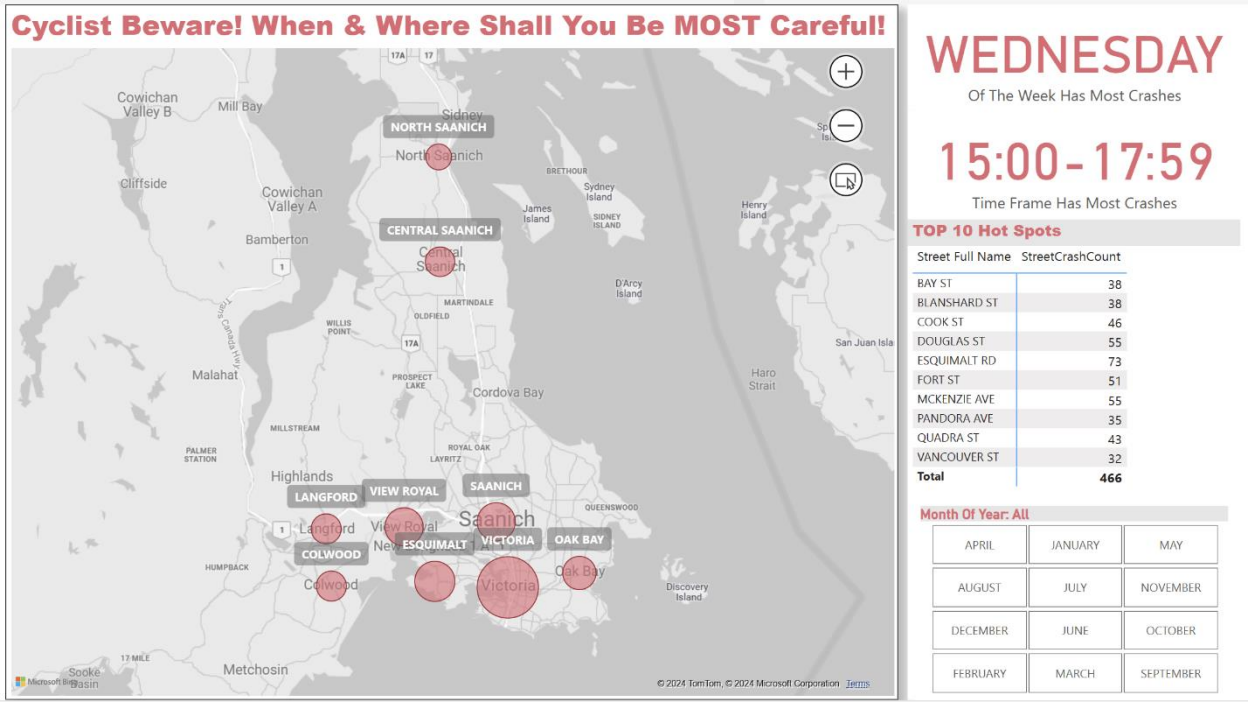
Project Assignment 4: Written Report [for graduate students]

Student name: Lepeng Zhou Student number: V01045967

1. Design Justification [2 pages max]

Choice of Visual Encodings

In my project "Cyclist Beware," I made several design choices to effectively communicate cycling safety data. I chose to use interactive maps with dots varying in size to represent accident rate. This choice was influenced by Tufte's principles of maximizing data ink and minimizing non-data ink, allowing for a clear and direct representation of accident frequency per capita. The use of dots vary in sizes serves as a visual metaphor for the magnitude of accidents, a technique widely used by many researchers and developers in corporations, although not the best to represent quantitative value, but it is the best I can have here for this map visualization. Dot map allows for immediate spatial recognition of patterns and anomalies, making it easier for cyclists to identify high-risk areas.



Layout

As for my project's layout, I decide to put the map on the left and information panels on the right. Such design is intentional, chosen for its simplicity and effectiveness in conveying detailed information without overwhelming the user. The design also has considered the flow of human attention is naturally from left to right, and top to bottom. With this in mind, this layout ensures that the user's attention is first captured by the overall spatial distribution of accidents before diving into specifics.

Interaction

Interactions such as pan, zoom, and the lasso tool were also used based on Shneiderman's "Taxonomy for Information Visualizations." [1] These elements allow active exploration of the data, enable users to uncover patterns and insights at their own pace and according to their interests. Interactions by intents such as select, explore, reconfigure, filter and abstract/details are also incorporated. Here user can select single or multiple municipalities by clicking on circle or using lasso tool. User can explore more by hovering mouse on circle to see more details about his municipality. User can reconfigure the crash count on each street by selecting and filtering. User can filter the data by month or municipality. The implementation of these features was inspired by the paper "Toward a Deeper Understanding of the Role of Interaction in Information Visualization", [2] which emphasizes the importance of interaction in information visualization for uncovering insights from data.

