Yelp dataset challenge

Big Data - Visualization Practical Work

Author: Yolanda de la Hoz Simón
DNI: 53826071E
Master EIT Digital Data Science - Course 2015/2016

Introduction

This report outlines the design and development of an interactive visual application designed to respond the most relevant questions in the small businesses domain taking advantage of the Yelp! business reviews.

As a visual analytic product, the development of this application follows an iterative cycle based on the following steps:

- **Analysis**: First, it is analysed the dataset provided, reviewing also the most common needs and questions present in this application domain. In order to define the basic components, interaction and tasks in this application, a brief review of the different solutions in both visual applications and analytics is also described.

- **Design**: In this section, the functional requirements are presented as a result of the analysis stage and the final general design and its components description.

- **Development**: In this section, it is described the algorithm development with the detail of each subtask and some R snippets.

- **Validation**: Finally, it is also described the validation of the product with an analysis of the functional requirements achieved (described in the analysis stage) and the non-functional requirements.

Analysis

Dataset description

The dataset provided by Yelp (http://www.yelp.com/dataset_challenge) is based on 61 million reviews with the aim to help people find the most relevant businesses for everyday needs.

The dataset is provided in JSON format and it includes some interesting features described below.

- **Business**: Localization, business category, reviews, starts and open hours.
- **Review**: Business, users, starts, review text, date and votes.
- **User**: User, reviews, votes, average starts, friends, antiquity, compliments and fans.
- **Check In**: Business and check in info (hours).
- **Tip**: Tip text, business, user, date and likes.

Application domain

In this section, I will identify some relevant questions that the application should solve with the purpose of have a user-centered design that better fix the user requirements.

The questions (with the associated underlying variables) that I will develop corresponds to the most common needs of two targeted users: business owners and customers.

Business owners
Need: Improve the quality of their services → Offer better services.
Questions:

1. Which is the average rating that users give to my business?
2. Is the number of hours that a business open affecting their ranking mark?
3. Is the location of my business affecting the ranking mark?
4. Is my business getting more users depending of the season of the year?
5. Is the number of services that my business offers affecting their ranking mark?
6. Predict when a business will be more busy.

Customers
Need: Look up the better business according to their preferences.
Questions:

1. What are the top ranked (restaurants, shops...) in my location?
2. What of these business open today?
3. Get the top ranking business according to: age of the reviewers, type of business, open_hours and localization.

In the challenge webpage it is also proposed some interesting questions related with some data science challenges and business needs, among which I've selected the following questions to be considered:

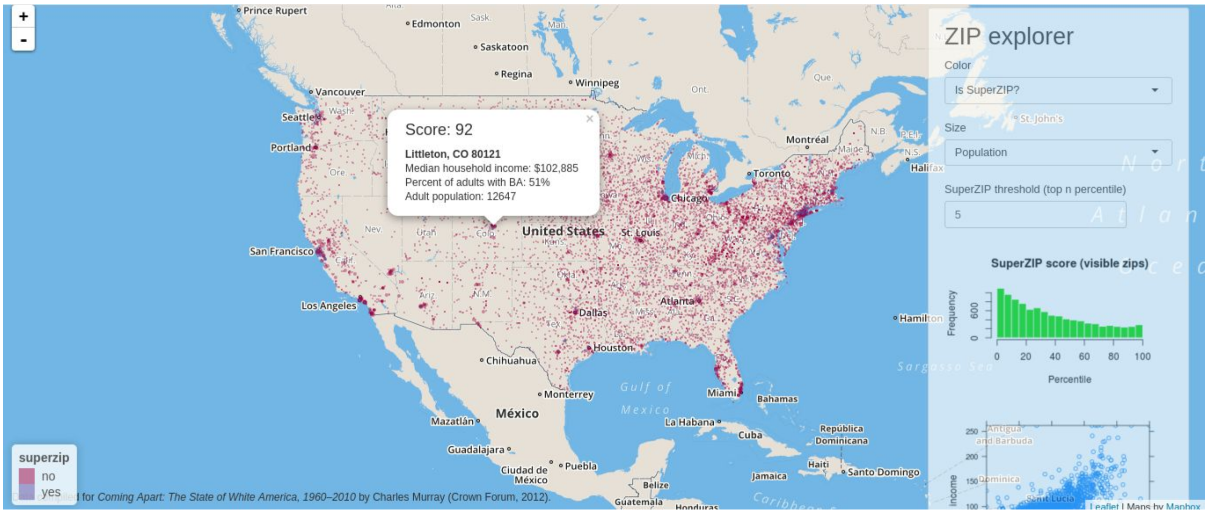
1. How much of a business success is really just location, location, location?
2. Are there more reviews for sports bars on major game days and if so, could you predict that?
3. How much influence does my social circle have on my business choices and my ratings?
4. What cuisines are Yelpers raving about in these different countries?
5. In which countries are Yelpers sticklers for service quality?

Research of existing solutions

In this section, it is analysed the different existing solutions attending the data and tasks abstractions in the case of visual analytics as well as to define the data mining goals to summarize data and extract better conclusions.

For the development and selection of the best visual analytic methods I have selected the following examples taken from the shiny gallery and D3.js examples.

The figure 1. presents a visualization of geolocation points in a interactive map. The points geographically distributed allows us to find trends and compare values to find outliers or locate clusters. The ZIP explorer that allows us to select two categorical key variables: color and size. It is also interesting that the interactive map also allow us to select one of the points located in the map and show the most relevant information for this location in a popup message.



Score: 92

Littleton, CO 80121

Median household income: $102,885

Percent of adults with BA: 51%

Adult population: 12647

ZIP explorer

Color: Is SuperZIP?

