CSCM27 - Human-Centred Visual Analytics

Suzannah Downie 2131572

November 2021

Data Set Features and Structure

The data set used is a time series data set containing 731 observations and 15 data columns spanning the following data types: [1]:

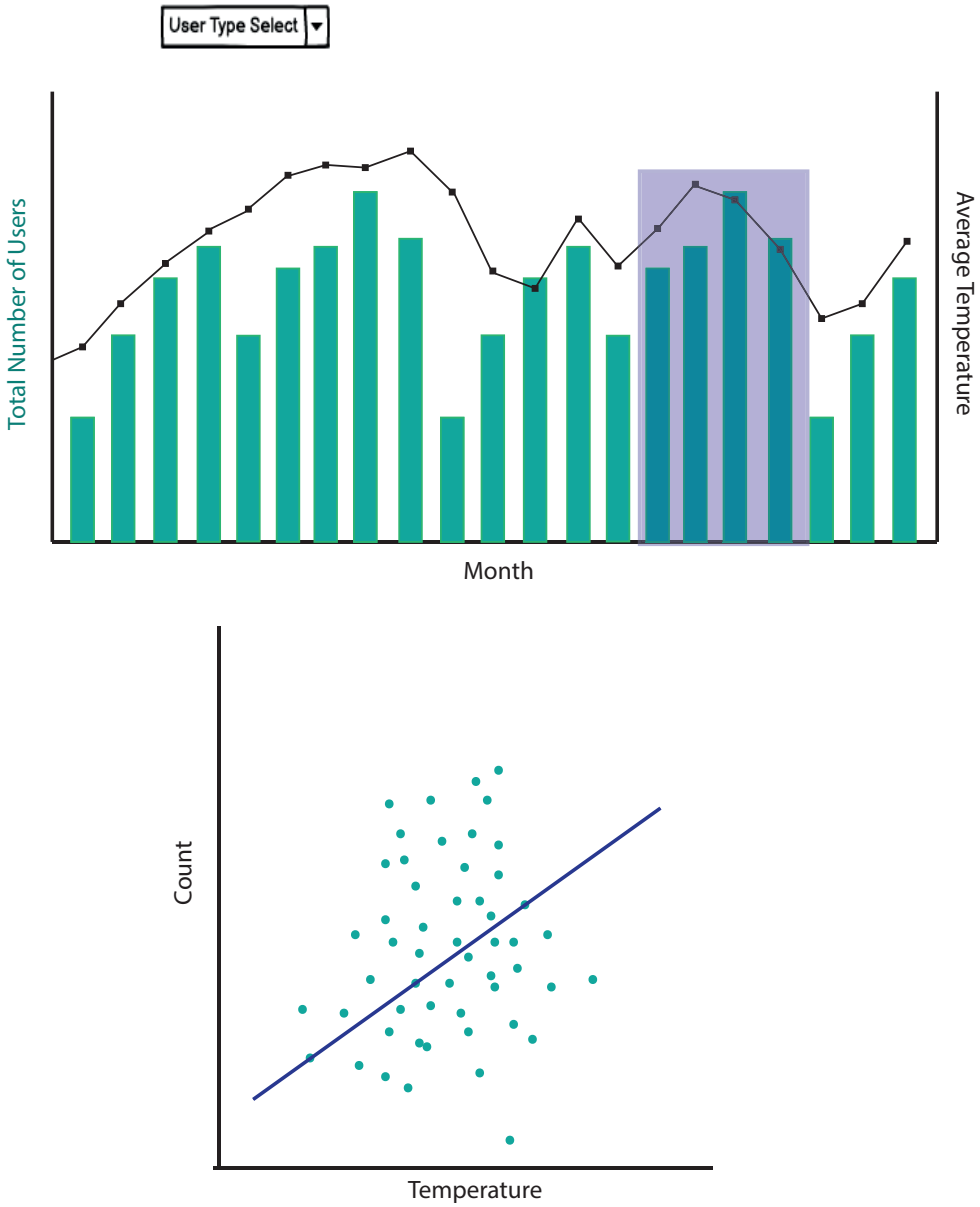
- Ordinal - day of the week, season
- Nominal - working day (yes or no), weather situation
- Quantitative - temperature, humidity, windspeed, count

https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

Proposed User Task

Our user wants to be able to understand the degree to which temperature affects bike rental usage so that they may attempt to anticipate demand during different temperatures.

