How to Reactivate Bogotá's Tourism Industry

DS4A TEAM 47 FINAL REPORT

Colombia, September 2021

Team members: Julian Santiago Tauta, Alejandro Soto, Vanesa Movilla, David Gonzalez, Juan David Arias

Partner Entity: Instituto Distrital de Turismo de Bogotá

INTRODUCTION AND BUSINESS CONTEXT

The Tourism sector in Bogotá contributed 3.2% for 2019 of the city's GDP and also contributed to the generation of 6.3% of the capital's employment. It is important to mention that the dynamics of the tourism sector is interrelated with other sectors, with which they have a direct relationship (Accommodation, Agencies) and a related relationship (Transportation, food, Shopping, recreation, well-being and culture) by what tourism represents important economic income in these other sectors; Thus, a crisis that affects the sector has a significant effect on most segments of the economy.

By 2020, according to the current situation generated by the Coronavirus Pandemic, Tourism worldwide has entered a crisis, this was due to travel restrictions imposed by different governments, affecting the free mobility of tourists. This scenario, together with the economic situation of travelers, has caused changes in travel preferences, so that for the sector to have an adequate recovery and return to pre-pandemic levels, indicators, characterization of trips and the profile of future travelers with the aim of generating policies and strategies to guide the appropriate decision-making in accordance with the UNWTO.

In this regard, the Bogotá Tourism Observatory periodically generates research, studies and measurements in relation to the information that is collected from surveys, polls and secondary information.

Otherwise, the information collected by these means is limited, therefore, it seeks to take advantage of the data not conventional to generate holistic analyzes of the sector, it is thus that techniques such as "web scraping", NLP and "analysis of feeling ”can play an important role, if applied to the different accommodation establishments and tourist attractions in the city, in order to determine the

characteristics, conditions of the offer, shortcomings, options of improve and even discover the mobility patterns of tourists in the city

However, work is still being done on the generation of methodologies that allow us to analyze and visualize the data in real time, in this way to be able to make the most of the information that is captured, to identify trends and guide the decision making. In the same way, to make up for the lack of unconventional data, different databases and platforms have been acquired from private organizations, which generates additional costs in the process of being able to fulfill the missionality of the observatory.

In this order of ideas, it is necessary to design an automated data visualization and analysis system (exploratory and predictive), for which it will be of great value for the entity to be able to have a prototype of a control panel (Dashboard) in real time analyzing the database in the cloud and accessible to the different analysts and users of the information.

And also take advantage of unconventional data to generate a characterization of the offer according to the conditions of the stay, quality of the service and subsequently generate a georeferencing of the different accommodation establishments for the 20 towns of Bogotá. At the same time, take advantage of the data from various review websites and / or social networks to generate a Sentiment analysis for the tourist attractions of the city according to the inventory of the Vice Ministry of Tourism.

APPLICATION OVERVIEW

The application that was created for the solution of the problem raised by the District Tourism Observatory of Bogotá, has the name Trip-City.

The application is built with the primary objective of generating knowledge valuable enough to be a support to make both public policy decisions, as well as decisions of private entities and people who are looking to increase their income by renting their properties to visitors and tourists. from the city of Bogotá.

Trip-City works in a very intuitive way, on the left side is the main menu, in which we find "Home", "Traveler Research", "Tourism Indicators", "Flights", "Lodging", "Tourist Sites" and "Lodging Prices".

The “Home” part gives a quick introduction to the most interesting and relevant parts of the application, the “Traveler Research” sections, creates through different descriptive graphics a profile of the people who visit Bogotá and its tourist attractions, the “Tourism Indicators” is a board that allows a constant update of different indicators of the city.

The “Flights” section, provides an important description of the origin of flights that land in the city of Bogotá, thus identifying which are the countries that contribute a greater number of tourists; “Lodging”,

is a description of the sector, mainly of proposals that emerge like airbnb, which have a great reception and generate important income, and finally two of the important adaptations of Trip-City, “Tourist Attractions”, which is a detailed description using Natural language processing of the opinions that visitors have of more than a hundred tourist attractions, which are in the city, which allows you to get an idea of how these places are and what type of interaction to do to improve it, and finally, “Lodging Prices”, which is a powerful model that allows to predict the price at which a property could be rented by putting certain amount of variables, this model gives us the price and behavior for each of the months of the year and also provides interesting data such as the number of tourist sites that are close to the property.

In addition, within each of these sections of Trip-City, there are subsections at the top, which allow us to delve deeper into the visualizations and know much more information that the application is generating.

