COMSM0089 Introduction to Data Analytics Coursework

QJ22919/ 22334033

Task 3: Tableau Visualisation of WHO Malnutrition Data

|  |
| --- |
| Task 3: Tableau Visualisation of WHO Malnutrition Data |
|  |
| 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?**’ refered 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 formatting of the data was challenging to use so I began by reshaping the data. By creating columns for sex, age, and wealth quintile on the appropriate spreadsheets I was able to reduce the overall number of columns by around two thirds. This was a fairly involved process achieved using the Python Pandas package.  Initially I created Regular Expressions to identify columns headings which indicated relevant data, I extracted these columns to a new dataframe and added the new feature columns ( for example sex) and populated it with the value common to all the data I’d extracted. This was repeated for each categorical value ( in this example male, female & 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 attribute dimensions would become useful fields for filtering.  Ultimately, this data pre-processing would reduce the cognitive challenge of the visulaisaton by applying a logical structure to the underlying data. |
|  |
| 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 an intuitive idiom that allowed the data to be navigated in a logical and Hiehachical way. The Hierachy 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* the regions  All individual malnutrition categories were available for exploration via a drop down as attempting to display all the information ( apart from the embedded detail) would have overloaded the user.  An *Algorithm* was used to generate the values for the dynamic *treemaps* whichdisplayed the best and worst performing countries for the current data categories and locations. The treemaps used *area, hue* and  *saturation*  to illustrate the *sequence* in *quantitative* values. Red to Blue was uised rather than red to green to ais accessibility for those with colour blindness.  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 so the cross filtering between visulisatons enhanced the user experience.  An average line was superimposed on the bar charts *common scale* so provide a visual reference which improves the *magnitude channel* accuracy. |
| The second dashboard maintained the same idiom as the first as this allowed the user to explore the data with minimal new skills. Rather that allowing the exploration of the original WHO dataset, as in the first dashboard, this dashboard used underlying algorithms to calculate the change in that 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. |
|  |
|  |
| **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.  A line chart is used to illustrate the change over time as this *mark* better communicates the change over time due to the *tilt* of the line indicating the direction of change.  Stacked bars were used to summarise the overall malnutrition figures as these used the identity channel (colour hue) to add additional detail (malnutrition type) to the overall values illustrated value given by the magnitude channel (hight of bar), |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
| Feedback & Questionnaire |
| The questionnaire used a mix of qualitative feedback in the form of grading aspects of the fvisualisation and also emotional responses such as feedback on the first impression of 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.  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- many aspects such as aesthetics, layout, choice of marks are encompassed in the first emotional response.  Questions aimed to gauge if the level of detail 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.  All feedback was positive with none of the confusing/ challenging options being selected. This implies that methods of visually encoding the data was successful.  Several question specifically asked for information to be retrieved from the vis so provided a concrete metric for the useability. These answers were generally correct but one of the three respondents had difficulty. Possibly this reflect the inability to create a solution which can translate complex information to all parties.  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 to the casual user the lables such as ‘negative overweight change’ was meant to illustrate a beneficial change.  The feed back ( 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 was appropriate as there was only very occasional slowing of the responses.  In general the visualiation was well received though hampered a little as I was unable to attend in person so could not give a guide on using the visualisation. |
|  |
|  |
|  |
|  |
|  |
|  |