



UCD Michael Smurfit
Graduate Business School

Business Intelligence & Visual Analytics - MIS41040

Scholastic Challenge - Data Manipulation & Data Visualization Assignment

Visualization Tool Used – Tableau 2020.1



Submitted by:
Aditya Nair
Yanwei Wang
Tristan Kenefick



Assessment Submission Form

Student Name and Student number	Aditya Nair 19201078 Yanwei Wang 19203623 Tristan Kenefick 15205050
Assessment Title	Scholastic Challenge - Data Manipulation & Data Visualization Assignment
Module Code	MIS41040
Module Title	Business Intelligence & Visual Analytics
Module Coordinator	Dr. Peter Keenan
Date Submitted	12-04-2020
Office use only Date Received	

A SIGNED COPY OF THIS FORM MUST ACCOMPANY ALL SUBMISSIONS FOR ASSESSMENT. STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED.

Procedures for Submission and Late Submission

Ensure that you have checked the University's procedures for the submission of assessments. **Note:** There are penalties for the late submission of assessments. For further information please see the *Assessment Guidelines* publication or the website at <http://www.ucd.ie/governance/resources/policypage-latesubmissionofcoursework/>

Plagiarism is the unacknowledged inclusion of another person's writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism are designed to impose sanctions that reflect the seriousness of the University's commitment to academic integrity. Ensure that you have read the University's *Briefing for Students on Academic Integrity and Plagiarism* and the *UCD Plagiarism Statement, Plagiarism Policy and Procedures*, (<http://www.ucd.ie/registrar/>)

Declaration of Authorship	
We declare that all materials in this assessment are our own work except where there is clear acknowledgement and appropriate references to the work of others.	
 Yanwei Wang	 Aditya Nair

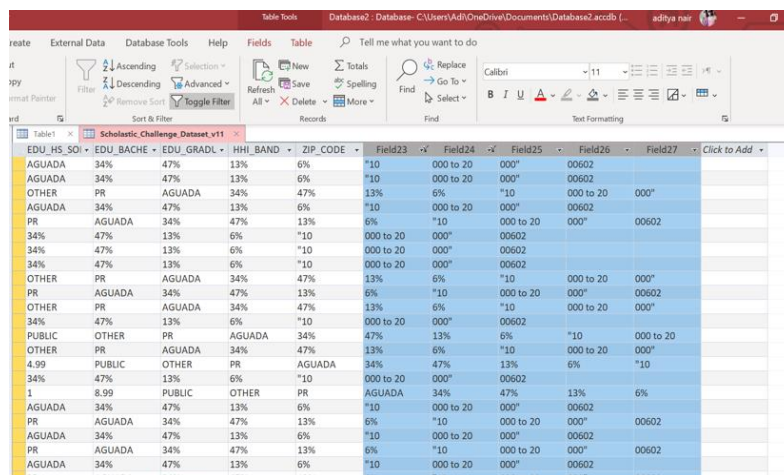
1. Motivation for the system and background

The motivation behind this Decision Support System via Tableau based visualisation is to aid the business stakeholders to arrive at a consensus regarding structuring a distribution mechanism that would leverage the capabilities of the existing system and eventually increase revenue. The two channels that Scholastic Corporation has deployed to cater to its customers in the US mainland, Alaska and in the adjacent territories like Puerto Rico are Channel 1 and Channel 2.

2. Data Cleansing and Preparation for loading

Tools used for data cleaning: Microsoft Access and Tableau Prep

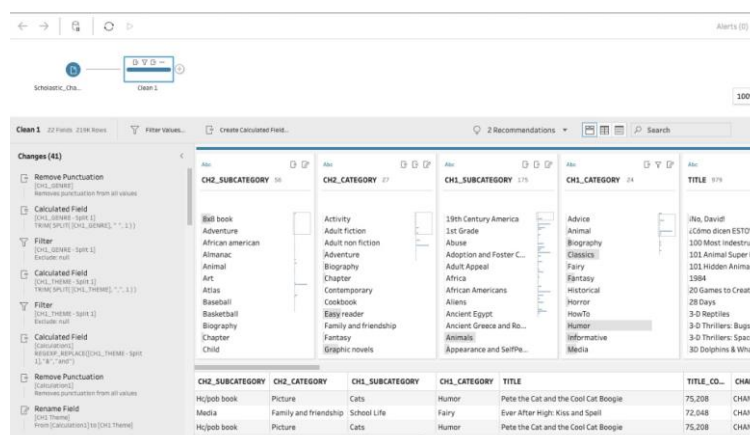
The Extraction Transformation and Loading engine that might be employed by Scholastic to gather data into a repository, transform data using a business logic and load data into a warehouse normally creates a CSV file at the output end.



