Visual Analytics Assignment - 998273

1. Dataset

The dataset that will be covered in this assignment is "Forest Fires" dataset. The dataset contains 13 attributes, 4 categorical attributes and 9 numerical. The dataset contains spatial information such as X and Y coordinates, time information like month and day, meteorological data such as wind, temperature, rain and relative humidity and lastly, four different indexes which is Canadian Forest Fire indexes that are related to the fire intensity, spread and potential. By visualizing this dataset, the user will be able to see the Montesinho park from a top view and based on their knowledge to comprehend the causes of fire, identify endangered areas that might need extra attention, evaluate the annual condition and predict potential fire places based on the factors that are available to them.

The dataset was provided by Cortez and Morais [1] and the source file with an explicit description of its features can be found in: https://archive.ics.uci.edu/ml/datasets/forest+fires

2. Visualization designs

2.1 First Design

Figure 1. First Design for the data set of Forest Fires

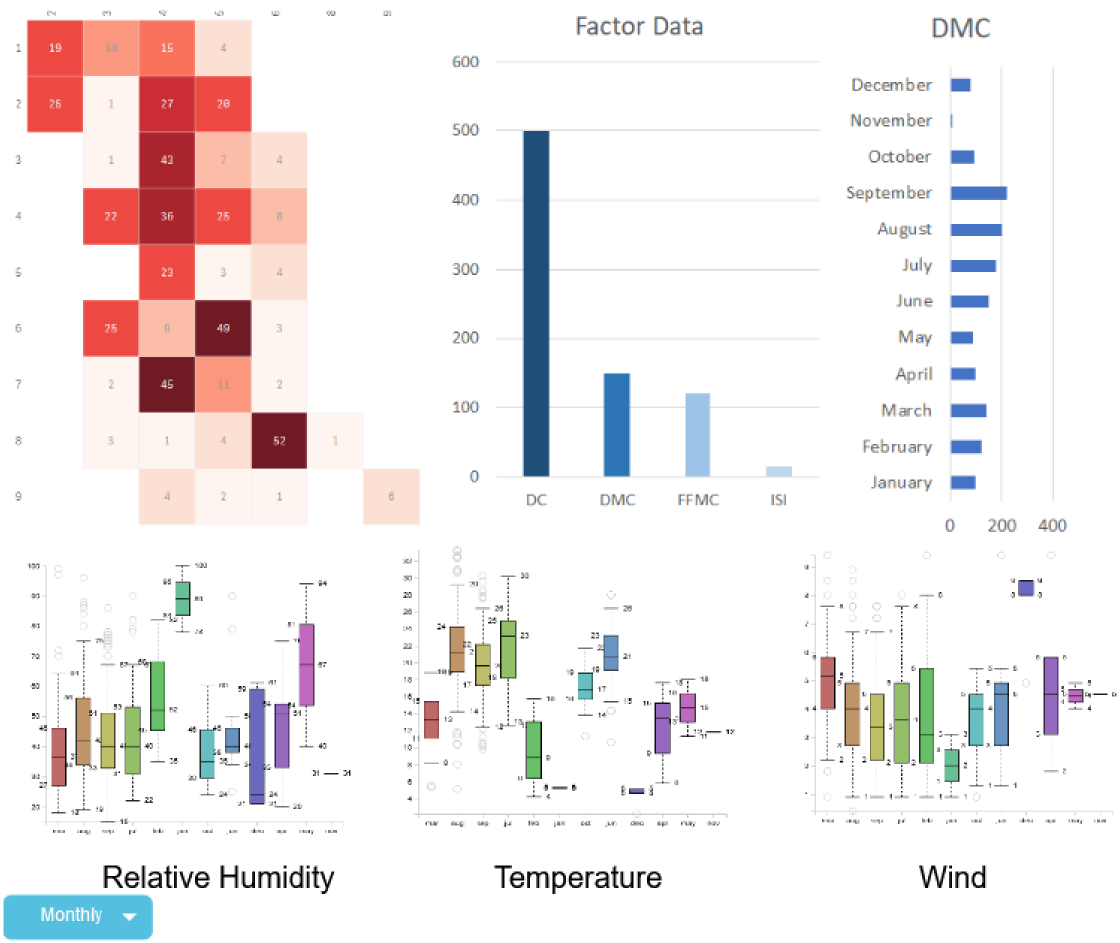


Figure 1 illustrates the prototype of the first design. The design is constituted by five graphs, one heatmap one bar chart and three boxplot charts. The heatmap is presenting the incidents on the park map from a top view. The luminance helps the user to identify areas with high fire frequency. Heatmap in combination with color luminance tends to perform adequately when spatial information is represented [2]. Moreover, the color hue has been chosen as red due to its intensity, as humans seem to connect it with caution and it attracts the attention of the user [2, 5]. Also, the burned area is now covered as text but in the final implementation it will be a diamond inside each block with different size. Although, humans are not so good at precepting area [2], only four different classes will be created between some thresholds. Thus, the user could comprehend the size of burned area without difficulties.

The other 3 boxplot charts have a similar axis which is the months of the year and there are 2 bar charts to illustrate the fire factors in different distributions. In top-right bar chart is shown one of the 4 different factors that are relative to the fire and their monthly distribution. The other 3 factors are not visible, and the user can see them by clicking the middle-top bar chart. Due to for both bar charts the X axis is a categorical one, whereas Y axis is a numerical one, the bar chart seems a decent choice for representation according to Munzer & Maguire, (2015) [2]. Furthermore, in the boxplots the user can see