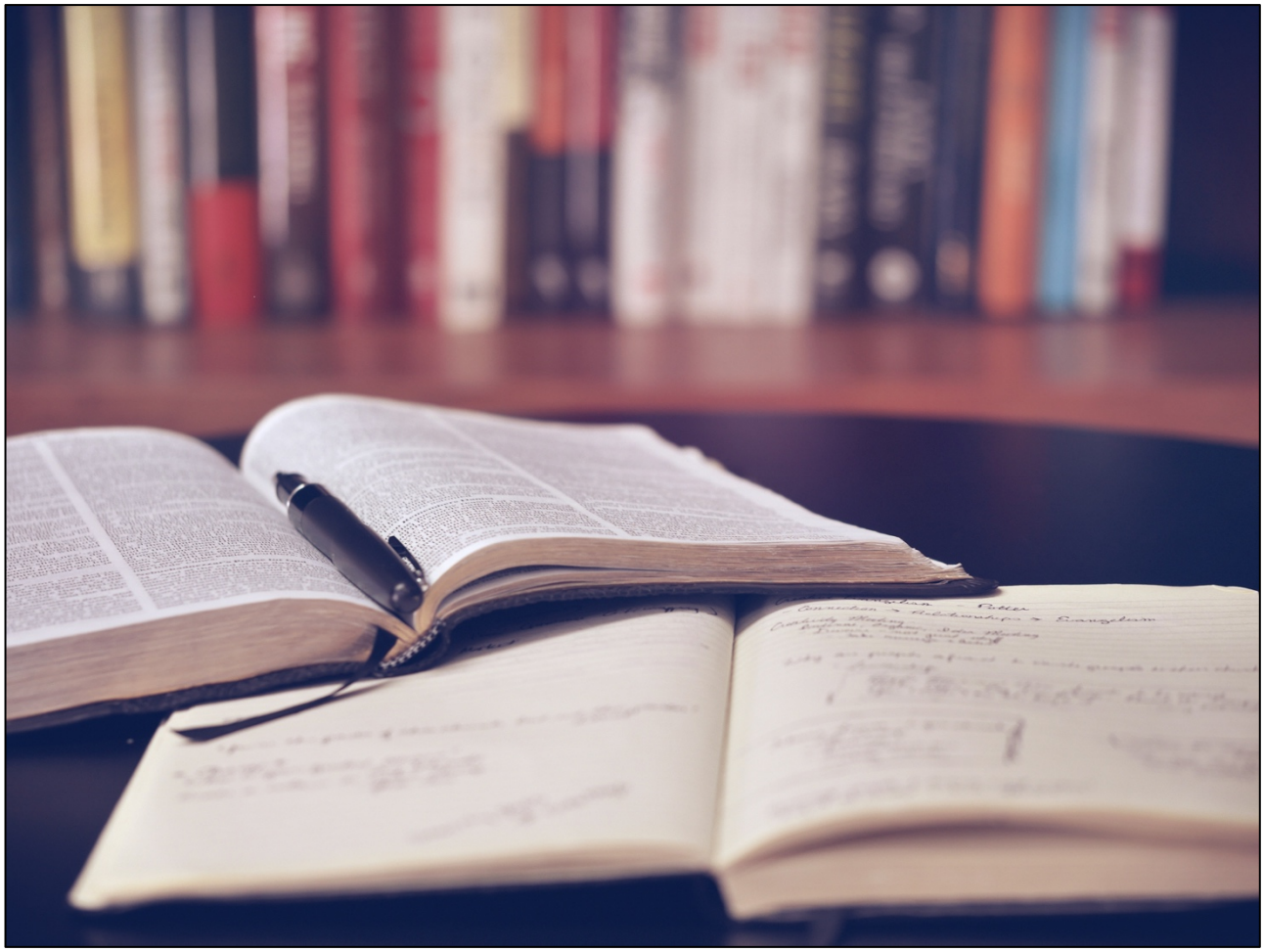
CS5603 Data Visualisation: U.S. Educational Finances



Issam Jaber
MSc Data Science and Analytics
Brunel University

Section 1
INTRODUCTION

The purpose of this project is to produce a visualisation of a chosen dataset, aimed at answering a specified problem. The report will begin by introducing the dataset before detailing the questions this report seeks to answer. The following chapter will discuss the rational and evolution of the visualisation design, from the initial paper prototype design to the final dashboard. In the third section, this report will describe the implementation process, it will outline the methods used to produce the final design. Section 4 will demonstrate through an illustrated walkthrough how the created visualisation sufficiently answered the planned questions. In the fifth and final section, there will be a critical assessment touching on the entire project.

1.1 The Dataset

The chosen dataset will be of "U.S. Educational Finances". This dataset is based on annual surveys, conducted by "The United States Census Bureau", assessing the finances of elementary and high schools (Kaggle, 2017). The dataset contains a summary of revenue and expenditure for the years 1997-2015, organised by state and year. Also added to the dataset is the GDP per capita for each state and the same period. The dataset consists of 13 columns and 913 rows. The variables are shown in Table 1.