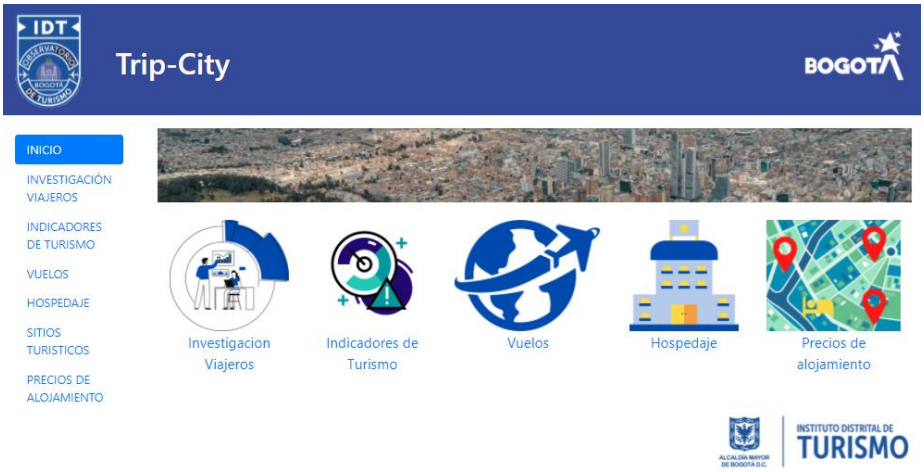
DATA ENGINEERING

Interactive Front-end

Trip-City is running in the following url: http://3.14.222.83:8050/

The interactive Front-end was built using the code hosting platform GitHub for collaborative development. We used plotly-Dash for it, based on a multi-page approach that helped us divide the coding work between all the team members. At the end we installed the last GitHub version of our code in an EC2 instance from AWS with elastic IP.

The result is a completely interactive Front-End for all the goals we accomplished: Visualizations of unified information for tourist profiles, tourism indicators, flights, lodging, tourist attractions’ sentiment analysism and prices for accomodation in Bogotá (predictive model).



Every tab in the side bar has its url and subcategories. The structure of the Front-End is the following:

Inicio (Home)

Investigación Viajeros (Travelers' profile)

- Información Geográfica (Geographical Information)
- Características Turistas (Tourist Profiles)

Indicadores de Turismo (Tourist Indicators)

- Indicadores Económicos (Economic Indicators) - PIB, Employment, TSPs
- Competitividad Turística y Turismo Sostenible (Tourism Competitiveness)
- Indicadores del Turismo Internacional (International Tourism Highlights)

Vuelo (Flights)

- Información por Contiente (Information by Continents)
- Información por Paises (Information by Counytries)
- Información por Ciudades(Information by Cities)

Hospedaje (Lodging)

- Características Económicas (Economic Features)
- Características Principales (Main Features)
- Distribución Geográfica (Geographical Distribution)

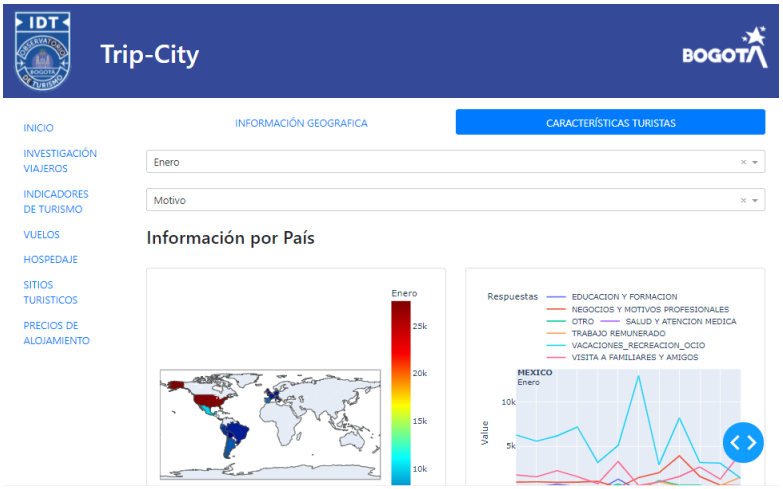
Sitios Turísticos

- Análisis por Tipos de Atractivos (Analysis by Types of Tourist Attractions)
- Análisis por Localidades (Analysis by Localities)
- Análisis por Origen de Visitante (Analysis by Visitor Origin)

