An analysis of Gentrification phenomena in Los Angeles

Salma Chabib[1] and **Lilia Grasso**[2]

[1]Master Degree in Data Science, ID 872519
[2]Master Degree in Data Science, ID 813210

5872 Words

Keywords: Gentrification, Los Angeles, Police, Calls service

20th June, 2021

Abstract

Gentrification is the process whereby the character of a poor urban area is changed by wealthier people moving in, improving housing, and attracting new businesses, often displacing current inhabitants in the process. [1]. To some, gentrification is synonymous with an inseparably interconnected web of violent acts, instead for others, though, gentrification is the simple mechanism by which we make our cities better, tied up in our most basic economic processes. But to many more, "gentrification" is a word that provokes anxiety and uncertainty, especially if we're people who hold some degree of economic, social or other power and we're not sure how best to use it. We might worry about our own role in gentrification when we scout for apartments or decide whether or not to support the new coffee shop down the street; we might consider gentrification when we make choices about who to vote for in local elections, or whether that shiny new development in a low income neighborhood is a good thing. And if we're disempowered people, we might think about gentrification when our landlord hikes up the rent, or when we see a street we've cherished suddenly and irreversibly change. The aim of this analysis is to examine Los Angeles Gentrification phenomena, thanks to data provided by Los Angeles Police Department (LAPD) and data acquired from a broadcast radio, that records all the calls for Service to the LAPD for everything related to vehicles such as grand theft auto, windows smashing. Our goal is to investigate gentrification in the four macro-division of Los Angeles ruled by LAPD, comparing the indexes such as percentage of white people, percentage of people with degree, households' annual incomes etc., in 2014 and 2000, in order to identify the gentrified macro divisions (or at least part of them). Once done, we analyze data acquired from Broadcastify, in order to examine whether the rate of crimes is high in the non-gentrified division. For this purpose, we use: Kafka to acquire data from the Broadcastify's archive; Offset Explorer (formerly Kafka Tool) that is a GUI application for managing and using Apache Kafka; Selenium library to scrape data from the website; Pandas methods to handle data for the integration; Tableau's graphics and Seaborn and Matplotlib libraries to plot some static graphs. Thanks to this analysis we are capable to identify some of the divisions as *gentrified* and some others as *being gentrified*. Moreover, looking at the crimes rate, the notable thing was that the rate of crimes is high both in the divisions gentrified and in the ones where the process is ongoing. In conclusion, we can state that the gentrification is a very complex issue that requires more investigations with powerful tools, with a lot more variables that need to be taken into account.

1 Introduction

Over the last two decades there has been an incredible change in the world's economics and society lead by the huge spread of powerful technologies to communicate among people in different countries. These technologies have been the fundamental hinge for the foundation of the globalization phenomena, leading to a situation where the borders among countries are mostly negligible and all the people feels to join a single community, where ethical differences are very tiny. Although the world has become more unite on the online world, in the real world it started to grow an increasing importance in the perceived image that countries wanted to give of them and of their cities, leading to restore most of the buildings owned and to renovate the districts perceived as "decayed". This behaviour could be considered "wealthier" on a hand, but it could be very frustrating for the people on the other side of the debate, who are the ones who actually live in the decayed districts. Furthermore, the debate is not as simple as we're presenting it, but it is very complex and it is called "Gentrification". In order to understand it, there are three key things to consider:

- The historic conditions, especially policies and practices that made communities susceptible to gentrification:
    - Redlining: a practice lead by federal government where people of color were denied access to loans that would enable them to buy or repair homes in their neighborhood.
    - White flight: housing and transportation policies of the mid-20th century increased the growth of mostly white suburbs and the exodus of capital from urban centers.
    - Urban renewal: Left behind in central city neighborhoods, low-income households and communities of color bore the brunt of highway system expansion and urban renewal programs, which resulted in the mass clearance of homes, businesses, and neighborhood institutions, and set the stage for widespread public and private disinvestment in the decades that followed.
    - Foreclosure crisis: it contributed to making places vulnerable to gentrification. In low-income communities of color, disproportionate levels of subprime lending resulted in mass foreclosures, leaving those neighborhoods vulnerable to investors seeking to purchase and flip homes.
- The way that central city disinvestment and investment patterns are taking place today as a result of these conditions. These ways are based on revitalization – cities are investing in some of these neighborhoods with improved transit access and infrastructure in part to draw in newcomers;
- The ways that gentrification impacts communities, causing displacement, which means that in some of these communities, long-term residents are not able to stay to benefit

from new investments in housing, healthy food access, or transit infrastructure and cultural displacement where even for long-time residents who are able to stay in newly gentrifying areas, changes in the character of a neighborhood can lead to a reduced sense of belonging, or feeling out of place in one's own home.

So we could summarize this phenomena as [2]:

> "The process of changing the character of a neighborhood through the influx of more affluent residents and businesses.It often increases the economic value of a neighborhood, but the resulting demographic displacement may itself become a major social issue.
>
> It often shifts a neighborhood's racial/ethnic composition and average household income by developing new, more expensive housing and businesses in a gentrified architectural style."

As said before, the process is typically the result of increasing attraction to an area by people with higher incomes spilling over from neighbor cities, towns, or neighborhoods. The gentrification phenomena is analyzed using these indexes:

- Displacement Pressure Index that captures the intersection between 2 classes that are [3]:
  - change measures that consider distance to current rail stations, distance to rail stations under construction/recently opened in 2016, proximity to Rapidly Changing Neighborhoods, distance to the closest "top tier" changing neighborhood as defined by the Los Angeles Index of Neighborhood Change and housing market. They indicate future revitalization due to growing businesses.
  - displacement pressure factors that capture areas with high concentration of existing residents who may have difficulties in absorbing massive rent increases that often accompany revitalization.
- Neighborhood Change Index that describes the processes of physical and socio-economic change within and in-between neighborhoods. The index scores are an aggregate of 6 demographic measures indicative of gentrification, that are:
  - Gross Rent: it shows the amount of rent stipulated in a lease and it refers to monthly payments in a year;
  - Household Income: it indicates the median gross income perceived by Los Angeles's households in a year;
  - Household Size: it illustrates the median number of component in a Los Angeles's family;

  - IRS Ratio: it indicates the Interest Rate Swap Ratio that is the change in the interest rate based on yearly gross income;
  - Population with Degree: it provides us information on the percentage of people older than 25 years old who have taken a Bachelor degree or more;
  - White Percentage: it indicates the percentage of non hispanic/latin white people.