Size: Population

SuperZIP threshold (top n percentile): 5

superzip

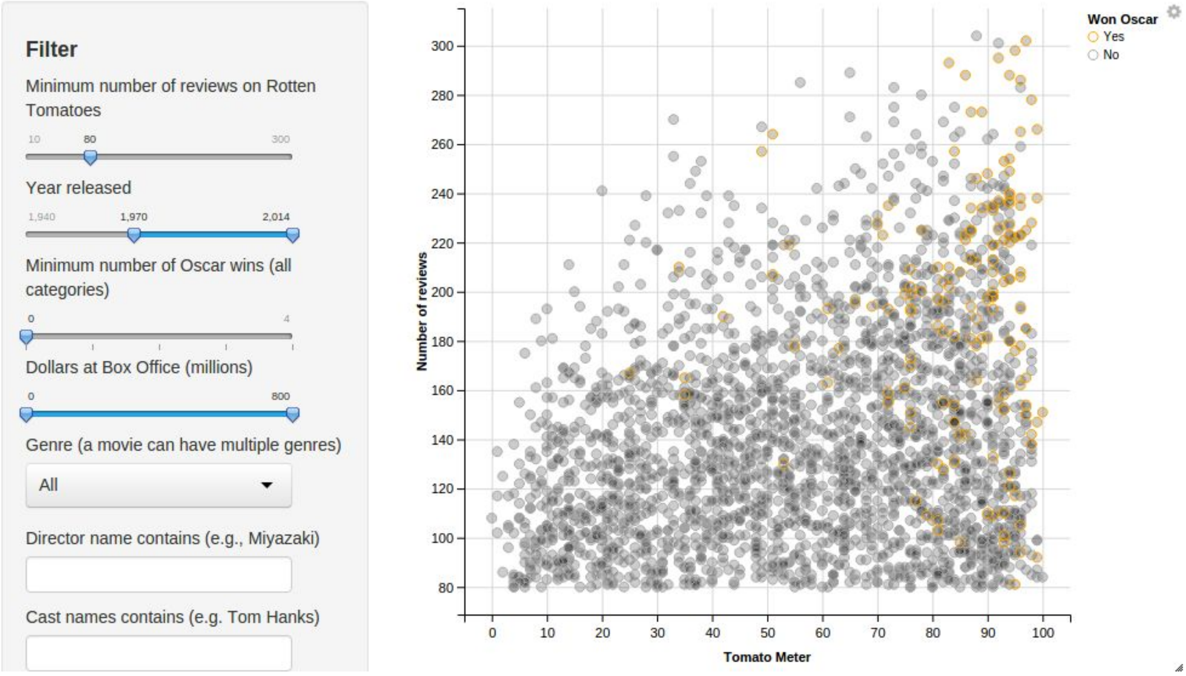
🟥 no

🟪 yes

...d for *Coming Apart: The State of White America, 1960–2010* by Charles Murray (Crown Forum, 2012).

Figure 1. Superzip - Shiny gallery

The figure 2. presents a visualization of a scatterplot for movie reviews, the most interesting feature of this figure is the filtering options contained in the widget showed in the left lateral. This widget allows the user to focus the attention only in the relevant variables that solve the question.



Filter

Minimum number of reviews on Rotten Tomatoes

10 80 300

Year released

1,940 1,970 2,014

Minimum number of Oscar wins (all categories)

0 4

Dollars at Box Office (millions)

0 800

Genre (a movie can have multiple genres)

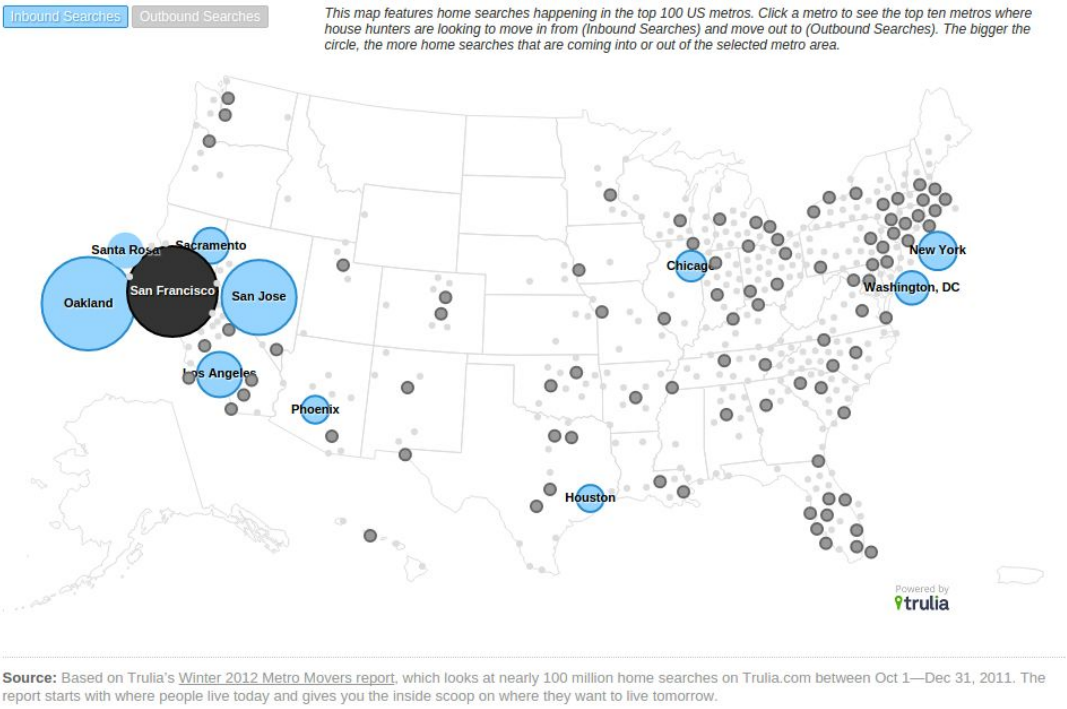
All

Director name contains (e.g., Miyazaki)

Cast names contains (e.g. Tom Hanks)

Figure 2. Movie explorer - Shiny gallery

The figure 3. presents also another usage application of a bubble scatter plot with the location positioned in an interactive map to show a report of where people live today and the inside scoop on where people want to live tomorrow.