VARIABLE	DESCRIPTION
State	Name of State
Year	Year of Data
Enroll	Fall Membership
Total_revenue	Federal_revenue + State_revenue + Local_revenue
Federal_revenue	Total Revenue from Federal Sources
State_revenue	Total Revenue from State Sources
Local_revenue	Total Revenue from Local Sources
Total_expenditure	Support_services_expenditure + Instruction_expenditure + Capital_outlay_expenditure + Other_expenditure
Instruction_expenditure	Total current spending for instruction
Support_services_expenditure	Total current spending for support services
Capital _outlay_expenditure	Total capital outlay expenditure
Other_expenditure	Total current spending for other elementary-secondary programs
GDP_per_capita	Gross Domestic Product per capita for each State and Year

Table 1. Dataset Variables

1.2 Planned Questions and User Type

Education plays an important role in the maintenance of a strong economy, and funding plays an even greater role in the maintenance of a good education. A study published in 2016 discovered that an extra 10% spending (for education) improved the potential salaries of pupils by 7% and lowered their probability of ending up in poverty (The Economist, 2017). Thus, this project will focus on the levels of funding afforded to educational institutions by state sources. The project will be aimed at state officials looking to contrast their levels of funding with other states and GDP per capita, it will also help them determine whether their level of funding was justified. The initial questions proposed during the presentation are as follows:

- *How much have State Sources provided per pupil over time across States?*
- *How much did schools spend per pupil over time across States?*

After the feedback that was received, relevant changes were made to the questions. The two questions that this report will be examining are as follows:

- How has the relationship between funding for education from state sources and GDP per capita changed over time?
- Have state sources provided a justifiable amount of funding towards education over the years?

Section 2
DESIGN

2.1 Initial Paper Prototype Design

The initial paper prototype design (Figure 1) that was proposed during the presentation was met with confusion. The first problem was the confusing titles, the titles were written from the perspective of the school. For instance, state revenue instead of state funding, this posed a problem since the target for the visualisation are state officials. Another issue was the similarity between the graphs, two maps were used alongside each other making the visualisation seem mirrored and difficult to understand. Moreover, the visualisation did not seem to convey enough useful information for the target audience. For example, the fourth graph, the bottom right of Figure 1, doesn't display necessary information for state officials.

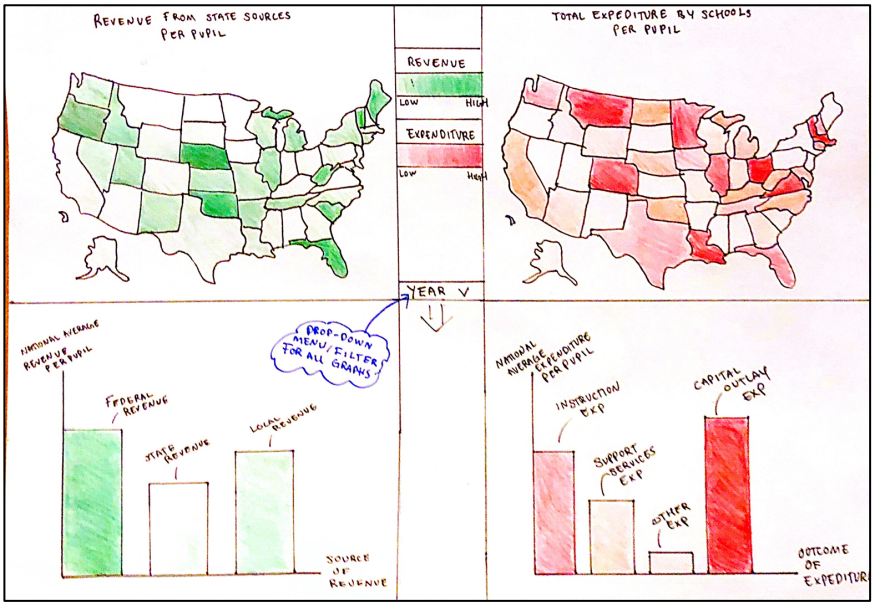


Figure 1. Initial Paper Prototype Design.

2.2 Final Paper Prototype Design

Due to the feedback received during the presentation, appropriate changes were made to the proposed questions and design.

The first graph, top left of Figure 2, is very similar to the one in the original design. It is a map demonstrating the average funding for education from state sources, the shade of the state corresponds to the value, the higher the value the darker the shade. An empirical study has revealed that in the majority of cases, map readers do presume that "dark means more and light means less" (McGranaghan, 1993). A colour legend will be placed alongside the map in the final visualisation. The map will also act as a universal filter, clicking on a state will filter all the other graphs to act in

terms of the chosen state. This will allow state officials to view the necessary information for their specific region.

The second graph, top right of Figure 2, is a scatter plot with a trend line. It depicts the relationship between GDP per capita and funding for education from state sources over time. By applying Gestalt's Law of Similarity, using a different colour and shape, the user can easily differentiate between the two variables. Also, Gestalt's law of continuity is taken into consideration through the addition of a trend line

The third graph, bottom left of Figure 2, illustrates the proportion of total school expenditure that us funded from state sources. In this graph, Gestalt's Law of Similarity is also used, different colours and shapes were used. The principle of similarity, states that elements tend to be combined if they are alike. This principle can be used with the law of proximity. Though, these two principles can also be contrasted individually to examine their combined effects on perceived grouping (Wertheimer, 1938).

The fourth graph, bottom right of Figure 2, portrays the z-scores for the funding for education from state sources over time. The orange bars represent negative z-scores, and the blue bars represent positive z-scores. Gestalt Law of Symmetry is applied here, where bars on the left are negative and bars on the right are positive.