monthly statistics about the meteorological data like wind, relative humidity and temperature that might help them to identify patterns and possible relations to the incidents of fire, a different color hue also corresponds to the count of monthly fire incidents. It worth mentioning that rain has been omitted from visualization because all the values are zero and the representation does not offer any practical value. The days also have been omitted because they create more noise in the charts without serving a useful purpose to the analysis. Additionally, to this the monthly statistics can be shown as daily distribution in order to make the identification of patterns easier by switching with the dropdown menu on the bottom left. The choice of boxplots has been done because it is a common practice in meteorological data [4]. Lastly, most of the graphs are interactive. Every time the user presses a box in the heatmap, they acquire the respective statistics to this box, and statistics is shown in the other graphs as well. Furthermore, the user can do a multiple selection of more than 1 box by holding Shift key so an inspection could happen to sub-areas of the park instead of specific places exclusively. Interaction not only does allow the user to inspect specific areas on the heatmap on demand but also as Kerren & Schreiber, (2012) indicate, it consists a fundamental tool to support the analysis process [6], because of this it consists a key element of the current visualization.

To sum up, the above design although it has some drawbacks, it could help both the exploration and the comprehension of the dataset which is equipped with proper color pallets and can cover almost all possible comparisons that might need to be done during the analytical process.

2.2 Second Design

Figure 2 Second design of visualization for the forest fires dataset

In the second design in Figure 2, visualization is quietly different. This design is based on timely distribution over spatial despite that spatial information is contained in it. First and foremost, the user can see both monthly and daily distribution of the meteorological factors, using the calendar view and the top-down bar chart which are saturated and desaturated properly. Moreover, the main component is the tree-map that shows the average burned area distributed by month. Interaction on the tree-map has changes over the heatmap, indicating the place where the fires happened that month with different hue based on the frequency of them. Lastly, on the bottom left there is the monthly distribution of the 'DMC' factor in a bar chart. The 'DMC' factor can be changed to 'ISI', 'FFMC' and 'DC' using the dropdown menu just above of the chart.

