Public Notice Analysis: Final Project Report

Krapi Shah (111464159)
Balraj Inder Kaur Mahal (111498697)

Introduction :

Visualization is instrumental in inferring the trends from the data, spotting outliers and making sense of the data points. In this project, we will be analyzing data obtained from mypublicnotices.com which provides online access to public notice advertisements from across the U.S.

As a part of this project, we have created two dashboards - overall and statewise, that are interactive and user-friendly. Both these dashboards have complete brushing and linking support. Apart from the dashboard, we have also used parallel coordinates to highlight the relation of public notices and socio-economic factors like population, income.

Dataset collection and preprocessing have already been explained above. Here we will focus mainly on the techniques used for visualizations and the analysis that we were able to produce.

Project Setup (MongoDB, Flask, Python, D3, jquery):

- MongoDB: As discussed previously we had set up a mongo database for storing and querying data.
- Python3.7: Python 3.7 has a library called PyMongo for connecting to MongoDB and querying the data. Apart from that it also acts as a service layer providing various web services calls that can be made from the web app.
- Flask is a microframework that comes with very basic stuff required to get a web app up and running as fast as possible.
- D3.js: A JavaScript library for controlling the data and building charts
- Dc.js : It allows highly efficient exploration on a large multi-dimensional dataset. DC.js acts as a glue between Crossfilter.js and d3.js allowing instant feedback on user interaction.
- Crossfilter.js: It allows exploring large multivariate datasets in the browser. Crossfilter supports extremely fast (<30ms) interaction with coordinated views, even with datasets containing a million or more records.

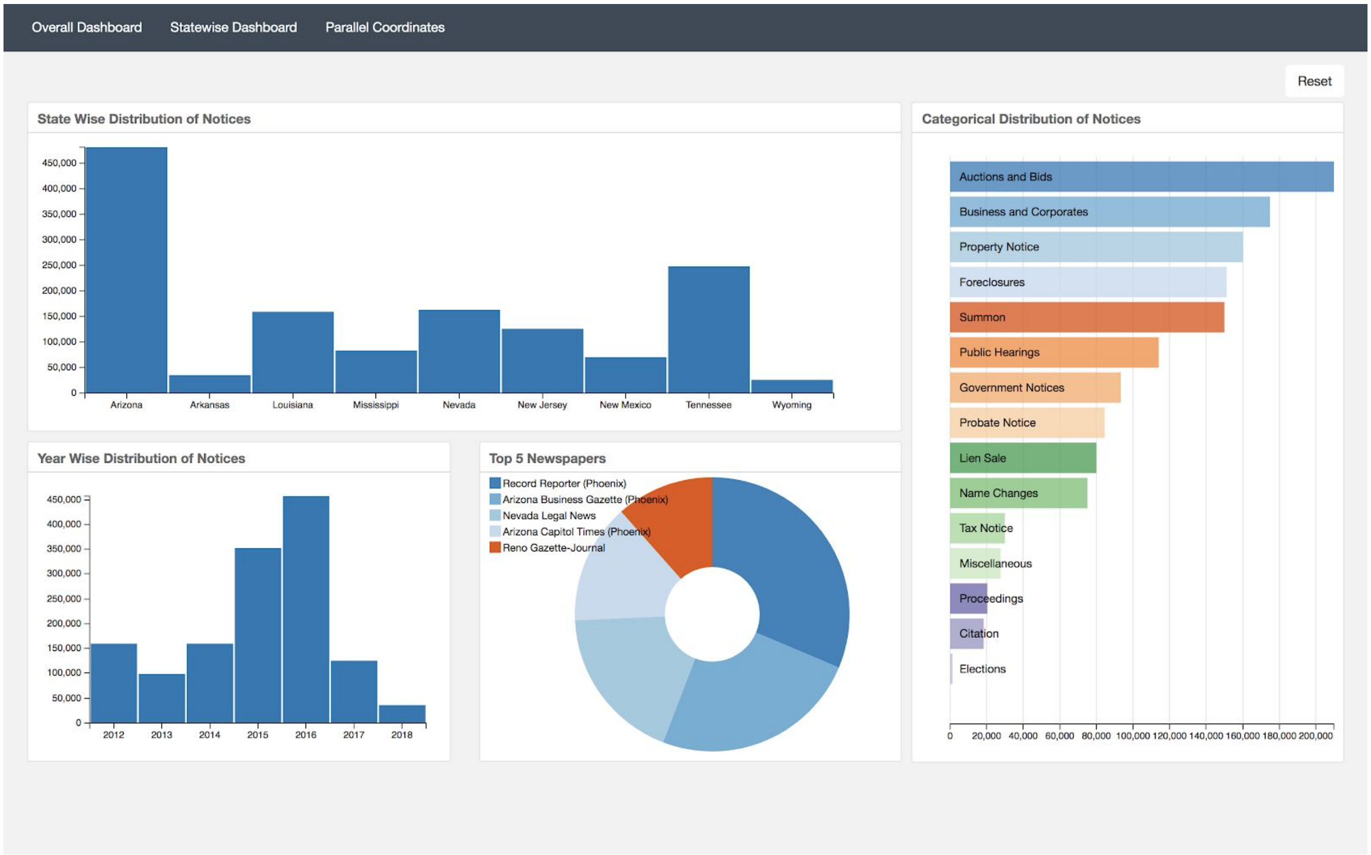
For the progress report, we didn't make use of dc.js or crossfilter.js and were facing issues with creating an interactive dashboard. Dc.js and Crossfilter.js allowed us to ease filtering the data and hence helped us in creating the dashboard in a much effective way.