This map features home searches happening in the top 100 US metros. Click a metro to see the top ten metros where house hunters are looking to move in from (Inbound Searches) and move out to (Outbound Searches). The bigger the circle, the more home searches that are coming into or out of the selected metro area.

Source: Based on Trulia's Winter 2012 Metro Movers report, which looks at nearly 100 million home searches on Trulia.com between Oct 1—Dec 31, 2011. The report starts with where people live today and gives you the inside scoop on where they want to live tomorrow.

Figure 3. Trulia trends visualization - D3.js gallery

Design

As the design of this application is focused to answer the questions proposed before; first it is defined a list of the specific functional requirements that the application should fulfil. This list is corresponds to the data and task abstractions and the different visualization methods and interaction levels.

Functional requirements definition

Data and task abstractions

The data that should be visualized in this application is extracted from the questions proposed before and ordered by the frequency of appearance in the questions:

(1) ranking mark
(2) average rating
(3) review counts
(4) localization
(5) business category
(6) number of hours open
(7) season of the year
(8) number of services

For the definition of the task abstractions have been considered the following main tasks:

(1) Express geospatial positions
(2) Navigation and interaction to show information from a specific position
(3) Find and locate outliers
(4) Offer different filtering options
    - Filter 1: Location area
    - Filter 2: ranking mark
    - Filter 3: review counts
    - Filter 4: users popularity
    - Filter 5: open /closed business
    - Filter 6: state, country, business name and zipcode.

(5) Information aggregation
    - Cluster 1: ranking mark
    - Cluster 2: review counts
    - Cluster 3: ranking mark difference between current stars and review stars

(6) Summarize to offer a general view of the current data
    - Summary: average rating by states

(7) Data exploration
    - Ordered table: show raking details

Interaction and visual encoding

The main visualization methods identified that best answer the proposed questions are:

(1) Bar chart
  - It helps to lookup and compare values.
  - It shows one quantitative value attribute and one categorical value attribute.

(2) Histogram
  - It helps to find trends, outliers, distribution, correlation and locate clusters.
  - It shows two quantitative values attributes: latitude and longitude to express location in a map. Two categorical key attributes:
      - Color: stars mark /review_count/ stars difference.
      - Size: stars mark /review_count/ stars difference.

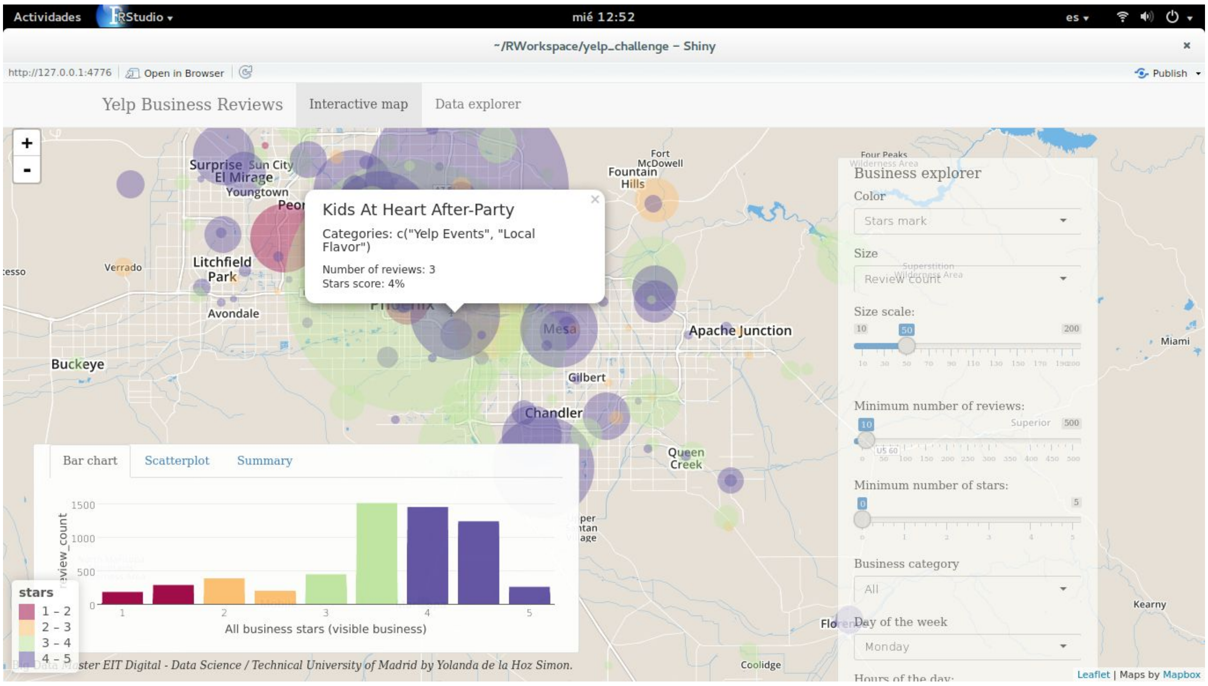
(3) Interactive map
  - Used to locate and filter per region data.
  - It allows to identify clusters, lookup and compare values, outliers and find trends.

The different interaction methods with each of the above visualization methods are enclosed within the following tasks performed in the main application area:

(1) Views synchronization
(2) Go task
(3) Selection
(4) Zoom
(5) Map navigation
(6) Chart navigation

General design

This section describes the general design of the application and its main components.



