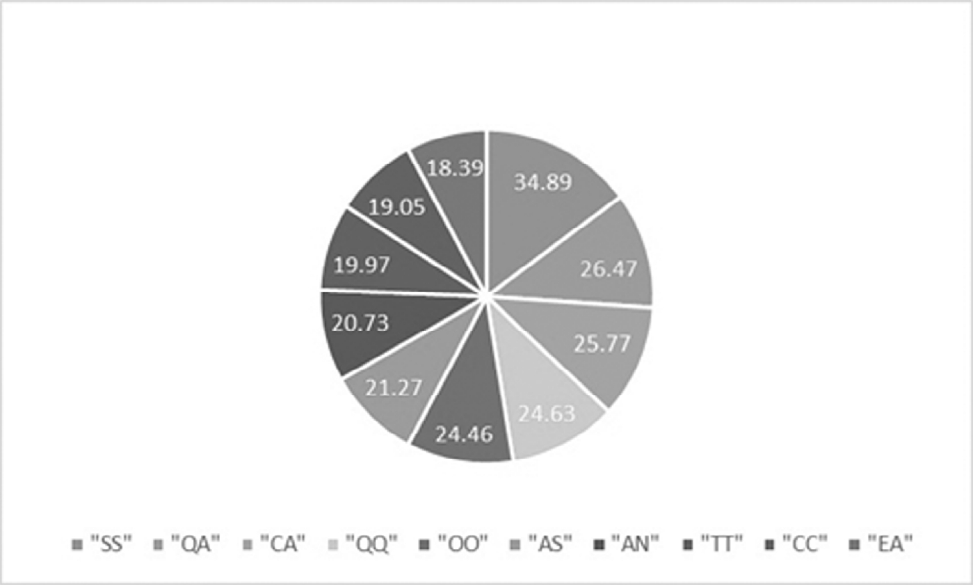

Figure 1.3: Percentage of students failed in subject S2

Figure 1.3 depicts an important result showing a remarkable increase in the percentage of failed students. While L1 and L2 highest percentage as 21.42 and 23.13 respectively, subject S2 has 34.89 as highest percentage of failed students.

2) Next we aimed at studying the count of absentees in the subjects during the examination and compare the results in various districts. Here, we have shown the graphs for the subjects L1 and S2. L1 has the highest number of absentees during the examination while S2 has the lowest. It can be observed that the number of students absent during the examination of L1 subject was highest in the district "QQ" (code for district GULBARGA). Similarly a large count of students was absent in districts "SS", "AS", "OO" (codes for districts BIDAR, BANGALORE SOUTH, BIJAPUR respectively) as well. These districts need to improvise the reason for such prodigious number of students missing examination in this particular subject and need to analyse and derive a solution for it.

(Note – Codes for districts refer to QQ- Gulbarga, SS- Bidar, AS- Bangalore South, OO- Bijapur, AN-Bangalore North, EE- Mysore, II- Chitradurga, GG- Mangalore, LL- Hassan)