Visualizations and Analysis:

We have created two dashboards: 1. Overall analysis for public notices
2. State-wise analysis for public notices

Overall Analysis :

This dashboard can be seen in the figure below :



This dashboard consists of 4 main charts :

1. State-wise Distribution of Notices : This is a simple bar chart explaining the distribution of notices across various states. The states out data consist of includes: Arizona, Arkansas, Louisana, Mississipi, Nevada, New Jersey, New Mexico, New Jersey, Tennessee, and Wyoming.
2. Year Wise Distribution of Notices: We have used a bar chart to see the distribution of notices across years from 2012 - 2018.
3. Top 5 Newspapers - We have used a pie chart to represent the top 5 newspapers that are used for publishing public notices
4. Categorical distribution of Notices: We used a horizontal pie chart to category wise distribution of notices.

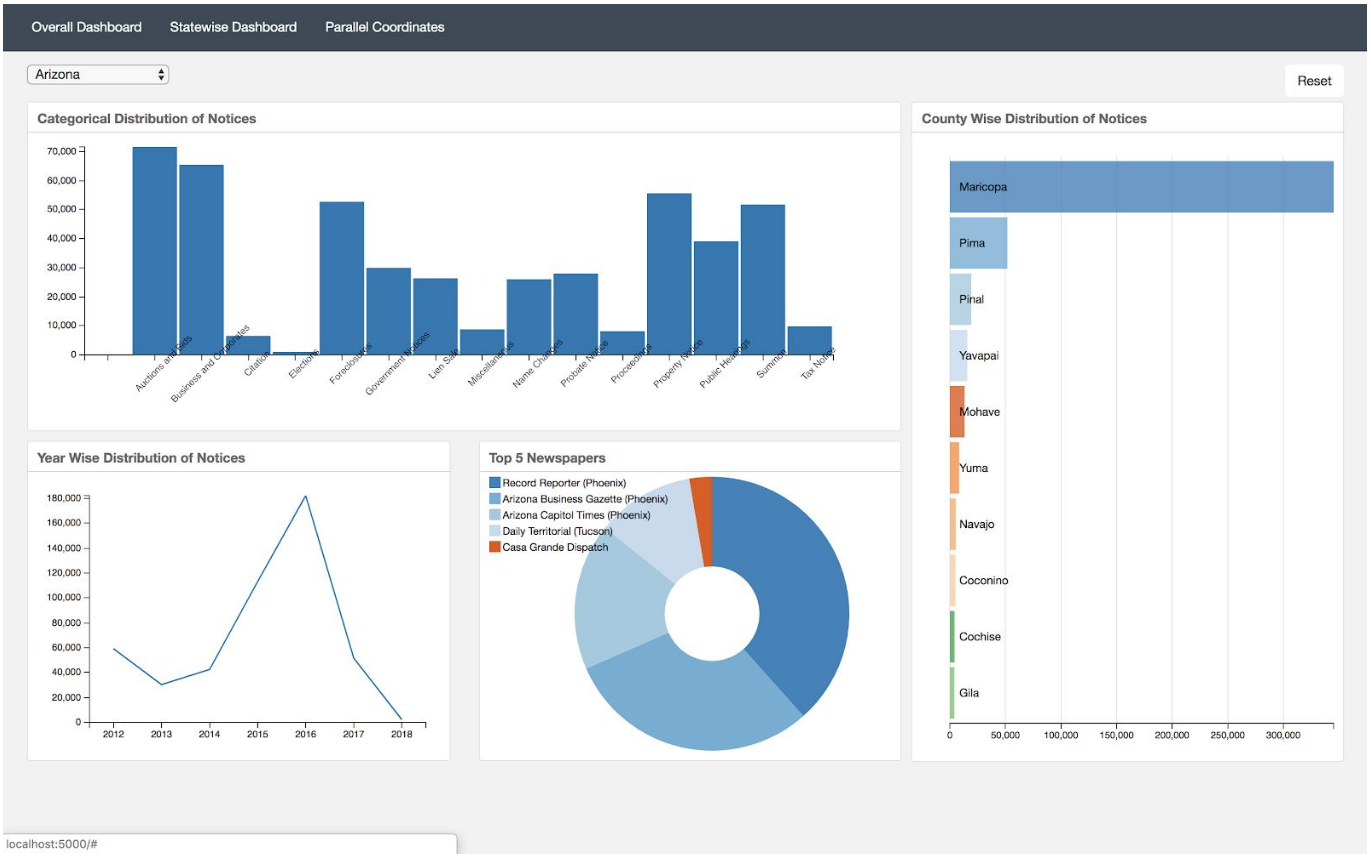
All the charts are created using dc.js. Clicking on a particular bar or slice on the dashboard, updated all other charts filtered on that particular bar or slice.