The screenshot shows a Microsoft Access database window with a table named 'Scholastic_Challenge_Dataset_v11'. The table has columns: EDU_HS_SOI, EDU_BACHE, EDU_GRADL, HHJ_BAND, ZIP_CODE, Field23, Field24, Field25, Field26, Field27, and Click to Add. The data rows contain various values, including percentages, state abbreviations (e.g., AGUADA, PR), and ZIP codes. Some cells contain trailing spaces or other junk characters, which are highlighted in blue.

Microsoft Access: Identified the trailing junk characters in the data

The provided CSV file was plagued with junk characters and information that was not relevant for our analysis.



The screenshot shows the Tableau Prep interface. On the left, the 'Changes (41)' pane lists various transformations applied to the data, including 'Remove Punctuation', 'Calculated Field', 'Filter', and 'Rename Field'. The main workspace shows a data preview with columns: CH2_SUBCATEGORY, CH2_CATEGORY, CH1_SUBCATEGORY, CH1_CATEGORY, TITLE, and CHANNEL. The data rows show various book titles and categories, such as 'Hcjob book', 'Picture', 'Cats', 'School Life', 'Fairy', 'Humor', 'Pete the Cat and the Cool Cat Boogie', '75,208', 'CHANNEL', '72,048', 'CHANNEL', and '75,208', 'CHANNEL'.

Tableau Prep: Split columns, exclude punctuations, filter null values and allocate roles

There were unknown and ambiguous values in the 'State' attribute. There were spelling mistakes of the State abbreviations and there were also null values. We created a second data

file, which we would use for appending publicly available data- State abbreviations and respective state names and census sourced data on the number of children in each.

Data_Scholastic_i4 (Multiple Connections)

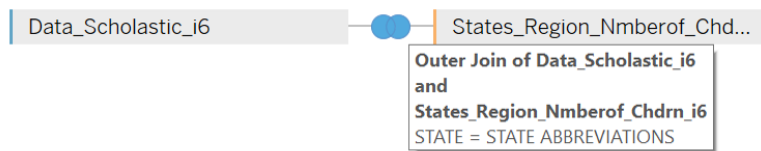


Tableau: Outer join between the attributes of the two tables

We used the *State* and *State Abbreviations* attributes in the two tables to create a full outer join on the two tables. We used Tableau Prep to split the category and theme columns, so that the visualisation only presents the main information. We also used it to remove punctuations, exclude null values and allocate data roles such as geographical and numerical data. There were multiple languages in the dataset, including English and Spanish. However, the Spanish only appeared in the book names- which does not affect data visualisation. Therefore, we did not make any changes to it.

Another problem in the dataset is null values in a few columns. We faced the dilemma of keeping the null value but possibly change the view or give incorrect information about the data, or we can exclude the null values before visualising it. Eventually we excluded the null values from the Channel 2 category attribute. We enumerated the attributes into four categories which helped us to narrow down the requirements of the dashboard.

Attributes if the Data_Scholastic_i6 table

Attributes Defining the book	Channel Descriptors	Attributes Defining Revenue	Demographics Descriptors
	CHANNEL	TOTAL_UNITS	SCHOOL_TYPE
TITLE_CODE	CH1_GENRE	UNIT_PRICE	REGION
PROD_TYP	CH1_THEME		STATE
SERIES	CH2_CATEGORY		COUNTY
LEXILE_11_DESC	CH2_SUBCATEGORY		EDU_NO_HS
			EDU_HS_SOME_COLLEGE
			EDU_GRADUATE_DEG
			HHI_BAND
			ZIP_CODE

Attributes if the State_Region_Numberof_Chdrn_i6 table

Demographic Descriptors	Appended Data
STATES	CHILDREN UNDER 18 YEARS
STATE ABBREVIATIONS	
REGION	

3. Analysis

The decision makers will likely want to know in which area the business is performing well so they can compare the similarities and differences between the good and bad performers, as well as allocate their resources and improve certain aspects accordingly. The DSS allows the user to see the total units sold and revenue made by each state as well as by region. It visually differentiates level of sales by using different shades of colours so that the user can see which states or areas stand out without checking the numbers. They can also check the exact number by hovering the mouse to the state they would like to see. From the data, we can see that two states- California and Texas, stand out the most with the units sold and revenue. Followed by a few other states including Florida, Miami, Illinois and New York. Overall, states in the Midwest and South perform better than states in the West with California as an exception. The representing colour of these states are darker than others.

Another area of interest to the user would be their product popularity so they can invest more into the well-selling ones. User will be able to find out the overall units sold by each book category and units sold by product type and compare across all categories. It also allows the user to compare sales by Lexile reading level to better know their customer groups. From the data, the most popular category is Picture books, which is almost three times the number of the second largest category, Instruction Resource. A few other popular ones fall into the categories of Easy reader, Graphic Novel and Humour. These popular types of books are easily associated with young children. By sorting the sales and revenue by Lexile level, which measures readers' reading level, we get the result that (Non-prose), BR (Beginning Reader), and GN (Graphic Novel) are the three most popular levels, which bring in the most revenue. There is a large demand for books in these categories and it would be beneficial to meet these needs of the market.

