**Do near venues affect students’ performance?**

Pedro Chacón

Applied Data Science Capstone

**Introduction**

Between Colombian parents there is the conception that schools nearby bars, clubs, stores and other recreational places are worst for their children, because it increasing the distractions and the probability of their children to skip classes to go to those places.

That’s why in the present project we will look if there is really a relationship the kind of establishments near schools in Bogotá, Colombia, and the results of its students in the national standardized academic test, ICFES, similar to the American SAT, which measure the competence of students in different academic areas.

**Data**

In order to archive our objective, we need 3 datasets: schools location, score average by school and venues near every schools.

**Schools location**

We will use the dataset of public and private schools from the government page *datos.gov.co* from the next page:

<https://www.datos.gov.co/Educaci-n/LISTADO-COLEGIOS-BOGOTA/qijw-htwa>

This dataset has the follow columns (The dataset is in Spanish):

* calendario
* sector
* nombreestablecimiento
* telefono
* nombre\_Rector
* correo\_Electronico
* direccion
* secretaria
* zona
* tipo\_Establecimiento
* etnias
* genero
* niveles
* jornadas
* caracter
* especialidad
* grados
* modelos\_Educativos
* capacidades\_Excepcionales
* discapacidades
* idiomas
* numero\_de\_Sedes
* estado
* prestador\_de\_Servicio
* propiedad\_Planta\_Fisica
* resguardo
* matricula\_Contratada
* internado
* estrato\_Socio\_Economico
* año

From those columns we only have to care about two, “nombreestablecimiento“ which means “Establishment name” and “direccion” which means “address”. Below is shown an example of this two columns’ information:

|  |  |
| --- | --- |
| nombreestablecimiento | direccion |
| INST PEDAG NACIONAL | AC 127 11 20 |
| COLEGIO CRISTO REY BOGOTA | AC 138 58 D 50 |

As you can see, in Bogotá addresses are given in number, therefore is relative easy to translate it to coordinates.

**ICFES score average by school**

Again from datos.gov.co we get the last ICFES scores dataset from the following page:

<https://www.datos.gov.co/Educaci-n/Saber-11-2018-2/m2nt-jw2h>

This is a rather large dataset with 83 columns and one row for each student that do the test, so let’s just look at the columns that we need to consider:

* COLE\_NOMBRE\_ESTABLECIMIENTO
* PUNT\_GLOBAL

The first row “COLE\_NOMBRE\_ESTABLECIMIENTO” means “Establishment name” and the second one means “Global score”, this is because the test is divide in several areas, so the global score is the final result of all of them.

Bellow you can see an example of these two columns.

|  |  |
| --- | --- |
| COLE\_NOMBRE\_ESTABLECIMIENTO | PUNT\_GLOBAL |
| INSTITUCION EDUCATIVA DISTRITAL PARA EL DESARROLLO HUMANO MARIA CANO | 305 |
| IE ANTONIO DE LA TORRE Y MIRANDA | 242 |

**Venues near every schools**

To get the venues near a school we would use the foursquare API, in which we send the location and the search radius, and it return us information about all the venues near that point. Bellow you can see an example of this API that may be found in the page https://developer.foursquare.com/docs/api/venues/search

{ "meta": { "code": 200, "requestId": "5ac51d7e6a607143d811cecb" }, "response": { "venues": [ { "id": "5642aef9498e51025cf4a7a5", "name": "Mr. Purple", "location": { "address": "180 Orchard St", "crossStreet": "btwn Houston & Stanton St", "lat": 40.72173744277209, "lng": -73.98800687282996, "labeledLatLngs": [ { "label": "display", "lat": 40.72173744277209, "lng": -73.98800687282996 } ], "distance": 8, "postalCode": "10002", "cc": "US", "city": "New York", "state": "NY", "country": "United States", "formattedAddress": [ "180 Orchard St (btwn Houston & Stanton St)", "New York, NY 10002", "United States" ] }, "categories": [ { "id": "4bf58dd8d48988d1d5941735", "name": "Hotel Bar", "pluralName": "Hotel Bars", "shortName": "Hotel Bar", "icon": { "prefix": "https://ss3.4sqi.net/img/categories\_v2/travel/hotel\_bar\_", "suffix": ".png" }, "primary": **true** } ], "venuePage": { "id": "150747252" } } ] } }

The principal attribute for our project is the “categories” which give us the category of the venue.

**Methodology**

First of all, we need to build a data set that combine the school, the location and it’s average ICFES score.

In order to archive this, we need to load the file “LISTADO\_COLEGIOS\_BOGOTA.csv” with the list of schools in Bogotá. From this dataset keep only the name and the addresses, which are the columns “direccion” and “nombreestablecimiento” of the dataset. After that we use the service “geopy” to translate the address to a geographic coordinate.

We have two problems we Geopy, the first one is that the service is not perfect, and therefore we won’t find every single address, there are some address that the service won’t be available to find; the second one is that there is a maximum amount of request you can send to the service, around de one thousand. This to problems have been taken into consideration and we will work with the data we are available to fetch (around 300 rows) which are completely random and should be a significant sample of the population. Any incomplete registry is dropped.