1.1 Description of the data sets

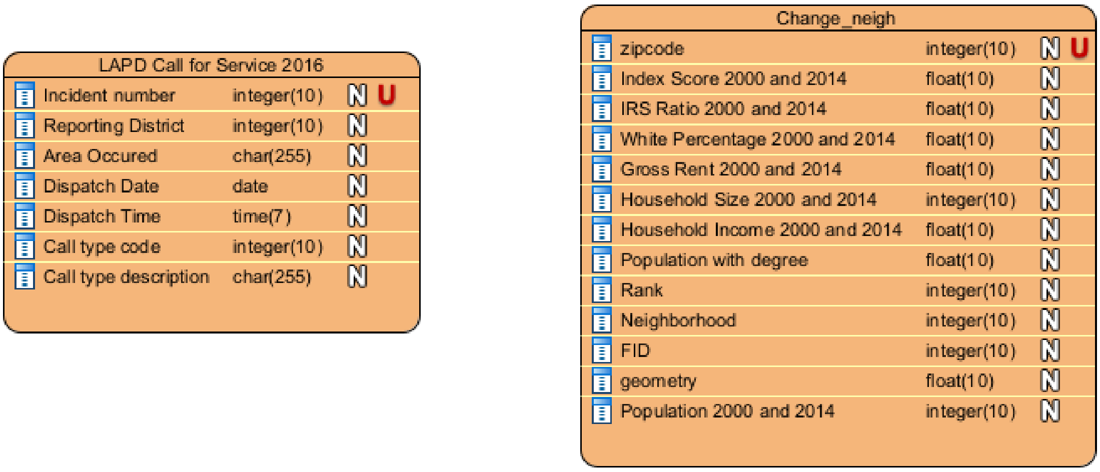
In order to conduct the analysis, we have utilized data sets provided by Los Angeles Police Department. We needed data sets that provide us information about both gentrification indexes for each neighborhood in Los Angeles and data sets that shows us data concerning calls for service to LAPD. The first one is a geojson file that is called *change_neigh.geojson*, later renamed as *street_map*. A GeoJSON file format is an open standard format that contains both geospatial data and attribute data (more information about an attribute). The geoJSON format is extended from the JSON (JavaScript Object Notation) standard format. and more information about an attribute. The columns of the geojson file are the indexes mentioned above both for 2014 and 2000, with other variables that identify more precisely the neighborhood, such as:

- FID: it indicates the district associated to the zip code;
- zip code: it indicates the zip code that refers to a specific neighborhood
- Rank: it shows how the neighborhood in ranked due to its life expectations;
- Neighborhood: it illustrates the name of the neighborhood taken into consideration;
- Population: the value of the population in 2014 in each area associated to the zip code;
- geometry: it is a vector with geographic coordinates that let us define each neighborhood's borders.

The second data set taken into consideration concerns calls for service to LAPD. We provide ourselves with a data set per each year from 2016 to 2020, where these csv files are structured as following (we show an example of the dataset from 2016 and the geojson file):

Where the variables are:

- Incident Number: numeric variable, it is unique identifier for incidents;
- Reporting Districts: numeric variable, it indicates the district to which the incident was reported;
- Area Occurred: string, it indicates the area where the incident has taken place;
- Dispatch Date: datetime variable, it indicates the date in which the incident has occurred;



LAPD Call for Service 2016			
Incident number	integer(10)	N	U
Reporting District	integer(10)	N	
Area Occured	char(255)	N	
Dispatch Date	date	N	
Dispatch Time	time(7)	N	
Call type code	integer(10)	N	
Call type description	char(255)	N	

Change_neigh			
zipcode	integer(10)	N	U
Index Score 2000 and 2014	float(10)	N	
IRS Ratio 2000 and 2014	float(10)	N	
White Percentage 2000 and 2014	float(10)	N	
Gross Rent 2000 and 2014	float(10)	N	
Household Size 2000 and 2014	integer(10)	N	
Household Income 2000 and 2014	float(10)	N	
Population with degree	float(10)	N	
Rank	integer(10)	N	
Neighborhood	integer(10)	N	
FID	integer(10)	N	
geometry	float(10)	N	
Population 2000 and 2014	integer(10)	N	

Fig. 1. Structure of dataset of calls to LAPD from 2016 and of the geojson file with indexes

- Dispatch Time: datetime variable, it shows at which time the incident has happened;
- Call Type Code: numeric variable, it illustrates the code that refers to that specific type of incidents;
- Call Type Description: string, it gives small description of what happened during the incidents, through a text code.

While the third data set refers to Divisions, in fact it has a column with the name of the district, a column with the number associated to each district and the geometry of each district. In order to succeed in associating each neighborhood to the district in which it is located, we have merged data sets providing us these correspondences. Since processing information about neighborhood and districts would not be efficient, we have decided to develop the analysis on a higher level of abstraction taking into consideration the four macro-divisions into which Los Angeles is divided. For this purpose, We have consulted reference table found on the Internet to implement a particular function that assigns each zip-code to the corresponding macro-division [4].

2 Data Management

We have decided to develop our project considering two V's: Variety and Velocity. Concerning the first one, we have integrated bunch of information from different data sets acquired in different formats such as: csv (comma separated values) and GeoJson files since the analysis required even geographic polygons in order to be more accurated as regards to the geographical location from which the calls for service to LAPD has been made.

2.1 Velocity

Regarding the Velocity, as we mentioned above, we have scraped data from a broadcast radio of Los Angeles. The website is called "Broadcastify" and it is a source pf public safety of radio audio live streams. The idea behind our velocity part is to split the data acquisition from the data usage, so that the website, searching by State

We have built a system, that whenever it is called, will send a message to Kafka cluster containing a json file, with the number of calls for service to LAPD of the day before and date of that day. There will be four topics, one per each macro-division, this choice has been made because there were 4 different Broadcastify websites, one per each divisions, where the calls, with both the start and end time, were archived. For this purpose, we have used Kafka [5], through a producer and a consumer. Apache Kafka is an open-source streaming platform that was initially built by LinkedIn. It was later handed over to Apache foundation and open sourced it in 2011. Apache Kafka is an open-source stream-processing software platform, initially built by LinkedIn, it was later handed over, developed and open sourced by the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. Think of it is a big commit log where data is stored in sequence as it happens. The users of this log can just access and use it as per their requirement. Kafka Concepts:

- Topics: Every message that is feed into the system must be part of some topic. The topic is nothing but a stream of records. The messages are stored in key-value format. Each message is assigned a sequence, called Offset. The output of one message could be an input of the other for further processing.
- Producers: Producers are the apps responsible to publish data into Kafka system. They publish data on the topic of their choice.
- Consumers: The messages published into topics are then utilized by Consumers apps. A consumer gets subscribed to the topic of its choice and consumes data.
- Broker: Every instance of Kafka that is responsible for message exchange is called a Broker. Kafka can be used as a stand-alone machine or a part of a cluster.

Some of the common use cases of Kafka are:

- real-time processing of application activity tracking, like searches;
- stream processing
- log aggregation, where Kafka consolidates logs from multiple services (producers) and standardises the format for consumers.

- another interesting use case that has emerged is the microservices architecture. Kafka can be a suitable choice for event sourcing microservices where a lot of events are generated and we want to keep track of the sequence of events (i.e. what has happened).

2.1.1 Producer

In the producer we have implemented three methods:

- *Connect Kafka producer()* that will give us an instance of KafkaProducer, connecting to the bootstrap server;
- *Publish message (producer_instance, topic_name, key, value)* takes in input the produce instance, the key and value and the topic name. It will publish a message with key and value converted into bytes, to a previous specified topic in kafka cluster.
- *Datepicker (url)* is a scraping function that uses Selenium, an open-source web-based automation tool, in order to scrape the archive of the website Broadcastify. The workflow is:

  1. It calls an instance of Chrome webdriver in order to open the url passed in the function input.
  2. In the url finds the source code and it finds elements by xpath, searching for the table in which the calendar is built, and saves the date into a list (elements).
  3. Once the calendar has been stored, it searches through the elements of the list and it clicks on the element that is equal to the date of the day before.
  4. Once that the dates has been clicked, it searches the xpath corresponding to the table with the starts and ends time of the calls, saving them into another list (l).
  5. It returns a json file with the day that the calls were made and the length of the list with all the calls of that day (len(l)).