The tree map is the main component of the design in which the whole visualization process is based on. In contradiction to a pie chart, a tree map can show areas with really small size, in this case "january", "may" and "november", and can make the visualization more neatly [8]. The bar chart even though is on a reverse row serves the same purpose as a decent option for visualization [2] and also gives to the user a sense of flow throughout the year regarding to average temperature or any other meteorological data. Nevertheless, the user cannot exploit all the available analytics like mean, first and third quartile and so on. On the other hand, the heatmap which has already been used on the first design as the main component, in this case is created only by the interaction of the user and it is non interactive besides the statistics that could be shown on it. The main reason that has been chosen twice is that spatial information is presented with the optimal way through a heatmap, especially when map is not an option [2]. The days are presented in a form of a clustered calendar view which serves the user task at this point even though it is used more on time series [7]. Finally, the fire factor data (i.e. 'DMC', 'DC' and so on) is illustrated through a bar graph in monthly sequence.

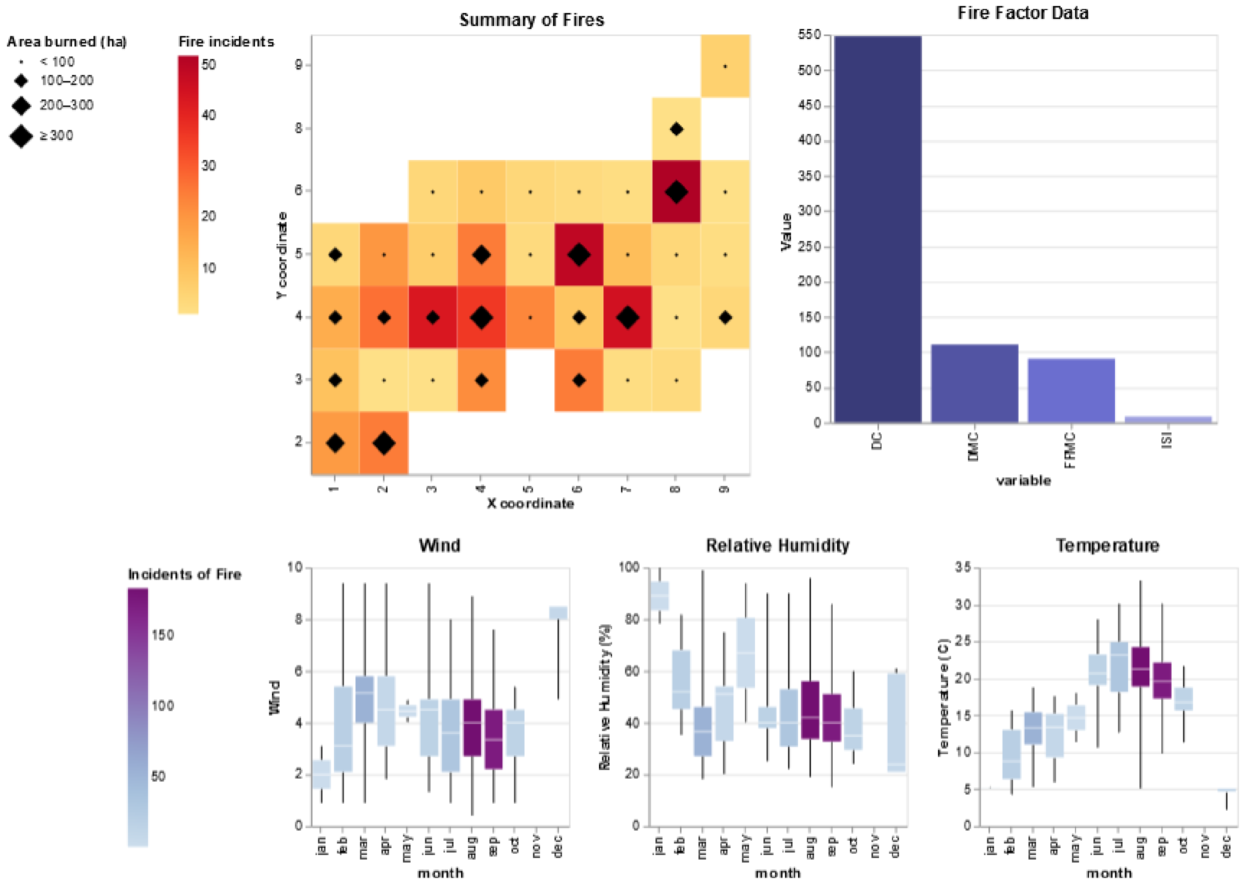
Ultimately, this design is more based on the monthly evaluation of the park and spatial information has a less significant role on it. The role of this visualization is the identification of patterns that have their core on specific periods of the year based on the meteorological data and other factors.