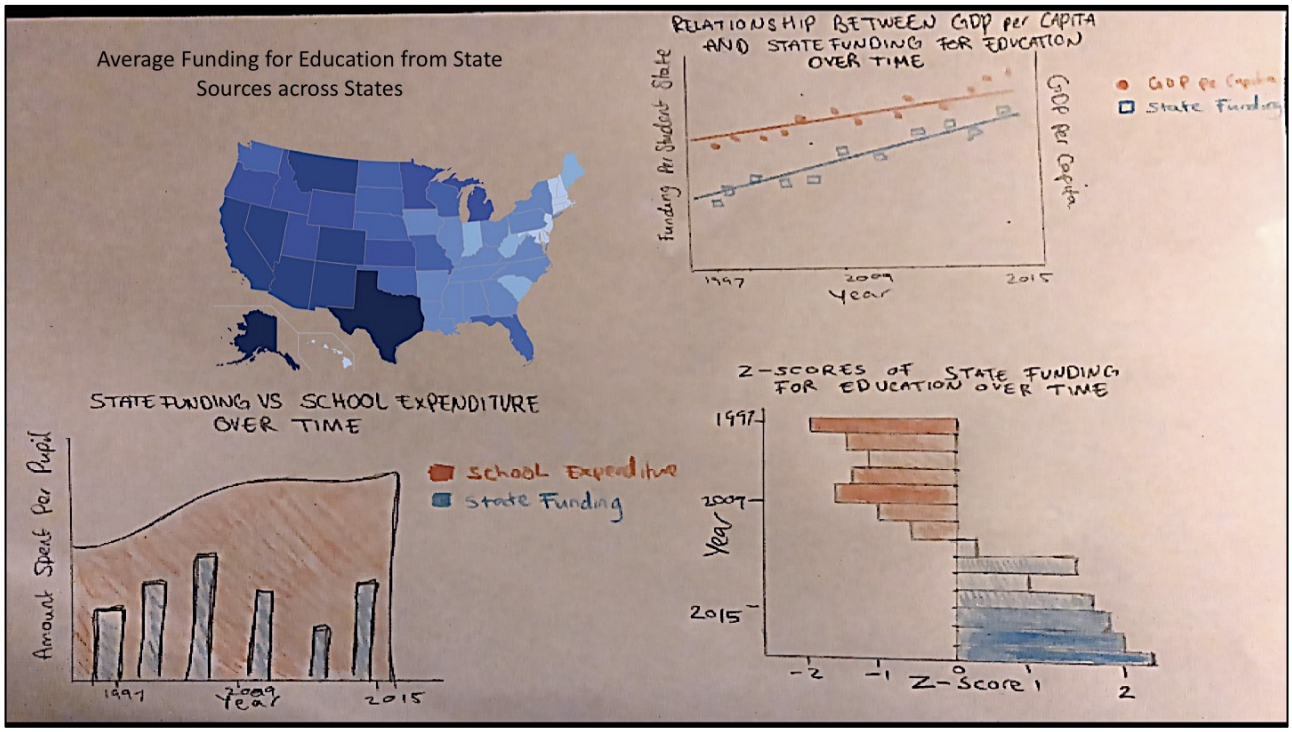


Figure 2. Final Paper Prototype Design.

Section 3
IMPLEMENTATION

3.1 Data Preparation

In order to adequately prepare the data missing values needed to be dealt with. To deal with this, R-studio was used to replace any missing values with the mean value (Figure 3). The next step was to remove Alaska and Hawaii, since trying to map them in Tableau without distorting the visualisation proved to be problematic. The years 1992-1997 were also removed, since the GDP dataset did not correspond to those years, (GitHub, 2012). The following step was to join the GDP dataset and a state abbreviation table to the main dataset. This was implemented using Tableau by conducting a left join to the data source (Figure 4). Subsequently, some variable names were adapted to suit the user type better, they were also divided by the number of students (Enroll) to provide a more accurate representation, this was done by creating a new calculated field in Tableau (Figure 5). The new variable names are as follows:

- State_Revenue was changed to Funding for Education per Student from State Sources
- Federal_Revenue was changed to Federal Funding for Education per Student
- Local_Revenue was changed to Local Funding for Education per Student
- Total_Expenditure was changed to School Expenditure per Student

```
# Taking care of missing data
dataset$TotalRevenue = ifelse(is.na(dataset$TotalRevenue),
                    ave(dataset$TotalRevenue, FUN = function(x) mean(x, na.rm = TRUE)),
                    dataset$TotalRevenue)
```

Figure 3. Missing Values, similar script was used for other variables.

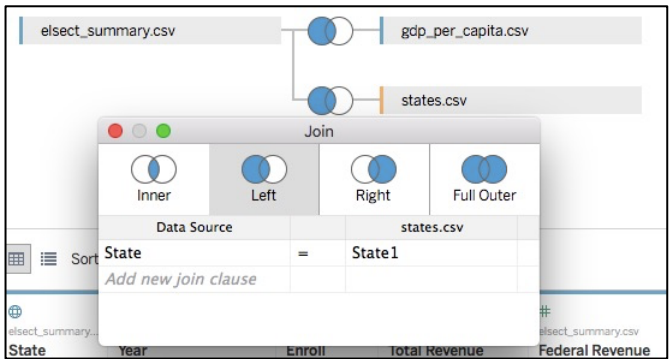


Figure 4. Joining using Tableau.

