Sales visualization

BOULESBAA Ishaq*
Data science student

ABSTRACT

Building construction is an ancient human activity. The total annual value of building construction in the various national economies is substantial. The visualization of sales data is an important tool for people of this domain like investors to understand and make a better decisions.
In this article we are interesting by the visualization of sales data and more precisely, the data of houses sale. It is a multidimensional datasets, so we will present related work in domain multidimensional datasets visualization, then we will describe our project and conclude with a summary and our plans for future work.

1 INTRODUCTION

Developing novel visualization techniques has long been one of the core activities in information visualization research. Each year new visual representations for new types of data are introduced, each one more complex than the next.
Visualizations differ according to the types of data, the domain and the audience. In this article we are interesting by the visualization of sales data more precisely, the data of houses sale, it is a multidimensional datasets and the audience are the investors in the domain of constructions building.
Building construction is an ancient human activity. The total annual value of building construction in the various national economies is substantial. In 1987 in the United States, for example, it was about 10 percent of the gross domestic product [2] witch illustrate the importance of this domain.

In this work, we are interested in preserving the simplicity and familiarity of line charts and Multi-level pie charts while addressing their shortcomings. We present a method based on those techniques for visual exploration of multidimensional datasets of houses sale using interactions and structured navigation in data dimension space, trading visual complexity for interactivity to reach the targeted audience.

We are going to talk first about related work, how does enterprise visualize their sales data, then we will give the project description where we talk about the purpose of the project, the choice of datasets, preprocessing and attribute selection and the visualization components. We will conclude with a summary and our plans for future work.

2 RELATED WORK

One of the important work interesting by the visualization of multidimensional datasets is SCATTERPLOT MATRIX.
The purpose of SCATTERPLOT MATRIX is to support structured visual exploration of multidimensional data using scatterplots. Instead of letting the user choose mappings for the axes of a single scatterplot, we create one scatterplot per every combination of dimensions and arrange them in a large scatterplot matrix.

*e-mail: bi_boulesbaa@esi.dz

The whole matrix serves as an overview of the dataset and also defines a visual space for navigation, turning the visual exploration process into a navigation task. Transitions from one scatterplot to another is performed using a 3D rotation that is consistent with the overall navigation metaphor and that provides more natural cues than standard interpolated animation.
Naturally, the order of columns and rows in the scatterplot matrix is significant, and can either be computed automatically as a function of the similarity between dimensions or the degree of visual clutter, or manually by the user through drag-and-drop of rows and columns in an interactive mode [1] .

An other work could be used in a similar situation is radar chart which is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point. The relative position and angle of the axes is typically uninformative.

We can't use these visualization because we have a lot of variables of different types.

3 PROJECT DESCRIPTION

3.1 Project purpose

The project is the visualization of sales data of the houses of King Country,USA.
The audience are the investors in the domain of building constructing, and people interesting in prices of the houses in King Country.
The purpose is to give visualizations clears and easy to understand.

3.2 datasets choice

Building construction is an ancient human activity. The total annual value of building construction in the various national economies is substantial. Like we see in figure 1, the construction output of the USA is 599 billion dollar, second after India, witch illustrate the importance of this field in USA.
King County is a county located in the U.S. state of Washington. As a result of the 2010 census its population was 1,931,249. King is the most populous county in Washington, and the 13th-most populous in the United States. The county seat is Seattle, witch is the largest city in both the state of Washington and the Pacific Northwest region of North America. [3].

For these reasons, we choose this datasets, which contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.
Attributes explication :

- Sqft_living, the total house square footage of the house.
- Sqft_basement, size of the basement.
- Sqft_above = sqft_living - sqft_basement.
- Sqft_lot, lot size of the house.
- Sqft_living15, the average house square footage of the 15 closest houses.

- Sqft_lot15, the average lot square footage of the 15 closest houses.
- Condition:
  - 1 = Poor- Worn out. Repair and overhaul needed on painted surfaces, roofing, plumbing, heating and numerous functional inadequacies.
  - 2 = Fair- Badly worn. Much repair needed. Many items need refinishing or overhauling.
  - 3 = Average- Some evidence of deferred maintenance and normal obsolescence with age in that a few minor repairs are needed, along with some refinishing.
  - 4 = Good- No obvious maintenance required but neither is everything new.
  - 5= Very Good- All items well maintained, many having been overhauled and repaired as they have shown signs of wear

3.3 Preprocessing and attributes selection

Before drawing visualizations, we must first of all to prepare data. We must make sure that data are clean(no missing values, no wrong value(wrong type ...).
Ones data are clean, we have a big matrix, raws are homes and columns are attributes. In our dataset, we have twenty one attribute, this a big number and it is difficult to visualize them all.
Beyond choosing a more complex visual representation, or using scatterplot Matrix Navigation [1] , the standard solution to this problem is to only visualize a subset of the dataset dimensions.
Using complex visual presentation isn't the best way to help investors in the domain of construction to understand sales data of King country. If we chose to use scatterplot Matrix Navigation, each visualization of two dimensions will be clear and easy to understand but we will have 21*21 visualization. So the best solution is to only visualize a subset of the dataset dimensions.