Kids At Heart After-Party

Categories: c("Yelp Events", "Local Flavor")

Number of reviews: 3

Stars score: 4%

Business explorer

Color: Stars mark

Size: Review count

Size scale: 10 – 50 – 200

Minimum number of reviews: 10 – 500

Minimum number of stars: 0 – 5

Business category: All

Day of the week: Monday

Hours of the day:

...ster EIT Digital - Data Science / Technical University of Madrid by Yolanda de la Hoz Simon.

Figure 4. General design

The Figure 4. shows a sample screen of the general design of the shiny application. Here it can be seen four main parts:

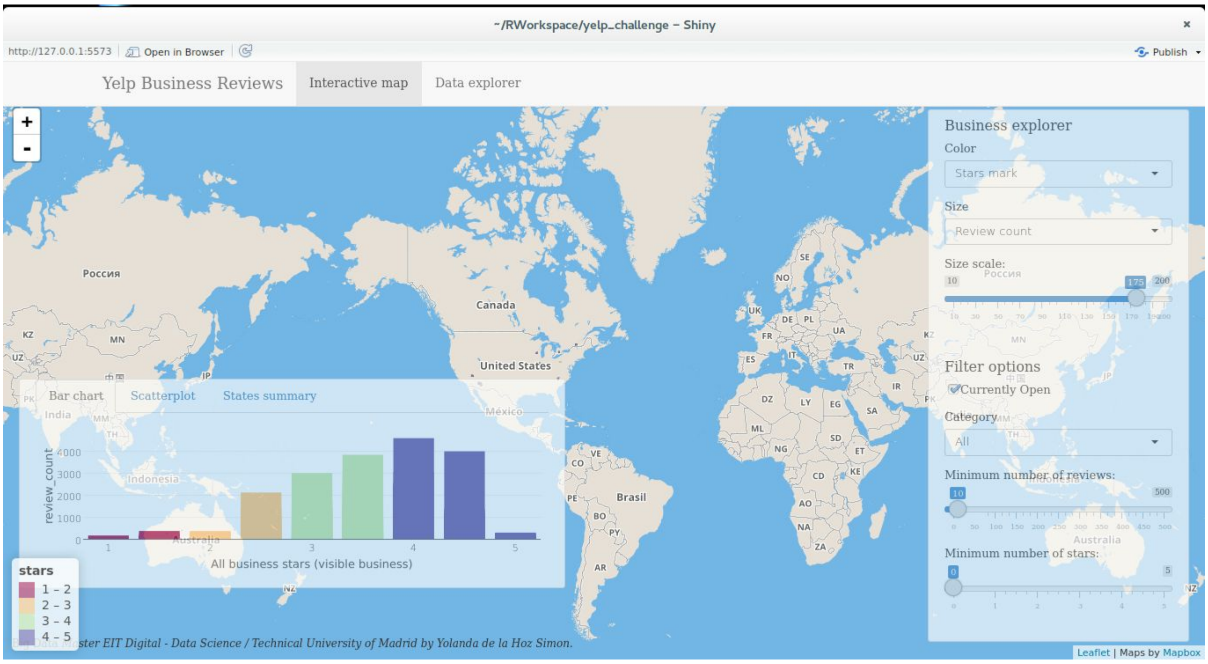
- **Interactive map**: It is the main layout of the application.
- **Business explorer**: On the right site, it is an absolute panel that gives the user different exploration options.
- **Graph summary**: On the lower left side, that shows the different chart options (Bar chart, scatter plot and chart summary).
- **Data explorer**: It can be accessed through the top navigation bar and shows a table with more detailed information.

Components description

In this section it is described the main components of the application and the different use cases and tasks that a user could perform to answer the proposed questions.

Interactive map

The interactive map shows a map with the data location. It can be navigated and zoomed to filter and show an interesting area.



~/RWorkspace/yelp_challenge – Shiny

http://127.0.0.1:5573

Yelp Business Reviews | Interactive map | Data explorer

Business explorer

Color: Stars mark

Size: Review count

Size scale: 10 – 175 (200)

Filter options

☑ Currently Open

Category: All

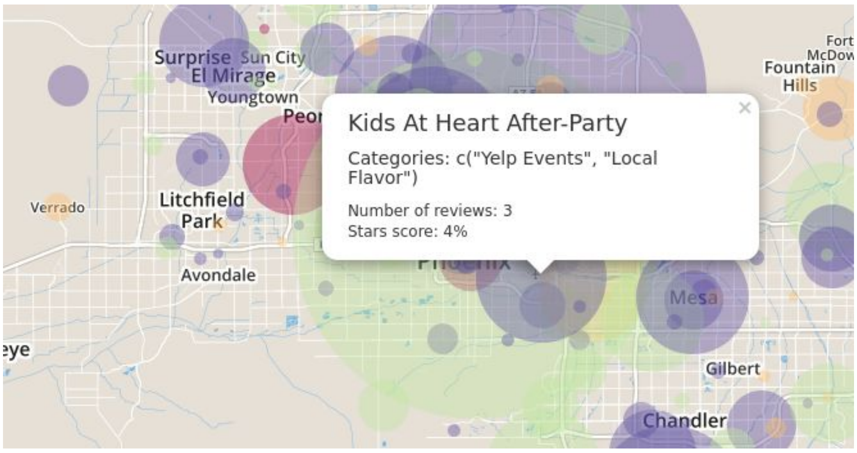
Minimum number of reviews: 10 (500)

Minimum number of stars: 0 (5)

Bar chart | Scatterplot | States summary

Figure 5. Interactive map

In the following picture it is shown an image with a zoomed region area showing in which the user has selected one of the business to show specific information such as the business categories that this business belongs to with the stars score and number of reviews.



