Investigating Customer Purchase Behaviour for a Toronto Medical Supply Store Chain

1. Research Question(s):

Stores selling specialist products such as medical devices need a targeted marketing approach to ensure they are constantly encouraging new clients to purchase their products and also retaining their loyal clients. Gaining an understanding of patterns of customer behaviour is an interesting and profitable venture for specialist companies. Data visualisation provides a useful tool for generating complex and hidden insights which are not obvious through domain knowledge or by simple numerical calculations on data tables. This background motivation was the key for the framing of the research questions for this project, which could provide useful insights into customer behaviour based on customer data provided for an anonymised medical supply store chain in Toronto. The following research questions were investigated:

- Over the three years, is there a seasonal trend in customers purchasing products from stores? This can be useful for stores looking to target their marketing campaigns at certain times of the year.
- Over the three years, what is the spatial distribution of products purchased by customers in stores across the Toronto metro area? Are there denser clusters in some parts compared to others?
- How are the clients being referred to do the store and where are clients who visit through referrals mainly located? An insight into this aspect can also help in increasing sales for the store.

2. Data Source(s):

The data for this project was available as csv file downloads as part of the Medical Store Geospatial Challenge from Databits, a community website setup for creating interactive data visualisations (http://databits.io/challenges/medical-store-geospatial-challenge)
These include data over three years on client purchases, referrer sources, location and coordinates of all cities and medical supplies stores. Additional csv files were created through parsing (slicing, merging, reformatting etc.) of these main data sources in Excel to produce it in the most simplistic structured format as required by the workflow stages of 'acquiring and parsing' specified by Fry (2008)

3. User Instructions:

The default visualisation sketch was produced in the Processing 3.3. programming language (Reas C et al, 2006, https://processing.org/) and can be generated by opening and running the 'Home.pde' file. It consists of four separate sketches transformed via scaling and translation (Wood.J, 2017, Session 7) and drawn on one window for ease of viewing (figure 1). The contents of the sketches are briefly described below:

- 'barchart.pde'- products bought at each city bar chart (top right of figure 1)
- 'torontomap.pde'- Toronto shapefile containing locations of the customers making purchases- data aggregated over 3 years (top left of figure 1)
- 'graph.pde' – Line plots of products purchased in 2012 (bottom left of figure 1)
- 'piechart.pde' -Pie chart showing the referrer sources that customers come from (bottom right of figure 1).

The basic method for dynamically changing the bar chart and map sketches in the visualization is through key press buttons. Pressing keys ‘7’, ‘8’, ‘9’ and ‘0’ allow the map to be changed to represent geospatial customer purchase behavior in years 2012,2013,2014 and back to the default setting (aggregated three years) respectively. Pressing key ‘6’ changes he map to a spatial overview of location of customers who visit the store through referrals. This map can also be activated by pressing the white button next to the pie chart labelled ‘Map Button’. Pressing keys ‘1’, ‘2’, and ‘3’ allow the user to navigate through the different plots of products purchased in different years (2014, 2013 and 2012 respectively). The user can also zoom in and out of the sketch by using ‘SHIFT’ and either rolling the mouse wheel or pressing the left mouse button and dragging. Pressing key ‘z’ after zooming in, resets the zoom to the default settings. Pressing the Canadian flag in the centre of the sketch or using the ‘TAB’ key resets the sketch to its default setting after any interaction.