Precios de Alojamientos

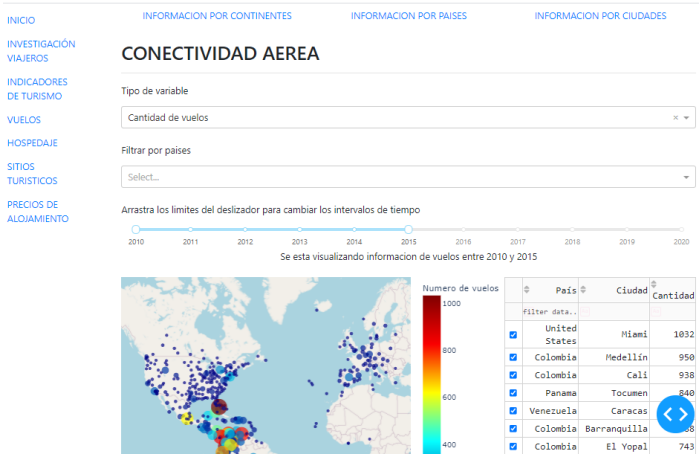
Next we will see some sections of Trip-City and its design on the front.

On the page of tourist characteristics, which is in travel research, it shows a line diagram and a map which complement each other and when you choose a country on the map, you can see the respective line diagram.

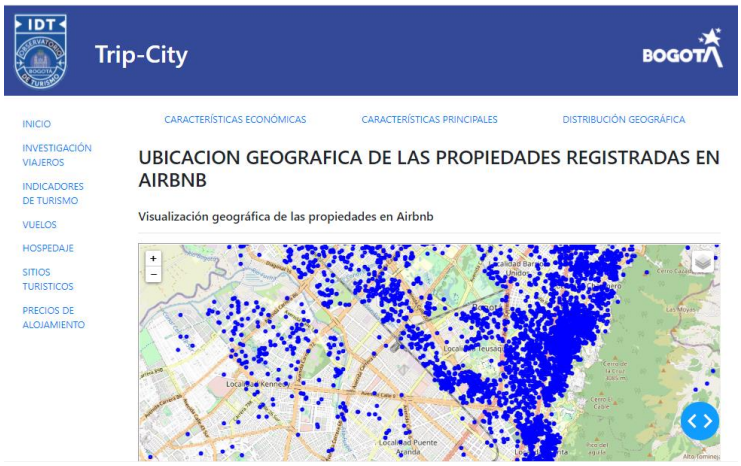


en el apartado de indicadores de turismo internacional, se visualiza la capacidad aerea para el año 2019 por pais y nos muestra la procedencia de las personas de las personas.

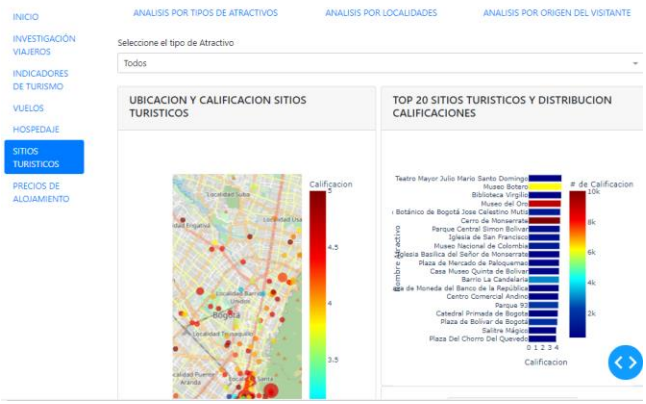
Air connectivity, shows all the countries in the world that have a direct flight to the city of Bogota, we can see this metry for different ranges of time, during the last decade.



In the following image, you can see the geographical distribution of Airbnb properties in the city of Bogotá, an interactive graph that allows us to navigate throughout the city and get closer to individualizing each of the locations



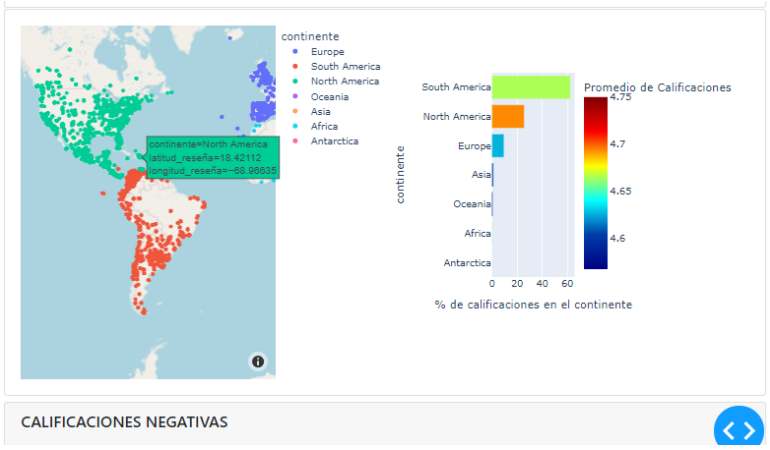
Then we go to the tourist attractions section, in this part a sentiment analysis is made of the opinions found in internet pages about many of the tourist attractions of the city, on the map we can each of the tourist attractions and where They are located and to the right a graph with the top 20 of the most rated tourist sites