Prototype Number 1



This prototype uses two superimposed layers within the same frame. One layer is a bar chart to demonstrate the total number of bike share uses each month over the time series period. A bar chart has been chosen because length is perceptually linear. With this in mind, users are able to understand magnitude channels like growth or reduction in bike rentals by comparing these lengths [7]. Studies have also shown that spaced bars can aid users' ability to process bar charts [11, 2]. The second layer within the layered frame is a line graph showing the average temperature for each month. Plotting data on a common scale is frequently claimed as one of the most effective means of visualisation [7, 4]. As temperature data has an inherent order, plotting this on a common scale supports the user to decipher the information as easily as possible. To help aid the user to distinguish between which of

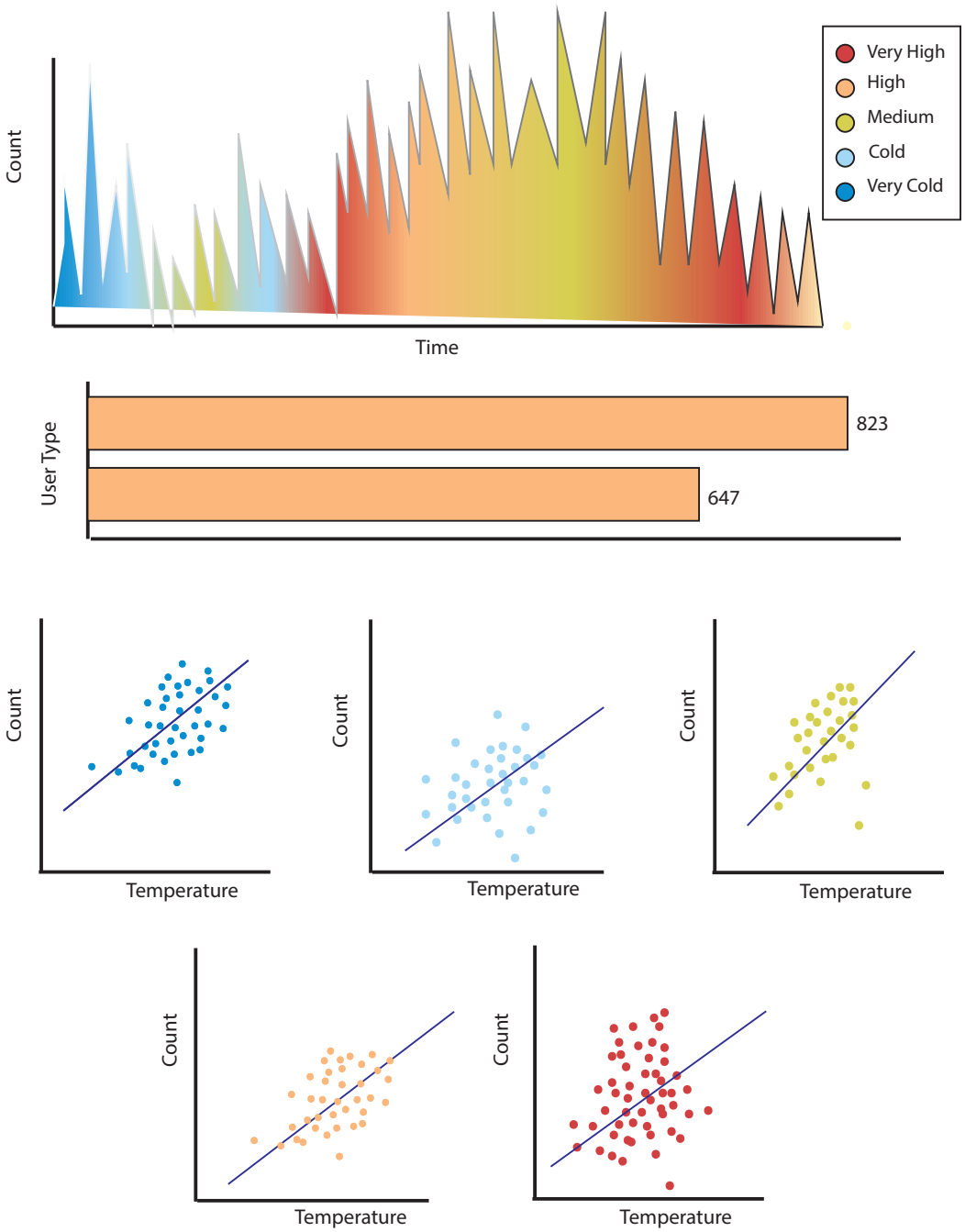
these layers uses which axis (left or right) we have utilised colour to help with ordering the information. The axis labels are coloured to match the layer of graphic that they relate to. Using colour in this way enables us to make use of preattentive processing, as users are able to group elements of a visualisation rapidly [5].

