Data Visualization using Tableau

Dharmik Trivedi
System and Computer Engineering
Carleton University
Ottawa, Canada
dharmiktrivedi@cmail.carleton.ca

Fateme Rajabiyazdi
System and Computer Engineering
Carleton University
Ottawa, Canada
fatemerajabiyazdi@cmail.carleton.ca

***Abstract*—This paper provides insights about how students in secondary education will perform in their 2 subjects: Mathematics and Portuguese. It is divided in 3 parts. Every part provides different insights. The visualization is made using tableau public.**

Keywords—data visualization, design optimization, insights, interactions

I. Introduction

The technique of presenting massive amounts of data in a visual format that aids in the discovery of important patterns in the data is known as data visualization. Some online data visualization tools, however, can present up to five dimensions of data at once, whereas traditional data visualization tools, such paper-based graphs, are fundamentally limited to presenting two dimensions of data.

The robust processing capabilities of human vision are largely responsible for data visualizations' success, with the best ones mapping data to perceptual processes in a way that makes it simple to identify significant patterns. Data visualization works well because it balances vision and cognition to better utilize the brain's capabilities. The visual cortex in the back of the brain, which is responsible for seeing (visual perception), is incredibly quick and effective. Without much effort, we automatically perceive. The cerebral cortex, located in the front of the brain, is primarily responsible for thinking (or cognition), which is significantly slower and less effective. The majority of the labor involved in using traditional data sense-making and display techniques requires conscious thought. By utilizing our powerful eyes whenever possible, data visualizations tips the scales in favor of more visual perception [1]

In this paper, a visualization has been created using Tableau to get meaningful insights for stakeholder.

II. Background

A. Dataset

Data is a crucial part of decision-making since it provides companies and organizations with valuable insights and a thorough understanding of the effects of their choices. Additionally, interactive dashboards and visualizations are crucial tools for everyone who wants to understand data and create stories using data, from CEOs to schoolchildren. The best sources to use when building visualization tool are open data sets. There are so many various datasets available at tableau public, Kaggle and Power-Bi. For this paper, the dataset has been selected from Kaggle.

In this case, the dataset taken from Kaggle has 30 various attributes/factors/features which can affect students' grades G1, G2 and G3 in Mathematics and Portuguese subjects. We may divide the features into numeric and categorical. Numeric can subcategorized into continuous and discrete features. Some features are related to **student's academic life** like study time, school, school support, extra-tuition, guardian; some are related to **family**, for example, family size, parents' education, parents' job, family support, relation between parents, relationship with family; some are **personal**, i.e. health, relationship status, age, daily and weekly alcoholic consumption whereas some feature also talk about non-academics. Like, travelling time to school, freetime, gender, extra-curricular etc. So, it's interesting to how these features affect and change overall student's performance at high school. Discrete features like parents' education, study time, free time, travelling time, alcoholic consumption are scaled on 0-4 in the data and categorical features are mostly binary(yes/no). The only thing which can make the dataset better is timeframe. There is no information about year, date, or month anywhere in it [2].

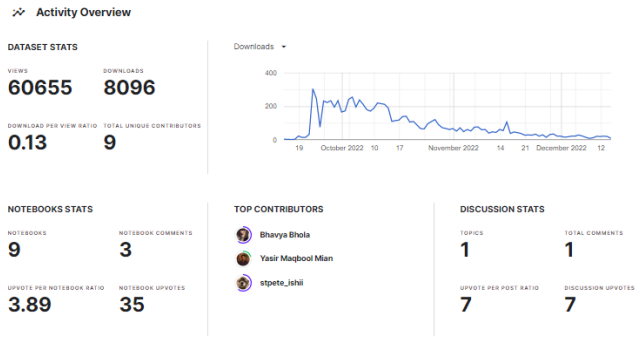

Fig 1. Overview of dataset statistics

B. Dataset Filtering

- For this paper, data filtering is done using Microsoft Excel

As the name implies, data filtering can assist you in removing extraneous data. Data filtering can be used, for instance, to eliminate all records that contain either type of field in order to determine the total number of records in a dataset with two different types of fields, such as numbers and strings.

