COMSM0089 Introduction to Data Analytics Coursework

QJ22919/ 22334033

Task 3: Tableau Visualisation of WHO Malnutrition Data

Task 3: Tableau Visualisation of WHO Malnutrition Data

3.1

Following Muntzners three-part analysis framework of What? Why and How? I constructed a Tableau visualisation of the World Health Organisations (WHO) malnutrition data.
The '**What?**' referred to the WHO data that spanned a 30-year period and most of the developing countries. This data was in the form of two spreadsheets, the first with a focus on the malnutrition metrics split by age and the second split by wealth quintile.

It was obvious that the original formatting of the data would make it challenging to use so it was reshaped. By creating columns for age and wealth quintile on the appropriate spreadsheets the overall number of columns was reduced by around two thirds. This was a fairly involved process achieved using the Python Pandas package.
Initially Regular Expressions were used to identify columns headings which indicated relevant data and these columns were extracted to a new dataframe. New feature columns were added (for example age) and populated with the value common to all the extracted data (e.g., '0-1 Years). This was repeated for each categorical value (in this example 0-1, 2-5 & all). These three dataframes were then merged on the country which resulted in a new complete dataframe with a third of the number of columns.
The process was time consuming but a necessary *abstraction* of the data as it meant the data changed from a single '*item'* data type (country) to an *item* and a series of *attributes* that would become *Dimensions* when imported to Tableau. The reshaping greatly reduced the number columns which would become *Measures* in Tableau and the new attribute dimensions would become useful fields for filtering.
Ultimately, this data pre-processing would reduce the cognitive challenge of the visualisation by applying a logical structure to the underlying data which facilitated navigation.

The primary '**Why?**' was to answer the three questions set out in the assignment, I decided at an early stage to create a visualisation that would allow the user to *Discover* the data by creating intuitive idioms that allowed the data to be navigated in a logical and hierarchical way. The hierarchy was essential as there was such a high volume of data.
Rather than working to address the coursework questions too specifically the idiom allowed the user to *consume* and *enjoy* the data whilst gaining an understanding which answered the specific questions.

How? Essential to allowing the user to discover the data was the ability to *identify* items of interest and *compare* the available data. The first dashboard in my story allowed the exploration of age-related data, by using a hierarchical grouping of countries – sub continents – continents the user could easily see trends in malnutrition by location. This grouping *encoded* the data by *separating* and ordering the geographical locations into Tableau groups with the resolution selectable through the 'Location Scale' drop down. This *superimposed* a simple *tree* structure on the country's *keys,* a *network* structure which aided the exploration of the dataset. Colour and hue were applied to the countries to illustrate the Location Scale the visualisation was currently displaying.

All individual malnutrition categories were available for exploration via a drop down as attempting to display all the information (apart from the embedded, Sub-Region Detail, summary plot for the currently selected geographical areas) would have overloaded the user.
An *Algorithm* was used to generate the values for the dynamic *treemaps* which displayed the best and worst performing countries for the current data categories and locations. The treemaps used *area* (magnitude channel), *hue* and *saturation (identity channel)* to illustrate the *sequence* in *quantitative* values. Red to Blue was used rather than red to green to aid accessibility for those with colour blindness. The treemaps showed the *similarity* in the *attributes* of areas selected for comparison.
A *reduction* in the cognitive load was achieved by *embedding* of all the malnutrition data for the selected regions within the *Sub-Region Detail.* The *filter* options and *aggregation* of countries into regions further facilitated this reduction of displayed information to a manageable level.
The confidence intervals were added as trendlines to the bar chart as these indicated the uncertainty which was important when comparing countries.