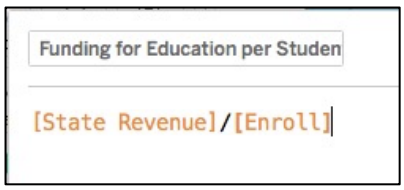


Figure 5. Creating a calculated field using Tableau, similar script was used for other variables.

3.2 Data Visualisation

Tableau was used to create the visualisations for this project. The paper prototype design previewed a prospective dashboard consisting of four visualisations, this section will detail how each one of those visualisations was created using Tableau.

3.2.1 Visualisation 1 – State Map

The first visualisation was to create a map demonstrating the average funding for education per student from state sources. The following steps were undertaken to create the visualisation:

1. Drag the State dimension to the sheet, this creates a map outlining the states
2. Drag the Funding for Education per Student from State Sources onto colour. Then change the colour to blue Teal. Click the dropdown menu on the mark just created and change the measure to Average.
3. Open up a new sheet, a right click the measures field to create a calculated field, call it Ln Funding for Education per Student from State Sources and type LN(SUM([Funding for Education per Student from State Sources])) into the calculation field and click apply. What this does is it takes the logarithm of the variable, this results in a more contrasted and saturated map so it becomes easier to spot the difference in funding between the states.
4. Repeat Step 2 in the new sheet but instead of dragging Funding for Education per Student from State Sources onto colour, drag the calculated field just created. Figure 6 shows the difference in contrast and saturation between the maps.
5. Now the problem with the new map is that it shows the logarithmic values in the tooltip and on the colour legend. To solve the tooltip problem, click on the dropdown menu for the mark and uncheck include in tooltip. Then drag Funding for Education per Student from State Sources onto tooltip. Afterwards drag the Abbreviation dimension onto label. The colour legend problem will be covered later.
6. Let's clean the map, open the map tab at the top and click map layers. Uncheck all the map layers.
7. Now finally the colour legend, first create a dashboard. Now drag the original sheet you created in step 2 onto the dashboard. And click layout on the left-hand side, check floating. Now change the size to 1x1, as you can see the map pretty much disappears but the colour legend remains. Now drag the second sheet onto the dashboard and remove its colour legend. Voila, it may seem like a long process but it makes a lot of difference to the overall appearance of the map. The final map is shown in Figure 7.

Figure 6. Original map (left), Logarithmic map (right).

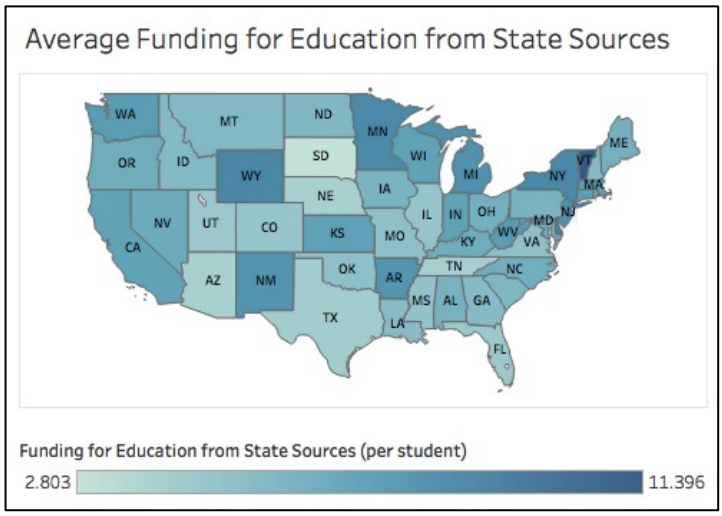


Figure 7. Final map visualisation.

3.2.2 Visualisation 2 – GDP and State Funding

The second visualisation was to create a scatter plot with a trend line. It depicts the relationship between GDP per capita and funding for education from state sources over time. The following steps were undertaken to create the visualisation:

1. Create a new sheet. Click the dropdown menu for the dimension Year and change the data type to date and time. Then drag Year to the column field.
2. Drag funding for education from state sources and GDP per Capita to the row field. This will create two plots stacked on top of each other. Using the dropdown menu for each variable on the row dimension, change the measure to average.
3. To merge the two plots into one, right click y-axis of the bottom plot and click dual axis. We won't synchronise the axis since it will be difficult to identify the relationship between the variables.
4. To change the graph from a line graph to a scatter plot, navigate to the marks dimension for all and click the dropdown menu labelled Automatic. Now choose shape.
5. Now go to marks menu for funding for education from state sources and GDP per Capita and change the shape to diamond and circle respectively.

6. Let's add a trend line and forecast, although a forecast was not proposed in the paper prototype design it seemed appropriate to add. Click the analytics tab on the left-hand side. Drag a linear trend line to the graph and do the same for forecast. Remove forecast indicator from the marks tab.
7. Right click each trend line and click format and change the trend line to dashed.
8. The final step is to change the colour, in the paper prototype design I used blue and orange, however to avoid confusion the orange was changed to green. So, choose green for the GDP and blue for State Funding. The final visualisation is shown in Figure 8.

