



## CS5803 - Data Visualisation

Coursework for 2020/21

### **Visualizing School Data of A level Students**

Md Abrar Fahim Jaki

Student ID 2027461

# 1. Introduction

We will try to inspect UK school data to generate some simple knowledge to answer some questions. The data is collected from the UK government [website](#) [1]. The dataset contains information about school's general information, positions in different index, location, students' grade, etc.

## 1.1 Data processing

For convenience, we only took data for 2019. We convert the postcode to latitude and longitude data with the help of a secondary dataset. We updated the schools' names to the latest name as the schools' names changed several times. We also updated the school's status and other information.

The school that provides A level studies is kept only and all the null points are removed. The percentage values for student's count converted to actual values.

The dataset was human-friendly but not friendly to Tableau plotting. So, for fulfilling the requirement all data goes through a transformation, original features and values are kept intact. All the preprocessing is uploaded on the remote server [GitHub Link](#).

## 1.2 Meta Data

| Column name       | Description  | Data Type   |
|-------------------|--|-------------|
| LAESTAB           | School unique identifier based on LA-Estab source AO | String      |
| SCHNAME.x         | Name of the school                                   | String      |
| LANAME            | Local authority name                                 | String      |
| TOWN              | School Town  | String      |
| REGION_NAME_1     | Region name  | String      |
| Latitude          | Latitude   | Float       |
| Longitude         | Longitude  | Float       |
| CS                | Number of Student in Computer Science                | Integer     |
| Biology           | Number of Student in Biology                         | Integer     |
| Chemistry         | Number of Student in Chemistry                       | Integer     |
| Math              | Number of Student in Math                            | Integer     |
| Further Math      | Number of Student in further Math                    | Integer     |
| Physics           | Number of Student in Physics                         | Integer     |
| Subject by gender | Student gender                                       | String      |
| Student_language  | Number of students in different language category    | Integer     |
| Language_category | Students' language category name                     | String      |
| Ofsted Grade      | Grades assigned to schools                           | categorical |

## 1.3 User type specification

The user is a Journalist, wants to make an article for A level study system in the United Kingdom. He wants to get an idea of what features he should focus more on in his article. He also wants to get some nice visualization to support his argument.

## 1.4 Questions

The user has the following questions:

1. At A level, which subjects are taken by most of the students, and is there any preference for choosing subjects based on gender?
2. The schools where English is not the first Language, are these schools located in a specific region (as they are clustered as different communities), and is language impact their results (e.g. grades)?

## 1.5 Requirement Specification

### Plots

To answer Q1 we need the columns that contain information about the number of students in different A-level courses. We also need the *subjects by gender* column.

**Plot 1:** A bar graph will be plotted. And The gender columns will be used as a filter.

To answer Q2 we need the *Students language, language category, latitude, longitude, and Ofsted grade* columns.

**Plot 2:** A map will be plotted based on the student's language column using latitude and longitude. The size of the pointers will be proportionate to the number of students so that the user can identify which region has what number of students.

To see students' schools grade based on students' language we need *Ofsted grade, Students Language and Language category* columns.

**Plot 3:** A boxplot will be plotted to see the students' grade distribution.

**For generic filtering (based on region), we need REGION\_NAME, TOWN, LANAME, LAESTAB columns.**

## Statistical analysis

We will try to cluster the school data based on numerical column to see if they are related and are fall into the same group based on similarity.