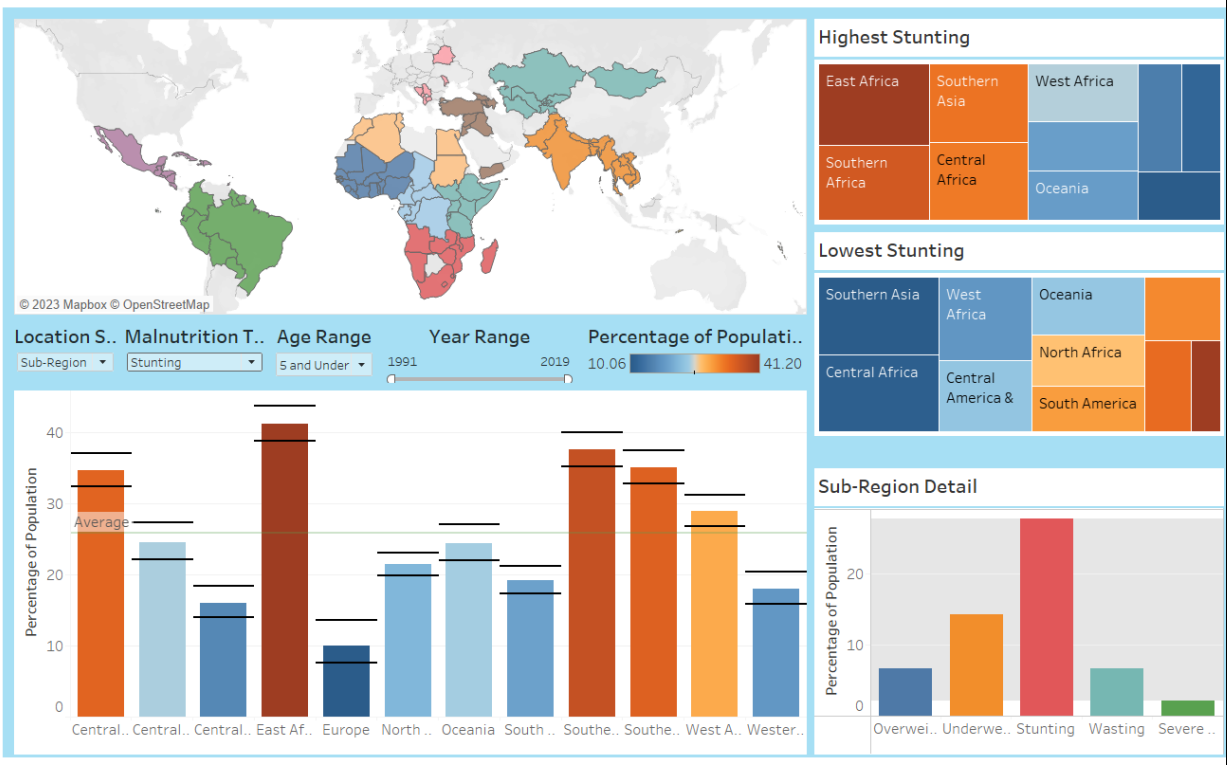

Fig1: Malnutrition by Age Dashboard

The map, location scale drop-down and bar chart could all be used to navigate the data as it is intuitive to click on features of interest. The cross filtering between idioms enhanced the user experience.

An average line was superimposed on the bar charts *common scale t*o provide a visual reference which improves the *magnitude channel* accuracy. Colour hue was used to reinforce the values given in the magnitude channel as the information density is quite high so further disambiguation is important.

The second dashboard maintained the same idiom as the first as this allowed the user to explore the data with minimal new exploration. Rather than facilitating the exploration of the original WHO dataset, as in the first dashboard, this dashboard used underlying algorithms to calculate the change in the chosen malnutrition type during the period. This specifically answers the first question in the task.