3. Report of Findings through the exploration

The design that has been chosen to be implemented is the first design as it serves the purpose of the visualization better due to the core task is based on spatial information instead of the monthly distribution. Moreover, the first design seems more neatly, and it might be more efficient as far as time is concerned.

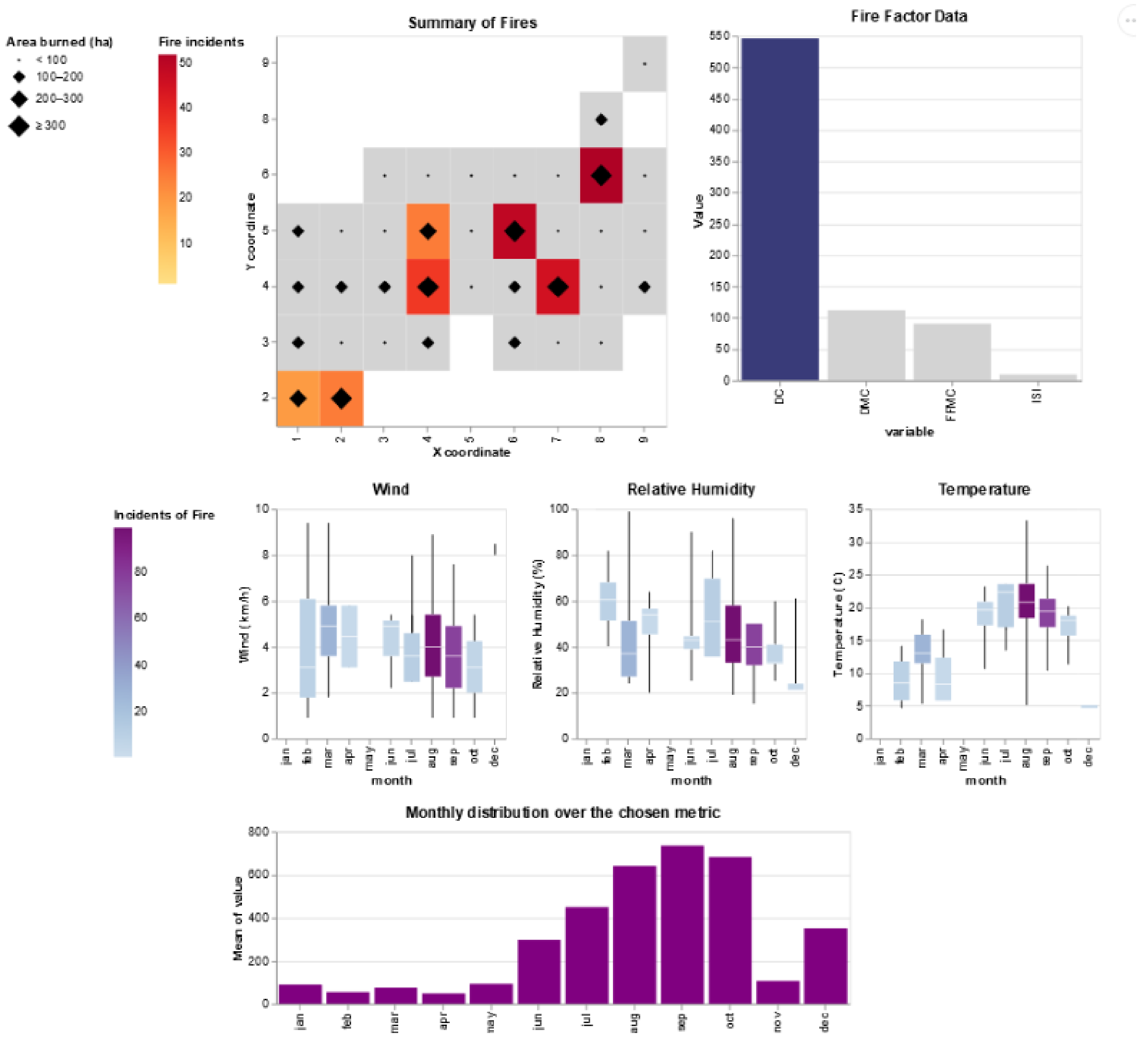
Overall, the findings show that the bottom-left side of the parked has suffered more fires and has the most burned area by now. The months which had the most fire incidents was September and August and simultaneously, those months had quietly low relative humidity and high temperature on average compared to the other months of the year as it is shown in Figure 3. The wind for this period was fluctuated between average values in comparison with the whole year.