In this part we can see a word cloud and one which shows us which were the words that ams were used

within the positive comments in all the attractions of the city

In this part of the front, we can see what is the geographical distribution of the positive comments that the tourist attractions of the city have

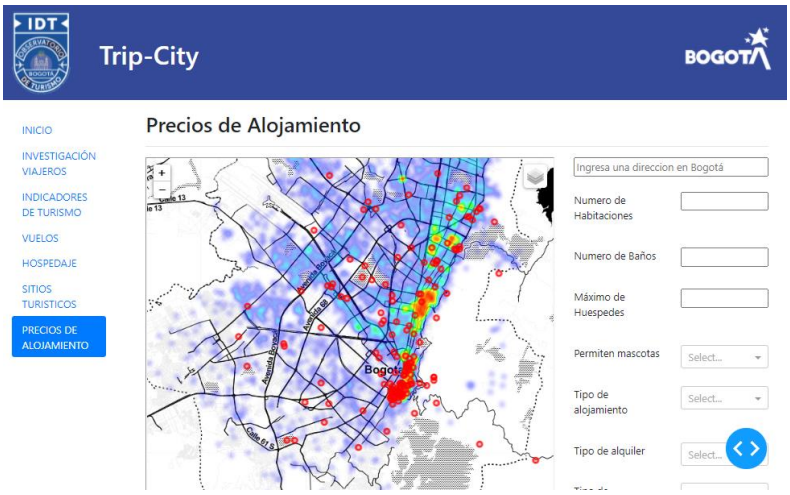


But not only can we, how positive comments are coported, we can also see negative comments, the words that are most used within those comments, as well as origin.

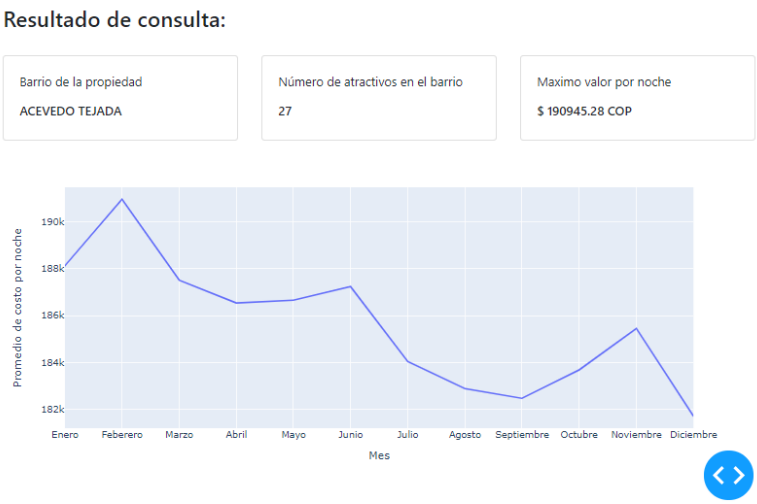
It should be noted that Trip-City allows us to observe all the part of tourist attractions in a general way or we can filter it by types of attractions and be much more specific in the search for insights

In the accommodation prices section, we find a model that allows us to make a visual comparison between the number of airbnb properties in the city (heat map) and the tourist attractions (Red dots), we can also locate the specific name of each one of the tourist attractions.

also on the right side we find the required fields for calculating the prediction of the accommodation price per night depending on the characteristics entered.



and finally we find the prediction made by the Trip-City model, which shows a timeline in which the price that should be charged per month in the accommodation is shown, it shows us in the neighborhood in which the property is located , the number of attractions near this property and the maximum price for one night



Each subcategory has its own interactive graphs and maps that can be changed by time, location, type, etc. depending on the kind of visualization that best describes the different types of information.