State-wise analysis:

A state can be selected from the drop-down option. Default selected is Arizona. The dashboard then plots various charts related to that state.

This dashboard consists of 4 main charts :

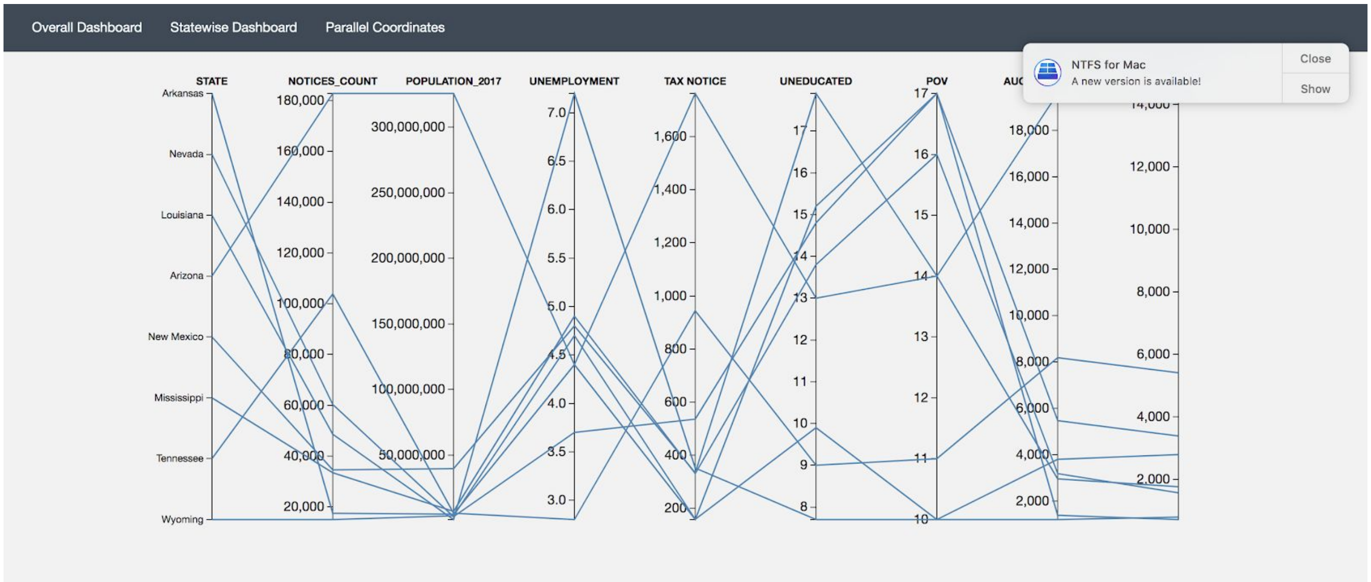
1. Category wise Distribution of Notices: This is a simple bar chart explaining the distribution of notices across various category for a particular state.