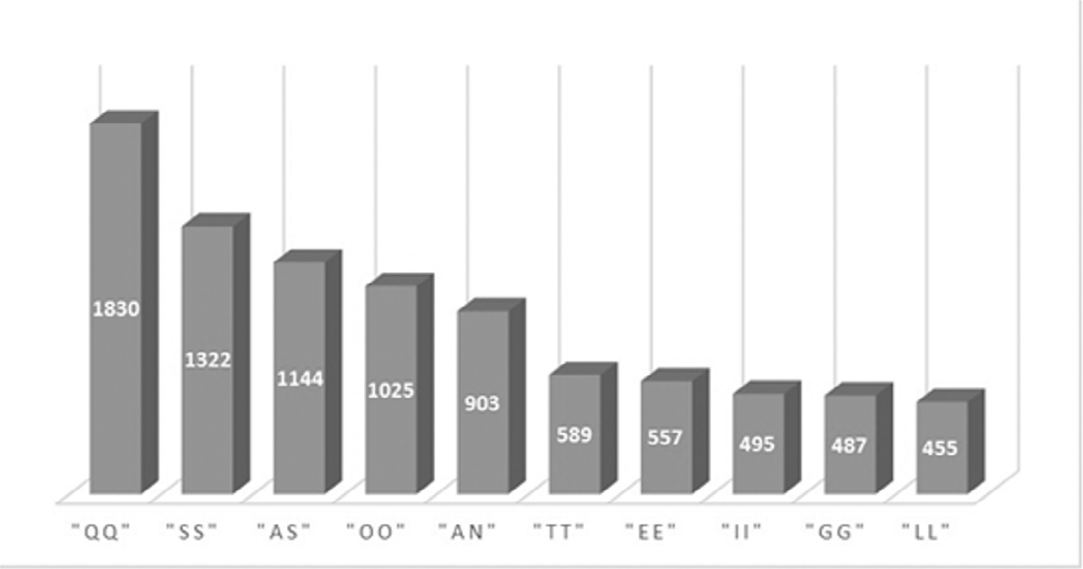

Figure 2.1: Number of Students absent for L1 in Various Districts

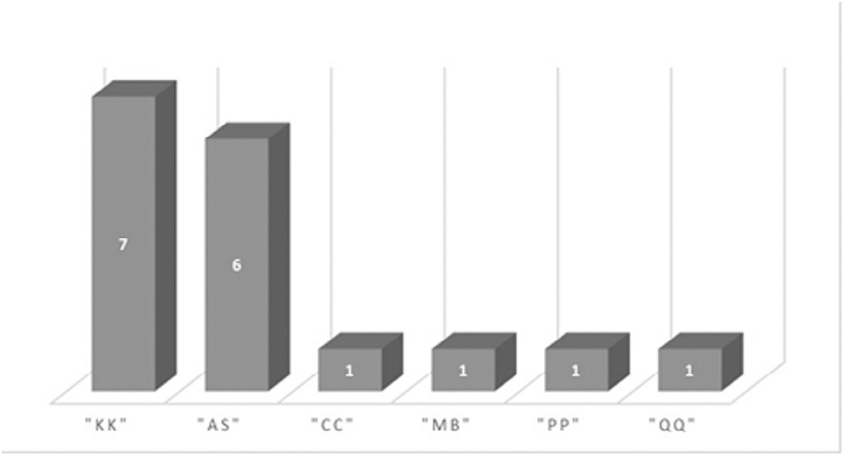

Figure 2.2: Number of Students absent for S2 in Various Districts

(Note – Codes for districts refer to KK- Shimoga, AS- Bangalore South, CC- Kolar, MB- Gadag, PP- Uttara Kannada, QQ- Gulbarga)

Also, it can be seen that S2 subject has the least number (maximum 7) of absentees in various districts and still has the largest percentage of failed students. Thus an immediate survey should be done in order to ascertain the cause behind the poor performance of the students.

3) The next task was to count the Percentage of Total number of students pass and fail in any district. Again, as the query required the results to be plotted against the districts, the results were grouped based upon the district. One of the prominent features that could be noticed from the graph was that the district with code as SS (BIDAR) had a 54 percent to 46 percent pass to fail ratio. To frame this in simple terms, it would mean that almost every second student belonging to that district failed. One of the important assumption that was taken during the calculation of this result was that if a student failed in any one of the given 6 subjects, he was treated as "fail".

   This means that the Bidar district requires immediate attention so as to curb the factors responsible for the poor performance of the students and it can be nipped in the bud by taking the appropriate measures as necessary.

   Along with this district, other districts could be ranked according to their pass to fail ratio so that the overall performance could be improvised.