Kids At Heart After-Party

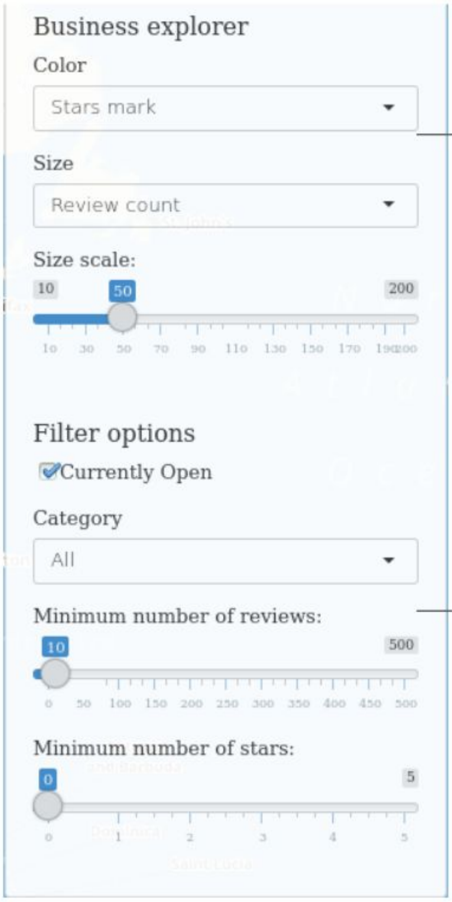
Categories: c("Yelp Events", "Local Flavor")

Number of reviews: 3
Stars score: 4%

Figure 6. Popup message

Business explorer

The business explorer allows the user to select interesting information and filter data. It is divided in two main parts: business explorer and filter options.



Business explorer

Color: Stars mark

Size: Review count

Size scale: 10 – 50 – 200

Filter options

☑ Currently Open

Category: All

Minimum number of reviews: 10 – 500

Minimum number of stars: 0 – 5

Business explorer

It allows to select the current variables that should be visualized. This is also it is the main component to allow the synchronization between all the views to visualize to focus the attention in the interesting data.

This component also includes a size scale selection to maximize the visualized data in case it is not currently showed.

Filter options

It allows to filter the current variables and show only the interesting information associated to the business data.

Figure 7. Business explorer

Between the different filter options, it is allowed to show the currently open business that could be very interesting if it is compared with the closed business, for example to know how many stars had a business already closed in this area.

This component also offers the possibility to restrict the data to show only the most ranking business with the minimum review number and the minimum stars sliders

Graph summary

The graph summary is an absolute panel displayed above the interactive map that offers different analytical visualization options such as bar chart, scatter plot and a bar chart summary of the current states.

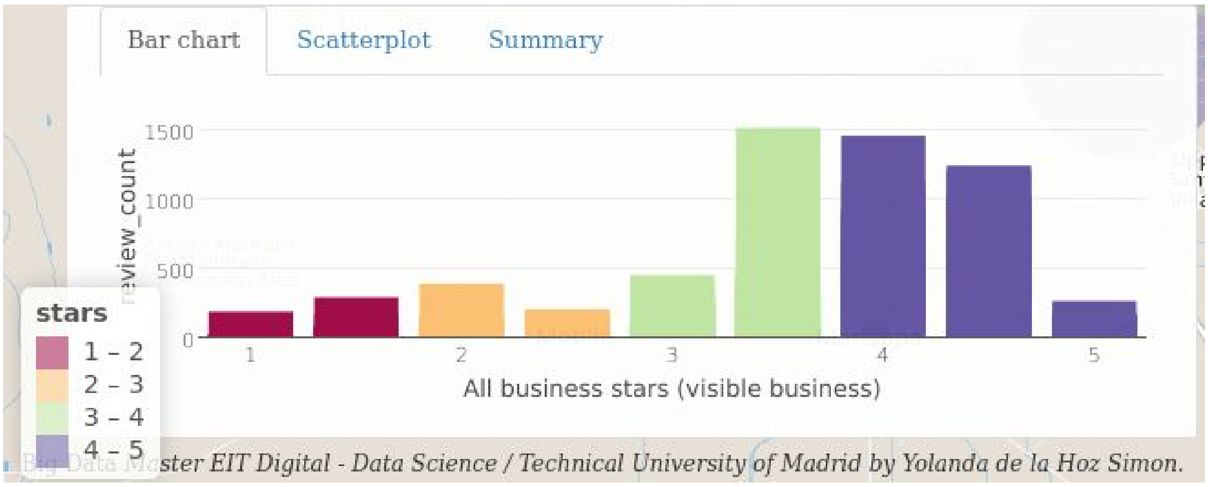
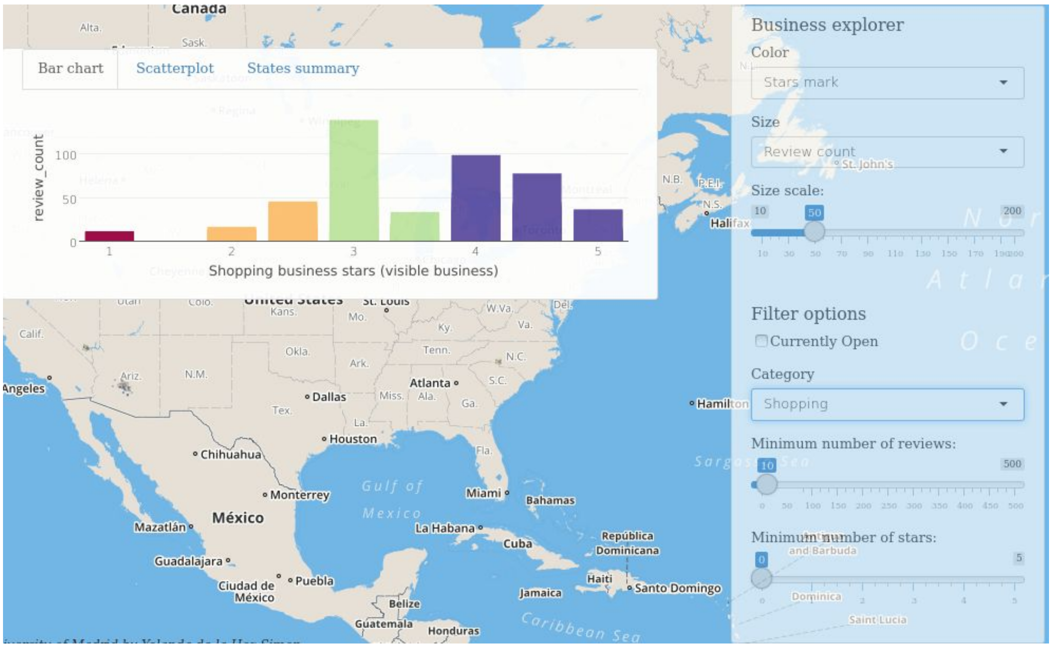