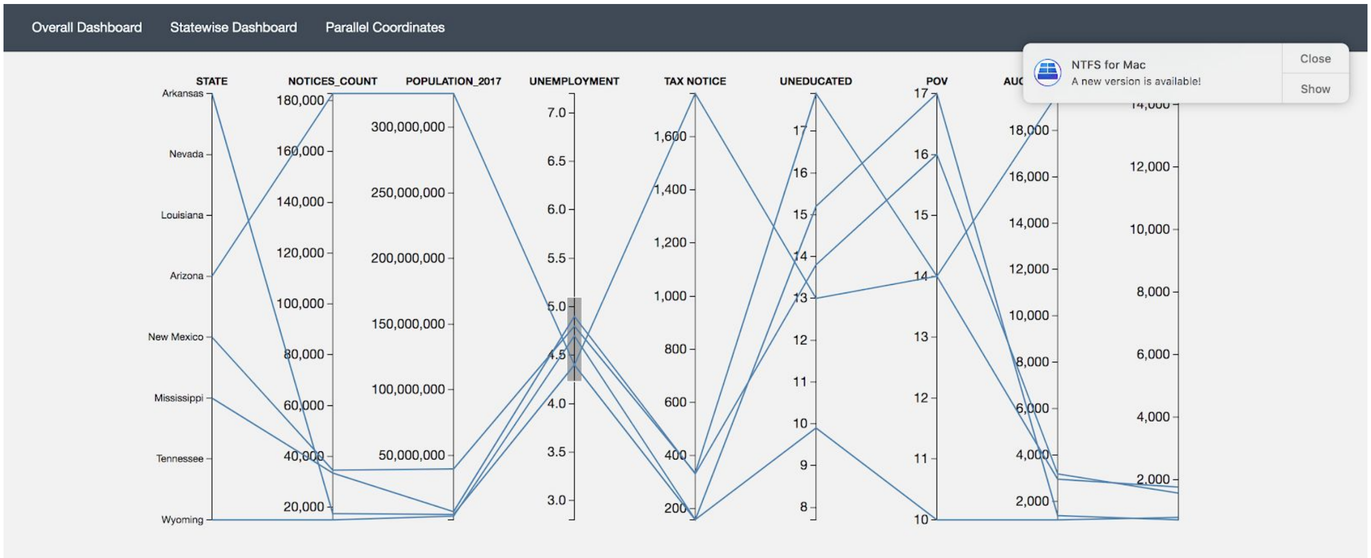


2. Year Wise Distribution of Notices: We have used a line chart to see the distribution of notices across years from 2012 - 2018.
3. Top 5 Newspapers - We have used a pie chart to represent the top 5 newspapers that are used for publishing public notices in that particular state.
4. County distribution of Notices: We used a horizontal pie chart to display county wise distribution of notices.

Parallel coordinates Plot:
We have included a parallel coordinate plot to view various relations between different factors like population, unemployment, poverty, education level and how it impacts the distribution of various types of notices.

We have added a filter option in the parallel coordinated plots to filter the ranges of different columns as can be seen in the figure below.





We performed LDA on the content of the notices using Python as we had reported in our progress report. However, it was not possible to convert the expected LDA analysis in D3. But we feel that as much as a visual analysis in D3 helped us in understanding various aspects of public notices, this LDA analysis revealed a lot of things as well. For eg: what were the most dominant words across notices, how content varies according to the type of notices. We tried a lot to represent this in D3 form, however, we could not come up with the interesting and interactive results as we could obtain using Python. (Gensim and pyLDAvis)

We did drop off on bubble chart and choropleth maps that we did make during mid-report progress. While they gave us relative information, that information wasn't relevant enough to reveal interesting observation or make real correlations as we wanted.

Interesting Observations :

1. Election notices were mainly seen in the year 2016. Furher investigation in the content of the notices revealed that the reason behind this would mainly be the Presidential election that took place in 2016. A few election notices were seen in 2017 as well. On closer inspection of actual notice content, we saw that these were for Mayoral Elections in 2017.
2. Arizona had the max number of notices overall and across each category as well.
3. Overall Auctions and Bids had the highest number of notices followed by Business and Corporates. Business and Corporates include notices like AOO and AOI. These notices are mainly related to the establishment of organizations and institutes. These notices show a general increase across the years, indicating the increase in the establishment of organizations and institutes. However, there were few notices in 2017-2018. We believe this may be due to these notices not included in the dataset. There is less probability that the establishment of organizations and institutes decreased in 2017-2018.
4. Each county has its own preferred publisher for notices even though there may be many local county newspapers.
5. Parallel Coordinates plot revealed some interesting relations :
    a. As expected, areas with greater population had a higher number of notices
    b. A positive correlation was seen between population and number of tax notices
    c. A negative correlation was seen between poverty rate and auction notices
    d. High poverty rate states show a lower number of foreclosure notices

Thus, we did see that the distribution for various categories of notices did somewhere depend on the factors like population and poverty in that state.

Conclusions and Lessons Learned:

While this project gave an immense experience in exploring the D3 domain and the interesting world of visualizations. Visualization makes it easier to identify trends, the relationship between entities. We humans find it difficult to imagine the multi dimensional space. However, with various interactions, and transformations applicable with visual analytics it becomes easier to view and relate this multidimensional data. Dashboard, brushing and linking, filtering, LDA analysis, help us bring out the relations between these multidimensional data in a much-informed way.

We did learn some lessons as well. This includes the importance of planned UI design. Sampling data too, has various impact on analysis. So it's important to understand how to sample the data. Data collection and preprocessing plays a very important role in any kind of analysis. The better the preprocessing of data, the better the results you obtain.

References :

- For interactive dashboard css templates - https://github.com/keen/dashboards
- Python LDA analysis on notices content - https://github.com/bmabey/pyLDAvis
- Parallel Coordinates - https://bl.ocks.org/jasondavies/1341281
- https://dc-js.github.io/dc.js/

Youtube link :