For example, a software company called xyz technologies creates and markets a project management solution. The business has assembled a small team to handle project management duties, which include compiling data on all of the organization's initiatives. The rest of the crew, excluding the project managers, is only involved in processing reports for recently finished or upcoming projects. The team decides to incorporate a fresh data filtering procedure that will enable all team members to exclude extraneous data from their reports.

The IT team initiates a data filter process that begins with adding a new field named "Status", which is assigned an integer value 1, 2, or 3. This field can fill out all reports created by one or more members of xyz technologies. The next step is to add a filter to the field named "Status". The filter replaces integer values with the strings "Recurring", "In Progress", or "Completed".Some Common Mistakes [3].

- Link: Dataset
- The dataset is about **student achievement in secondary education** of two Portuguese schools for the subjects: Portuguese language and Mathematics. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.
- Size :
  1) Portuguese.csv: Column: 33, Rows : 650
  2) Mathematics.csv: Column: 33, Rows: 396
- The two csv files Maths.csv and Portuguese.csv are merged into a single excel file in order to accommodate 'subject' filter in the visualization
- Excel operations like sorting, filtering and transposing were done in order to sort the data to see the filter option in visualization in a better way.

III. Sample Visualization in Literature

After the dataset has been chosen, a visualization found in literature is used to critique.

A. Source

Data source includes extensive data of energy statistic of Canada (Sample Visualization) which is being generated and collected every second. The data from data source includes electricity consumption by every state in each year, Usage of electricity generation by various resources, types of use (residential, industrial, commercial, transportation). From this data, they predict various scenarios for electricity consumption of every state in future.

B. Data Attributes

Data attributes in this visualization are the following: Energy Unit (PJ-Petajoule), year, region, energy resource, scenario, electricity generation, oil production, gas production, total demand etc. There are quantitative (nominal, ordinary, binary) and qualitative (numeric, discrete, continuous) attributes [1]. For example, scenario is a binary attribute as it has only two possibilities (past, future). Nominal attributes: Energy resources, Discrete attributes: year. Likewise, data attributes can be further classified into their subcategory. However, at this point of discussion, it does not have much importance.

C. Data Dimentions

Data dimensions constitute categorical data, such as year, sectors, states, and year range. Categorical data is also known as nominal data and is used for discrete values. The dimension called states may include the values of ON, MB, PE, QC, etc. In interval data, each value in the dimension represents a range of values. The dimension PetaJoule can be categorized into the following values of salary ranges: 3000PJ-4000 PJ, 4000PJ-5000PJ,…,11000PJ-12000PJ.

D. Marks and Channels

Connection Marks: rectangles for state names, connecting lines for energy resources, connecting small circles for the category-scenarios, small squares to show main 5 category's name of energy future Containment Marks: dotted lines for the category-by sectors, dotted lines for the category-by sectors Identify channels: simulation to show motion, color hue for different types of energy resources, shape ratios to show usage of particular resource, horizontal positions and vertical positions of bar charts in by region category Magnitude channels: area of circle to show respective usage, time scale to represent year, symbols to identify energy resource, social media link symbols

IV. Task Identified

Visualization is about "Student Grades in Secondary Education" and the target audience, using the visualization, would be able to:

-Compare the grades between St. Mary and St. Thomas Schools

-Compare the grades among Male and Female students

-Observe trend of student habits having particular grade/grades' range

After using this visualization, one would be able to comment about "School Performance" and "Gender wise Student Performance", and "Habits of top performing/least performing" students.

Examples

- Apply Age,Subject, Gender filters

-Compare St. Thomas and St. Mary based on Average Marks of Students

Possible Insight:

-Which School is better for Mathematic, age =18 students

- Apply family, personal, academic filters

-Compare Male and Female Student Marks

Possible Insight:

-Which school's girls are performing poor in Portuguese?

- Filter grades range, gender and subject

-Check the trend of Student habits for different grade ranges

Possible Insight:

-Check the study time, health, going out frequency, weekly alcoholic consumption of students scoring less than 13(out of 20) marks.