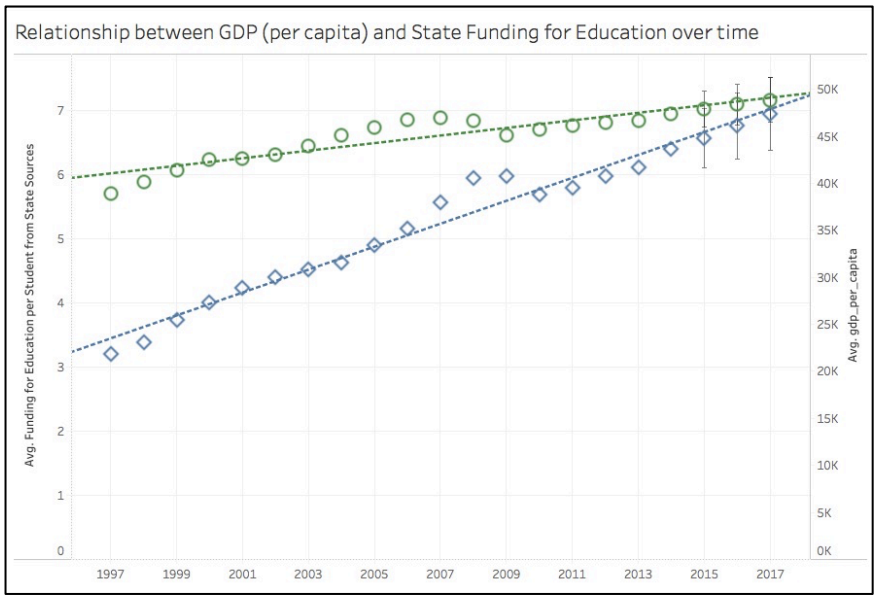


Figure 8. Final GDP and State Funding visualisation.

3.2.3 Visualisation 3 – School Expenditure and State Funding

The third visualisation was to create a graph depicting the proportion of total school expenditure that is funded from state sources. The following steps were undertaken to create the visualisation:

1. Create a new sheet. Drag Year to the column field, then drag School Expenditure per Student and funding for education from state sources to the rows field. Using the dropdown menu for each variable on the row dimension, change the measure to average.
2. To merge the two plots into one, right click y-axis of the bottom plot and click dual axis and also click synchronise axis.
3. Change the mark for School Expenditure per Student to shape then colour it orange and change the mark for funding for education from state sources to bars then colour it blue and add a black border. Then whilst pressing ctrl drag funding for education

from state sources from rows onto colour then change the size to manual and make it largest.

4. Now we want to label the bars with a percentage to show the proportion of total school expenditure that is funded from state sources. Create a calculated field called Percentage of School expenditure funded by the State and type [Funding for Education per Student from State Sources]/[School Expenditure per Student] into the calculation field. Now drag the new measure onto label under the funding mark.
5. Click the dropdown menu for the label and change the measure to average then open the dropdown again and click format. Change numbers to percentage with no decimals. You're done, the final visualisation can be seen in Figure 9.

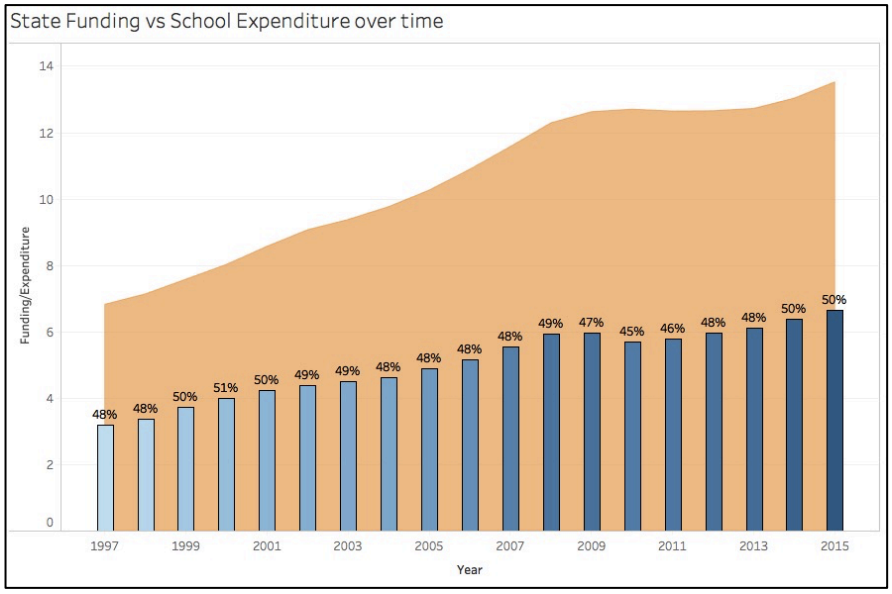


Figure 9. Final School Expenditure and State Funding visualisation.

3.2.4 Visualisation 4 – Z-Scores

The fourth visualisation was to create a graph portraying the z-scores for the funding for education from state sources over time. The following steps were undertaken to create the visualisation:

1. Create a new sheet, then create a new calculated field titled Ave-Z-score. Type WINDOW_AVG(SUM([Funding for Education per Student from State Sources])) into the calculated field.
2. Create another calculated field titled Sd-Z-score and type WINDOW_STDEVP(SUM([Funding for Education per Student from State Sources])) into the calculated field.
3. Create a final calculated field titled Z-scores and type (SUM([Funding for Education per Student from State Sources]) - [Ave-Z-score]) / [Sd-Z-score] into the calculated field.