Databases

The databases with which the project works were provided for the most part by the Instituto Dsitrital de Turismo de Bogotá and these were:

- Microdatos Encuesta viajeros en Bogotá: Manual filled survey with information about international travelers to Bogota , it includes the profile of the traveller, purpose of the travel, activities done in the city etc.
- Tablero de Indicadores de Turismo: Main Kpis of interest for the “Observatorio de Turismo de Bogotá” , these KPIs are collected from multiples studies and data available on the web .
- Histórico de Conectividad Aérea Aero Civil: Data related to paid air traffic of passengers and cargo , it includes origin and destiny , and passengers quantity for each route in a given period.
- Establecimientos AirBnb en Bogotá: This database collects important information about the properties that are part of Airbnb during various periods of time.
- Atractivos Bogotá con los enlaces de los sitios web de reseñas:List of attractives places in Bogota with the web links of google and tripadvisor to get information about rates and comments of visitors.
- Web Scrapping: this database was obtained from websites that collect reviews from touristic places

Databases external to those provided by the Bogota District Tourism Institute were also used, to improve the interpretation given to the data and have a more general overview of them, some of these databases were:

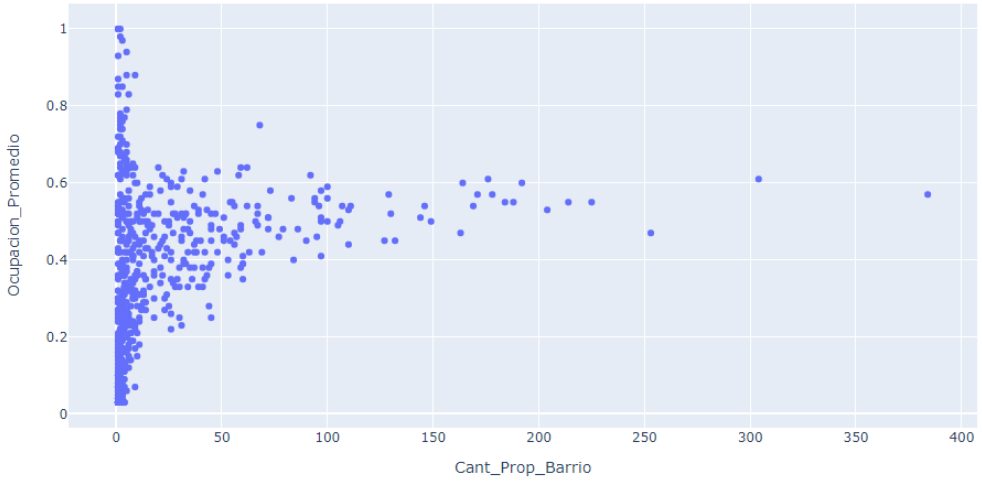
- DANE. Update date for 2020 and 2021 in the “Tourism Indicators Database” and in the “Travalers Investigation Numbers” Database.
- Aeronáutica civil: Update date for 2020 and 2021 in the “Tourism Indicators Database” and in the “Travalers Investigation Numbers” Database.
- DANE. Maps for Bogotá (“Localidades”). Info to create choropleth maps in .shp format

DATA ANALYSIS & COMPUTATION

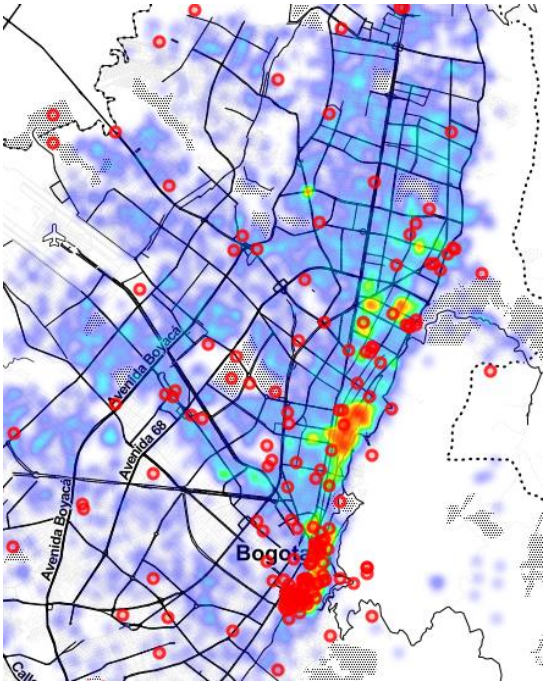
Exploratory Data Analysis

During the Exploratory Data Analysis we took all these sources of information and cleaned them, dropping missing data that could not be interpolated nor extrapolated. We unified names of countries, cities and travel purposes, among others. We mainly prepared the data for visualization in the dashboards.