After that we load the file “Saber\_11\_\_2018-2.csv” which contains data from every student that did the test in the last round. This is a large data set (83 columns) therefore we take only the columns “COLE\_NOMBRE\_ESTABLECIMIENTO” and “PUNT\_GLOBAL” that contains the name of the school and the final score of each student. With the resulting dataset we calculate the average score by each school.

Finally, we merge the two datasets by the school’s name and drop any registry that is incomplete.

After doing this the dataset is rather reduce (less than 30), therefore we look for differences in the schools’ names and fix some common mismatches. The resulting dataset is over 90 registries.

Now we have the data have the schools, scores and geographical coordinates, we only need the list of venues nearby every location. For this we use the Foursquare API to look for the list of venues within a 1 km radius and normalize this data intro binary rows

Finally, we use the library “feature\_selection” to select the 10 venue categories that are more significant to the outcome and calculate the correlation of each one with the score.

With this result we create a linear regression model that predict the score of the school depending these 10 venue’s categories and we evaluate the accuracy of the model.

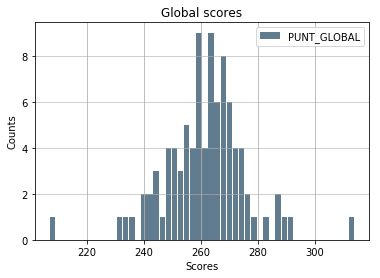
**Results**

After organizing the dataset we end up with a dataset like the following head:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | nombreestablecimiento | Latitud | Longitud | PUNT\_GLOBAL |
| 0 | COLEGIO HELADIA MEJIA (IED) | 4.694550 | -74.120600 | 264.188034 |
| 1 | COLEGIO ROBERT F. KENNEDY (IED) | 4.507666 | -74.114071 | 270.012048 |
| 2 | COLEGIO ALEJANDRO OBREGON (IED) | 4.595061 | -74.116695 | 274.000000 |
| 3 | COLEGIO FLORIDABLANCA (IED) | 4.653520 | -74.057848 | 264.391304 |
| 4 | COLEGIO VEINTE DE JULIO (IED) | 4.682009 | -74.037129 | 247.060606 |

This dataset columns correspond to the name of the school, the latitude, the longitude and the ICFES score.

The next histogram corresponds to the distribution of the scores in 50 bins



The statistical characteristics of this data are a mean of 261.447387 and a standard deviation of 14.361094

After analyzing the venues, we get that the top 10 venues that have the most impact on the outcome are the following:

1. Car Wash
2. Chinese Restaurant
3. Comfort Food Restaurant
4. Flea Market
5. Food
6. Gastropub
7. Gay Bar
8. Roof Deck
9. Shopping Plaza
10. Wings Joint

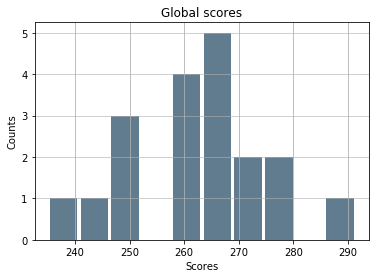
We calculate the correlation of each one of those features and we get this:

|  |  |
| --- | --- |
| Venue Category | Correlation |
| Car Wash | 0.145766066257 |
| Chinese Restaurant | -0.0606510097239 |
| Comfort Food Restaurant | -0.127732430463 |
| Flea Market | -0.08288025501 |
| Food | 0.0182476942593 |
| Gastropub | 0.0590529981259 |
| Gay Bar | -0.0117665688057 |
| Roof Deck | 0.0896350974092 |
| Shopping Plaza | 0.0896350974092 |
| Wings Joint | -0.140851116428 |

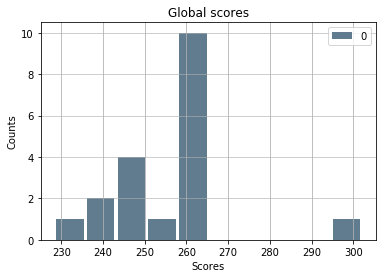
We split the data set into a training dataset with the 80% of the data and a test set with the 20% of the data, train a linear regression model using the sklearn library and generate a prediction, archiving the following metrics.

|  |  |
| --- | --- |
| Mean Absolute Error | 17.2136571016 |
| Mean Squared Error | 518.503733809 |
| Root Mean Squared Error | 22.7706770608 |

For the test set we have the following histogram with the actual results



And for the predicted results we have:



**Discussion**

Based on the results we can say that there is not any strong correlation between the venues categories and the average ICFES students’ scores, and that’s why the regression model has a high error, bigger than the standard deviation.

Therefore, we could recommend parents that they should look for other attributes of the school of their children and not for the venues nearby.

**Conclusion**

After analyzing the average scores of around one hundred schools in Bogotá, Colombia, we can conclude that, at least in the aforementioned city, there is no relationship between the category of the venues nearby a schools and the school’s student’s performance.