Several statistical methods have been designed to find the most interesting 2D planes and to create 2D views that best summarize a high-dimensional dataset; they are globally called dimensional reduction methods. Among these methods, Principal Component Analysis (PCA) is the simplest and most popular, PCA suffers from several pitfalls, as described by Koren : it is very sensitive to outliers and to artifacts in the data ,give no guarantee that all the interesting planes will be found, nor that all planes found will actually be interesting. due to these reasons , we can't use theses methods.

Our method consist of 5 steps :

- Consulting experts of the domain, then try to find the most significant attributes, for an investor, it's clear that the price, the superficies and the condition are very important, then we find number of floors, superficies of the basement that are important, we find too number of bathroom and bedroom which less important and there is attribute not important like the identifier of the house.
- We noticed that there is some attributes have the same value for majority of the houses like waterfront and view, so we conclude that these attributes are less important than the others.
- We noticed too that some attributes are deductible from others like: sqft_above = sqft_living - sqft_basement.
- Ones we found the attribute the most important, we can also apply some modifications, like :
  - Replace the superficies of the basement with an attribute indicate that there is a basement ou not.
  - Replace the sqft_living and sqft_lot with the attribute sqft_li_lot = sqft_living / sqft_lot .
- After the past steps to select and create the following attributes :
  - price.
  - Date.
  - floors.
  - sqft_above.
  - yr_built.
  - sqft_li_lot.
  - avec_sans_bas.
  - condition.

  , the final step is to divide the dataset to multiple sub-datasets, because for example we will not compare houses in perfect conditions with houses in bad condition, but we will compare houses in similar conditions.

3.4 The visualization components

Line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.
A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.
Multi-level pie charts are a set of concentric circles which is used to visualize hierarchical relationships.

The visualization contain four components, three line charts and a multilevel pie.
The three line charts are interactive, and the multilevel pie is linked to the line chart that visualize the price according to the factor : living / lot.
Next we will explain each visualization and its importance.

Figure 2 give the price of house according to the portion the house living surface divided by the lot surface according to the number of floors, this visualization too give a lot of interesting informations like the number of house per portion living/lot surface. This multi lines chart is connected to multilevel pie chart, by choosing a point from a line, the pie change.

Figure 3 is multilevel chart illustrate the division of houses following the presence or absence of the basement and the number of floors.
The first level show the portion of houses having or not a basement. The second level show for each partition of the first level, the partition of houses following the number of floors.
This visualization is connected to the one in figure 2, when the user selection a price , the multilevel pie chart change,the new chart illustrate the repartition of houses following the presence or absence of the basement and the number of floors, the prices of this houses are around the price selected (+/-) 10% and around the portion li/lot selected (+/-) 0.2 .

Figure 4 visualize the price of sale of houses in time, this is very importance, because it give a first idea about houses sales, a clear idea understood by everyone. As explained by Matthew Ericson during his keynote address at IEEE InfoVis 2007, scatterplots are

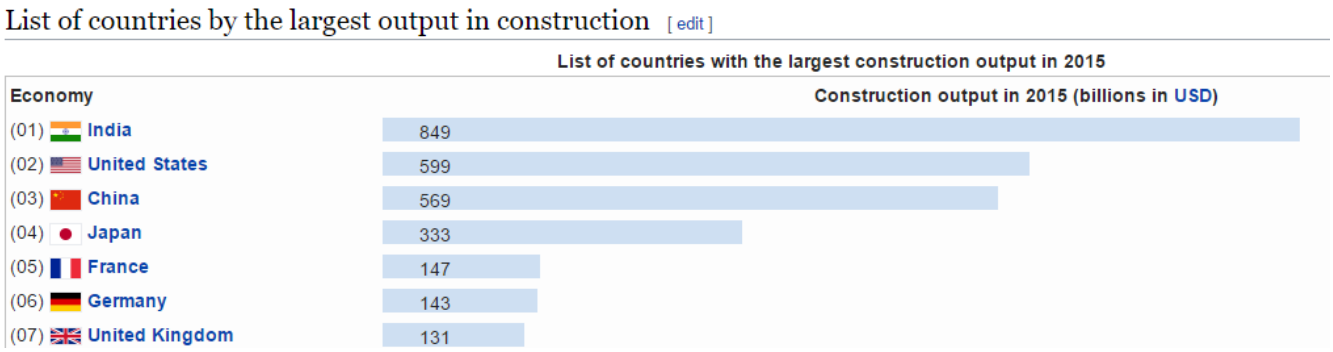

Figure 1: List of countries by the largest output in construction (2015). The image is from [4] and is in the public domain.