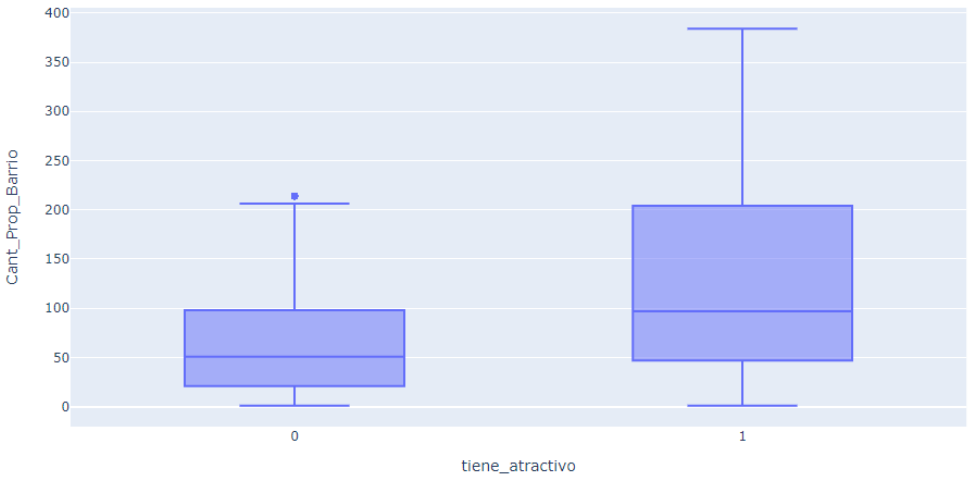
During EDA we also checked for relationships between airbnb occupation, price, location, and the location of tourist attractions (see notebook “EDA_airbnb_vs_atractivos.ipynb”). We analysed, for example relationship between average occupation rate and amount of properties per neighbourhood as can be seen in the following scatterplot that shows a positive correlation (0.31), with some neighbourhoods that have small amount of properties and high occupation rate :



However, the main findings from the Airbnb data analysis were the ones related to the location of the tourist attractions. Visually by comparing a scatter-geoplot of the attractions with a heatmap of Airbnb properties, it could be seen that they are related.

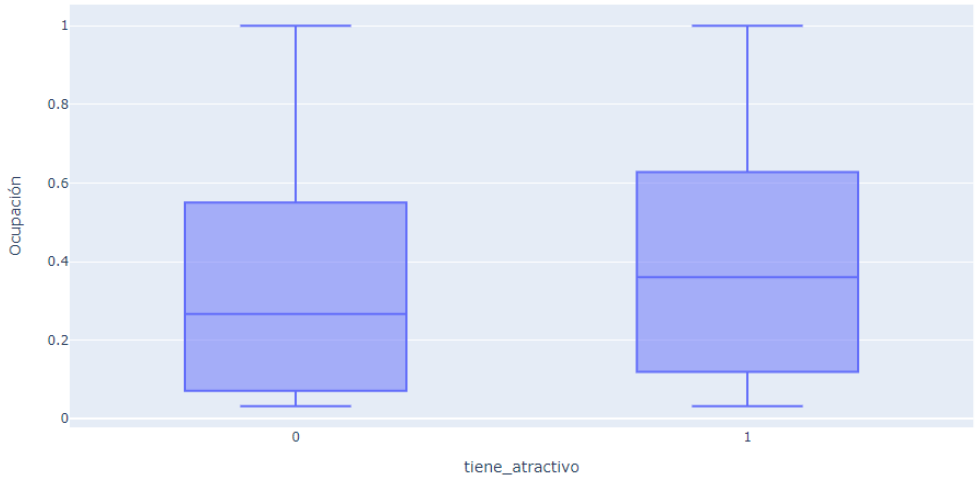


By using box plots we found that neighbourhoods with tourist attractions have an overall higher number of Airbnb properties (average of 67 properties in neighbourhoods without tourist attraction vs 130 properties in neighbourhoods with at least one of them). Both distributions are statistically significant (t-test).



	T	dof	tail	p-val	CI95%	cohen-d	BF10	power
T-test	-39.926238	6401.366979	two-sided	0.0	[-66.77, -60.52]	0.836898	inf	1.0

The mean occupation rate is also affected if the property is found in a neighbourhood with tourist attractions or not. Again using a box plot we see a clear difference. The means are 0.34 for neighbourhoods without tourist attractions and 0.39 for neighbourhoods with tourist attractions, again statistically significant (t-test).



	T	dof	tail	p-val	CI95%	cohen-d	BF10	power
T-test	-10.84706	9646.764727	two-sided	2.957472e-27	[-0.06, -0.04]	0.188451	5.172e+23	1.0

We also tested a cluster (K-means) model on the Airbnb data, but didn't obtain good insights from it (see notebook "Cluster_Kmeans_Airbnb.ipynb").