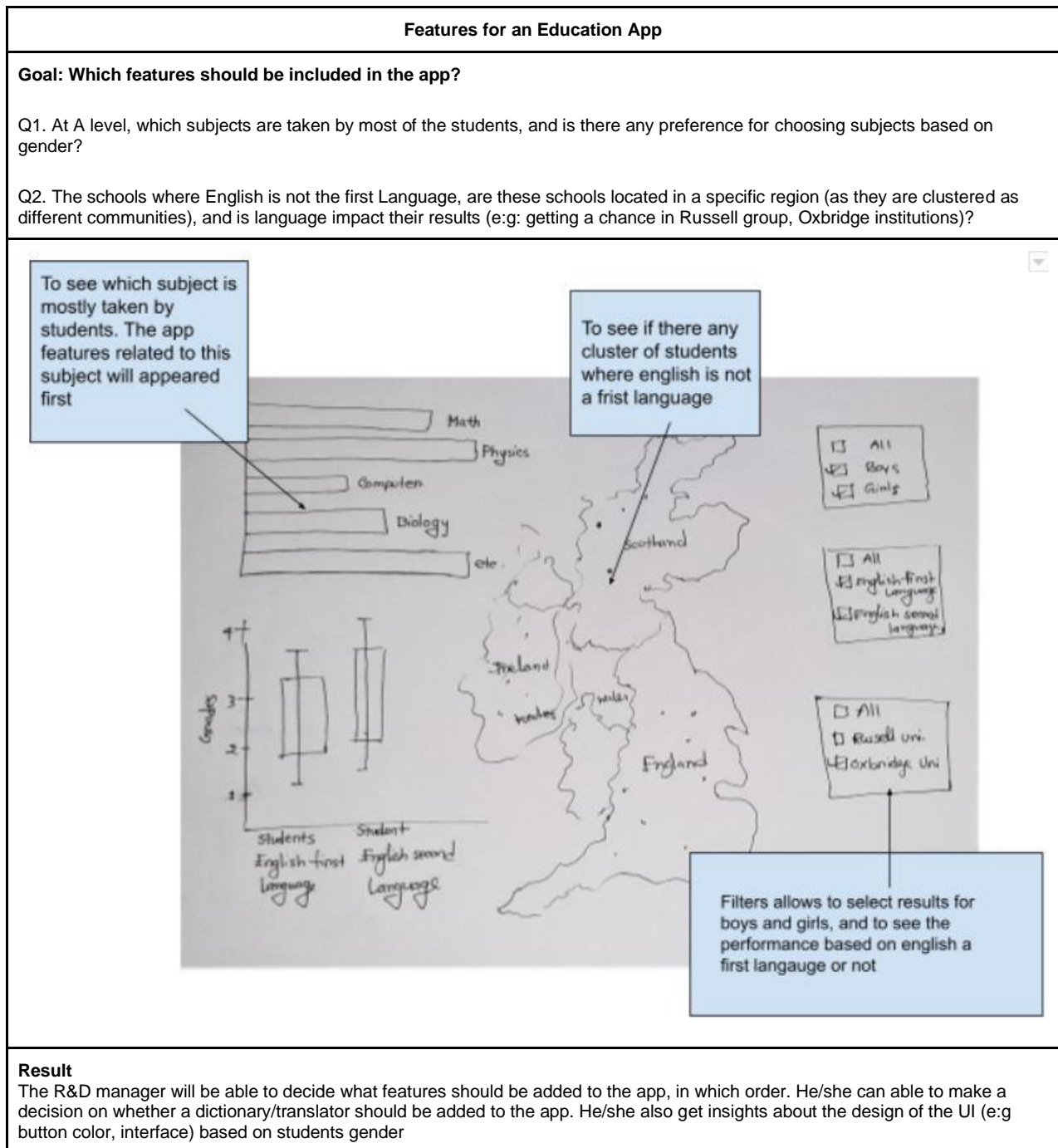
## 2. Design

From our proposed design we made a little change. The changes are based on feedback.