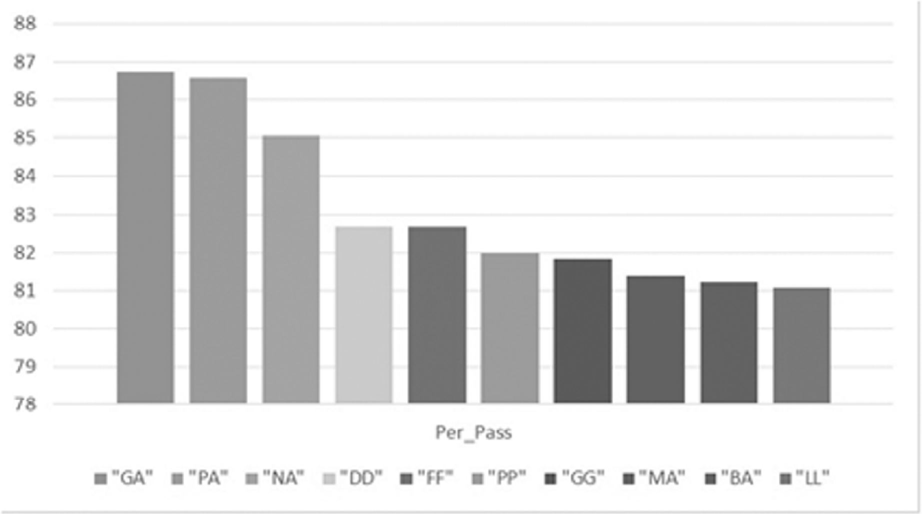

Figure 3.1: Percentage of students pass district wise

(Note – Codes for districts refer to GA- Udupi, PA- Sirsi, NA- Chikodi, DD- Tumkur, FF- Mandya, PP- Uttara Kannada, GG- Mangalore, MA- Haveri, BA- Ramnagara, LL- Hassan)

It can be observed from the above graph that the districts GA and PA (codes for UDUPI and SIRSI respectively) students' are outperforming others with pass percentage and thus other districts also need to perform on similar benchmarks and motivate their students to achieve better results in the examination.

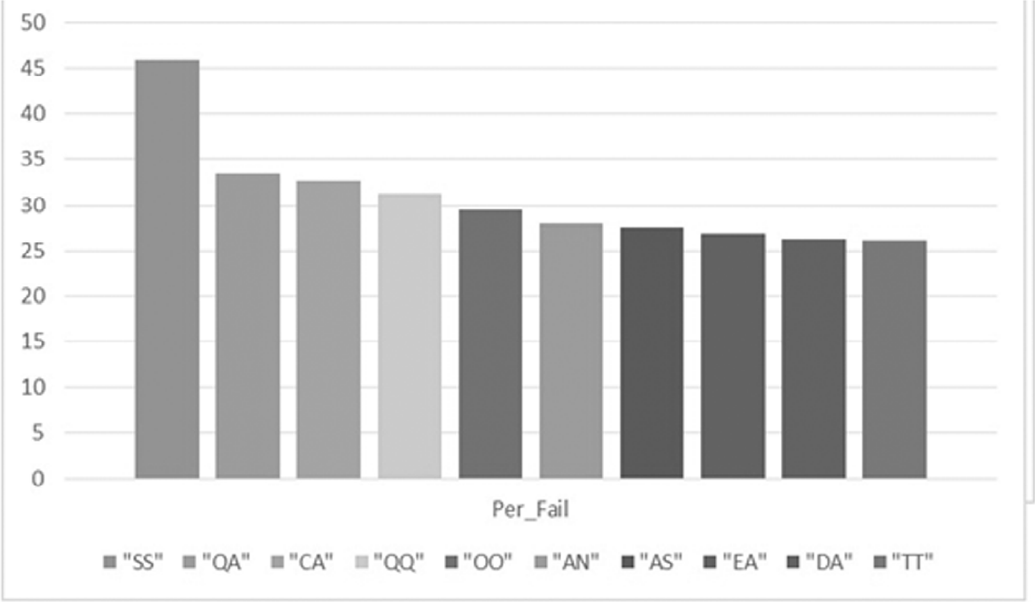

Figure 3.2: Percentage of students fail district wise

(Note – Codes for districts refer to SS- Bidar, QA- Yadgir, CA- Chikkaballapur QQ- Gulbarga, OO- Bijapur, AN- Bangalore North, AS- Bangalore South, EA- Chamarajanagar, DA- Madhugiri)

4) Further we wanted to provide a comparison between the reserved category students and the unreserved category students. For this two parameters provided in the given database were utilized and the graph showing the comparison was plotted. First, percentage of students pass in the reserved category versus the percentage of students in the unreserved category. A similar graph was plotted to depict the percentage of failed students in both the earlier mentioned categories.[8]

   Both the queries grouped results based on the district codes. This was done to segregate the results so that each district could easily recognise the performance of students in their respective areas. Here, we have also provided a partial snapshot showing the results of this query in the form of bar graphs.