Trip Advisor Reviews

Reviews of trip advisor were obtained using web scraping, lead by the team of the IDT , also we received a list of attractions , after receiving this information we had to de work in several task to clean the database and be able to produce insights, some task were:

- Based on the data of the list of attractions we had to build a new database with the list of attractions , this was time consuming because we needed to be able to merge this data with the trip advisor review database , so an extended cleaning task was performed in both tables.

- To convert direction in geographical coordinates for each attraction we use geocoder and

google api.

· We also need to identify the origin (country/ continent) of each review , this was time consuming because this can be an open question and the format could have multiple variations , so after a cleaning process we could get geographical coordinates to identify country and continent for each review using geo polygons.

· Date of each review was also formatted.

To analyze the reviews , we generate a " clen text for each review this was the text after processing ,in the processing task we lower all the sentences, remove all punctuation , remove stop words and lemmatize using spacy Spanish library , below and example for one review :

Original review:

"asistí para ver una exposición de fotografíca bogotá 2015 y quede impactada del gran lugar que este espacio en donde se protegen los recursos documentales de bogotá y donde tambien dan espacio al arte y la cultura"

Clean review:

['asistir', 'parir', 'ver', 'uno', 'exposición', 'de', 'fotografíca', 'bogotá', '2015' 'y', 'quedar', 'impactar', 'del', 'gran', 'lugar',

'que', 'este', 'espaciar', 'en', 'donde', 'se', 'proteger', 'lo', 'recurso', 'documental', 'de', 'bogotá', 'y', 'donde', 'tambien',

'dar', 'espaciar', 'al', 'arte', 'y', 'lo', 'cultura', '.']

After this standardization we did an analysis of the positive and negative reviews based in the frequency of each word, and the frequency of each word in positive vs negative reviews , also the word close to positive or negative adjectives, all this to generate the summary of associated words to positive and negative sentiment :

Also we perform some models to classify negative or positive reviews, this application for future use in

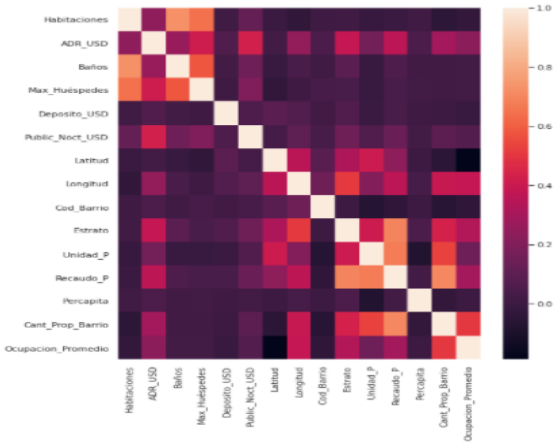
applications where the customer can write reviews (ex twitter) and the model can predict if it is positive or negative, the best model was with **Term Frequency-Inverse Document Frequency (TF-IDF)** method with logistic regression , having 97 % of accuracy.

Also we performed a clusterization of the positive and negative reviews to identify reviews that are similar in topics , with this analysis we created 9 clusters for positive and 6 clusters for negative using TFID (with bi-grams).

Statistical Analysis & Machine Learning

The model found in Trip-city, allows users to properly price properties that they already have or choose a new one to invest in according to the income they are looking for. We started the design process with a database that contains around 60 thousand entries for 15 thousand properties all over the city. Using that information, we tested 3 different machine learning models: linear regression, decision tree and random forest.

First we checked for correlations in order to avoid redundancy and get a better performance. However, the results were not significant enough to remove any variable yet.