Once that this method has been implemented, we called it in the main of the producer in order to publish a message into the topic of a division, where the value of the key "bu" is the json with calls and date.

2.1.2 Consumer

Once that all the messages were sent to the Kafka cluster through broker -i.e. when the days of acquisition ended- we called the consumer through an instance KafkaConsumer, a client that consumes records from a Kafka cluster, in order to read all the json files in a topic. Per each message we appended it on a list, in order to create then a dataframe per each division. Finally, once disconnected, the consumer stream is closed by calling consumer.close(), sending an ACK to the Kafka broker. The dataframes created have three columns: "Divisions" that indicates the name of the division, "Date" and ""number_of_calls".

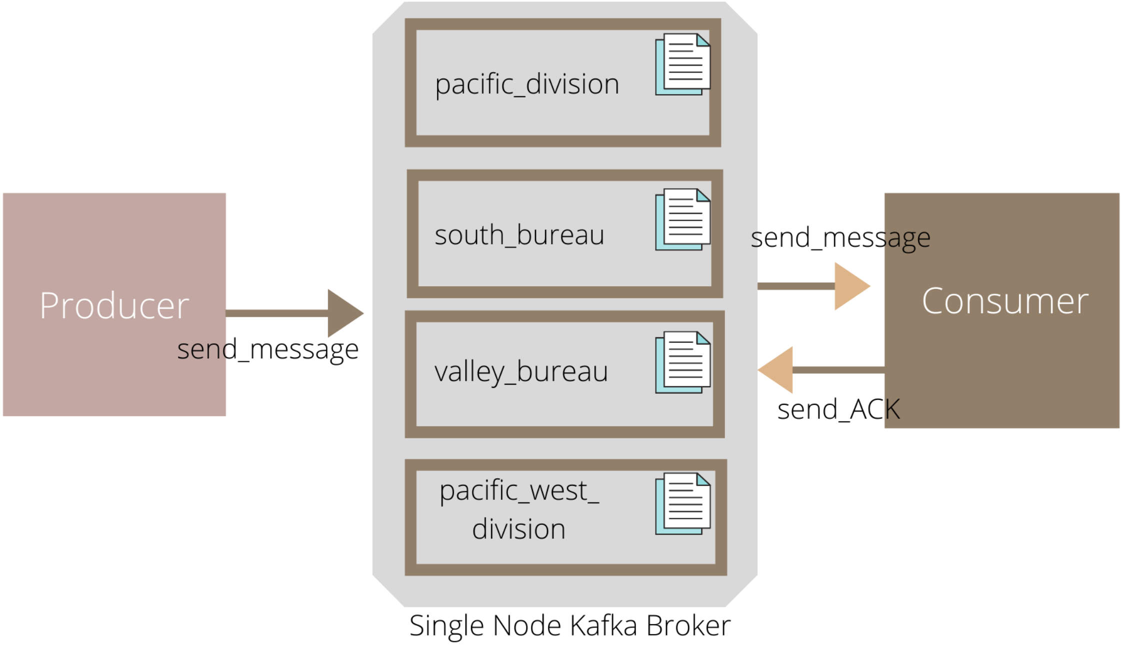


Fig. 2. Kafka workflow

2.2 Variety

First of all we have verified data quality [6] of all the sets of data analyzing these metrics:

1. Accuracy: it's the degree to which data has attributes that correctly represent the real value in a specific context of use. It could be both syntactic and semantic: syntactic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered syntactically correct; semantic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered semantically correct.
2. Completeness: it's the degree to which an observed phenomena is correctly represented from sets of data; it could refers to tuple's completeness, attribute's completeness and table's completeness
3. Consistency: it refers to an overall coherence that data should show, for example there could be consistency of data with integrity's constraints ecc.
4. Temporal Dimensions: it's the currency with respect to change in data.

2.3 Integration Part 1

2.3.1 Dataset concerning gentrification indexes

The first dataset that we have used is the geojson file concerning the gentrification indexes. We've verified the table's completeness with the following formula:

$$\frac{\text{Numbers of null values in table}}{\text{Number of rows} \cdot \text{Number of columns}} \tag{1}$$

Considering the fact that the NaN and null values in the table were 0, the ratio was equal to zero, denoting the presence of a complete table. Our goal was to categorize the zipcode into 4 macro divisions corresponding to the Los Angeles Police Department macro divisions, and for doing that we built a method in python called "create" that while reading the whole zipcode's column, returned the macro divisions to which that zip-code was associated. In order to guarantee a correct association we used a reference table provided by LAPD [4], showns in the A appendices, obtaining the group mentioned above. Subsequently, to ensure consistency, considering that the data set had a column with the geographic coordinates for the area of each zip-code, we have computed the sum of all the polygons whose zip-code referred to the same divisions. For this purpose, we have implemented a function called "geometry" that takes as parameter the name of the division and returns per each macro-division, its geographic polygon. Thanks to this analysis we have obtained two dataset:

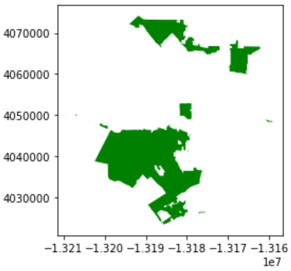


Fig. 3. Total polygon of West Division

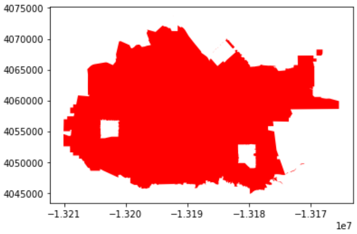


Fig. 4. Total polygon of Valley Bureau

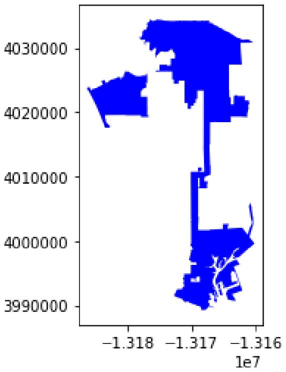


Fig. 5. Total polygon of South Bureau

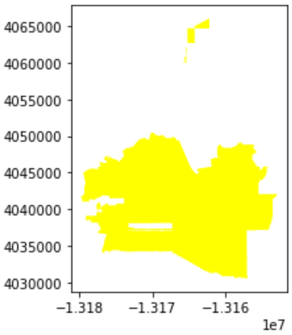


Fig. 6. Total polygon of Central Bureau

- *prova.geojson*, a so-called "enriched dataset" intended to be used for producing data visualization, whose columns are the same as the original dataset plus the divisions columns and the geometry columns with polygons of the divisions.

- *divisions.geojson*, intended to be used for the second part of the integration process, whose rows are just 4, one for each divisions, and the columns are the name of the division, the geometry and the amount of population. This last column was obtained summing the population per each zip code associated to a division.