Figure 3 Overall statistics for the park



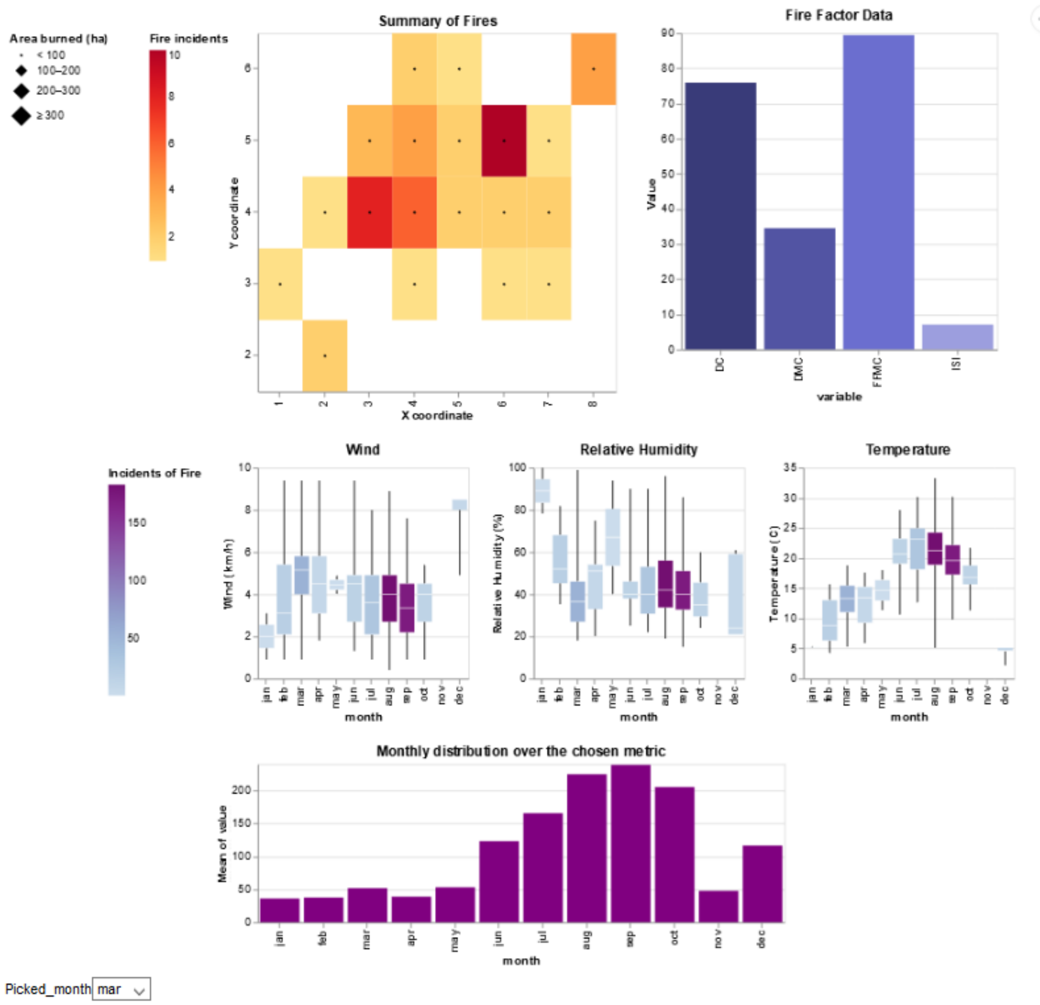
However, the wind seems to influence more the spread of the fire instead of the initial start of it in comparison to the temperature which is known that increases the chance of a fire in an area [1]. Additionally, if only the areas with the highest amount of burned area are chosen (Figure 4), it is shown that average relative humidity for those areas are significantly lower than the total average of them. Especially in August and September which the most incidents took place.