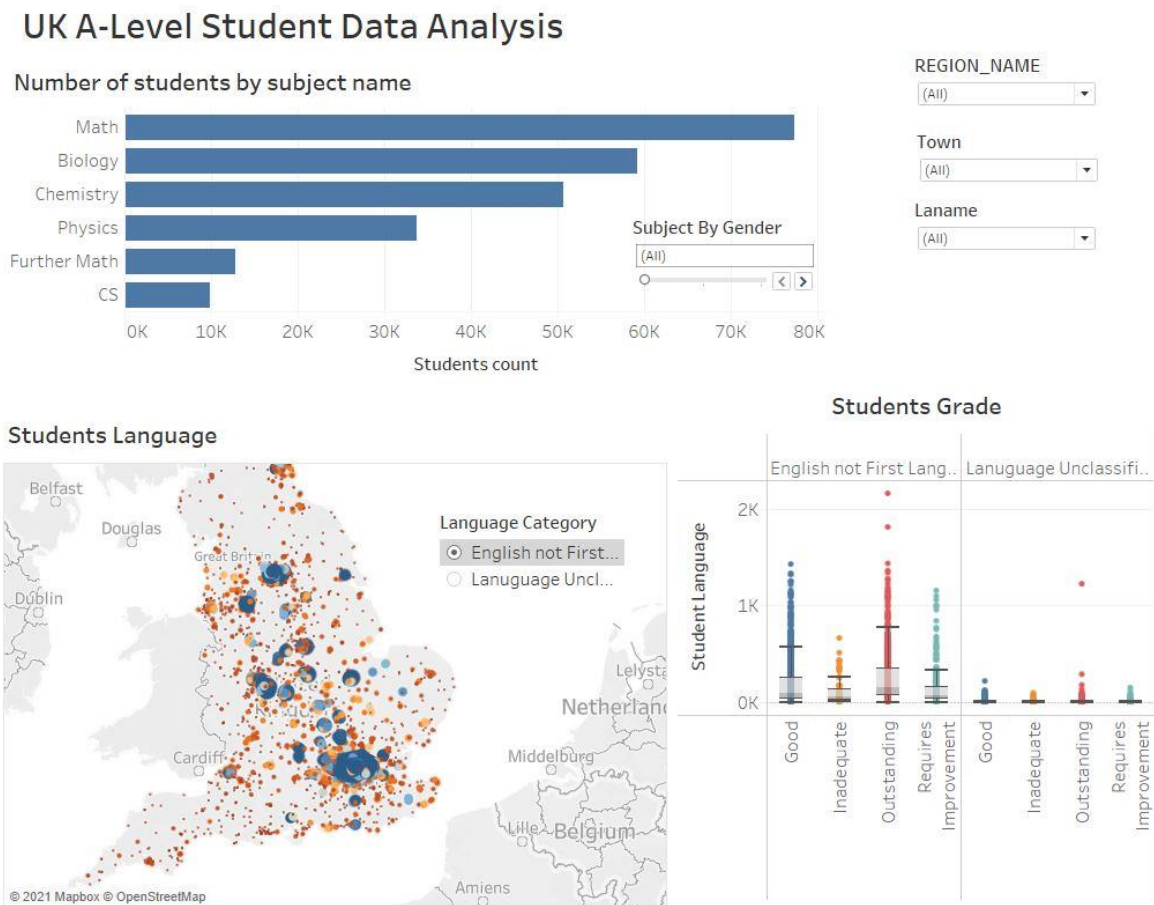
### 2.1 Changes

1. The number of students who get admitted to Russell group university and the Oxbridge university is deleted, as it is a little bit far from the question the persona may ask.
2. The filters are now connected to all the sheets of the Dashboard.
3. Brushing and linking are added (suggested in proposal feedback). The user can filter the location both manually (selecting the town, local authority name, etc) or can select using brushing from the map and the changes will be reflected in all the figures of the dashboard.
4. The persona role changed from Research and development manager to product Journalist. So that the questions become more relevant to the persona (suggested in proposal feedback)