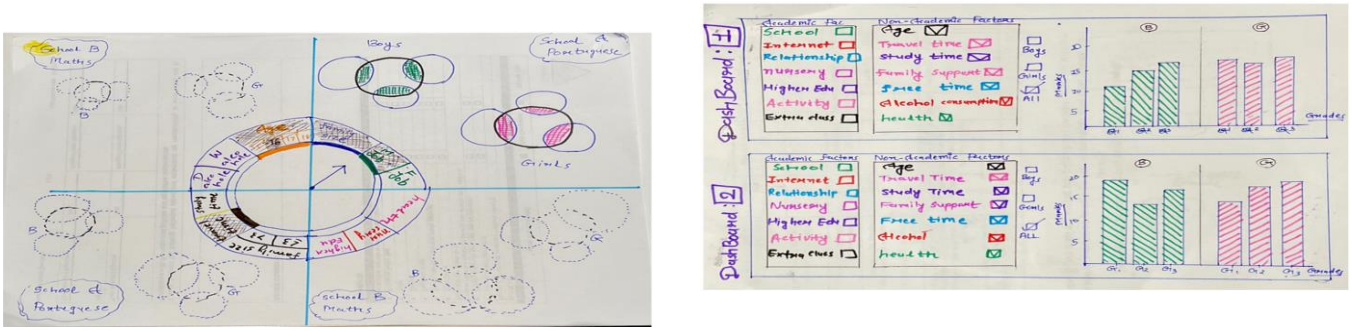

Fig 2. *Brainstorming* Visualization with Pen and Paper before making it

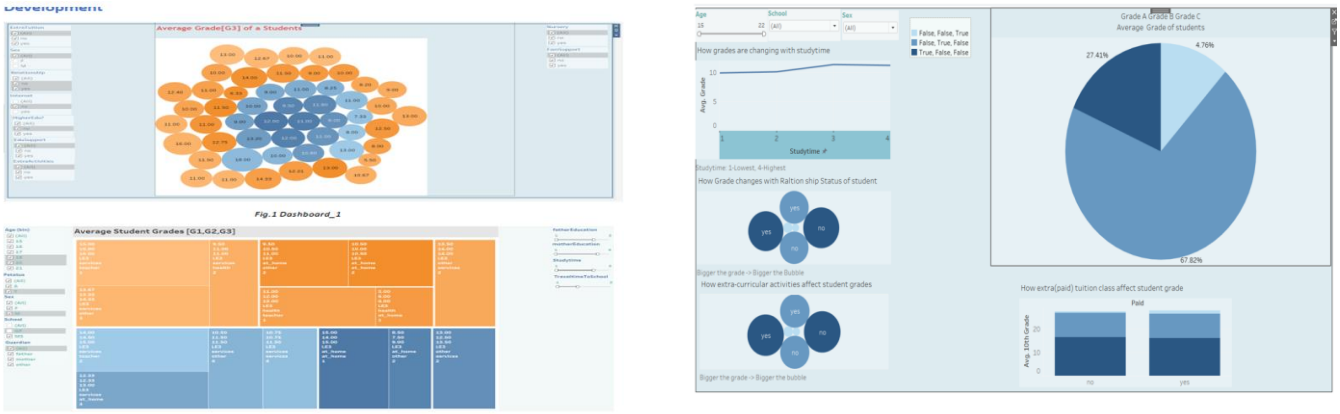

Fig 3. Initial Designs made in Tableau

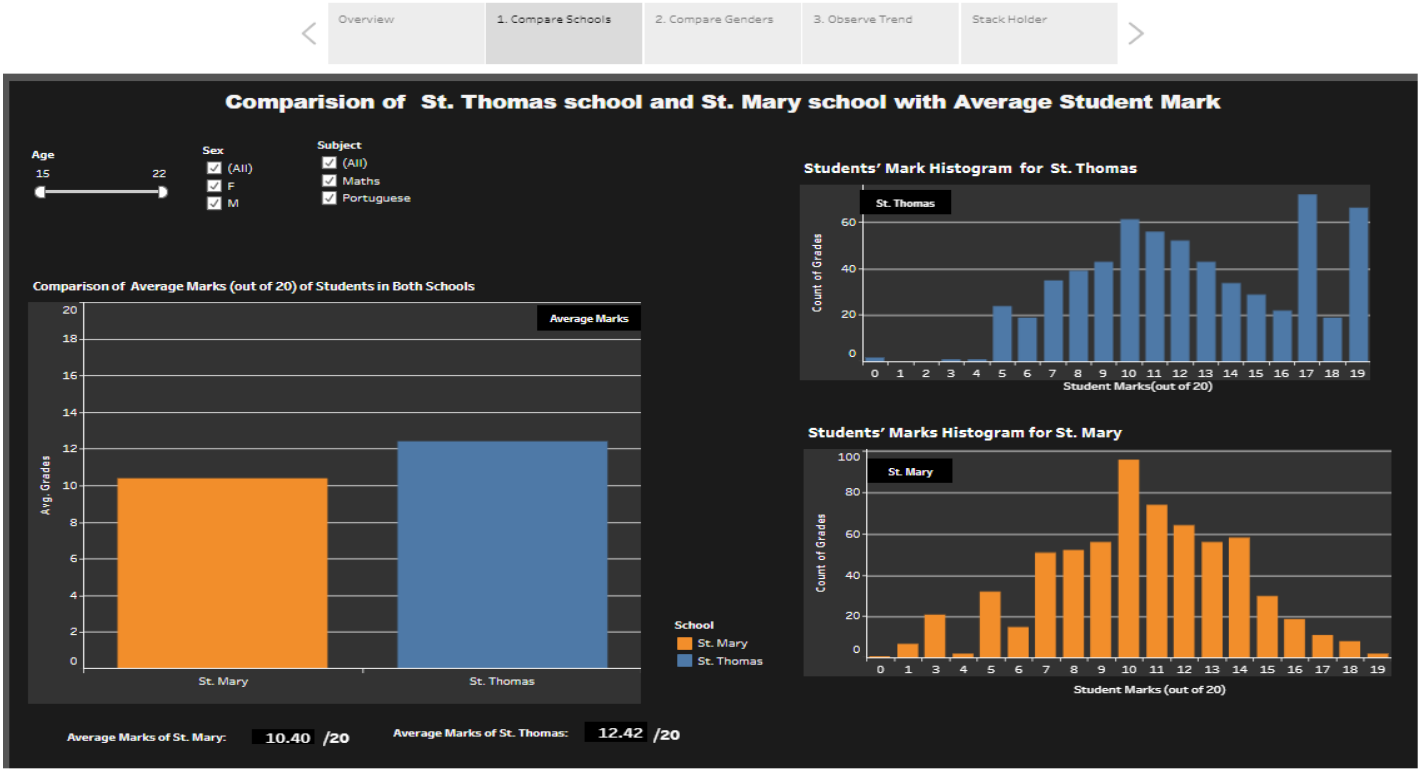

Fig 4. Final Design in Tableau

V. DESIGN AND ITERATIONS

Initial design with pen and paper was bubble chart to compare average grades of students. To compare the grades between male and female students, pie chart was finalized as a center chart of dashboard. Third dashboard was specially designed to see how student performance affect with family factors like student-parent relationship, parents' education, parents' education background and so on. The comparison was proposed using bar chart at that time.

INTERACTION IN DASHBOARDS

Dashboard	Description	
	Title	***Design Intention***
1	Overview	When user comes first time, an overview is given on how to get most out of the visualization
2	Compare Schools	Filters are provided in a such way that using them, gives a comparison between school
3	Compare Grades	Filters are provided in a such way that using them, gives a comparison between grades of Male and Female students
4	Observe Trend	Gives a chance to user how students having various range, have various habits right from study time alcoholic consumption
5	Stake Holder	Information for stake holder, how the visualization could benefit to them

Table 1- Dashboard overview in visualizaiton

Dashboard 1

It gives overview, what is visualization is all about, how to use it to get useful insight. This is the first tab where user comes.

Dashboard 2

Why needed?

This is the first dashboard where user comes after reading the overview of the visualization. The main intention to develop this entire dashboard is to have a fruitful comparison between St. Thomas and St. Mary high schools. The comparison is done in a measure of average marks of students. The chart for that is kept biggest for good visual output for users. Other two charts are histograms for St. Thomas and St. Mary Respectively. A separate color coding is used to see difference between them. Histogram tells how many students scored a particular mark number.

Why is a particular chart used?

Bar chart is used for comparing school. Initially, bubble chart and pie chart were proposed. However, bubble chart was not making sense according to the data; pie chart was hard to understand and think more insight. After the evaluation study and discussion with guide, a conclusion was achieved 'to go for the easiest encoding which can be understood'; bar chart is used for comparison.