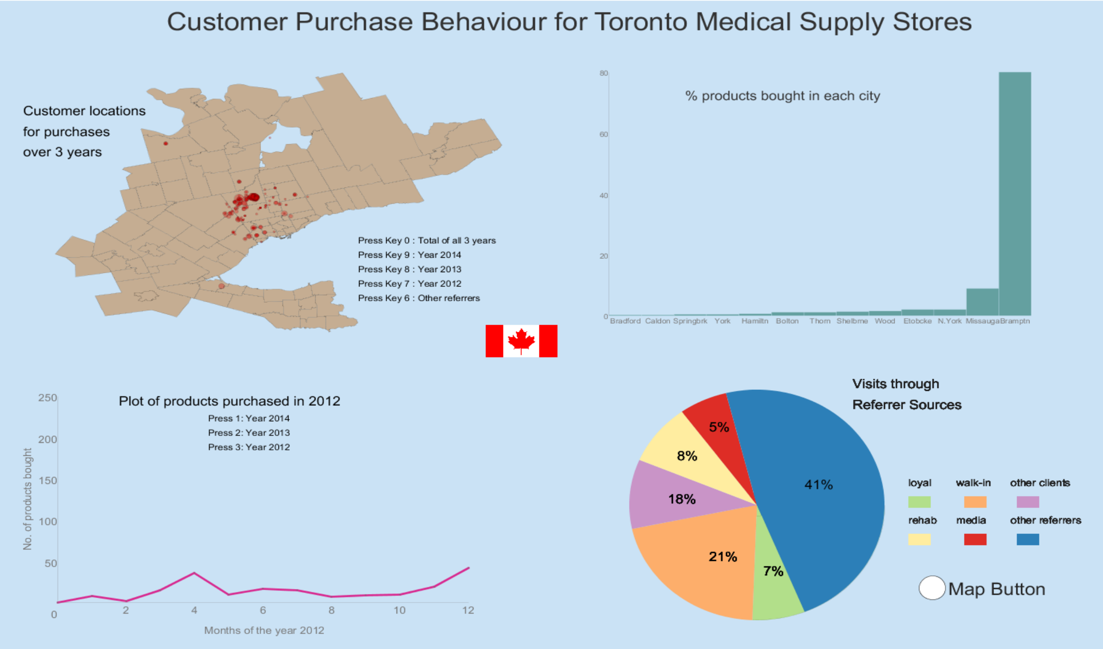

Figure 1: Working sketch design in Processing 3.3 with the four sketches scaled down and transformed to fit on the same window

Implementation of the tooltip feature allows the user to move over the piechart, geo map, barchart and line graph to gather more information about the raw data values and axis labels. In addition, by highlighting the bars in the barchart under which the mouse is hovering over, the user is able to interpret which bar he is looking at a lot more easily.

Some features could not be implemented due to technical difficulties and time constraints.

A filter option would have been useful as per the guidelines in Shneiderman, B. (1996) for interactivity design, that enables the user to click on the a bar in the barchart and filter only the data for the city for which the clicked bar represents. This would have helped the further delve into the purchasing behavior of the popular cities (highest number of purchases) independently without interference from the other data points in the map. A few more button features (in addition to those already available) to compliment the ‘key press’ option to dynamically change the graphs, which would have made navigating the sketches easier without having to remember the keys which need to be pressed. The difficulty in implementing the filtering design was mainly due to the scaling of the different sketches on the same screen and the subsequent complications arising from mouse position calculations for implementing interactions. This could have been easily solved by having the sketches on different tabs instead. However, I felt having the layout sketches all in one window was a very important design choice (Fry, 2008, Sedlmair et al., 2012) which made the process of finding insights to the research questions a lot less cumbersome.

4. Design Justification:

The design process for building this visualisation sketch was motivated by steps in Fry's 7-stage workflow (Fry,2008) - 'acquire, parse, filter, mine, represent, refine, interact' for constructing effective data visualisations and the visualization framework described by Munzner (2015).

4.1 Visual Encoding:

A bright background colour (light blue) was selected without any other changing combinations to avoid over complicating the design (Stone et al., 2006). The colours of the bar charts and the line chart were chosen so that it gives a good contrast between the background and the chart. Inspiration for the colour choice of the map was taken from the giCentre GeoMap documentation (https://www.gicentre.net/geomap/using) where light brown was used to mimic the natural colour of land with the default blue background fitting in well with the natural colour of the ocean. The data points in the map were given different shades of red depending on the number of purchases made from customers at different post code locations. A smaller circle (indicating fewer purchases) was encoded dark red whilst a larger circle (indicating larger number of purchases) was encoded light transparent red. This double encoding scheme not only helps to reinforce the difference in values of each data point relative to the others but also allows points which are overlapping to be viewed easily i.e. smaller sized circles would have been masked by the larger size circles if the latter was not made transparent (Kirk, 2016, Chapter 9, Wood, 2017, Session 3). The colours in the pie chart were chosen to be bright colours and colour blind safe using ColorBrewer (Brewer et al, 2003). Percentage values were included in the pie chart, to make it easier to accurately perceive and compare the quantities in the pie chart, as described by Ware (2012). Without these labels, pie charts are not usually a good option as it is difficult to decode the visual representation of quantities accurately. (Tufte, 2001). The barchart was designed so that the bars are ordered (Munzner, 2015) from high to low value to reduce cognitive load, also known as preattentive processing (Ware, 2012). The only symbolisation used was the traditional Canadian 'maple leaf' in the middle of the sketch, which besides functioning as a 'reset' button also immediately alerts the user to the fact that the tool is designed for data related to Canada.

4.2 Interaction:

The main interactions implemented were details on demand, key presses and zooming, following the guidelines in Schniederman (2006). Details on demand were implemented to enable more information to be made available to the user when the mouse hovers over all graphs and the geo map. This enabled each sketch to look cleaner with less clutter and removing the requirement for having labels included in the sketch (Tufte, 2001). For the bar graph, a bar highlight option was also implemented to enable the user to know which city the value in the tooltip corresponds to (Wood, 2017, Session 8). The use of keys to change the sketches B and C to a different year in each window allows the user to interactively compare different combinations of customer purchase behavior in different years on the temporal and spatial plots. The keys were purposely selected to be on the same row on the keyboard so the user can easily manipulate without requiring too much focus (Munzner, 2015, Schniederman, 2006). They are also spaced apart conveniently so that one could manipulate the line graphs with the left hand and map with the right. Zooming and panning can be used to see any text information on the graphs that may not be immediately visible in the default sketch due to the scaling applied. This may be also useful to see the spatial distribtuion fo data points in the graph when comparing the density of clustered points from one year to the next. Buttons were also designed to allow for tedious tasks like resetting the sketch to its default value.