   The following four graphs depict the results of the query mentioned. The first two graphs (Figure 4.1 and 4.2) show the comparison between percentages of failed students. One of the results derived was that the percentage for Unreserved category students falls more steeply (which is the ideal case) whereas for the Reserved category students it is not as steep. Thus the reason behind the poor performance may be attributed to the Reserved category students' performance and necessary steps should be taken to improve their performance.

   In the next two graphs (4.3 and 4.4) a similar result is evident. The highest percentage of students pass in Reserved category is 41.54 whereas for Unreserved it is 73.32. This adds to the conclusion drawn above that the overall performance of various districts may be declining majorly because of the negligence of Reserved category students towards academics.

   Thus from the graphs it is clear that specific measures should be taken to improve the performance of Reserved category students more than Unreserved category.

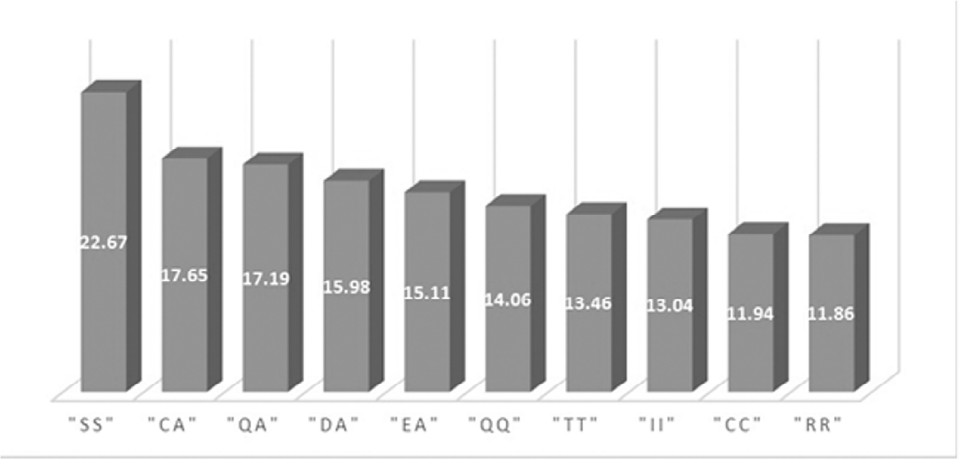

Figure 4.1 Percentage of students fail in Reserved category District wise

(Note – Codes for districts refer to SS- Bidar, CA- Chikkaballapur, QA- Yadgir, DA- Madhugiri, EA- Chamarajanagar, QQ- Gulbarga, II- Chitradurga, CC- Kolar, RR- Raichur)

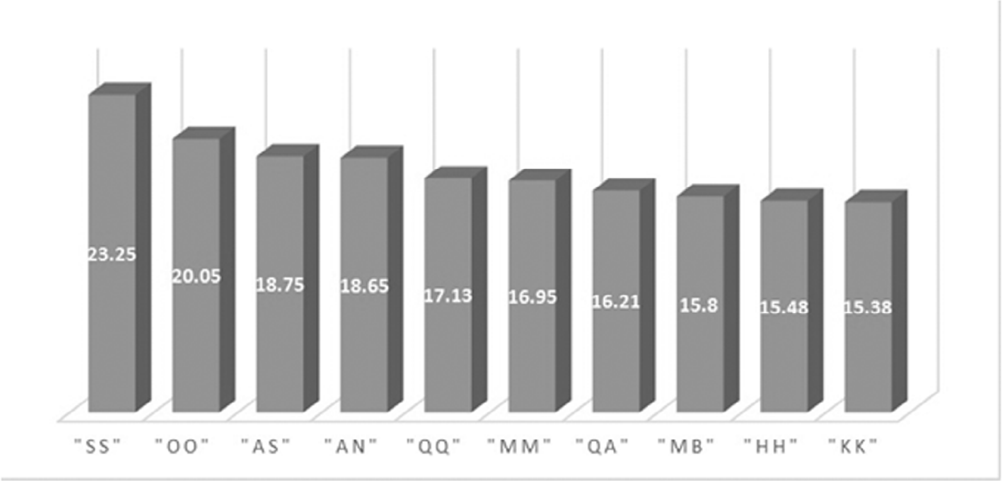

Figure 4.2 Percentage of students fail in UnReserved category District wise

(Note – Codes for districts refer to SS- Bidar, OO- Bijapur, AS- Bangalore South, AN- Bangalore North, QQ- Gulbarga, MM- Dharwad, QA- Yadgir, MB- Gadag, HH- Kodagu, KK- Shimoga)