4. Drag Z-scores into columns and Year into rows, then select compute using years for the Z-scores using the dropdown menu.
5. Change the mark to bar, then use ctrl to drag Z-scores from columns onto colour and then drag it to label. Change the colour to red-blue. The final visualisation can be seen in Figure 10.

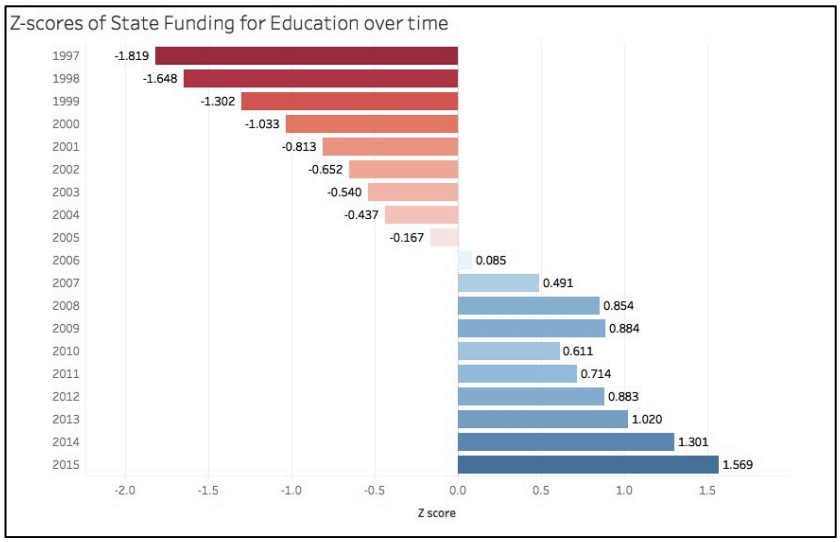


Figure 10. Final Z-Scores visualisation.

3.2.5 Dashboard

The final part of the implementation involved combining the visualisations to create a dashboard. The first visualisation was added to the dashboard in 3.2.1, the remaining visualisations were added to the dashboard in the same order as demonstrated in the paper prototype design. The final step was to make the first visualisation act as filter by clicking the icon on the sheet. The final dashboard can be seen in Figure 11.