In addition, users can use the DSS to categorize revenue by household income band for get further insight about what type of family are more willing to spend money on children's book purchasing. Families with Household Income Band between 40k and 60k are the main driver of revenue, followed by band between 70k and 80k. Surprisingly when the income band is over 80k, not as much revenue is generated from those families. It will be useful to find out what the reason is behind this phenomenon. When the income band is below 30k, hardly any revenue is generated.

Scholastic wants to better understand the children's book market, therefore we thought it is natural for the decision makers to want to know the demographics of children in different states in order to see if there is any relationships between children density and book sales. The DSS is connected to the data of number of children under 18 by state from the 2000 census. Users can see the number of children of each state and the sales of the corresponding state to find out the correlation. The trend of this data is generally in accordance with the trend of units sold by two channels. The top two states with the most children under 18 are California and Texas, and these two states also rank the top 2 in book units sold. Overall, states with a higher density of children such as New York, Miami, Illinois, etc outperform

states with less children such as Montana, Wyoming, North Dakota and South Dakota in terms of book units sold. Considering this, it would be worthwhile to place stronger marketing strategies in states where the number of children is high, for example, the Midwest and the south.

In the last dashboard, we have added the trend which shows the number of deliveries attempted. This field is a calculated field which is the count of the *Channel* attribute.

Another calculated field that we have used is for the number of children under 18 years of age in each state. As the states are recurring, we had to use the below calculated field.

Distinct Children

Data_Scholastic_i4 (Multiple Connections)

```
{FIXED [CHILDREN UNDER 18 YEARS] : MIN([CHILDREN UNDER 18 YEARS])}
```

4. Scholastic's distribution channels and distribution strategy

Another function of the system is that user can compare and evaluate the performance details of two distribution channels graphically and numerically, for instance, overall and each state's units sold, revenue, product type, region etc. Channel 2 has an obvious advantage over Channel 1 in terms of units sold and revenue. Channel 2's sales is more evenly spread among all regions with the most units sold in Midwest, followed by the South, West, and Northeast. Channel 1 has the most sales in the South and performs weakly in other regions compared to Channel 2.

In terms of product type, total units of paperbacks sold is roughly 9 times the number of hardbacks, since the average price of hardbacks is over double the price of paperbacks. With total units sold far behind Channel 2, Channel 1, however, accounts for about 63% of the hardbacks sold. With more hardbacks sold percentagewise, Channel 1's product average price is about 8.5 dollars and Channel 2's is slightly less than 7 dollars.

5. Marketing strategy & strategy to prevent competition between the two channels

The DSS allows user to see and compare two channels' sales performance in each state and region in the form of a map presentation. From the map view, we can see that the two channels have a lot of overlap in sales, especially in the best performing states including CA, TX and FL, and in the region of the south. However, Channel 2 outperforms Channel 1 on almost every parameter. One of the advisable options is to cease Channel 1 and allocate the resources used to operate and maintain Channel 1 to the support and development of Channel 2.

Note that Channel 1 takes up almost half of the market in a few states in the south, such as Florida, Georgia, and Tennessee. The revenue that Channel 1 makes is almost a quarter of the total revenue. Plus, considering the fact that Channel 1 sells more hardbacks of its total sales, and the average price of each unit is over 1.5 dollars more expensive than that of Channel 2, it is worth considering the marketing strategy of strengthening Channel 1's expertise in selling hardbacks, and only operating Channel 1 in the south region, instead of all over the country. It would save the cost of maintaining channels in unnecessary location and avoid much of the competition with Channel 2 in most part of the country.

6. Instructions to Operate the System

The tableau 2020.1 based Decision Support System is self-explanatory. We have added linked filters visualisations so that the end user can have a better understanding of the reach of each channel, reach of the channels in states, regions, the units sold and revenues generated in each of the scenario.

We have added our analysis of each dashboard in the right top corner of each dash. This analysis will be visible only when the end user hovers near this area.



7. Link to the Tableau Public Server

https://public.tableau.com/views/ScholasticSalesSurveyAnalysis/Story?:display_count=y&origin=viz_share_link

8. References for the additional data used

www2.census.gov. n.d. [online] Available at: <https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf> [Accessed 15 March 2020].

Census.gov. 2004. [online] Available at: <<https://www.census.gov/population/socdemo/hh-fam/tabST-F1-2000.pdf>> [Accessed 15 March 2020].

SuburbanStats.org. n.d. *Current Population Demographics And Statistics For Puerto Rico By Age, Gender And Race..* [online] Available at: <<https://suburbanstats.org/population/how-many-people-live-in-puerto-rico>> [Accessed 16 March 2020].