Thanks to the dataset *divisions.geojson*, we were capable to plot a choropleth map of Los Angeles Macro divisions' population in figure 7, in order to have an overall idea of which division was the most populated in 2014. We can notice that the division with the largest population was in South Bureau.

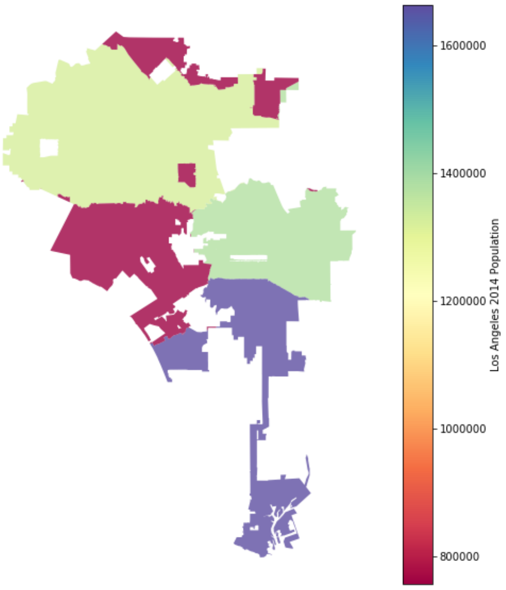


Fig. 7. Los Angeles Police Department Macro-Divisions Population

2.3.2 Dataset concerning calls for service

In order to obtain the FID number associated to each district, we have merged a csv file provided also by LAPD called *divisions.csv* with the file that contains effectively the calls for service (-i.e. the ones called *LAPD_Calls_for_Service_2016*). The data set "divisions" is structured as following, with ArcGIS geoprocessing tool that have added an shape_area and a shape_length field:

- FID: numeric variable, it indicates the district associated to the zipcode;
- APREC: string, it indicates the name of the district;
- PREC: numeric variable,
- AREA: numeric variable, it illustrates the dimension of the area;
- PERIMETER: numeric variable, it indicates the perimeter of the district taken into consideration;
- SHAPE_Length: numeric variable, it is always in the units of the output coordinate system specified by the Spatial Reference parameter; and it is the planar lengths of the polylines;

- SHAPE_Area: numeric variable, it is the planar area of the polygons and it is always in the units of the output coordinate system specified by the Spatial Reference parameter.

In other words, since the data set, containing the calls for service, does not have the districts' number, we have merged it with the division data set (exploiting this last one like a reference table), using as foreign keys "Area Occurred" (for *LAPD_Calls_for_Service_2016.csv*) and APREC (for *divisions.geojson*). To understand better the dynamics of this merge: Considering the join between the two variables, in order to ensure a complete linkage among them, we have computed the edit distance between "APREC" and "Area Occurred" values, obtaining this output 1:

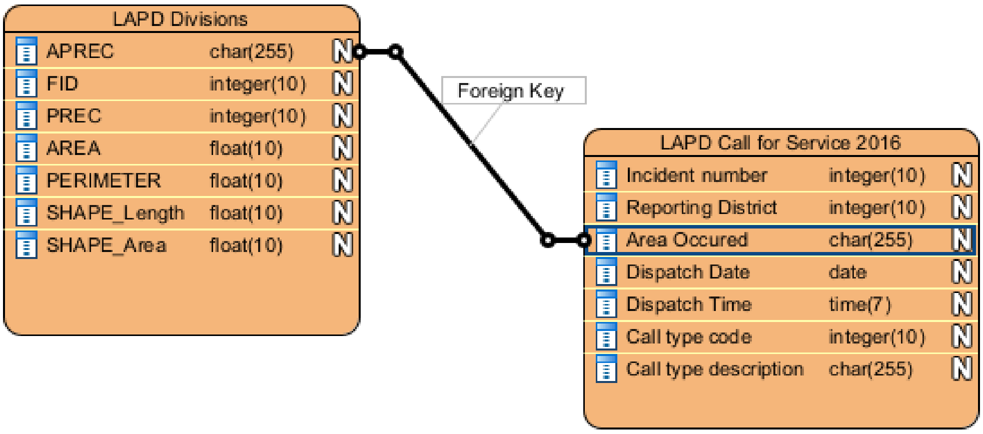


Fig. 8. Inner join among Divisions and calls for service

Area Occurred	APREC	Edit Distance
77th Street	77TH STREET	7
Central	CENTRAL	6
Devonshire	DEVONSHIRE	9
Foothill	FOOTHILL	7
Harbor	HARBOR	5
Hollenbeck	HOLLENBECK	9
Hollywood	HOLLYWOOD	8
Mission	MISSION	6
N Hollywood	NEWTON	10
Newton	NORTH HOLLYWOOD	14
Northeast	NORTHEAST	8

Area Occurred	APREC	Edit Distance
Olympic	OLYMPIC	6
Pacific	PACIFIC	6
Rampart	RAMPART	6
Southeast	SOUTHEAST	8
Southwest	SOUTHWEST	8
Topanga	TOPANGA	6
Van Nuys	VAN NUYS	5
West LA	WEST LOS ANGELES	12
West Valley	WEST VALLEY	8
Wilshire	WILSHIRE	7

Table 1
Structure of the table with the comparison among the columns and edit distance

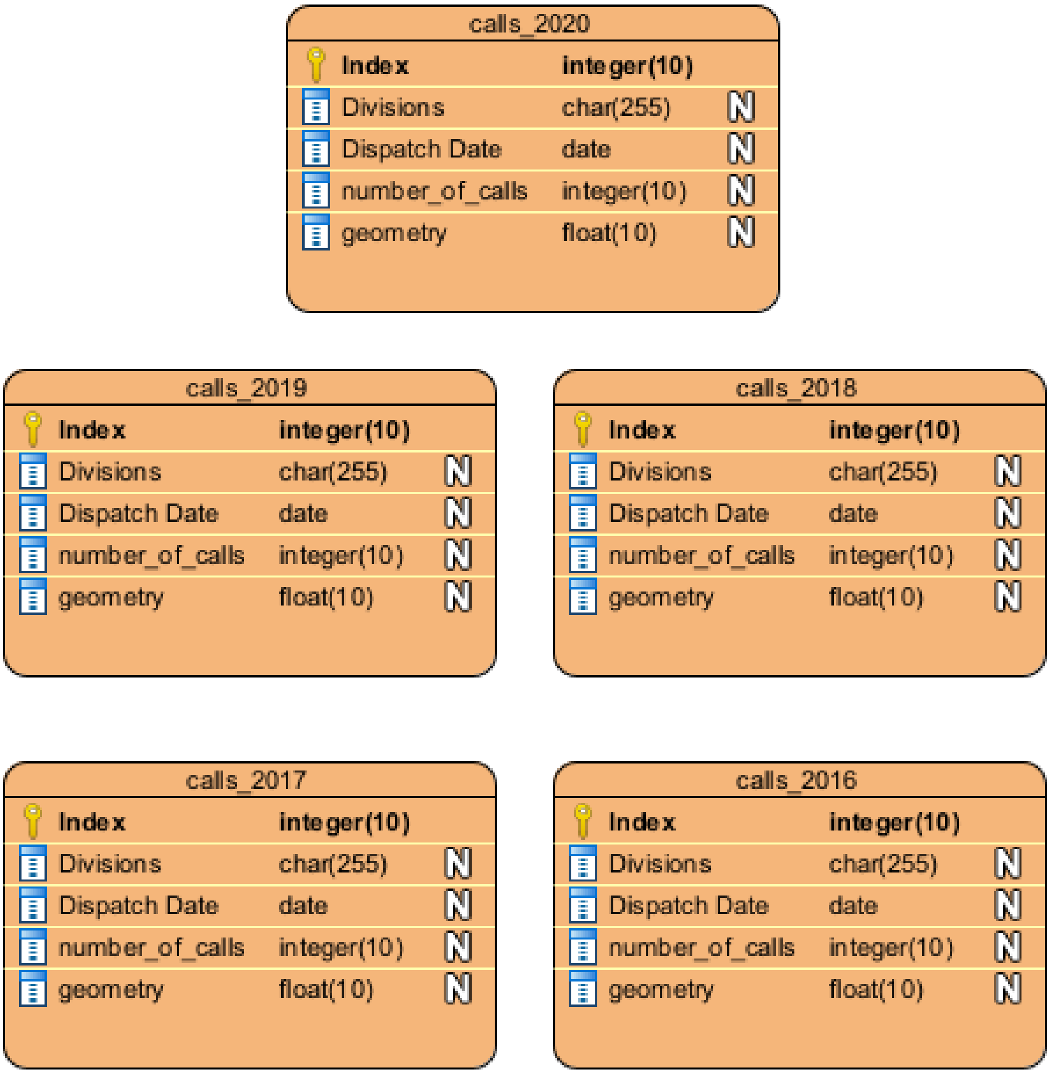
In order to nullify the edit distance we have lowered all the values both in the APREC and in the Area Occurred column; furthermore we've noticed that the district "North Hollywood" and "West LA" were written in two different ways in the columns, generating a name conflict. This has led to a different alphabetical order, in fact the edit distance function was comparing two values referring to different districts (N Hollywood vs Newton). To solve this problem we have standardized the name of the districts, in the following way:

- n hollywood ⟶ north hollywood;
- west la ⟶ west los angeles.

Furthermore in the dataset of 2019 and 2020 there were another category in the "Area Occurred" variables that stated "Outside". In order to uniform all the years from 2016 to 2020 we have decided to not consider crimes from outside LA areas. Once that the data set has been created, we have used the FID columns (with number of the districts) to aggregate in macro divisions; furthermore, we have filtered per each year only street's related crimes, keeping the ones that involved traffic stops, GTA (Grand theft auto) and car stripping. This decision was done because the data acquired in the velocity part were street's related. Once that filtering was done, we grouped by divisions and dates and we counted all the number of calls from the same divisions in different date in order to obtain the following geodataframe 9, called *calls_2016.geojson* (one for each year).

2.4 Integration with datasets of 2021

In this part we concentrated on the creation of two dataset: a csv file and a geojson file, using datasets both from velocity part and from the first part of the variety. We concatenated the four data sets obtained during the velocity part, concerning data on number of calls to LAPD for street related crimes, from the 1st of January of 2021 to the 28th of February of 2021. Furthermore, since Pacific is a district of the West, we computed the sum among all the calls from Pacific and West divisions. Moreover, in order to standardize the columns' names of the acquired data based on the data provided from LAPD, we renamed them and we converted the dates to datetime format. In order to create the csv file, we concatenated the dataset related to 2021 (obtained through the velocity phase) to the others obtained above (integration part 1) concerning calls from 2016 to 2020. We've kept only the columns: "Divisions", "Dispatch Date" and "number_of_calls", obtaining the final dataset called *calls.csv*. On the contrary, the idea behind the geojson file was different: our goal was to obtain a column per each year (between 2016 and 2020) with the total amount of calls form each divisions, with the relative geometry. To do that, we have taken all the dataset of calls created before and we have extracted the month and the day from the column "Dispatch Date". In order to create the geojson file, we merged by pairs each yearly dataset on Divisions and month-day, summing yearly the total amount of calls per each division, obtaining the following datasets 2 called *total_calls.geojson*.



calls_2020

Index	integer(10)	
Divisions	char(255)	N
Dispatch Date	date	N
number_of_calls	integer(10)	N
geometry	float(10)	N

calls_2019

Index	integer(10)	
Divisions	char(255)	N
Dispatch Date	date	N
number_of_calls	integer(10)	N
geometry	float(10)	N

calls_2018

Index	integer(10)	
Divisions	char(255)	N
Dispatch Date	date	N
number_of_calls	integer(10)	N
geometry	float(10)	N

calls_2017

Index	integer(10)	
Divisions	char(255)	N
Dispatch Date	date	N
number_of_calls	integer(10)	N
geometry	float(10)	N

calls_2016

Index	integer(10)	
Divisions	char(255)	N
Dispatch Date	date	N
number_of_calls	integer(10)	N
geometry	float(10)	N

Fig. 9. Structure of the built datasets of yearly calls

Divisions	2016	2017	2018	2019	2020	geometry
Central Bureau	3902	6161	7273	8494	7716	MULTIPolygon
South Bureau	3906	5243	7136	7398	6988	MULTIPolygon
Valley Bureau	10095	10331	14112	14507	12790	MULTIPolygon
West and Pacific	7607	9376	11675	13543	12002	MULTIPolygon

Table 2
Structure of "total_calls.geojson"

2.5 MongoDB

NoSQL database system such as Mongodb [7] is document oriented database into which data is organized as key value pairs across lightweight Binary JSON documents: JSON is formatted as name/value pairs. In JSON documents, fieldnames and values are separated by a colon, fieldname and value pairs are separated by commas, and sets of fields are encapsulated in “curly braces” ({}). Thereby this facilitates flexibility in schema design which is a contributing factor in offering high performance against processing massive volume of data. This is the main reason why we have decided to store the dataset ”calls.csv” into mongo db, in fact, while a table might seem like a good place to store data, there might be fields in the data set that require multiple values and would not be easy to search or display if modeled in a single column. Our dataset have multiple dates referring to different divisions and in order to facilitate the query, this storing was the most efficient.

Example of one document:

```
    _id:ObjectId("605a5a9152836ffce1d64d12"): 0
    Dispatch Date: "2016-01-01 00:00:00"
    Divisions: "Central Bureau"
    number_of_calls: 16
```

For instance we can ask to mongo the following queries:

- number of documents in the collection:

```
        query1_result = calls.estimated_document_count()
        print(query1_result):

        7477

```

- the number of calls from South Bureau of 2nd February of 2018

```
        query2_result=calls.find({"Divisions":"South Bureau", "
    Dispatch Date":"2018-02-02 00:00:00"})
        print(list(query2_result)):

        [{’_id’: ObjectId(’605a5a9152836ffce1d65a09’), ’’: 395, ’
    Dispatch Date’: ’2018-02-02 00:00:00’, ’Divisions’: ’South
    Bureau’, ’number_of_calls’: 23}]

```

- counting number of days where number of calls is between 33 and 45:

```
        days=calls.find({’number_of_calls’ : { ’$gt’ :  33, ’$lt’ :
     45}}).count()
        print(f"the number of days that have calls between 33 and
    45 is {days}"):

        the number of days that have calls between 33 and 45 is 767

```

Besides queries we have examinated data on MongoDB through several graphs that could be found on this link Data Analytic and Exploration.