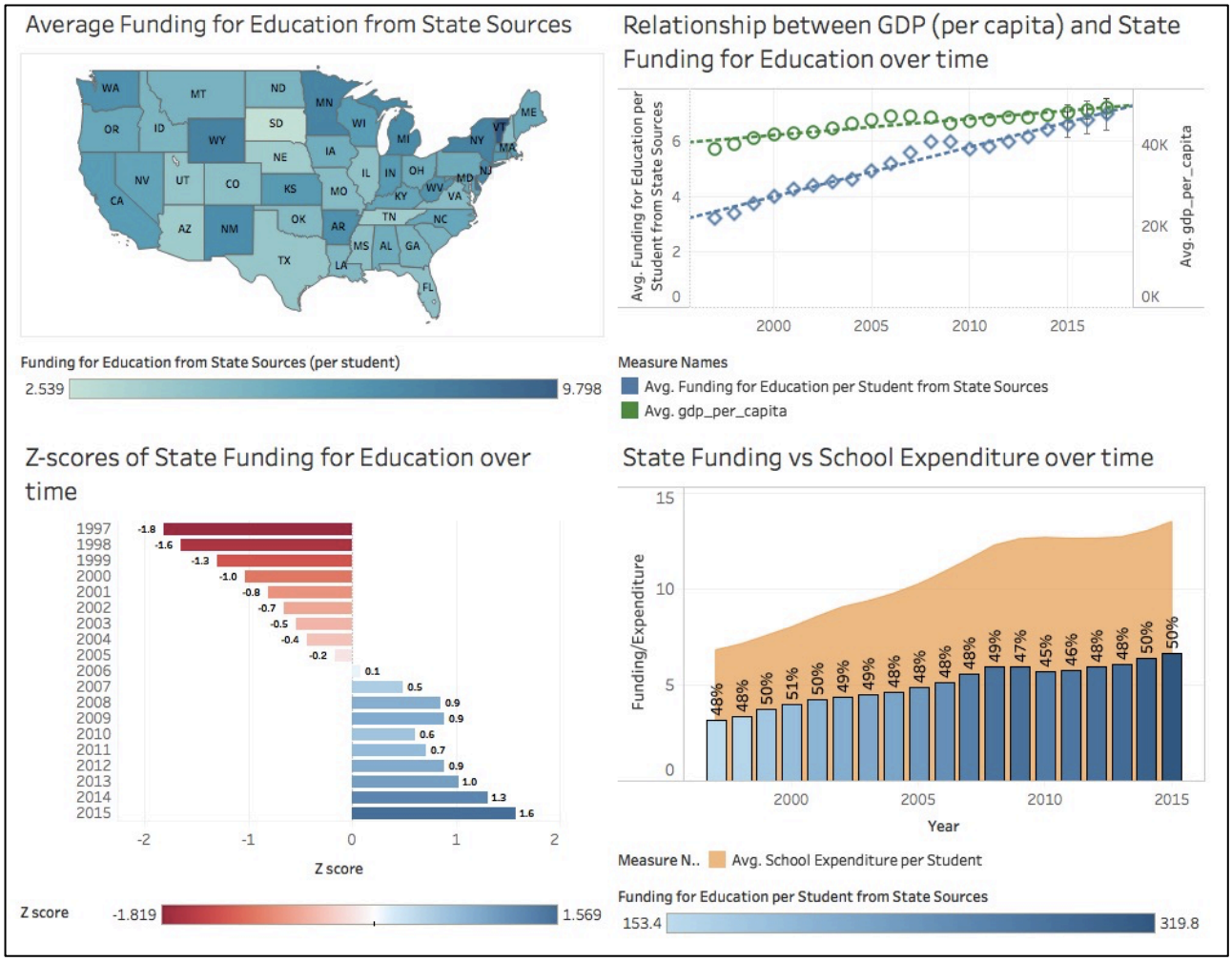


Figure 11. Final Dashboard.

Section 4
WALKTHROUGH

This paper will demonstrate how, using the visualisations, the two proposed questions have been answered. Although the model is designed for all State Officials to use, it will be difficult to answer the questions for officials from all the states. Thus, the walkthrough will focus on a State Official for the state of New York.

4.1 Question 1

The first question is "How has the relationship between funding for education from state sources and GDP per capita changed over time?". This question is easily answered using the implementation, by clicking the state of New York on the map, the scatter plot showing the relationship between GDP and state funding adjusts to show the values specific to the state (Figure 12). From Figure 12, we can see that there's a positive linear correlation between GDP and State funding for education, as funding increases over time so does the GDP per capita in the state of New York. This is positive information for state officials since it shows spending more on education has a positive effect on the economy. Also, from the graph we can see a forecast for the future indicating that GDP and funding is likely to increase if the current trend continues, this is a discovery which was facilitated by visualisation.

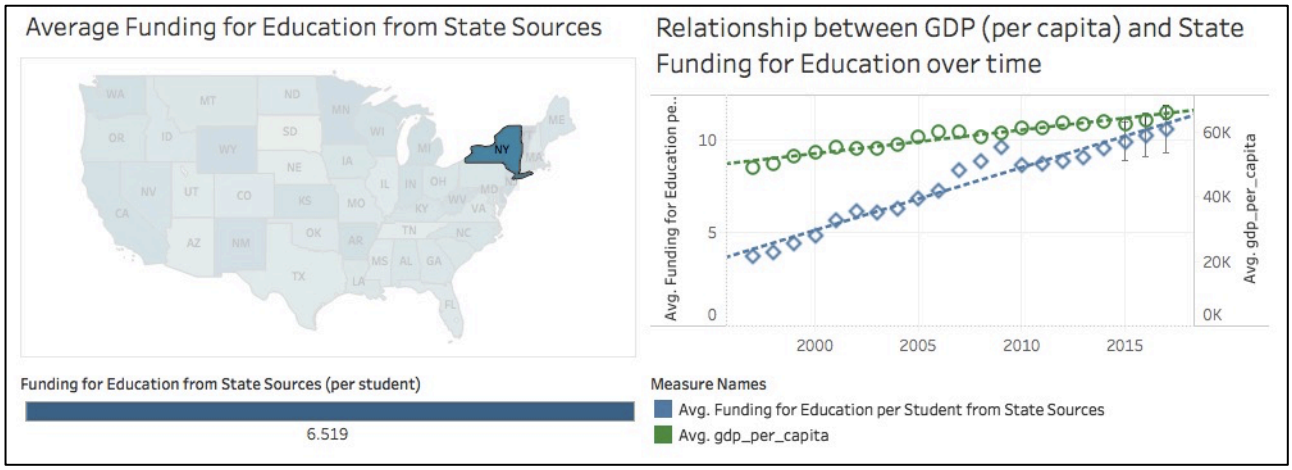


Figure 12. Question 1.

4.2 Question 2

The second question is "Have state sources provided a justifiable amount of funding towards education over the years?". This is a much more difficult question to answer, since it's somewhat subjective. However, using the visualisation it becomes a lot less subjective and more data driven. The question makes use of all the graphs to solve the problem. Drawing from the first question, we can determine that the funding is justified from an economic perspective.

From Figure 13, we can see how the Z-Scores of state funding for education have changed over time. The z-score of a data point represents number of standard deviations that it is greater than or less than the population mean. As a general rule, z-scores are deemed unusual if they are less than -1.96 or greater than 1.96. In this case the state of New York did not spend significantly higher or lower than their average spends on education throughout the years. From the graph we can determine that their level of funding has been pretty much consistent from the years 1997 to 2015.

Additionally, Figure 13, we can see the proportion of total school expenditure that was funded by the state of New York. From the graph we can see that over the years the state has gradually increased the amount it funds per student and schools have also gradually spent more, so in that respect the level of funding by the state is justified. We can also see that the lowest proportion of total school expenditure that was funded by the state was 37% in 1997 and the highest was 45% in 2007 and 2009, and that the proportion stayed pretty much consistent throughout. Since the schools have received a significant amount compared with their spending and that it's been consistent, the proportion of funding is also justified.