Figure 8. Bar chart

In the figures 9 and 10 are shown the synchronization between the business explorer, interactive map and the graphi summary panel. As it is selectionend one of the available categories it is changed the title and the data associated and as a result and update of all the views of the application.



Business explorer

Color: Stars mark

Size: Review count

Size scale: 10 – 50 – 200

Filter options

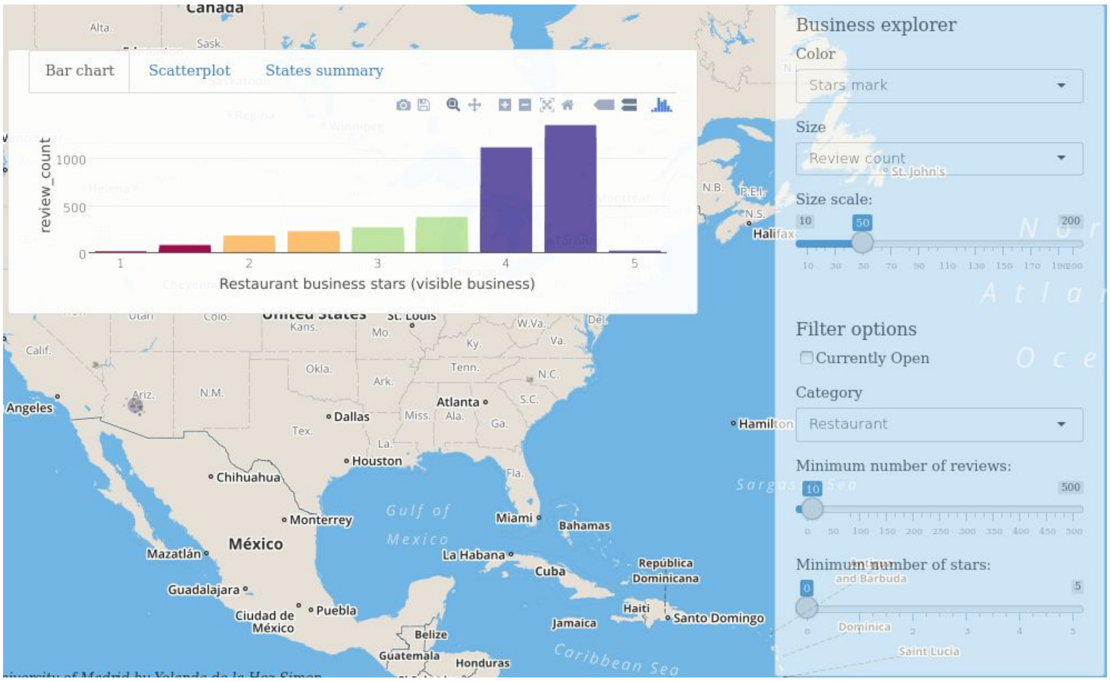
□ Currently Open

Category: Shopping

Minimum number of reviews: 10 – 500

Minimum number of stars: 0 – 5

Figure 9. Bar chart - Filtered Shopping



Business explorer

Color: Stars mark

Size: Review count

Size scale: 10 – 50 – 200

Filter options

☐ Currently Open

Category: Restaurant

Minimum number of reviews: 10 – 500

Minimum number of stars: 0 – 5

Figure 10. Bar chart - Filtered Restaurants

The figure 11 shows a scatter plot of the current selected data, this visualization allows to find trends and outlier easily. For example in this figure we can visualize the stars and number of reviews for that each stars has with some inline gaps that indicates that most of the rankings are done with no more of 2500 reviews.