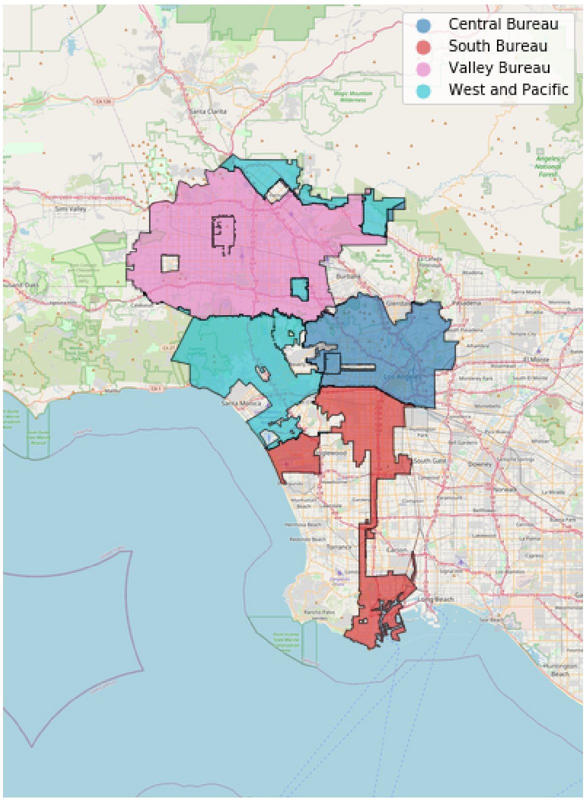
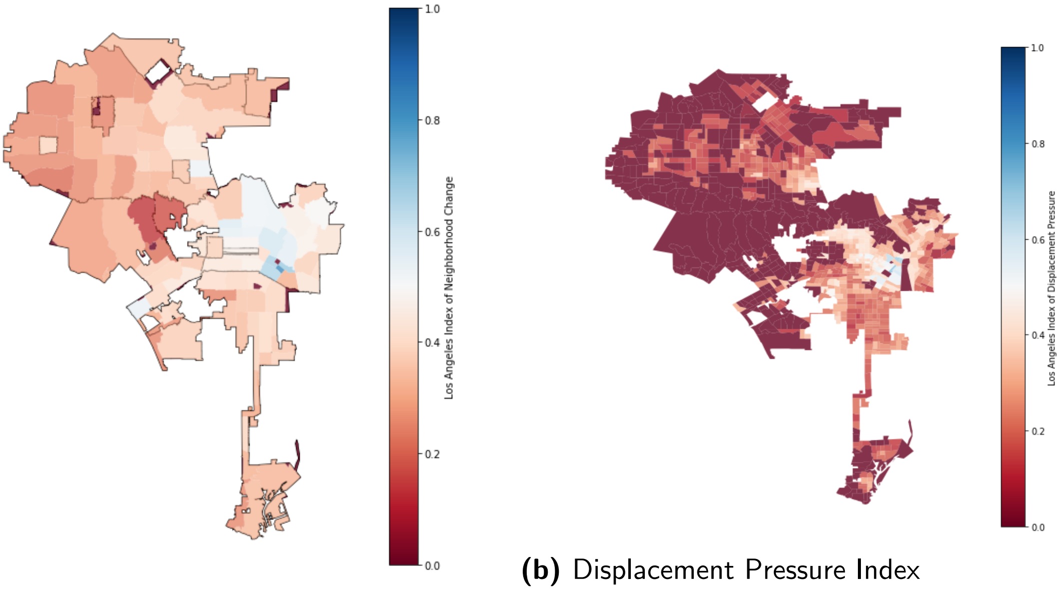


Fig. 10. Los Angeles Police Department Macro Divisions

3 Data Visualization

The idea behind the first group of infographics that we have developed in Tableau was to show the change in the gentrification indexes during 14 years, from 2000 to 2014 here: Analysis of Gentrification Indexes. What has happened during these last years is unknown because we have the data only for 2000 and 2014, not for the years in between. Because of this, we have decided to build some choropleth maps showing just the indexes for the 2014, in order to have a generalised idea of how the situation was in that year. First we started showing the macro divisions' map thanks to the aggregation of the zip codes into South Bureau, West and Pacific, Central Bureau and Valley Bureau, using reference tables in the appendix, as shown in the figure 10. We, then, decided to plot also the Neighborhood Change Index and the Displacement Pressure Index of 2014 in order to have an overview of which division had some indexes that meant to be noticed. First of all, we have to understand whether an index is significant or not, for this purpose thresholds have been defined. Considering the Neighborhood Change Index, there has been a very high change if it's greater than 0.8; a high change when it is between 0.6 and 0.8; a medium change when it is between 0.4 and 0.6; a low change when it is between 0.2 and 0.4 and finally no/minimal/reverse change when it is smaller than 0.2. Looking carefully at map 11a, it is worthwhile to notice that the division which has the highest Index is Central Bureau, where the Neighborhood Change varies among zip codes from about


1.0
0.8
0.6
0.4
0.2
0.0
Los Angeles Index of Neighborhood Change
1.0
0.8
0.6
0.4
0.2
0.0
Los Angeles Index of Displacement Pressure

(a) Neighborhood Change Index

(b) Displacement Pressure Index

Fig. 11. Gentrification Indexes of 2014

0.31 to 0.68, while the other divisions' indexes lie in smaller intervals with some exceptions. Concerning the Displacement pressure in figure 11b, it is considered very high when the index is greater than 0.377; high when it is between 0.284 and 0.377; medium when it is between 0.203 and 0.284; medium/low when it is between 0.162 and 0.203 and finally low when it is smaller than 0.162. Some of the zip codes were excluded from the analysis because they were considered not eligible for being gentrified. Looking at the map, it can be noticed that, again, Central Bureau has the highest index among divisions, with a variability from 0.17 to 0.67, and it also has most zip codes that are eligible for being gentrified, while South Bureau and Valley Bureau have zip just some zip codes with a medium/high displacement pressure index. West and Pacific, instead, is the division were almost every zip code has been considered not eligible for being gentrified.

Diving deeper into our analysis, we have investigated also the correlation between gentrification indexes in the 2014, plotting a correlation chart in figure 12. We can notice that *Household income Ratio* is positively correlated, with a correlation greater than 0.85, to *Gross Rent*, *Population with Degree* and *White Percentage*. This means that if the Household Income of a district increases (most of the families earn more), we could expect an increase also in the gross Rent and it is likely to have more white people, with degree. Hence, among this variables there is always a positive correlation, a symptom of multicollinearity.