Style and Aesthetics

The style and aesthetics were deliberately chosen to be clear and non-distracting 2 color design, focusing the user's attention on the data itself. I am a believer in minimalism, therefore the map's minimalist design and the prominent red title, were made to ensure clarity and to draw attention to critical information. The use of varying dot sizes and only red colors was informed by the principles of pre-attentive processing. Ensure the fast accumulation of information from the environment, as red often associate with danger.

Alternatives and Justifications

Initially, I considered employing a heat map and filled map but ultimately chose the dot map. There are several reasons: 1, the heat map has vague boundary, hard to distinguish. 2, filled map requires different colors to make the boundary clearer, but this is against the minimalist design. 3, filled map can not offer visual variables like size, and each municipality has different area and fixed. This can associate with the accident rate, which is not we want here.

2. Reflection on Learning [1 page max]

For the "Cyclist Beware" project, I'd like to conclude that this has been an excellent exercise in applying visualization principles and a comprehensive learning experience. From creating ideas, drawing sketches, and receiving feedback, to learning the tools, building prototypes, and eventually finalizing them, each phase has been an enjoyable and creative way to learn visualization. I will summarize my learning as follows.

Regarding teamwork, I worked alone. This had its pros and cons. On the positive side, I could decide how much progress I wanted to make each day without needing to coordinate with others. This gave me full control over the design and ensured that all ideas were my own, allowing for greater freedom, creativity, and efficiency. On the downside, it meant more work for me, fewer opportunities for brainstorming, potentially missing out on better advice, and losing a chance to learn cooperation with others.

The ideation and prototyping phase involved learning about interactive visualization tools from extensive resources such as YouTube and websites that teach you how to use tools like Tableau and Power BI. Initially, I used Tableau but found its interface inconvenient and complicated, with many limitations on layout choices. Then, I switched to Power BI, which was much easier for Windows users to pick up, as it is designed by Microsoft and features a similar UI across all Microsoft products. For prototyping, I quickly experimented with all the tools offered and chose the best for my needs—in this case, a dot map over the heat and filled maps. I then added some widgets and filters to the side and wrote some DAX queries to calculate insights from the data. Of course, this process involved a lot of trial and error, but I overcame each challenge and chose the design that best fit my theme.

Choosing data from Vancouver Island published by ICBC was crucial. I am grateful to them for formatting the data and making it easy to preprocess, as cleaning the data was almost effortless. What I learned is that documenting your company's data is important; it can be used for many social goods, and good documentation can help others understand it faster, saving time that would otherwise be spent guessing each column's meaning.

3. Discussion and Reflection [2 pages max]

Discuss the strengths and weaknesses of your visualization solution [~0.5 page]

I would argue that my visualization solution has more weaknesses than strengths, to be honest. Many final designs result from compromises.

First, the software's limitations restrict how I originally envisioned visualizing the data. I wanted to mark the details of crashes on each street and, ideally, have the map highlight the entire length of each street. However, this goal was unachievable because the software only offers three types of maps: dot, heat, and filled. Choosing the dot map allows for marking each street with a dot, but since streets are long and narrow, the dot must be placed at the center of each street, making it difficult to identify which street the dot represents. Furthermore, overlapping streets and dots make the map very visually distracting and not appealing at all. The lack of glyphs is another limitation; in Power BI, the dot map can only use dots to represent points, varying only in size and color. As for street names, many in Victoria appear in the USA, so when importing data into the software, many streets are mistakenly marked in the USA.

The second weakness is that the zoom-in and zoom-out function does not meet my expectations. In my ideal scenario, the zoom feature would follow a classic overview+detail design, where zooming out provides municipality information and zooming in provides street information. Unfortunately, this is not achievable in Power BI software. There is no function associated with zoom level to set a threshold; before the threshold, a schema of municipalities could be shown, and after the threshold, another, but unfortunately, this is fixed so only one can be shown at a time.

The third weakness is that filtering by crash rate is not implemented due to the lack of screen space. After minimizing the size of the panels, they have all reached the smallest size they can be without needing to fold the filter. For example, the month filter is at its smallest; making it any smaller would turn it into a scrollable window where only 9 months are displayed, requiring the user to scroll down to see the remaining 3 months. Thus, there is no more space available.