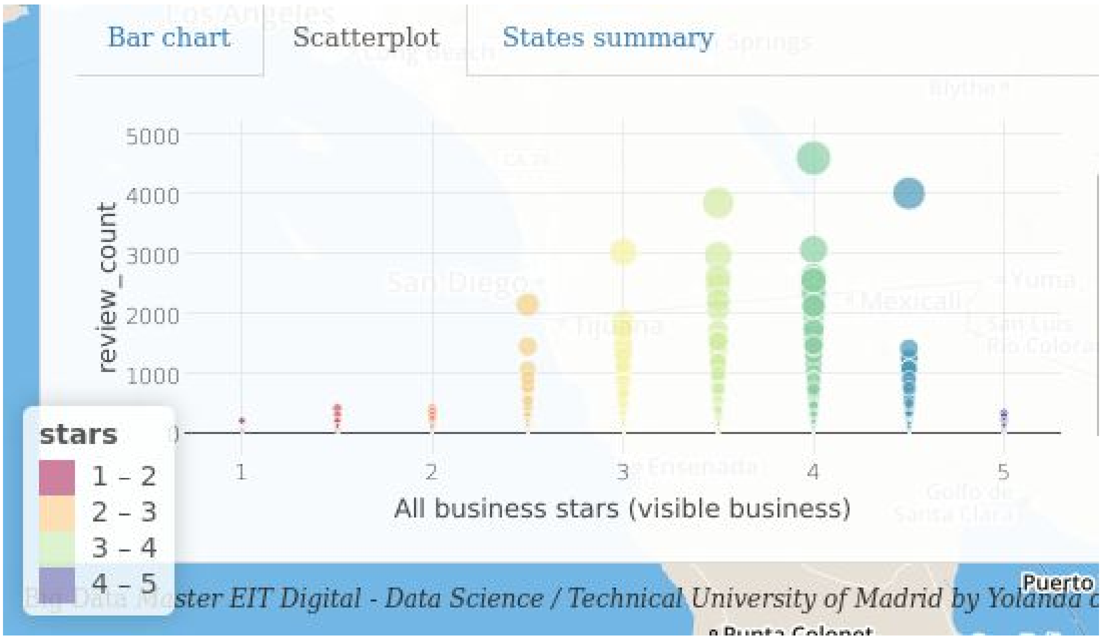


Figure 11. Scatter plot

The figure 12 shows the states summary tab of the absolute panel, this view shows a summary of the selected variable in the business explorer aggregated by each the state that are within the visualization area. It is very interesting because it help us to focalize the attention for example only the areas with more number of reviews.

In this figure it is shown also one the filtering options that allows the chart selecting the interesting region area to zoom and visualize this area.

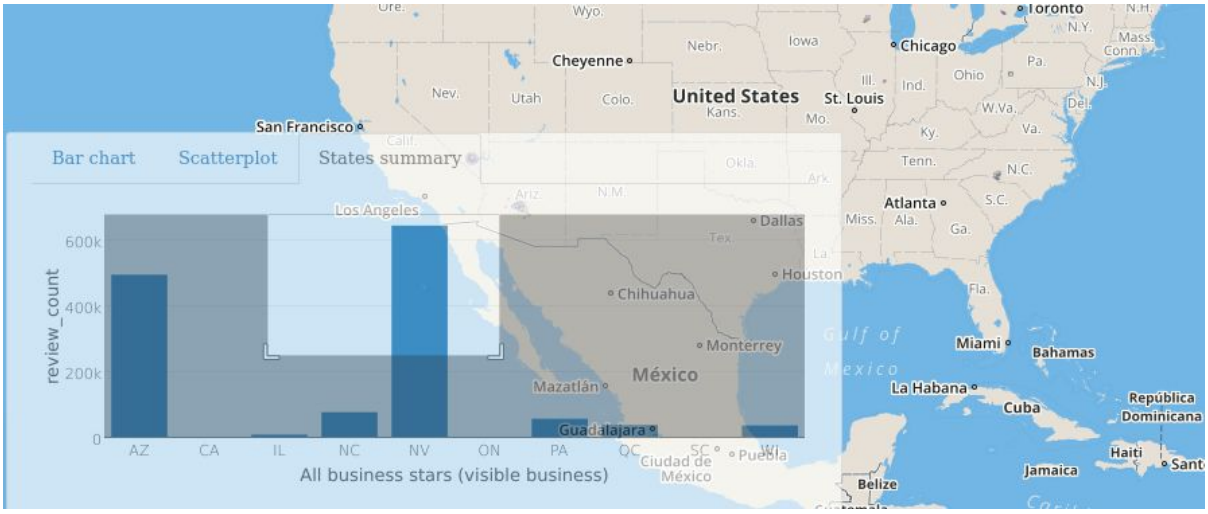
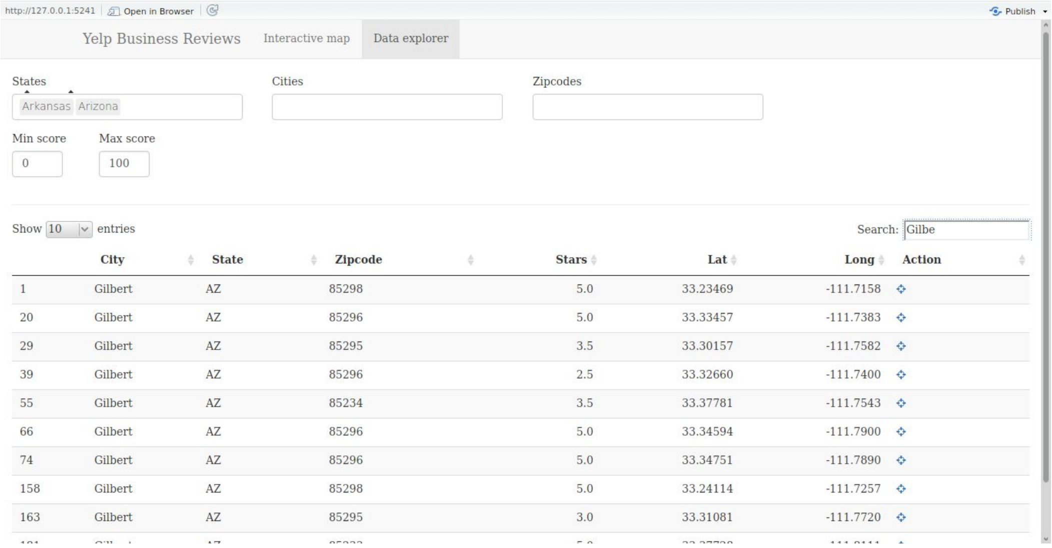


Figure 12. States summary

Data explorer

The data explorer is a table that show more detailed information of each of the states, it allows also to filter information by location and state or given the maximum and minimum score.