## 2.2 Paper Landscape from proposal



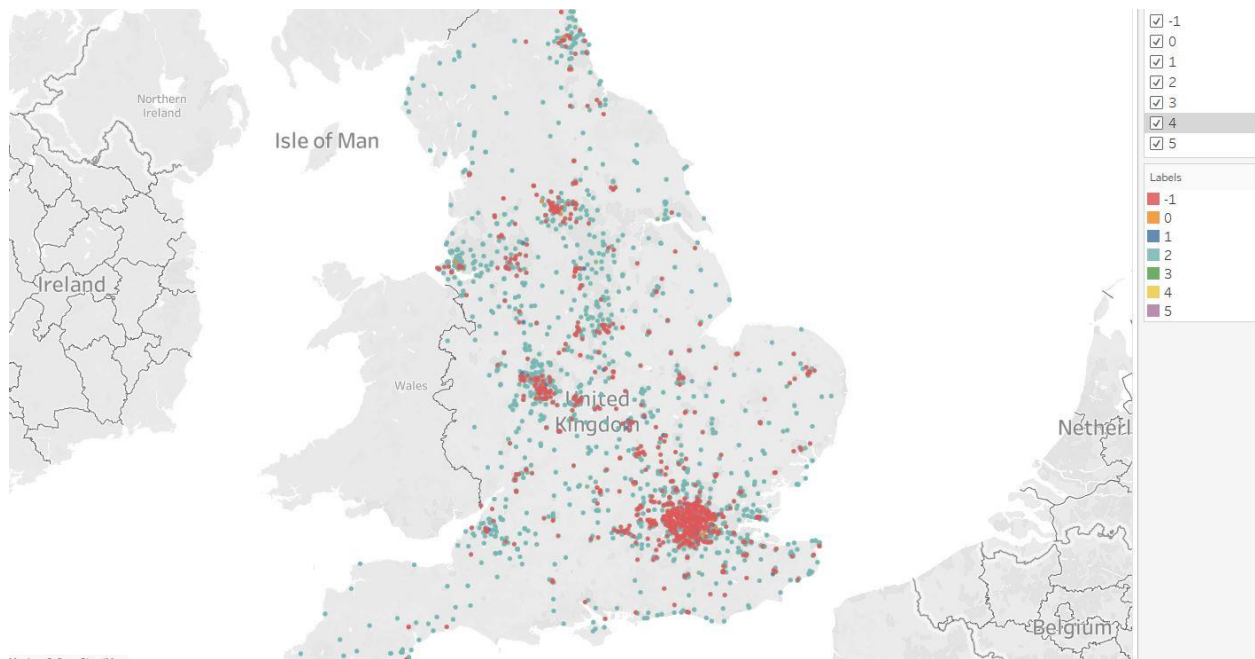
## 2.3 Screenshot of Final Dashboard



## 3. Implementation

### 3.1 Exploratory data analysis

We will try to do clustering on school data. We choose DBSCAN for clustering as it is very good at not including outliers in clusters. As we have scattered data and we want to know only what schools are very closely related to others. For DBSCAN we only selected the numerical columns and do one-hot encoding for Ofsted grade(categorical). DBSCAN assigns a label to each cluster and -1 to outliers. The python notebook for DBSCAN is uploaded on this [GitHub link](#). After that, we visualize our schools on Tableau Map,



The red dots here are detected as outliers, they are somehow different from the rest of the data. So, all the outliers are looking like in a single cluster. And it looks like all the outliers are in major cities. The evenly scattered green dots are in other clusters.

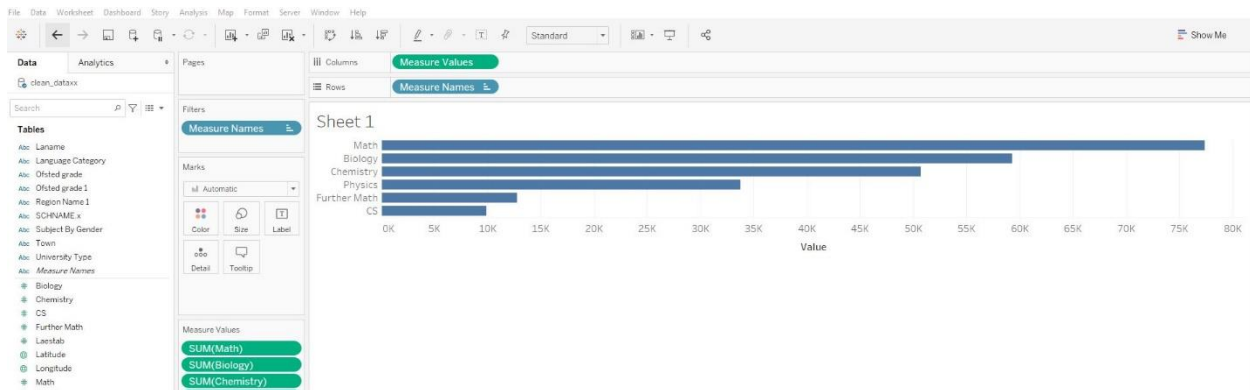
There are some other clusters too, but these clusters are not prominent.


### 3.2 Tableau Implementation

First, we imported the dataset from our local machine to Tableau software. While developing the visualization we will follow the classic guideline from the 1985 paper that is “create the simplest graph that conveys the information you want to convey” [2]. We will also keep in mind Richard W. Hamming's quote: "The purpose of computing is insight, not numbers."