While too many elements within one frame might be difficult to interpret, this frame contains only two graphs. These two graphs also have a shared axis. Owing to these factors, it is not overwhelming for the user [7]. The rationale for having two visualisations within the same global frame is because the user is then able to switch between the two quickly using their eyes rather than their memory, thus reducing the cognitive load [7]. The user is therefore able to track how the total number of daily bike share users changes with the monthly average temperature, supporting the user goal of understanding how temperature affects bike share usage.

Underneath the line/bar layered frame is a scatter plot. In order to support the end goal of being able to anticipate demand for bike share services during different temperatures, we need to be able to predict what usage is likely to be. To support this, we utilise a regression scatter plot to show the connection between temperature and bike rental count. Scatter plots help to make predictions and are particularly effective for spotting correlation and for using levels of high correlation to make predictions [3].

The interactivity of this visualisation takes in to account Ben Schneiderman's mantra: "Overview first, zoom and filter, then details on-demand" [9]. The visual first presented to the user is static, with 3 graphs to consider. This is the overview. This layout also helps the users with data "chunking" by allowing them to process a chunk of 3 graphics (a bar plot, a line plot, and regression scatter plot) with no other visual distractions. The user is then able to zoom and filter by using the "User Type Select" drop down menu. This drop down is intended to allow the user to filter the data in the bar plot according to whether someone is a casual or registered user of the bike share service, allowing them to begin to understand the effect that the average temperature in a given month has on user types. Similarly, a selection tool allows the user to select specific months by brushing over them in the bar chart, this would then dynamically update the scatter plot allowing the user to see the effect that the average temperature within a given month has on usage predictions made in the scatter plot. The aim here is to give the user control over the jump cuts between two frames (overview and user selection) to support effectiveness [7]. Finally, an interactive tool tip would enable the "details on demand" portion of the Schneiderman mantra.

Prototype Number 2



To support our users' end goal, we must consider three things: how many bikes are rented, the temperature, and whether there are patterns over time. As we have established with our discussion of Prototype 1, data plotted on a common scale is perhaps most effective

for ordered information [7, 4]. We therefore show the time series data along an x axis scale, and the count of the bike rentals on a y axis scale. However, to further understand the impact of the temperature, we need to introduce another graphical element to encode this. Colour is a wise choice when we consider that it lends itself well to preattentive processing, reducing the cognitive load for users [5]. To allow us to plot on a common scale, but to provide enough colour to support preattentive processing, we have used an area map.

Whilst rainbow colour maps can be used to show temperature data [12], these have more recently been seen as less effective for many visualisations, particularly those spanning larger areas [10]. As the area we wish to colour is quite large, we have therefore avoided rainbow colour maps. As an alternative option, we have used a divergent colour scheme as scholars have argued that the middle points of these highlight extremes as well as themselves [6]. This allows us modify colour to see temperature as a spectrum of extremes without being as intrusive as a rainbow colour map over a larger area. As our data set has vast amounts of continuous temperature data spanning two years, we propose binning the data to 5 percentile clusters. This is to prevent overwhelming the user with too many differing colours. We therefore map our divergent colour scheme to these 5 percentiles. To select exact shades most appropriate, we used ColorBrewer suggestions.

Just as in Prototype 1, we make use of the regression scatter plot's ability to demonstrate any connection between rental count and temperature. However, in contrast to Prototype 1, we make use of small multiples, which are effective at showing changes using static images [8]. By comparing images side by side, the user is able to determine the degree to which temperature impacts upon bike share usage; in addition to this these scatter plots are able to predict how future temperatures might increase or decrease demand for the service.