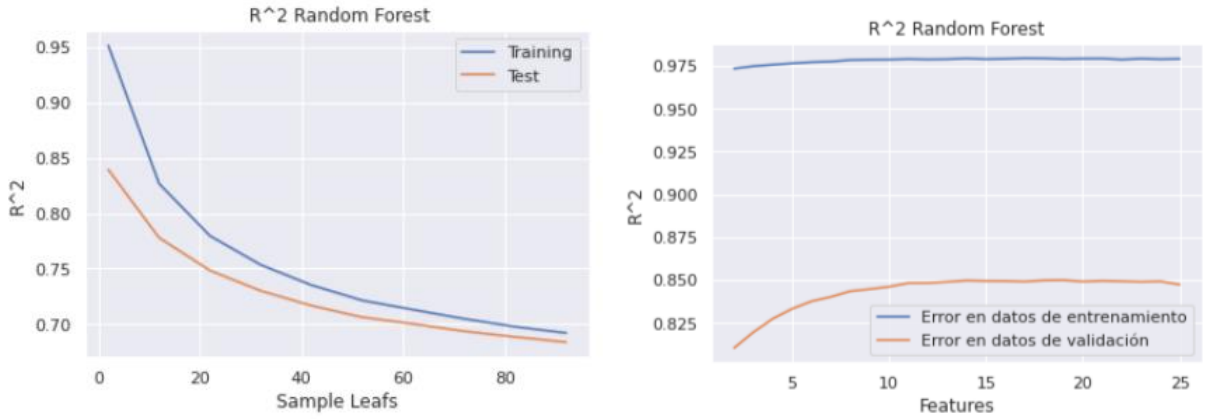


After that, we look at the summary of the linear regression model, which gives us a 0.536 R^2 at estimating the Log of the average daily rate for each property. Even if the model overall is not performing as well as we would expect, the coefficients for each of the variables give us an idea of what to adjust. For example, excluding variables such as “Recaudo_P” and “Unidad_P”

For the Random Forest option we obtained an initial R^2 of 0.83 but after hyperparameter tuning (Sample_leafs, number of estimators, depth, max_features, etc.) we managed to increase the value to

0.87, choosing the best option for each parameter. some metrics for the final version of the model are:

$$R^2 \rightarrow 0.8744$$
$$MAE \rightarrow 1.1428$$
$$MSE \rightarrow 0.056$$
$$RMSE \rightarrow 0.2366$$



One of the theoretical advantages of the implemented model is that it has a better performance in generalizing it, that is, it is consistent when going from training to validation and subsequent prediction.This improvement in generalization is achieved by compensating for errors in the predictions of the different decision trees

Another big advantage that the model has, is that it considers the number of tourist attractions in the neighborhood as well as the number of Airbnb properties that are already available there taking into account competitiveness and giving estimates for every month of the current year. When searching through the dashboard, we can go to the tab 'Precios de alojamiento' where a heat maps is showing the different tourist attractions (red markers) and a heat map with the Airbnb properties. Next to it we can find the variables necessary to predict the price. After filling the information, we are provided with a plot of the average daily rate for each month, the name of the neighborhood and the number of landmarks there.

CONCLUSIONS AND FUTURE WORK

During our Exploratory Data Analysis we found a relationship between the tourist attractions in the city and the average occupation rates in Bogotá's neighbourhoods. We developed a unified, easy-to-use source of tourist information for the city. Including a sentiment analysis tool to manage and improve

tourist attractions, as well as a price prediction tool for investors in the lodging business.

Products such as Trip-City, help cities to detect shortcomings and focus resources, with the goal of improving and enhancing the different sites of interest, in addition to showing the potential to give tools to different entities to make investment decisions in an industry that has been little exploited in Colombia, but that have so many possibilities given the characteristics of our country.

With Trip-City, local Government and investors can take better, data supported decisions to promote the tourist sector in this new era.

As future work, in addition to the already mentioned adaptations and implementations of the application in many other Destination Management Organizations throughout the country, Trip-City has great options, in the coupling of new data sources that enrich the visualizations and the possibilities of using more advanced techniques of data science, different types of models, neural networks, graphs, among many others.

One additional feature that can be added to the tool is a new model to estimate the monthly occupation rate of each property, in order to have more information and get closer to what the real income would be.

REFERENCES

Observatorio de Turismo de Bogotá: https://www.idt.gov.co/sitbog

Asociación Hotelera y Tuística de Colombia- COTELCO, ICTRC: https://cptur.org/ICTRC/

Departamento Administrativo Nacional de Estadística- DANE, Cuentas Nacionales: https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales

Cámara de Comercio de Bogotá- CCB: https://www.ccb.org.co/observatorio/Entorno-para-los-negocios/Entorno-para-los-negocios/Turismo

Ministerio de Comercio, Industria y Turismo- MinCIT: https://www.mincit.gov.co/

International Tourism Highlights (2012-2019): https://www.e-unwto.org/doi/book/10.18111/9789284421152

Geojson Bogotá's neighbourhoods: https://gist.github.com/john-guerra/ee93225ca2c671b3550d62614f4978f3#file-bogota_cadastral-json

Geojson Bogotá's divisions: https://datosabiertos.bogota.gov.co/dataset/856cb657-8ca3-4ee8-857f-37211173b1f8/resource/497b8756-0927-4aee-8da9-ca4e32ca3a8a/download/loca.geojson