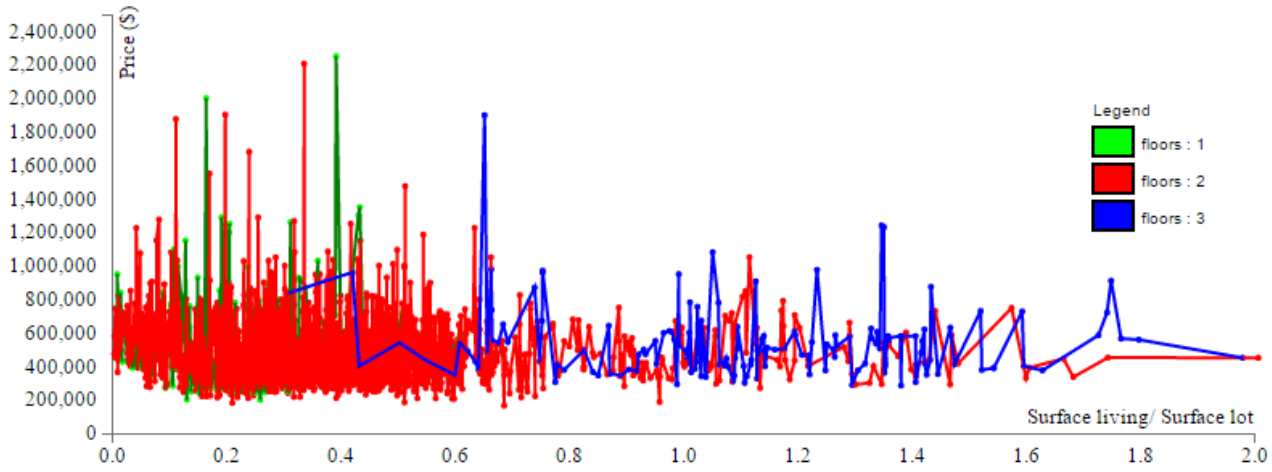

Figure 2: house prices according to the factor sqft_living/sqft_lot and le number of floors.

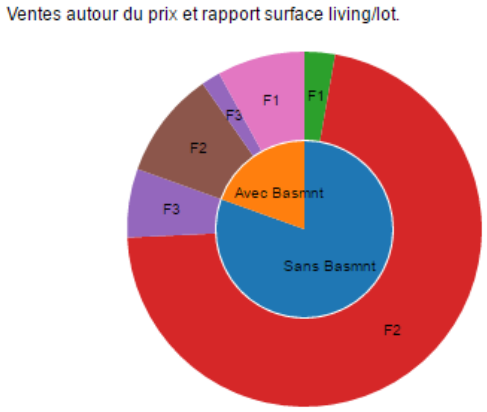

Figure 3: Percentage of house according to the presence or not of basement and le number of floors.

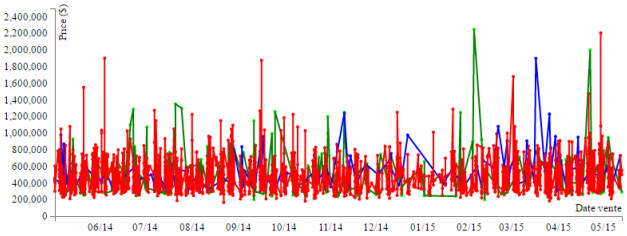

Figure 4: Price of houses according to de date of sale

Figure 5 visualize The average of house prices according to the presence or not of a basement and le number of floors.
Seeing this visualization as an investor, give you a lot of conclusions, for example: in average,the price of a house of one floor with basement is less than the one without basement, this information should be interesting when you want to build houses of one floor.

considered too difficult to understand for readers of the New York Times, except when one of the axes is time [1], the same thing for line chart which explain the need of these visualization.

4 Discussion

The visualizations need to be more interactive, so we must give to the user more options, like the possibility to select some attributes

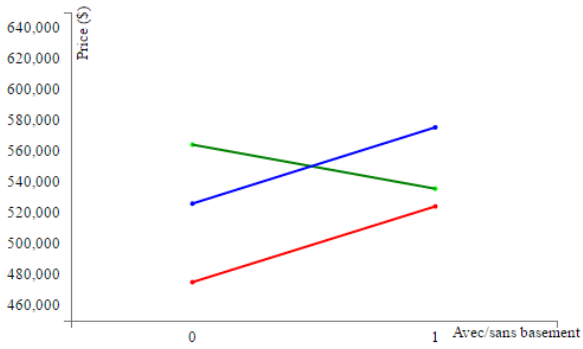

Figure 5: The average of house prices according to the presence or not of a basement and le number of floors.

or zooming or specify interval of the values of an attribute.

5 Conclusion

In this paper, we have presented a visualization of sales data. We used multi lines chart and multilevel pie chart to make the visualization. and we illustrate how to select important attributes.

In the future, we envision validating this design and new components that give more interactivity and more informations to the audience.

References

[1] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE transactions on Visualization and Computer Graphics*, 14(6):1539–1148, 2008.

[2] global.britannica. The economic context of building construction, January 2017.

[3] wiki. King country, January 2017.

[4] wiki. List of countries by the largest output in construction (2015), January 2017.