#### 3.2.1 Subject by Gender (Bar Graph)

We directly go to Sheet1. Here we can see all the column names of our dataset are in the left portion of the screen. We can see Tableau detect all the numeric values containing columns as measure values and color-coded as green. We drag all the subjects (e.g: Math, Physics, etc) to the Rows section of the sheet. We can see that Tableau calculated the aggregated sum automatically. If not, we will click the **Analysis** option from the top navigation bar and will tick the Aggregate Measures option. Then from the Right top of the screen, we click the “**Show Me**” button and select the Horizontal bar option. We will see that Tableau Assign all the rows into a new Value named “Measure Values” And crates a filter named “Measure Names”.



If we look at the second row of Tableau’s top navigation bar, we can sort  the bars ascending/descending order. Then we will drag “REGION”, “LANAME”, “TOWN” column to the Filter option. We select all the values for filters and show the filter.

As there are a lot of options in the new filters’ options showed on display, we will select a single column dropdown to makes the dropdown looks better. Before and after view of selecting single column dropdown is given below,



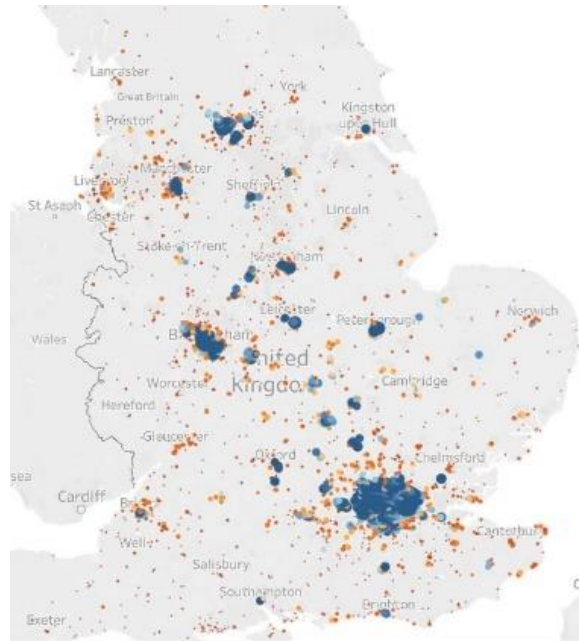
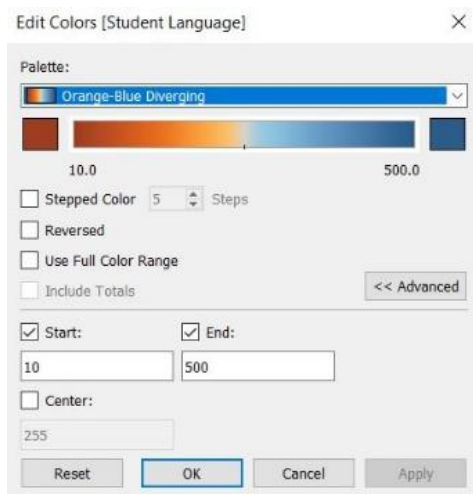
If we look at the filters, we can see that they follow a hierarchy. First, Region then Town and LANAME. If we select a Region, Towns that fall in that specific region should be shown only. Same things apply for LANAME. But here we can see all other Towns and LANAME from different Region also shown. To deal with this we will select “**Only relevant values**” option from the dropdown from each filter card to show only the valid value for a specific selection.

After that, we will drag the “Subject by Gender” column to Filters.

### 3.2.2 Students Language (Map)

In our second sheet, we will drag latitude and longitude data to the Rows and Columns section. Tableau will generate a map for us for each school. We will then drag the “Students Language” column to the Size option from the Marks section. We will see that the markers/dots on the map are of different sizes proportionate to the number of students. Then we will drag the “Language Category” column to filters. After showing the filter, from that card We then go to **customize option** and deselect the “**All**” option from our filter as it does not make sense in this context. We also select a Single value list from the options so that only a single option from the filter can be selected.

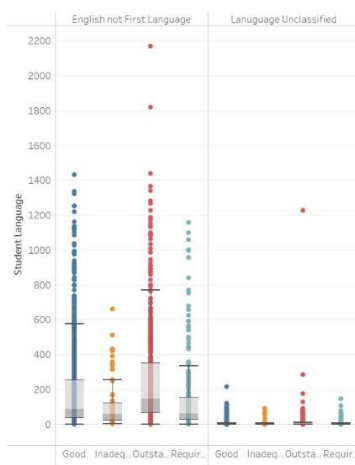
We will also drag the student language column on color so that based on students’ numbers we can show the dots of different colors. We customize the colors in Orange-blue-Divergence and select start and End value from the **advanced** option. We range the value between 10-500 after seeing different combinations of ranges.