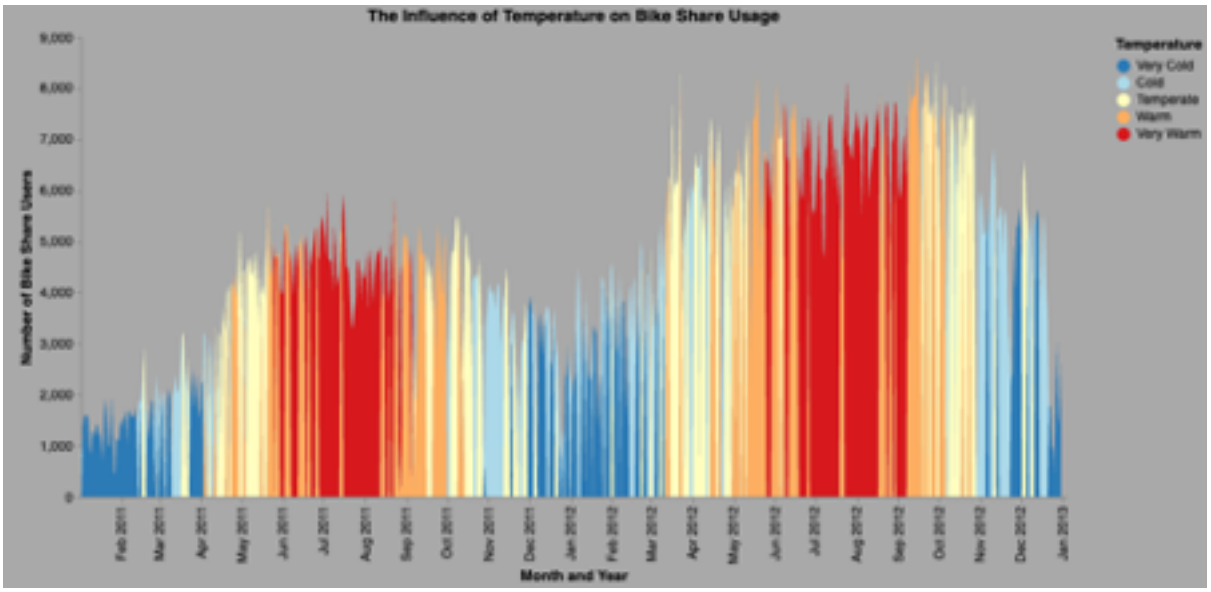
Interactivity in Prototype 2 is primarily driven by an interactive legend. By selecting this legend the user is able to filter the area chart to only display the selected binned data. This filter updates the user type bar chart so as to show differences in usage by specific user types based on temperature. It is also possible to scroll in and out of the area chart to get a zoomed in view of the data. Given this data set spans two years, there are potential occlusion problems given the amount of data shown. Giving users the ability to zoom in, alongside implementing an interactive tool tip for both the area chart and the scatter plots helps to alleviate this problem.

Finally, it is worth mentioning that this second prototype uses more individual graphical elements than Prototype 1, however these still remain within the generally accepted 5-9 "chunks" that human working memory is able to deal with.