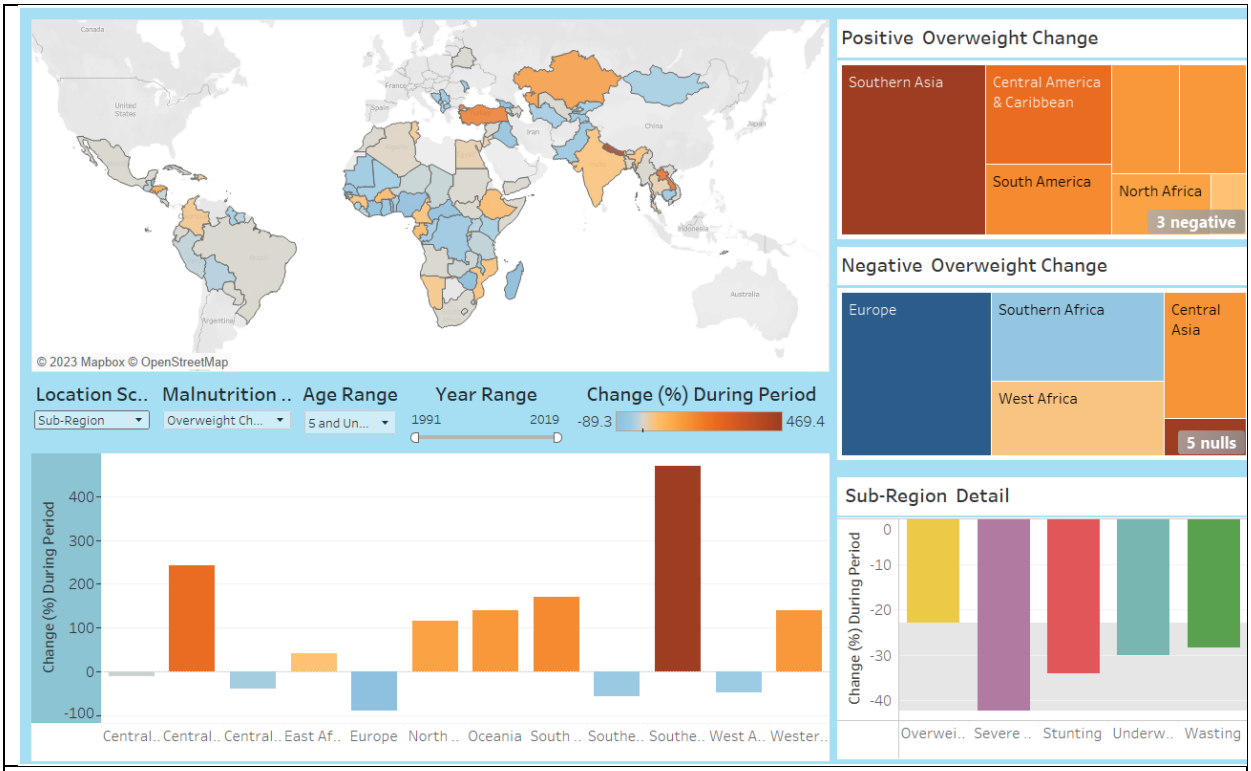

Fig 2: Change in Malnutrition Dashboard

The bar chart illustrating change uses a common scale which uses the magnitude channel of perception along with hue and saturation to reinforce the information using the identity channels. The information from the bar chart was re-enforced by the application of the corresponding colour hue to the map. This aided the identification of *spatial* trends within regions.

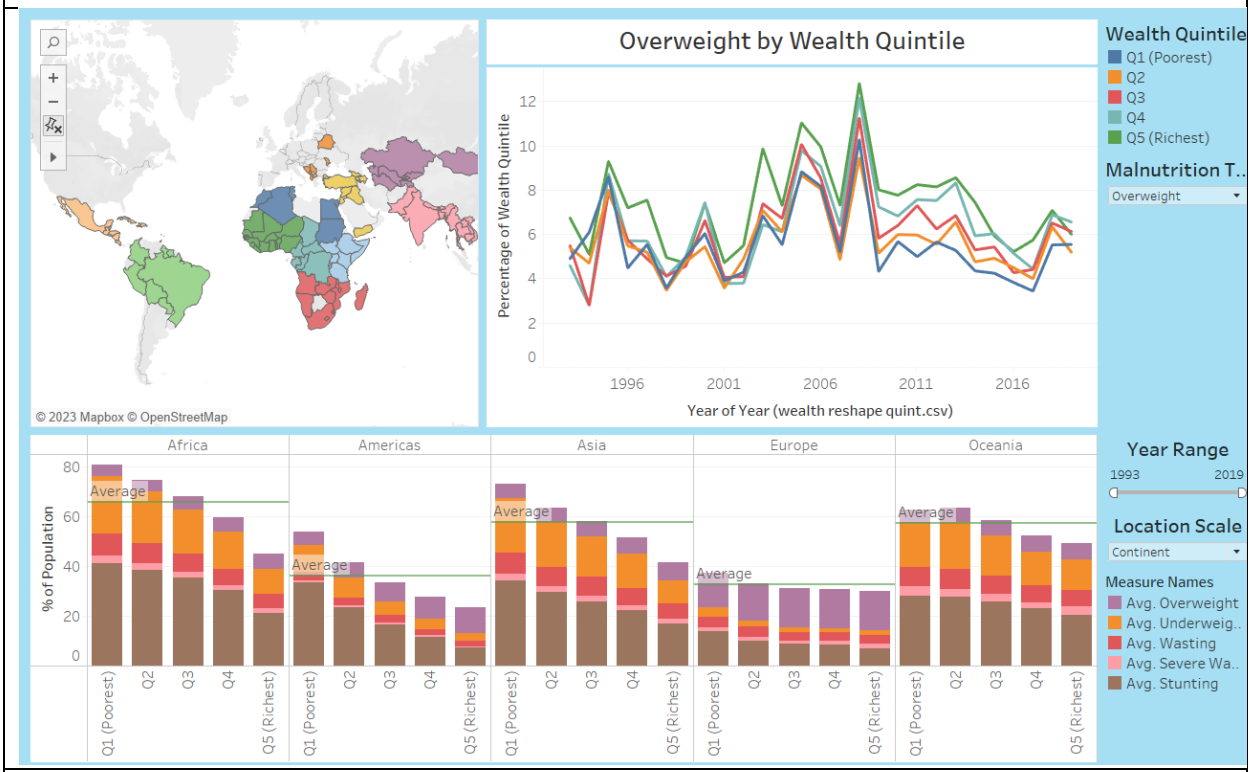

Fig 3: Malnutrition by Wealth Dashboard

The final dashboard changes the idiom slightly as there is significantly more detail to present due to there being five categories for each measure (the wealth quintiles).