5. Data Insights:

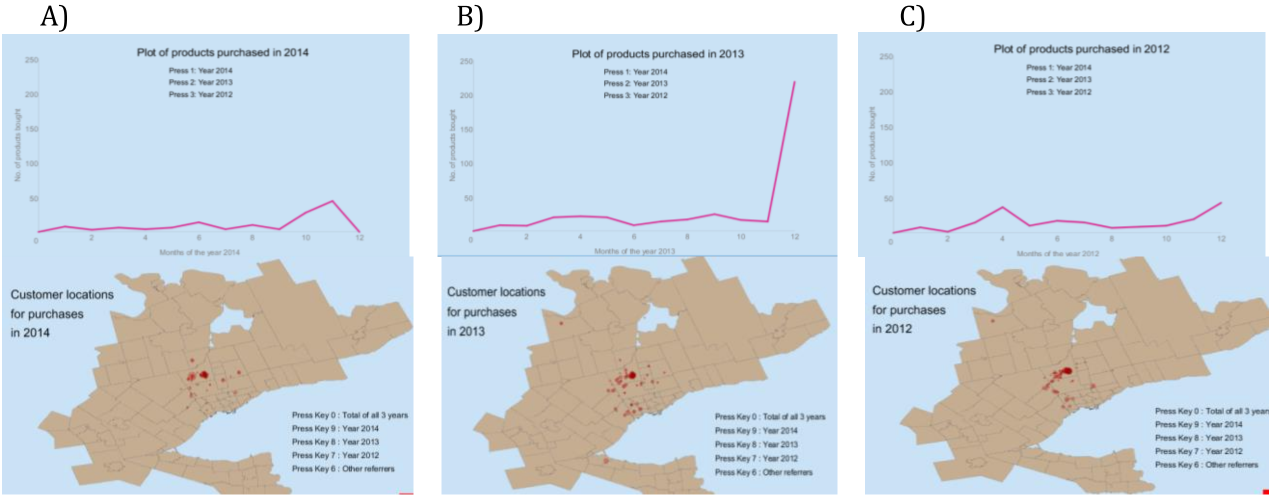

Figure 2: Screenshots of Sketches 2 and 3 produced by key press interactions, enabling the user to choose sketches from a common year for comparison on the same window. A-C are from years 2014, 2013 and 2012 respectively.

To answer the first two research questions, it was necessary to use the key press options described in sections 3 and 4.3 to produce both spatial and temporal sketches in the same year for comparison. Figure 2 (A-C) are cropped screenshots of the pair of sketches produced in the window after using the key press option, representing the years 2014, 2013 and 2012 respectively. To evaluate the seasonal trend in product purchases, we can see a small sharp spike in purchases in year 2012 at around April and then again in December, a large spike in December in year 2013 and a small spike in purchases at around November in year 2014. Hence, there appears to be some evidence of customers purchasing products towards the end of the year during the more recent years.

When looking at the respective spatial plots for the given years, we can see a trend in the number of customers purchasing products mainly from the 'downtown' (central) Toronto area, with a few customers living in the outskirts (Shelburn and Hamilton) purchased products in 2013 and 2012. Further support to these insights can be lent from the bar chart in figure 1. It is easy to see that the customers purchasing the majority of products (around 80%) came from the Brampton area downtown, which corresponds to the high density of points (large dark red cluster) in the maps across all three years. The next highest was Missisauga which corresponds to the cluster of points just south of Brampton, although this is less pronounced in year 2014. Customers from boroughs in the east of Toronto (York, Springbook) also seem to be purchasing less across the time span. Generally, there seems to be an increase in purchases spatially in 2013 relative to years 2014 and 2012, with the highest density focused in Brampton. Without extra information, it is difficult to deduce whether this is due to the population size being larger in this borough, the proximity of store location to this area or due to other factors like age, ethnicity which predisposes this group of people to certain diseases requiring special medical products.

The final research question can be answered through viewing the pie chart (sketch D) in figure 1. Almost half of the customers were referred by a range of 'other anonymized referrers' whilst the next majority were 'walk-in' customers. Only 7% were classified as 'loyal clients' i.e. having made repeat purchases over the three-year period. The map in figure 3, shows three obvious clusters of client groups based in Brampton, Caledon and Missisauga who have made visits through these referral sources.