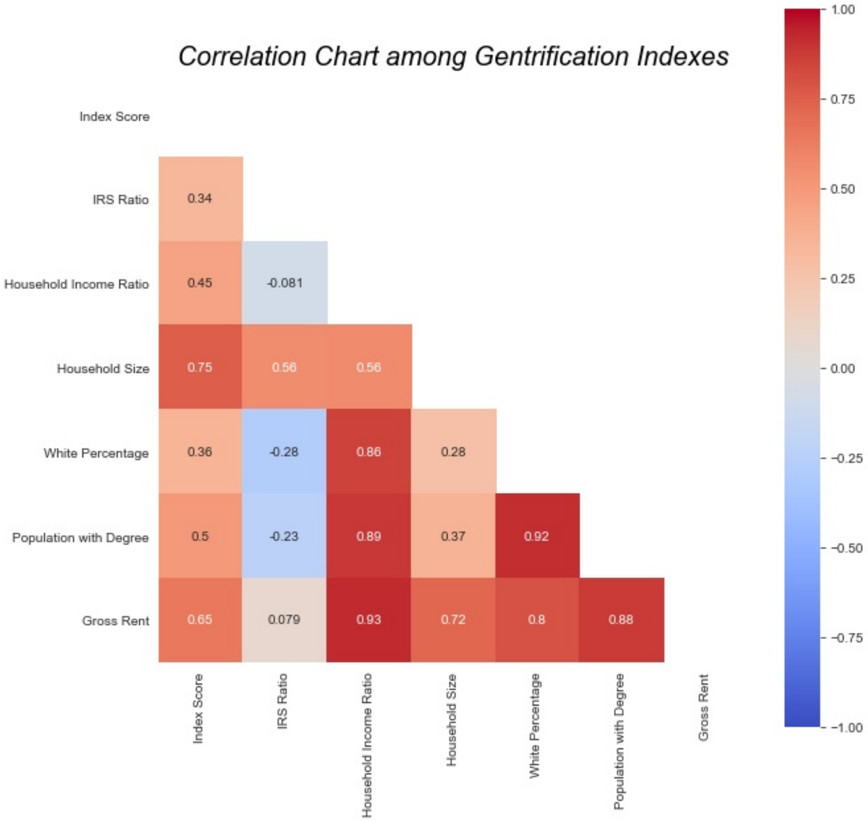


Fig. 12. Correlation between Indexes

Considering that the Index Score is a linear combination of all the six indexes, looking at the correlations [8] of this variables with the others indexes, it can be noticed that the *Household Size* and the *Gross Rent* are the factors that influenced it more, denoting that if you have a large family and a high gross rent, it is more likely to see your neighborhood change.

In order to analyze the change in gentrification measures, we have computed the ratio between the measures' values of 2000 and the ones of 2014.

$$\text{Gross Rent Change} = \frac{\text{Gross Rent 2014}}{\text{Gross Rent 2000}} \tag{2}$$

$$\text{Household Income Change} = \frac{\text{Household Income 2014}}{\text{Household Income 2000}} \tag{3}$$

$$\text{Household Size Change} = \frac{\text{Household Size 2014}}{\text{Household Size 2000}} \tag{4}$$

$$\text{IRS Ratio Change} = \frac{\text{IRS Ratio 2014}}{\text{IRS Ratio 2000}} \tag{5}$$

$$\text{Population with degree Change} = \frac{\text{Population with degree 2014}}{\text{Population with degree 2000}} \tag{6}$$

$$\text{White Percentage Change} = \frac{\text{White Percentage 2014}}{\text{White Percentage 2000}} \tag{7}$$

If a ratio is greater than 1, then the value of the index taken into consideration in the 2014 is greater than the one of 2000, and this indicates an increase of the index. As shown in the graph on Tableau (accessible clicking on the link in the Appendix B) it is possible to see which measures for the Divisions has increased/decreased. Overall, according to the indexes evaluated before, Central Bureau has most of the indexes smaller than 1, and the 68% of the observations fall in an very large confidence interval, this is due to a large standard deviation associated to each ratio. Considering that the IRS Ratio has decreased, indicating a poorer society and that the white percentage has decreased in some districts and increased in others, and that the population with a degree change drastically among zip codes, we suppose that this division might be a victim of gentrification. Furthermore, we can make the same assumptions for South Bureau, even though the variability in not such as significant as Central Bureau. Thanks to this analysis we might identify as gentrified West and Pacific Division and as being gentrified both Central Bureau and Valley Bureau. Regarding South Bureau, it could not be done any statements because the indexes were contraddictory among them.

After having individuated as gentrification victims Valley Bureau and Central Bureau, we wanted to investigate whether the rate of crime from these divisions have increased or not over the next year. In order to do this we have developed the last infographic on a Tableau's dashboard and the idea behind was to show the evolution of calls from 2016 to 2021 in the four macro-divisions: Number of Calls from 2016 to 2021 to to LAPD per Division. Analyzing this area chart, it emerged that overall the divisions with the more number of calls are West and Pacific and Valley, and that the peaks of the calls are concentrated in 2019 and on the start of 2021. (It is possible to notice that for the first two months of 2021 Central Bureau's data are missing because we haven't found any radio from which acquire data). Diving deeper into our analysis, we investigate the peak of June 2019 that we can point out from the previous area chart. It is worthwhile noticing that the central part of Los Angeles including some districts of South and Central Bureau are the most dangerous. Moreover, doing further researches and plotting a stacked barplot, we can state that the day with the highest number of calls in the period taken into account (that is june 2019 peak) seems to be Tuesday. Last but no least, the most frequent type of crime in 26th week of 2019 is traffic related incidents, followed by GTA and car strip.

3.1 Infographics Evaluation

In order to evaluate our infographics, we used a qualitative evaluation, through Schneidermann Heuristics and a quantitative evaluation using a Cabitza-Locoro Questionnaire and a User test (the summary of the evaluation could be found in the last link in the Appendix B).

3.1.1 Heuristic Evaluation

For this purpose we have asked 6 people to read carefully the following Schneidermann Heuristics [9], taken by the following website [10]:

1. Consistency at all costs: in a system, similar action sequences are needed for similar situations. The terminology must be identical in the prompts, menus and help screens. The use of colors, layouts and lettering must be consistent throughout the system. Standardizing the way information is conveyed allows all users to quickly become familiar with the environment.
2. Universal usability: recognize the needs of different types of users, considering users of all ages and with different technological backgrounds. For example, adding guides for inexperienced users and special functions for experienced users improves the perceived quality of the system.
3. Offer informative feedback: each user action must correspond to a response from the interface so that each user knows what is happening at all times in a clear, appropriate and legible way. For frequent and minor actions, the response may be modest, while for infrequent and important actions, the response should be substantial.
4. Dialogue with users: the sequences of actions must be organized with a beginning, an intermediate point and a conclusion. Informative responses to the completion of a group of actions to users satisfaction of completion, a sense of relief, an indicator to prepare for the next group of actions. E-commerce sites, for example, convey users from product selection to checkout and end with a confirmation page that completes the transaction, adding a "Thank you" at the end of the purchase gratifies the user but clearly describes the conclusion of the process.
5. Prevent mistakes: as far as possible, it would be advisable to design the interface so that users are not led to make mistakes. The interface should provide simple, constructive, and solution-specific instructions. For example, users shouldn't need to reset an entire module in case they enter wrong data, but they should be guided to correct only the faulty part. Incorrect actions should leave the interface unchanged, instructions on restoring operations.
6. Ensure reversibility: as far as possible, the actions should be reversible. This feature relieves anxiety, as users know that mistakes can be undone and encourages the exploration of new options. Survivors' units should be a single action, a data entry activity or a complete group of actions.
7. Give users control: power users want to be in control of the interface and want it to respond to their inputs without hesitation. They hate surprises or changes and are annoyed by data entry sequences, difficulties in obtaining key information and the inability to achieve the desired result.
8. Reduce short-term memory load: man's limited ability to process information in short-term memory requires designers to avoid interfaces in which users

> must store information to be carried over from one screen to another. It means that forms, especially on small displays, should be compacted to fit on a single screen.