Sample Insight

Filter the age group 15-18 from age filter, keep all gender and filter Maths category from subject filter. The visualization provides, the average mark in mathematics subject for all students between 15-18 age, in St. Thomas and St. Mary school.

Dashboard 3

Why needed?

The main intent to develop this dashboard is to have a academic comparison between male and female students. Here, filters are selected from various situation in student life. Not only academic factors (like study time) affect the student mark, non-academic factors (like extra-curricular activities), too. Four category is selected to draw 2 factors from each. For a particular factor, how it affects average marks of male and female student is presented here.

Why is a particular chart used?

Only one big bar chart, in order to make it simple and easygoing for users. Filters are more of a focus.

Sample Insight

Filter Portuguese and 'no' in non-academic factor; observe how average mark differs in both genders

Dashboard 4

Why needed?

To watch trend of habits of high-performing and low-performing student, this final dashboard is made. It is most important for schools, to watch habits of low-performing student in various age groups and guide them to change it; to improve their academic success. The comparison can be done school wise, too, so Edu Tech companies can target particular school, particular age group of specific gender.

Why is a particular chart used?

To observe the trend, line chart is used. Not only too see the trend in individually but also, it is easy to compare looking at four various line charts at a time.

Sample Insight

Filter female students at St. Mary school of grade range 0-9; see what the study time is, alcoholic consumption of majority of girls. Observe the same, for grade range 13-20. Note the difference in their habits.

Dashboard 5

This is designed to see how the visualization is helpful for stake holders.

VI. EVOLUTION STUDY

Why is it needed?

The evolution study was done to know how visualization will perform for stake holders. In order to do that, a comprehensive study was done. The way of conduction is presented below:

Evolution Study

- Type of Evaluation: in- person
- Relation: Friend
- Time: 20 minutes
- Location: Home
- Connection as a target audience: To some extent!
- Way of collecting feedback: Oral
- Type of Consent: Signed consent form (Electronic)

General Questions

- How long it took to understand each and every functionality correctly?
- Were filters easy to play around and useful?

- Were you able to find insights at the end or able to comment about your conclusion?
- Can it be proposed to general audience/target audience?
- Was it easy to navigate through different graphs and understand the relation between them?

Factors for rating	Rating out of 5
General Que 1(3 Minutes – 5.0, 5 Minutes-4.0, 7 Minutes-3.0, More than 7 minutes: 1.0)	3
General Que 2 (Very Easy and Useful-5.0, Complex and Useful-4.0, Complex and Not Useful-3.0, Not understood-1.0)	4
General Que 3 (2 Conclusions-5.0, 1 Conclusion-4.0, No Conlcusion-1.0)	4
General Que 4 (Yes without changes-5.0, Yes with Minor changes-3.0, Yes with Major changes-3.0,No Conlcusion-1.0)	3
General Que 5 (Very Easy and Useful-5.0, Complex and Useful-4.0, Complex and Not Useful-3.0, Not understood-1.0)	5
	Avg = 3.8

Fig 5 Evolution Results

Results

The visualization is improved as per the user experience. The drastic change was design itself. Instead of using pie charts, the comparison was done using bar charts.

VII. Discussion and Future Work

So far, the visualization has been improved a lot since the initial design. However, there is always a room for improvement. Here are the few suggestions to improve the overall efficiency of visualization. With more data of subject, school, comparison level can be improved

1. More filters which affect student life can be introduces.
2. Visualization is not evaluated on visually blind users yet. Respective changes in color coding can be made.
3. More optimal design, which can navigate new users more easily, more conviction of relationship among graphs can be proposed.

VIII. Conclusion

There are lot of academic competition among schools. A healthy environment of competition is fruitful for all the students. The visualization successfully compares performance of students at 2 different schools in 2 subjects, Mathematics and Portuguese. This is useful for edu-tech companies to target students at particular school of which, male or female students poorly performing in a specific subject. The policymakers of school can see the habitual difference between high performing and low-performing students; guide them; accordingly, introduce a policy which is optimal for all the students.

References

[1] Center of teaching excellence-University of Waterloo (uwaterloo.ca/centre-for-teaching-excellence/teaching-resources/teaching-tips/educational-technologies/all/data-visualization)

[2] Official Documents of Tableau (link)

[3] Official Documentation of Kaggle (link)

Appendix

1. Link to the visualization : Data Visualization using Tableau