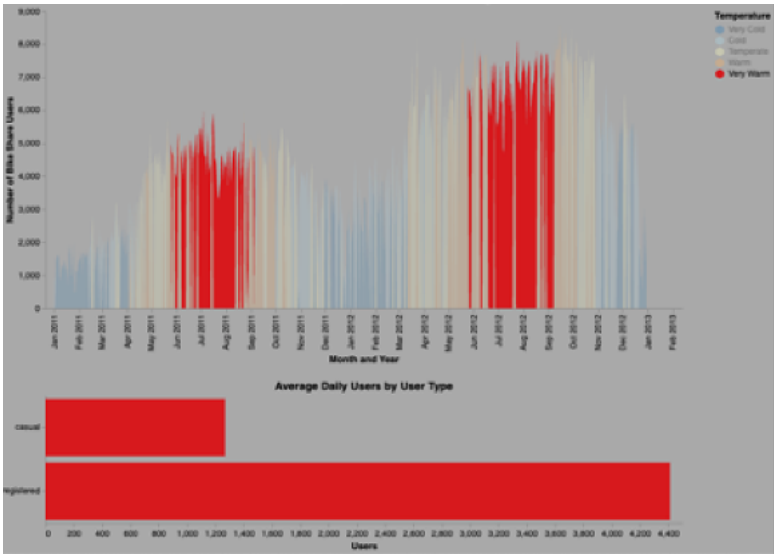
Data Discoveries

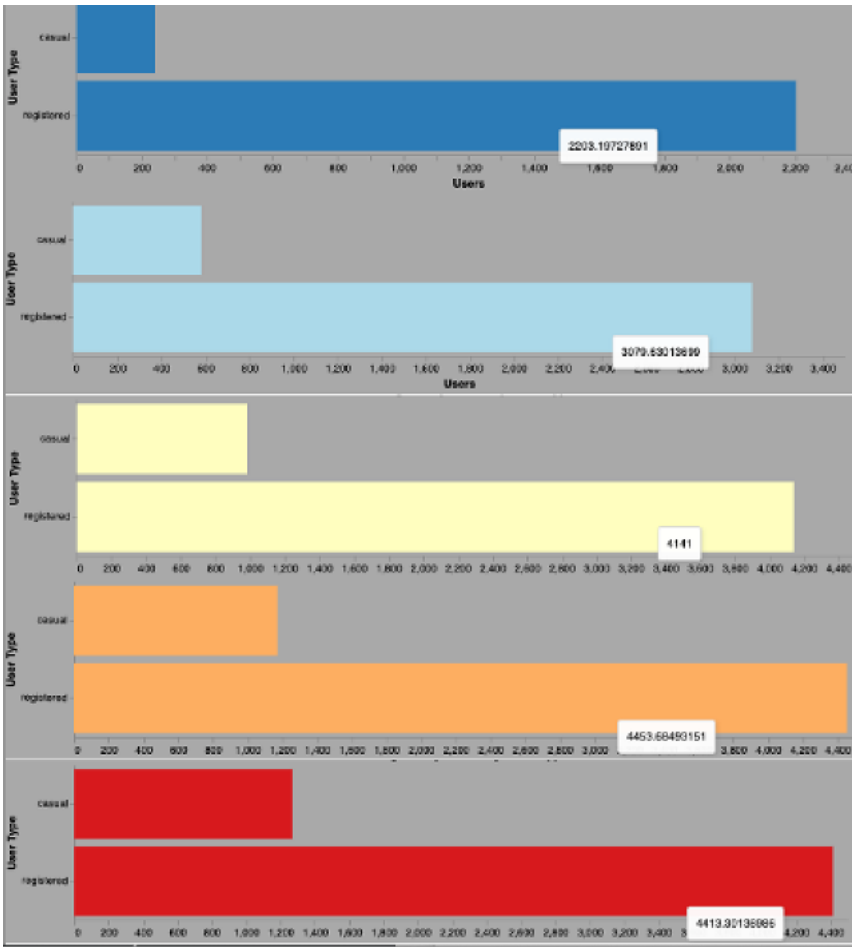
When looking at the data through the lens of the implemented prototype (Prototype 2), we can begin to gather various insights in to the data itself. The image below shows that by binning the quantitative data so that we have fewer and larger clusters, we are able to see that temperature does appear to influence usage in some way. Higher usage according to the area chart coincides with mid to high-level temperatures, coloured yellow, orange and red within the divergent colour scheme. Likewise lower levels of usage appear to be associated with colder temperatures coloured dark blue and light blue. Though it is evident from this visualisation that there is growth in overall usage over the course of the two years shown in the data set, looking at the time series area chart from this perspective we are

able to spot peaks and troughs relatively over time so that we understand them in context. Had the data not been binned, the resulting colour scheme would have been more chaotic and these trends would have been more difficult to spot.