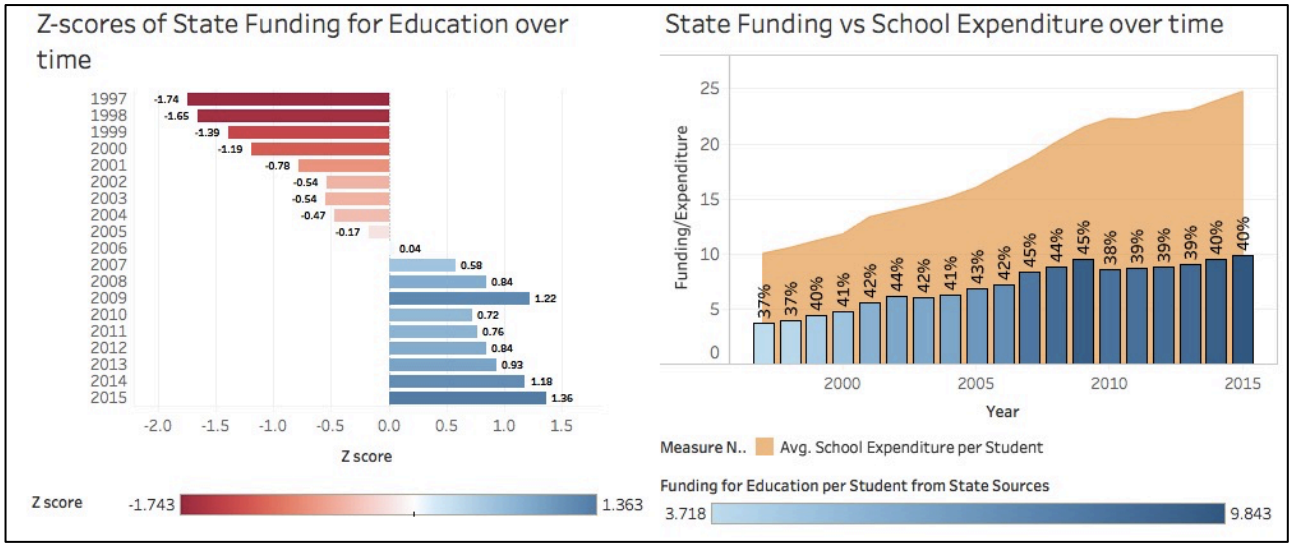


Figure 13. Question 2, filtered for NY.

Section 5
EVALUATION

5.1 Effectiveness of the project

The questions have been answered reasonably well using the visualisations, as demonstrated in section 4. The use of Gestalt theories to aid in the design have significantly improved the overall effectiveness of the visualisations. However, there is room for improvement. Take for instance the map in Figure 7, it's rather difficult to compare the states, it may look pretty but many of the colours are very similar. It takes a much closer look to identify the highest and lowest, even with the application of logarithmic values. Some researchers have also gone on to say that graphic design for data analysis and presentation is very much unscientific, (Cox 1978).

Another potential problem lies with the use of Z scores, although effective in describing the deviation, this may be a little too statistical for the layman.

Also, looking back on figure 8, it may not have been the best choice to use shapes where a simple line graph would have sufficed, "chart junk" as Tufte would say, (Tufte, 1983). A line graph would have described the connectivity better especially since a forecast was added.

5.2 Effectiveness of the Tableau and looking forward

Tableau, is not only a great tool for producing beautiful visualisations, it's remarkably easy to use. The drag and drop method is refreshing compared to the tedious coding involved with programs such as python and R. However, like all tools it does come with its various drawbacks. For instance, if one wanted to do a linear regression they would have to connect to R using rserve, and even then, coding the regression is a little different to the way you would do it with R. It seems when you want to perform some little more advanced statistical concepts, using Tableau becomes a little tiresome. The one major problem I faced with Tableau was regarding mapping Alaska and Hawaii, it proved to be the bane of my existence. If I chose to include them I had one of two choices, all the other states will appear incredibly small, or I could plot them separately on different sheets and have them float on the dashboard. Both seemed unappealing, so they were withdrawn from the data altogether.

This module has helped prepare me for a career in data science, learning to use tools such as Tableau and JMP has very much broadened my skillset. the visual theory learnt during this module, however, proves to be the most important. One can know how to use every tool on the market but without the theory and understanding of what makes a good visualisation, the potentially useful tool becomes somewhat "useless".

Looking forward, I would like to develop my understanding of visual theory further rather than learn new tools. The concepts I have learnt so far have been eye-opening, and I imagine I have only touched on what is possible.

References

- Cox D R 1978 Some remarks on the role in statistics of graphical methods. Applied Statistics 27: 9
- GitHub. (2012). jasonong/List-of-US-States. [online] Available at: https://github.com/jasonong/List-of-US-States/blob/master/states.csv.
- Kaggle.com. (2017). U.S. Educational Finances | Kaggle. [online] Available at: https://www.kaggle.com/noriuk/us-educational-finances/kernels.
- McGranaghan M 1993 A cartographic view of data quality. *Cartographica* 30: 8–19
- The Economist. (2017). America's school funding is more progressive than many assume. [online] Available at: https://www.economist.com/news/united-states/21732817-how-states-and-federal-government-offset-effects-local-inequality-americas.
- Tufte, Edward R. (1983). The Visual Display of Quantitative Information. Cheshire, CT
- Wertheimer, M. "Laws of Organization in Perceptual Forms", 1938.