As for strengths, quite obviously, it is minimalistic, visually appealing, easy to understand and read, and easily deployed to a website for cross-platform sharing.

Discuss its relation to existing visualization literature/related work [~0.5 page]

I would like to say that 2 papers I have read is the most helpful for my design, they are "A Framework of Interaction Costs in Information Visualization" by Heidi Lam [3] and the paper "Toward a Deeper Understanding of the Role of Interaction in Information Visualization" [2] by Ji Soo Yi. After reading them, the limitations I faced is also challenges in the visualization field. Lam's framework, particularly, sheds light on the decision costs involved in visualization design, highlighting the trade-offs between system capabilities and user goals. This framework suggests that visualization tools need to consider not only the visual representation but also the interaction costs associated with different design decisions. My project's limitations mirror the gulf of goal formation, where the decision costs in establishing a data analysis focus were constrained by the available software options. Also, the “Physical-motion costs to execute sequences” mentioned in the paper also informs me, to design a filter easy to use, avoided the complex scrolling feature. The filter selection in my project supports the notion that effective visualization tools should provide users with "rapid, incremental, and reversible operations that impact the visualization immediately, allowing users to explore alternative hypotheses and understand the

dataset in a hands-on manner" [2]. By enabling users to quickly adjust filters, the tool empowers them to get insights without the need of knowledge for complex query languages.

Discuss the usefulness of visualization to uncover insights about data in general [~1 page]

Compared to old ways like looking through big tables of numbers, making graphs out of data is a great leap forward. The less automated way takes a lot of time and makes your brain tired because you have to think so extensively about what all those numbers mean. It is easy to miss something important. And this is exactly what Lam has mentioned in his paper “A Framework of Interaction Costs in Information Visualization”. Although in his paper he talks about the cost in interaction, but here we can think the less automated way is also an interaction, but just involves much more cost than a more automated way, like my visualization.

On the other side, sometimes we let computers do all the work to find patterns in the data. This is strong but sometimes too mysterious. We might not understand how the computer decided something. This is where the visualization of the data helps again. It can show us what the computer found in a way we can understand. This helps make sure the computer is doing its job right and helps us trust what it says.

A relation to my career as a data scientist, machine learning is very extensively applied to uncover data insights, the result of machine learning is often the information we want to visualize but during the machine’s learning process, we often need to visualize the tuning and experimenting of different models and parameters. For a positive example, something called the Mean Square Error (MSE) which can tells whether a model is overfitting (data is too little and model is too complex) or underfitting (model is too simple). Just looking at the MSE from a table we might not uncover any thing interesting but only if when we plot it against the model complexity as x-axis, i.e. MSE as Y-axis, we can see the clear correlation and find the sweet point for model complexity. Another example is that automation (machine learning) is mysterious. In statistic, one of the main reason linear regressions is still popular today, even it has so many constraints, is its explainability. There are so many models outperform linear regression such as Support Vector Machine (SVM), Neural Network. But for many scientists it is like a black box, you put something in and get something out, tuning the black box in someway until you get what you want. Nobody really can explain in math equations of the principle of how it works. Linear regression, however, can be explained in statistic. You can design a F-test to prove model A is better than B, to prove the increase of bias will decrease variance, and to argue one data entry is an outlier with statistic support. Thus, when I need to explain to people that this model is the best, I have a lot of data to visualize, I can select a subset or all of them to tell the story. Visualization also helps the cleaning and preprocessing of data at early phases. I can plot out all data and select and remove outliers. using Fishers Linear Discriminant (FLD) to optimally reduce the dimension. But all these are at a premise that I plot the data first to visualize the nature of data, like variance, distribution.

Looking ahead, visualization has much potential waiting me to discover, whether I want to work in health care, financial, there are always places interactive visualization can play a big role, just don’t make the interaction too complicated and follow the cost guidance by Lam [3] and I am good to go.

4. Appendix [Not included in page limits]

References

[1] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," Proceedings 1996 IEEE Symposium on Visual Languages, Boulder, CO, USA, 1996, pp. 336-343, doi: 10.1109/VL.1996.545307.

[2] J. S. Yi, Y. a. Kang, J. Stasko and J. A. Jacko, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp. 1224-1231, Nov.-Dec. 2007, doi: 10.1109/TVCG.2007.70515.

[3] H. Lam, "A Framework of Interaction Costs in Information Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 14, no. 6, pp. 1149-1156, Nov.-Dec. 2008, doi: 10.1109/TVCG.2008.109.