Now we can easily see, where the biggest cluster of students situated. From the top navigation bar of the Tableau software, we selected the **Map > Background Maps** to light. From the Map layers option, we Washout the map a little bit so that the background of the map goes a little bit out of focus. From selecting Orange-blue-divergence to Washout map a little bit, we followed the principle of color section techniques, based on color distance, linear separation, and color category [3]. We tried different sets of colors and sizes for scattered points on the map so that the schools that have a larger number of non-English speaking students can be separable from the schools that have little or no non-English speaking students.

### 3.2.3 Boxplot

For the boxplot, we drag the Students Language column in the Row field and disaggregate it so that Tableau does not sum up the values. Then we click on the boxplot option from the **Show me** button. After that, we drag "language category" and "Ofsted grade" columns to the Column option of Tableau. Now we can see the boxplot divided in more categories. We can see there is a boxplot created named **Null**. This one because of the schools where Ofsted grade is not assigned. We will filter the null values by dragging the Ofsted grade on the filter and deselect the Null option.





### 3.2.4 Dashboard

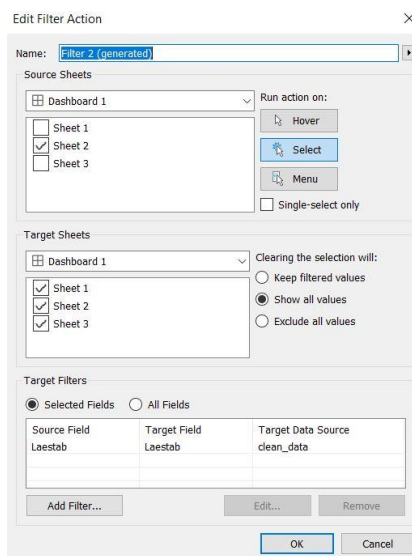
For designing Dashboard, we tried our best to show the useful information. Stephen Few [4] in his article argues that although visually appealing, a dashboard can be truly useless as it cannot give the actual information. We choose priority of views from left to right. First, we will add the bar plot, as this plot capable of showing most of the information and easy to understand. Then the filters will be added as filters are mandatory for selecting specific things. The map is added on bottom left, and boxplot at the end. Boxplot is not prioritized as it requires a little statistical knowledge to understand boxplot.

We select dashboard organizing as Floating rather than tile from the bottom left options. We then drag sheets one after another and adjust. If there any resize issue, we resize in the main sheet, so the same change occurs in the dashboard. From the filters “Region” “LANAME” and “TOWN” dropdown arrow we selected these filters for “**Apply to worksheet > All using this data source**” so that these filters can be used to all figures of the dashboard. After that, we give a nice name to our Dashboard.

### 3.2.5 Brushing and Linking

We will use LAESTEB for the unique identifier of all our sheets. So, in sheet1, 2, and 3 we will drag LAESTAB to filter.

Now from the top navigation bar of Tableau Dashboard, we will select **Dashboard > Actions** (or Ctrl+Shift+D) to select the action filter (also known as brushing and linking). We will click on add action option. We will use sheet 2 as the source and all other sheets as target sheets. We added LAESTAB as a specific source and target field as this column is the unique identifier for all the datapoint of the dataset we are using. After adding the action filter, if we select some points on the map, the changes are reflected on all dashboard sheets.



We will also select Sheet3 (boxplot) as the source Sheet and Sheet 1 and 2 (barplot and map) as the Target sheet. So that if we make any change in sheet3 it will be reflected on sheets 1 and 2.