A line chart is used to illustrate change as this *mark* better communicates the change due to the *tilt* of the line indicating the direction of change. Colour hue was used to distinguish each quintile on the line chart and, although a little cluttered, does well to illustrate both trends over time and relationships between quintiles.
Stacked bars were used to summarise the overall malnutrition figures as these used the magnitude channel (area) and identity channel (colour hue) to add additional detail (malnutrition type) to the overall values. A trendline illustrating average value was used to aid comparison between quintiles and regions.
A quirk of the way the WHO collects data is that individuals are double counted if they correspond to several malnutrition types (i.e. Wasting and Severe Wasting). As such, the '% of Population' metric is misleading when summarising all malnutrition categories. This is a limitation of the dataset and the relationship between countries holds true.

3.1 Feedback & Questionnaire

The questionnaire used a mix of qualitative feedback in the form of grading aspects of the visualisation and also emotional responses such as feedback on the first impression of the visualisation. Some quantitative evaluation was included by asking for specific information to be found using the visualisation.

Domain validation was not possible as there was no current user group to consult, some research was done into WHO malnutrition reporting to ensure I had planned my visualisation in a contemporary way. Questions regarding the layout and fit of the visualisation to the individual's device were used to assess whether the product was at a professional level.

Though still a qualitative opinion, as I felt first impressions would be telling- in particular as to whether there was too much or too little information included. This is an example of a dimensionality reduction task- the evaluation of many aspects such as aesthetics, layout, choice of marks is encompassed in the first emotional response.

The questions aimed to gauge if the level of detail was appropriate were all positive but two of the three had caveats about the ease of navigation. The visualisation was targeted at a high level with all the original dataset being available (including confidence intervals). As such, a certain learning curve could be expected.
In hindsight, removing the ability to show every country's values simultaneously in the main bar chart would have improved the user experience. When visualising at the highest level of granularity the plot becomes cluttered and difficult to interpret. It would have been better to restrict to showing all the countries in a continent or sub-continent as the highest granularity.

All feedback was positive with none of the confusing/ challenging options being selected. This implies that methods of visually encoding the data were successful.

Several questions specifically asked for information to be retrieved from the vis as to provide concrete metrics for the useability. These answers were generally correct but one of the three respondents had difficulty. Possibly this reflects the inability to create a solution which can translate complex information to all parties.
In a commercial environment the visualisation would expand on existing methods for visualising the data and undergo iterative testing by the users. Considering the evaluated features would equate to the 'Minimum Viable Product' (which would then undergo refining in response to user needs) I consider the feedback very positive.

There were issues when respondents were asked to identify which 'Which country had the worst increase in Severe Stunting over the total period?' The way I had phrased the question was ambiguous due to the double negative and the answer relied on the treemaps which had issues.
The treemaps used area, hue and saturation to illustrate the sequence in quantitative values. There was an attempt to use heuristic associations of red being bad and blue good to illustrate positive and negative.

Unfortunately, though applied correctly, the chart could be confusing. This was in part due to the choice of labels - ‘negative overweight change’ was meant to illustrate a beneficial change but could easily be misinterpreted.

The feedback (or anecdotal, downstream evidence) generally reassured me that I had overcome the threats to validity- the problem was correctly addressed and the data abstraction method was well received.
The idiom used is questionable as the key visualisation to answer the ‘most/ least’ type question was easy to misinterpret. This is an example of an *immediate* threat as it the final presentation of an algorithm which spoils the idiom rather that the underlying (*upstream)* implementation/ choices.

The algorithms used and volume of data were appropriate as there was only very occasional slowing of the responses.

In general, the visualisation was well received though hampered a little as I was unable to attend the evaluation session in person. As I could not give a guide on using the visualisation, I feel the feedback is very positive considering the density of the available information. An opportunity to explain the features and methodology would have allowed the users to become comfortable with the interface and possibly resulted in better buy-in (read patience in exploring the features).