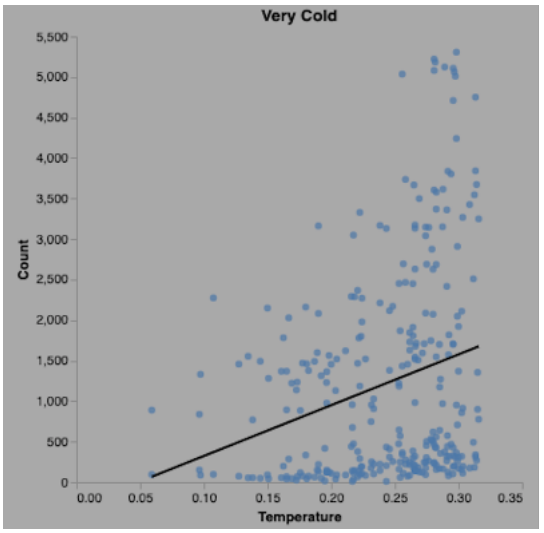


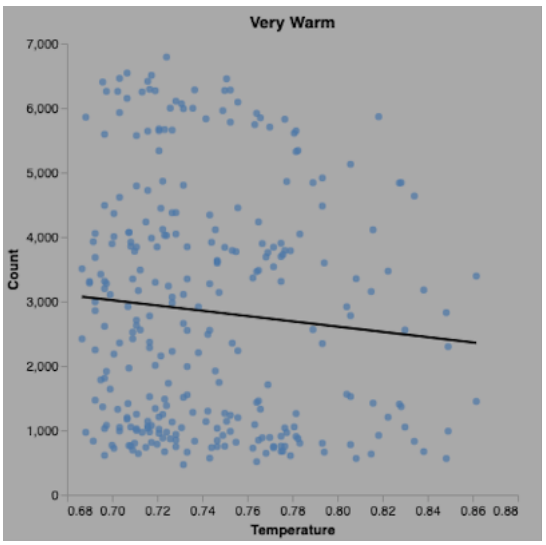
By using the interactive legend selection tool, we can see a number of key pieces of information. This dynamically updates the below bar chart so that it shows us the average bike rental usage for that temperature cluster across the two year period. This is also further broken down by customer type - those who are registered users with the service or those who are casual users. What is interesting about these averages is that they grow consistently as we move through from the lowest temperatures upwards. This is the case across both user types until we hit the highest temperatures. At this point, we see a slight drop off in registered users using the service.





Finally, the scatter plots show that, despite the previous findings, the relationship between the number of users and the temperature is less correlated than we might at first think, especially as the coloured area plot is the first thing that draws our attention. While there is a degree of correlation across some of the binned data, it is minimal. Perhaps surprisingly, the strongest positive correlation can be seen in the data from the coldest temperatures. This is particularly stark when the scatter plots for each cluster are viewed side by side, as we can determine that the steepest incline for a line of best fit with regard to usage is as we move through the "Very Cold" temperatures. Lines for "Cold" and "Temperate" weather could be considered as very minor inclines, to the extent of being nearly flat. This changes as we move in to "Warm" and "Very Warm" temperatures. The regression analysis for these shows that we see a reduction in usage as the temperatures get too hot. This is shown on the axis as negative correlation.





Implementation Notes

There are a few areas where it was not possible to fully implement some of the proposed features of the prototype. The first area is that we were not able to implement the bar chart text counts at the top of the bars, as outlined in the prototype. We believe that it was necessary to create a separate graphic for this and to layer this on top of the bar chart. However, when we did this it did not appear. This might have something to do with the labels potentially being hidden underneath the bars themselves, thus it might well be resolved by changing the colour of the text and nudging it along using dx or dy properties within the encoding itself. Similarly, we were not able to change the colour of the scatter plots or add tool tips. The initial intention was to map the scatter plots to the colour of the binned data to try to ensure consistency of colour. However, once transform regression was applied using Altair, it was not then possible to encode colour or tool tips directly within the graphics themselves. This again could be caused by layering constraints within Altair, and it may therefore be necessary to stipulate colour properties at another level of the graphic, be it local or global.

References

[1] Bike sharing data set. https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset.

[2] Cleveland, W. S., and McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association 79*, 387 (1984), 531–554.

[3] Cook, R. D. *Regression graphics: Ideas for studying regressions through graphics*, vol. 482. John Wiley & Sons, 2009.

[4] Ghanbari, M. Visualization overview. In *2007 Thirty-Ninth Southeastern Symposium on System Theory* (2007), IEEE, pp. 115–119.

[5] Healey, C. G., Booth, K. S., and Enns, J. T. Visualizing real-time multivariate data using preattentive processing. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 5*, 3 (1995), 190–221.

[6] Moreland, K. Diverging color maps for scientific visualization. In *International Symposium on Visual Computing* (2009), Springer, pp. 92–103.

[7] Munzner, T. *Visualization analysis and design*. CRC press, 2014.

[8] Robertson, G., Fernandez, R., Fisher, D., Lee, B., and Stasko, J. Effectiveness of animation in trend visualization. *IEEE transactions on visualization and computer graphics 14*, 6 (2008), 1325–1332.

[9] Shneiderman, B. The eyes have it: A task by data type taxonomy for. In *Proceedings of IEEE Symposium on Visual Languages*, vol. 96.

[10] Silva, S., Santos, B. S., and Madeira, J. Using color in visualization: A survey. *Computers & Graphics 35*, 2 (2011), 320–333.

[11] Talbot, J., Setlur, V., and Anand, A. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 2152–2160.

[12] Van Wijk, J. J., and Van Selow, E. R. Cluster and calendar based visualization of time series data. In *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis' 99)* (1999), IEEE, pp. 4–9.