### 3.3 PowerBI Implementation

We also tried to generate the same dashboard with the help of PowerBI. We can see that the filter options in powerBI cannot be dragged on an individual plot. The plots in PowerBI are not as pretty as Tableau. In PowerBI we do not need to add the action (linking) manually. All the graphs are connected automatically and if one point of the graph is selected then the filter is automatically applied to the other graphs. Another advantage of PowerBI is

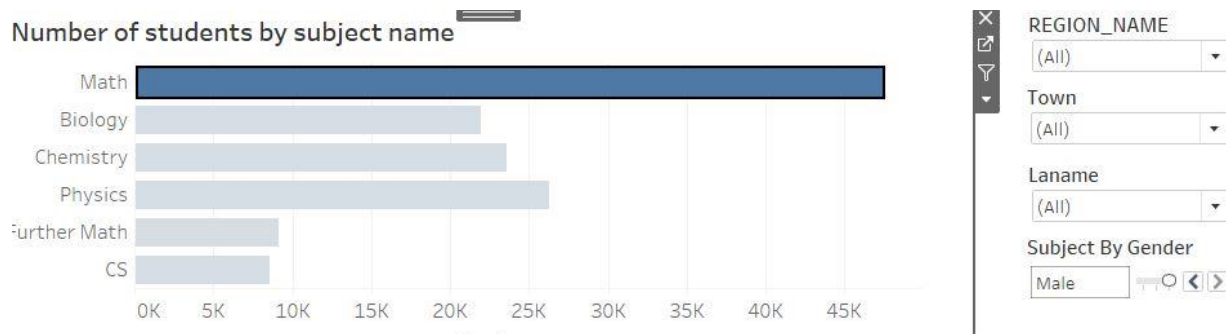
custom-made visualization from third-party software can be imported. So, It has more graph options as anyone can upload his/her visualization tool to PowerBI. E.g. Our boxplot is imported from a third-party software provider. The problem with PowerBI is, a single point of a map cannot be selected. PowerBI uses Bing Map for its map visualization. So, an internet connection is required for the map and it takes time to load Bing Map.



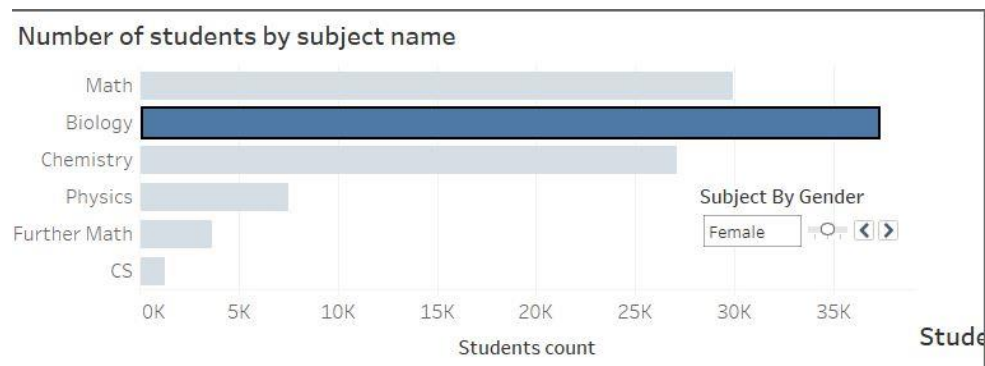
## 4. Evaluation

From the two tools (Tableau and PowerBI) we choose Tableau for the user. From the bar plot, the user can see the number of A-level students in a different subject in 2019. The user can filter an area based on region name, town, and local authority name. The number of boys and girls in different subjects also can be shown from another filter. So, based on the number of students taking a different subject, the user can decide how his article will be written. Suppose most of the students take mathematics, then the user can use the statement like, "Math is the most popular subjects in A Level studies in UK". He also can prioritize his focus on specific local area and compare the statistics. Using a gender filter, the user can see if there any preference for taking certain types of subjects based on the student's gender. Maaik van der Vleuten et al [5] found that gender ideology affects educational choices via competence beliefs, occupational values, and subject preferences. Ann colly et al also found English and Humanities are preferred by girls, and Physical studies and mathematics are preferred by boys [6]