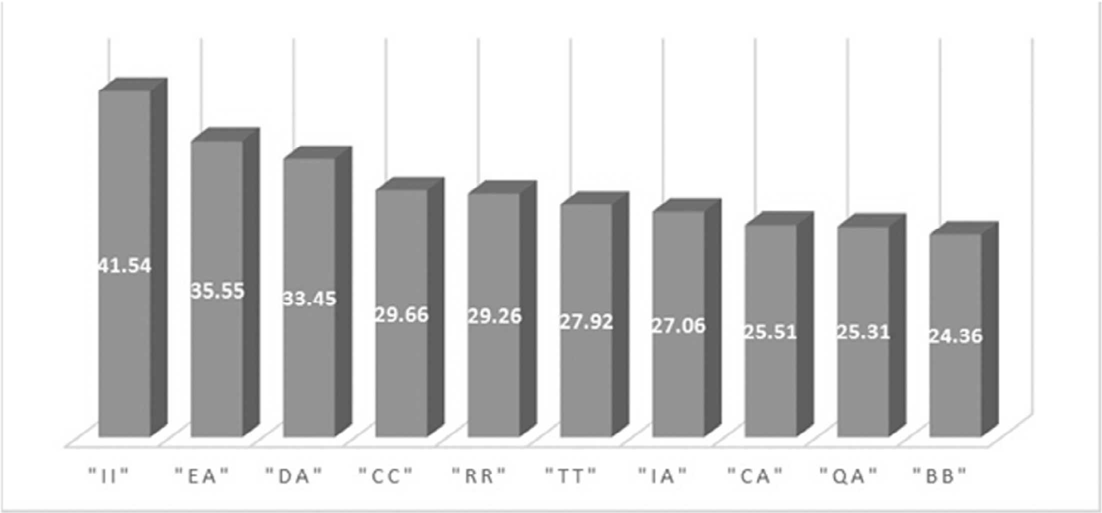

Figure 4.3 Percentage of students pass in Reserved category District wise

(Note – Codes for districts refer to II- Chitradurga, EA- Chamarajanagar, DA- Madhugiri, CC- Kolar, RR- Raichur, IA- Davanagere, CA- Chikkaballapur, QA- Yadgir, BB- Bangalore Rural)