Once that the heuristics were read, we asked them to analyze our infographics looking for something that doesn't fit the points read. In the first infographics, the remarks noted by the interviewers were mainly about technical issues with Tableau, in fact reversibility was not always perceived, since most of the people were not familiar with Tableau's user interface. At some points users faced a lack of clarity in the measures and indexes explanation due to their poor knowledge about the topics involved. Overall, the second infographic was appreciated for its simplicity, immediacy, coherency and easy usability.

3.1.2 Quantitative assessment

For this assessment we led two interviews:

- Cabitza-Locoro questionnaire [11], proposed to 25 people, requires them to rank from 1 to 6 both story and dashboard, according to the following parameters: beauty, clearness, informativeness, usefulness and an overall idea. It's worthwhile noticing that almost 90% of the people has considered our story clear and informative, with a rank from 3 to 6. While the dashboard reaches higher score in almost every parameter, leading to a better appreciation from all interviewers.

- a user test, proposed to 12 people, where we asked them to time themselves while answering these questions:

  1. Which is Valley Bureau's IRS Ratio Change? Answer: 0.91;
  2. Which are the measures with the highest correlation index (cor = 0.8)? Answer: Household Income Ratio Population with degree + White Percentage + Gross Rent;
  3. Which division has the lowest Neighborhood Change Index? Answer: West and Pacific Division;
  4. How many calls the LAPD received in September 2018 in Central Bureau? Answer: 795

  The optimal duration was measured by letting users run each task 3 times, quickly and slowly and then calculating the average time. Moreover we have decided to choose simple questions in order to verify base infographics' comprehension.

4 Conclusion

In conclusion, six measures are not enough to make clear statements on gentrified divisions, but we could notice that the West and Pacific division is gentrified, in fact the percentage of white people has increased over 14 years and also the households' income, denoting a richer population, with few zip codes eligible for the study. Nevertheless, the rate of street related crimes is very high, so we cannot make further statements. On the other hand, we have Valley Bureau and Central Bureau whose indexes are increased, but the variability associated to each one is very large, denoting areas with wealthier people and areas with poorer people. Probably, analyzing the infrastructures in the poorer neighborhood we could find decay buildings and low income schools, but we don't have data to state it with certainty. Anyway, the street related crimes from Central Bureau until December 2020 were very high too, (it's really a pity that Broadcastify don't provide data for this division), so the high rate is both from gentrified division and the ones being gentrified, characterizing two sides of a coin. To wrap up, when talking about gentrification, we have to ask ourselves rigorous questions about the many processes we're likely untangling, seek more evidence, and consider all sides. We'll find out that we're talking about something far bigger than can be encapsulated in a single word. The only way to really be aware of the consequences that this current phenomena can bring with is keeping talking and working to learn more about it.

5 References

1. *Oxford Languages* https://languages.oup.com/.
2. Lees Slater, W. The Gentrification Reader. (English) (2010).
3. *Los Angeles Index of Displacement Pressure (English)* https://www.arcgis.com/home/item.html?id=70ed646893f642ddbca858c381471fa2.
4. *Los Angeles Zip Code Map, [GUIDE TO LOS ANGELES ZIP CODES AND NEIGHBORHOODS]* https://www.usmapguide.com/california/los-angeles-zip-code-map/.
5. Garg, N. *Apache kafka* (Packt Publishing Ltd, 2013).
6. Fan, W. & Geerts, F. Foundations of data quality management. *Synthesis Lectures on Data Management* **4,** 1–217 (2012).
7. Győrödi, C., Győrödi, R., Pecherle, G. & Olah, A. *A comparative study: MongoDB vs. MySQL* in *2015 13th International Conference on Engineering of Modern Electric Systems (EMES)* (2015), 1–6.
8. Spearman, C. Footrule for measuring correlation. *British Journal of Psychology* **2,** 89 (1906).
9. Craft, B. & Cairns, P. *Beyond guidelines: what can we learn from the visual information seeking mantra?* in *Ninth International Conference on Information Visualisation (IV'05)* (2005), 110–118.
10. *Schneidermann's Heuristics* https://www.interaction-design.org/literature/article/shneiderman-s-eight-golden-rules-will-help-you-design-better-interfaces.
11. Locoro, A., Cabitza, F., Actis-Grosso, R. & Batini, C. Static and interactive infographics in daily tasks: A value-in-use and quality of interaction user study. *Computers in Human Behavior* **71,** 240–257 (2017).

A Reference Tables

The four Los Angeles macro-divisions with the districts and the FID numbers are:

- South Bureau, it includes Southwest 15, Southeast 19, Harbor 20 and 77th Street 18;
- Central Bureau, it includes Newton 16, Northeast 8, Hollenbeck 11, Central 21 and Rampart 12;
- Valley Bureau, it includes Devonshire 2, Foothill 3, Mission 1, North Hollywood 6, Topanga 4, Van Nuys 7 and West Valley 5;
- West and Pacific Division, it includes Hollywood 9, Wilshire 13, West Los Angeles 10, Pacific 17 and Olympic 20.

For the sake of completeness, we report also the correspondences between zip-codes and macro-division that we have dug up:

- South Bureau: '90247','90248','90293', '90001', '90002', '90003', '90007', '90008', '90011', '90016', '90018', '90037', '90043', '90044', '90047', '90059', '90061', '90062', '90089', '90305', '90045', '90245', '90275', '90301', '90302', '90304', '90501', '90504', '90717', '90502', '90710', '90731', '90732', '90744', '90745', '90802', '90810', '90813', '90058', '90262', '90280';
- Central Bureau: '90029', '90031', '90032', '90041', '90042', '90065', '90004', '90005', '90006', '90012', '90013', '90014', '90015', '90017', '90019', '90021', '90026', '90027', '90028', '90036', '90038', '90039', '90046', '90048', '90057', '90068', '90069', '90071', '90023', '90031', '90032', '90033', '90063', '91030', '91801', '91803', '91101', '91103', '91105', '91202', '91204', '91205', '91214';
- Valley Bureau: '91311', '91321', '91326', '91040', '91304', '91306', '91307', '91316', '91324', '91325', '91330', '91331', '91335', '91340', '91343', '91344', '91345', '91352', '91356', '91364', '91367', '91401', '91402', '91403', '91405', '91406', '91411', '91423', '91436', '91504', '91505', '91506', '91522', '91601', '91602', '91604', '91605', '91606', '91608';
- West and Pacific Division: '91607', '91206', '91302', '90095', '90211', '90265', '90290', '90024', '90025', '90034', '90035', '90049', '90056', '90064', '90066', '90067', '90073', '90077', '90094', '90210', '90212', '90230', '90232', '90272', '90291', '90292', '90402', '90403', '90404', '90405', '91042', '91342'.

B Tableau's Infographics

Analysis of Gentrification Indexes
Number of Calls from 2016 to 2021 to LAPD per Division
Infographics Evaluation Dashboard
Data Analytic and Exploration