States: Arkansas Arizona

Cities:

Zipcodes:

Min score: 0

Max score: 100

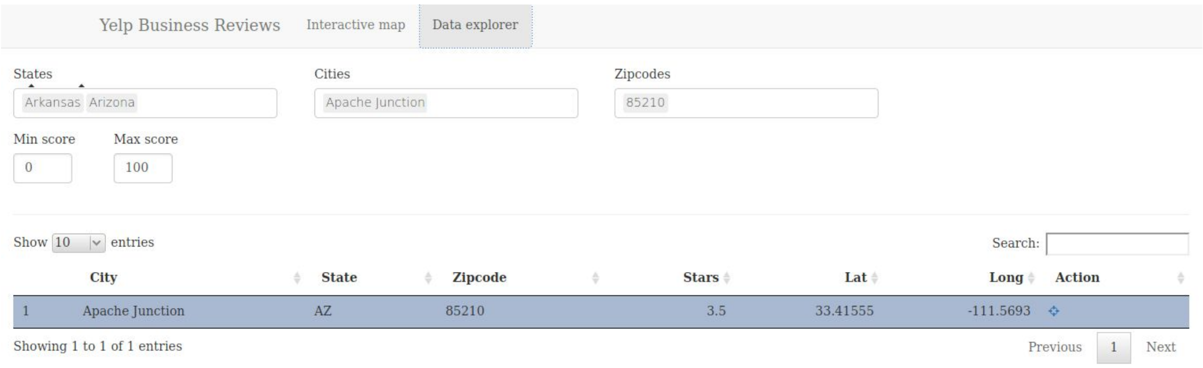
Show 10 entries

Search: Gilbe

	City	State	Zipcode	Stars	Lat	Long	Action
1	Gilbert	AZ	85298	5.0	33.23469	-111.7158	✥
20	Gilbert	AZ	85296	5.0	33.33457	-111.7383	✥
29	Gilbert	AZ	85295	3.5	33.30157	-111.7582	✥
39	Gilbert	AZ	85296	2.5	33.32660	-111.7400	✥
55	Gilbert	AZ	85234	3.5	33.37781	-111.7543	✥
66	Gilbert	AZ	85296	5.0	33.34594	-111.7900	✥
74	Gilbert	AZ	85296	5.0	33.34751	-111.7890	✥
158	Gilbert	AZ	85298	5.0	33.24114	-111.7257	✥
163	Gilbert	AZ	85295	3.0	33.31081	-111.7720	✥

Figure 12. Data explorer

The figure 13 shows the possibility to search in this data explorer the most ranked city and then navigate to this city through the *Action* column.



States: Arkansas Arizona

Cities: Apache Junction

Zipcodes: 85210

Min score: 0

Max score: 100

Show 10 entries

Search:

	City	State	Zipcode	Stars	Lat	Long	Action
1	Apache Junction	AZ	85210	3.5	33.41555	-111.5693	⌖

Showing 1 to 1 of 1 entries

Figure 13. Go task

Development

The development of this application has followed the following stages:

0- Research of the available tools to develop the project

The project has been developed in the R programming language with the *shiny* library. Nevertheless, this application has made use of other libraries such as *leaflet* library in the implementation of the interactive map with the clustering option and the *plotly* library in the development of the charts of the graphics summary absolute panel.

1- Data Preprocessing to get the data model

The source code that corresponds to this part it is located in the *DataPreprocessing* folder and it has to be launched before the application starts.

As the most of the application it is done with the business data, for testing purposes it recommended only run the part associated to clean and load this data. Only it is required to load the reviews for the calculate the ranking difference.

The phase could be subdivided in the following stages:

1.1. Data collection

The data used for the development of this application is freely downloaded from the webpage http://www.yelp.com/dataset_challenge.

1.2. Transform json and load the data in R

For this purpose I used the library *jsonlite* to read and save in R the data and convert it into data.frames objects. Nevertheless, in the case of the reviews data set it is also required a compression and reduction of the dataset due to huge volume of text data.

In some cases, depending of the machine, it could be also required to allocate more memory *RAM* to launch the application.

1.3. Clean and filter the data

In the cleaning data process, I transformed most of the variables with *NA* value into 0 or 0:00 in case of hours. This allows me to visualize these data without losing the lost of information knowledge in some cases.

To visualize the data and locate it in the map, it is also extracted the zip_code from the business dataset.

2- UI - View development

1.1. Composition and development of the different widgets in the ui.R file

The UI is composed of two main views separate with a navbar and the tabPanel object, organized hierarchically with the composition of the following elements:

- Data Explorer
    - selectInput (states, cities, zipcode, min and max score)
    - dataTableOutput
- Interactive Map
    - leafletOutput (Interactive map)
    - absolutePanel (Graphs summary)
    - absolutePanel (Business explorer)

3- UI - Controller development

This phase correspond to the development of the server side of the graphical interface server.R, it is responsible to perform the next two main tasks:

1.1. Development of the logic associated to each view
1.2. Coordination of each view to create an integrate and reactive design

The source code of this part it is composed of the following shiny elements to maintain synchronized the different views:

1. A reactive expression that returns the set of business that are in bounds right now
2. A reactive expression that filters the business, returning a data frame
3. Render plots to show a bar chart, a histogram and a scatterplot
4. An observer responsible for maintaining the circles and legend, according to the variables the user has chosen to map to color and size
5. An observer to show a popup with business info when the map is clicked
6. Observers to update the data explorer according to the filter options
7. Common functions that performs common calculations

Validation

In the validation and test of the application it has been considered the following main features:

(1) The user is capable of answer most of the questions proposed
(2) The application fulfill the general design guidelines
(3) Other features related to performance and responsiveness