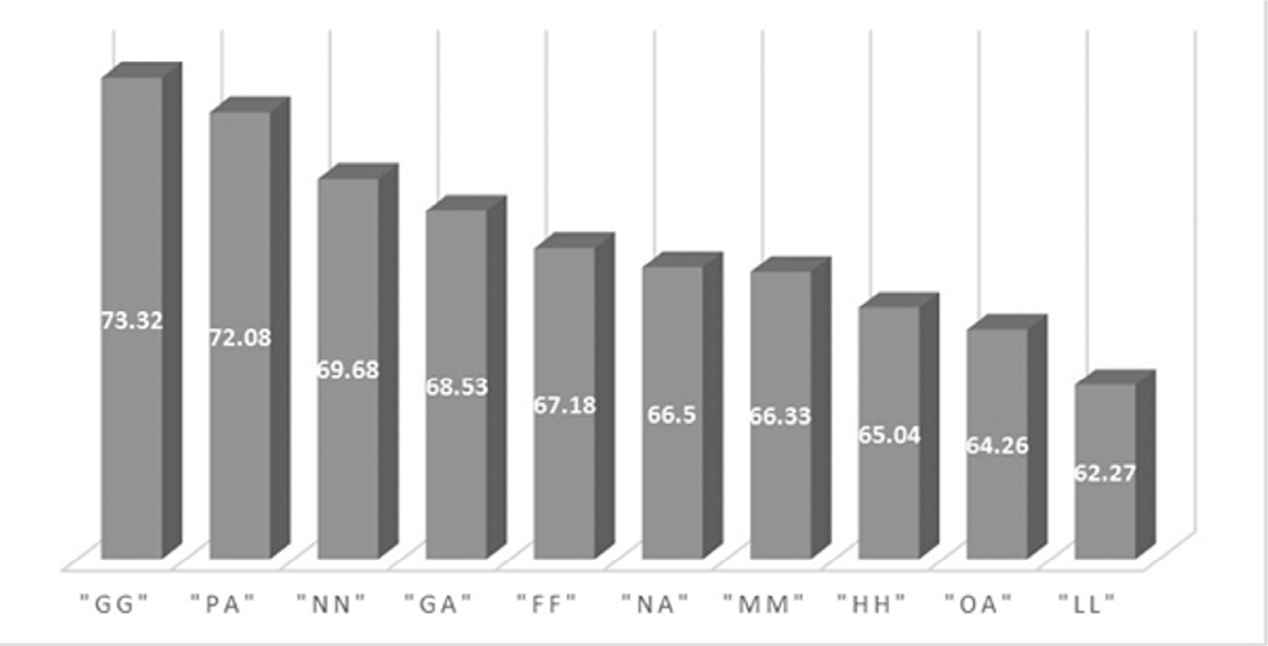

Figure 4.4 Percentage of students pass in UnReserved category District wise

(Note – Codes for districts refer to GG- Mangalore, PA- Sirsi, NN- Belgaum, GA- Udupi, FF- Mandya, NA- Chikodi, MM- Dharwad, HH- Kodagu, OA- Bagalkote, LL- Hassan)

The 2 graphs (Figure 4.1 and 4.2) show the percentage of failed students. The next ones, (Figure 4.3 and 4.4) similarly show a comparative graph between various districts for the percentage of passed students. Only 10 bars are shown in each graph to maintain the limpidity for the readers. The total number of districts present in the original database were 34 and all of them were compared against each other in order to calculate the required result.

4. CONCLUSION

The work presented here shows the performance of various students from a number of views – district-wise pass to fail, district wise number of absent students in each subject, a comparison of reserved category students to unreserved category students and taluq wise pertaining results. The goal was to utilize the various parameters of the provided database and help develop a better education system for the students.

The overall performance can only be improved by providing the relevant results to the authorities and motivating them towards the remedial measures that need to be taken. Through this work, we have tried to present an optimal statistical measure to give a summarised insight into the state board's high school examinations. The results show some realistic problems, like subject S2 being the one with the poorest performance and at the same time having least number of absent students, that should be addressed at the earliest.

5. FUTURE WORKS

Through this study, we have tried to present the sustaining problems in the state board's education system so that they can be addressed at the earliest and the students can make the most of their education. While some research has been initiated, similar works can be undertaken to address the problems of other state board examinations as well. At the same time, it is also important for the concerned people to consider the results and introduce the required changes and norms to address the pertaining issues. Other similar useful information can also be extracted to raise similar issues in the examination process. Apart from the work presented here, other methodologies for data mining and processing can be used to develop better and more useful outcomes from the same database.

REFERENCES

[1] V Cloud News, "Every day big data statistics - 2.5 Quintillion bytes of data created daily -," in *Big Data*, 2015. [Online]. Available: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/. Accessed: May. 19, 2016.

[2] M. A. Beyer and D. Laney, "The importance of 'big data': A definition," Gartner, 2012. [Online]. Available: https://www.gartner.com/doc/2057415/importance-big-data-definition. Accessed: April 30, 2016.

[3] U. Fayyad, *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 115–119, 1998.

[4] Ahmad, Fadhilah, Nur Hafieza Ismail, and Azwa Abdul Aziz. "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques." *Applied Mathematical Sciences* 9.129 (2015): 6415-6426.

[5] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." *IEEE Data Eng. Bull.* 23.4 (2000): 3-13.

[6] Coffey, Amanda, and Paul Atkinson. *Making sense of qualitative data: complementary research strategies*. Sage Publications, Inc, 1996.

[7] Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A prediction for performance improvement using classification." *arXiv preprint arXiv:1201.3418* (2012).

[8] Pal, Saurabh. "Mining educational data to reduce dropout rates of engineering students." *International Journal of Information Engineering and Electronic Business* 4.2 (2012): 1.