Figure 4 Chosen places with the most burned area



Also, average DC seems to be above 550 which based on the rating system for Canadian Forest Fire indexes is characterized as "extreme" chance of fire [9]. Nevertheless, the other factors do not consent to "extreme" rating but could be characterized as "high" based on the other metrics [9]. In comparison with those places those with the least burned area presented different results which are somehow expected based on the outcome. Another thing that could be mentioned is that the 3rd month with fire incidents was March with approximately 100 incidents. Despite that fact, as it is shown in Figure 5 March had one of the lowest burned area and the respective metrics for this month tend to be quietly low in comparison with August and September, being almost one order of magnitude lower than the latter. This is contradicting to the other studies that showed high frequency and burned area in March [9].

Figure 5 Metrics and areas that suffered fire in March



Ultimately, in this visualization design user is able to demonstrate both spatial and time information to comprehend better the possible causes of fire incidents and to have aggregated statistics about specific places of the park regarding to all possible metrics that are needed to be measured.

4. Bibliography

[1] P. Cortez and A. Morais. (2007). *A Data Mining Approach to Predict Forest Fires using Meteorological Data*. In J. Neves, M. F. Santos and J. Machado Eds., *New Trends in Artificial Intelligence*, *Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães*, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9

[2] Tamara Munzer, Éamonn Maguire, *Visualization Analysis and Design*, 2015, AK Peters Visualization Series, CRC Press, Taylor & Francis Group, ISBN- 9781466508910

[3] B.D. Amiro, K.A. Logan, B.M. Wotton, M.D. Flannigan, J.B. Todd, B.J. Stocks, D.L. Martell, 2004, "*Fire weather index system components for large fires in the Canadian boreal forest", International Journal of Wildland Fire*, 2004, 13, pp 391-400

[4] Banacos, Peter. (2011). Box and Whisker Plots for Local Climate Datasets: Interpretation and Creation using Excel 2007/2010.

[5] Fugate JMB & Franco CL (2019) "*What Color Is Your Anger? Assessing Color-Emotion Pairings in English Speakers*". *Front. Psychol*. 10:206. doi: 10.3389/fpsyg.2019.00206

[6] Kerren, Andreas & Schreiber, Falk. (2012). Toward the Role of Interaction in Visual Analytics. Proceedings - Winter Simulation Conference. 420. 10.1109/WSC.2012.6465208.

[7] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, San Francisco, CA, USA, 1999, pp. 4-9. doi: 10.1109/INFVIS.1999.801851

[8] Tarrek Azzam, Stephanie Evergreen, "*Data Visualization, Part 1*", Number 139, Fall 2013, American Evaluation Association, A Publication of Jossey-Bass and the American Evaluation Association, pp 39-42

[9] Ullah, Mohammad Rahmat & Liu, X.D. & Al Amin, Prof. Dr. Mohammed. (2018). Spatial-temporal distribution of forest fires and fire weather index calculation from 2000 to 2009 in China. Journal of Forest Science. 59. 279-287. 10.17221/8/2013-JFS.