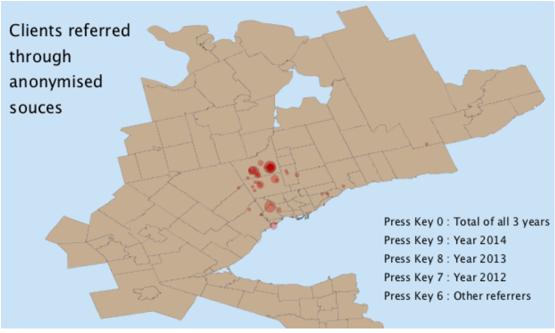
Clients referred through anonymised souces
Press Key 0 : Total of all 3 years
Press Key 9 : Year 2014
Press Key 8 : Year 2013
Press Key 7 : Year 2012
Press Key 6 : Other referrers

Figure 3: Spatial distribution of clients who made visits through anonymized referral sources.

6. Critical Evaluation and Reflection:

By following good visualisation practice from the literature (Fry, 2008, Munzner, 2015), this visualisation tool has enabled me to decode complex data to answer the research questions effectively. Implementation of further enhancements to the sketch such as filtering (discussed in section 2) would have also allowed me to derive even more insights and adhere to step 3 of Fry's (2008) 7-stage workflow. The strength of this visualisation design was the layout of the sketch, which is all on one screen and not complicated to navigate and the information is as easy to find as possible (Tufte., 2001, Fry, 2008). The suggestion by Fry (2008) of choosing a basic visual model of representations was also consistently adhered to, through the use of bar chart, line graph and pie chart. Additionally, all graphics were designed to include the highest possible data-ink ratio, a concept introduced by Tufte (2001), through the elimination of grid lines, borders around graphs and keeping a common background colour for all sketches. A critique of this tool may be that it is not possible to see the visuals in real time (using an 'autoplay' button or similar) or viewing all the dynamic maps on the same screen at the same time. It is also easy to skew or overcomplicate the visual design and convey a false message if there are limitations in the amount of data available. For this reason, the data was aggregated and into years and presented rather than using a 6 or 3-month cycle. The 45 'other referrers' were aggregated and represented as a single section in the pie chart (blue coloured section in figure 1), rather than representing 45 separate arcs. Although, this strategy was sufficient to answer all the research questions, it may not be applicable for other situations where details down to individual client level or a non-aggregated information is required.

If I was to repeat this exercise again I would definitely implement more interaction options as it would have been a useful addition to this design. Scaling the sketches to include an all in one window proved a challenge when implementing interaction between them (due to calculation of the mouse positons). I would have definitely planned this particular aspect a lot better if I had to repeat my implementation again and likely to be more successful in designing the interactions. Implementing buttons instead of keys would have made it easier for the user to move back and forth between sketches without having to remember the key which needs to be pressed. Analysing repeat customer visits in time and space to determine loyal customers and merging with demographic/census data are other avenues which could be explored.

7. References*:*

1) Brewer, C., Hatchard, G. and Harrower, M. (2003) ColorBrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science, 30(1), pp.5-32*

2) Databits.io, (2016). Medical Store Geospatial Challenge. [online]. Available at: http://databits.io/challenges/medical-store-geospatial-challenge [Accessed 15 March 2017].

3) Fry, B. (2008). *Visualizing Data, 1st ed. Beijing: O'Reilly.*

4) giCentre.net. GeoMap library documentation. [online]. Available at: https://www.gicentre.net/geomap/using [Accessed 1-25 April 2017]

5) Kirk, A. (2016). Data Visualisation: A Handbook for Data Driven Design, Sage.

6) Munzner, T. (2015). *Visualization Analysis and Design.* A K Peters Visualization Series, CRC Press

7) Processing.org. (2004). Processing Website Tutorials. [online]. Available at: https://processing.org/tutorials/ [Accessed 1-25 April 2017]

8) Reas, C. and Fry, B. Processing: programming for the media arts (2006). *Journal AI & Society, volume 20(4), pages 526-538, Springer*

9)Sedlmair, M., Meyer, M. and Munzner, T., (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*, *18*(12), pp.2431-2440.

10) Shneiderman, B. (1996) The eyes have it: A task by data type taxonomy for information visualization, Proceedings of the IEEE Symposium on Visual Languages, pp.336-343.

11) Stone, M. (2006) Choosing Colors for Data Visualization, www.perceptualedge.com/articles/b-eye/choosing_colors.pdf

12) Tufte E.R. (2001). *The visual display of quantitative information*. Cheshire, Conn, Graphics Press.

13) Ware, C., (2012). *Information visualization: perception for design*. Elsevier.

14) Wood, J. (2017). City University Data Visualisation Module Lecture Notes. Accessed from Moodle, Jan-Apr 2017