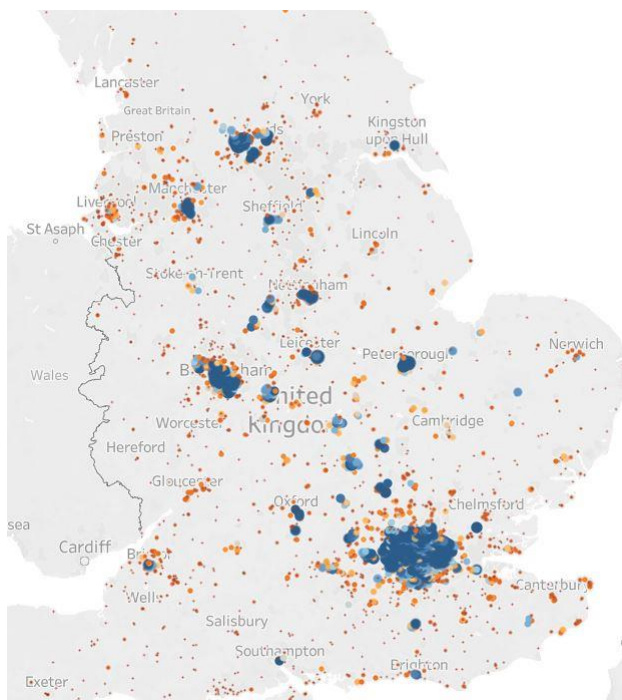
From our dashboard, we also can see that Math is the most popular subject for male,



And Biology is the most popular among girls,



So, the user can design his argument based on previous article and support the argument from the findings of the dashboard. The user also can see non-English-speaking students' numbers on the geographical location. Is that a random distribution or non-English speaking students live in a specific region? We expected it could be random. But we found that rather than scattered randomly. Non-English speakers live very close region, looks like different clusters. Most of them live in 4 or 5 specific cities.



This dashboard is also capable of showing if language affects the grade. We can see that the school has non-English-speaking students most of them are good and Outstanding on average. So, language is not that much barrier to achieve a good grade.

## 5. Reflective Discussion

The main challenge of the project was data cleaning. The data was human-friendly. But for plotting in Tableau, goes through a lot of transformation and took most of the time.

The overall Tableau experience is good. It is easy to use, like drag and drop. For map options, the single country map cannot be selected. The alternative way is, drag the city name on color and wash out the area. This method only works if it has all the data for all the locations, and in filled map. So, this seems a limitation.

The PowerBI is a good tool for visualization, but still has a lot of room for improvement. For linking it is perfect, but on the map, several points cannot be selected. It could improve its graphs to looks prettier.

From 2918 schools we derived simplified knowledge. Within only three plot we tried to generate a snapshot of A level study system of the United Kingdom. We tried our best so that our dashboard does not end up with a junk chart [7].

The persona can get insights from single clicks rather than looking at the whole dataset. We found that most diversity is in the big cities where non-English-speaking students are more, that makes perfect sense. And the DBSCAN clustering was also able to separate the Non-English-speaking community from others.

**Limitation:** In the dashboard while using boxplot as filter, the “subject by gender” filter of barplot should be in “ALL” selected. Otherwise (if male or female) selected the filtering will be misleading. As the gender column (male/female) is not related to the Ofsted grade column.

**Future learning objective:** Google AI experiment with high dimension data visualization [8] can visualize high dimension data along with animation. Learning this type of data visualization is the future goal. To achieve this, transformation (e:g PCA) for projecting high dimension data to low dimension should be learned. Then Using a 3D visualization tool will help to achieve the goal.

## References

1. gov.uk(2021) *Search for schools and colleges to compare*. Available at: <https://www.compare-school-performance.service.gov.uk/download-data> (Accessed: 1 March 2021)
2. Tufte, E.R., 1985. The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, 7(3), p.15.
3. Healey, C.G., 1996, October. Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization'96* (pp. 263-270). IEEE.
4. Few, S. and Edge, P., 2007. Dashboard confusion revisited. *Perceptual Edge*, pp.1-6.
5. van der Vleuten, M., Jaspers, E., Maas, I. and van der Lippe, T., 2016. Boys' and girls' educational choices in secondary education. The role of gender ideology. *Educational Studies*, 42(2), pp.181-200.
6. Colley, A., Comber, C. and Hargreaves, D.J., 1994. Gender effects in school subject preferences: A research note. *Educational Studies*, 20(1), pp.13-18
7. Kaiser Fung(2021) *Junk Charts* Available at: <https://junkcharts.typepad.com/> (Accessed: 20 April 2021)
8. AI Experiments(2021) *Embedding Projector*. Available at: <http://projector.tensorflow.org/> (Accessed: 1 March 2021)