Jcp_Inm Dataset

Pablo Edronkin, 2019

https://orcid.org/0000-0001-8690-7030

Abstract.

This is a dataset based in Sqlite3 database format containing data on real estate values for the downtown area of the city of Jose C. Paz, Buenos Aires, Argentina, gathered in 2019.

It is presented in Sqlite3 database format and contains values in US dollars per square meter averaged (built and not built), plus data on the number of pedestrians strolling thorough the streets of the commercial district of the city.

Keywords.

Real estate, database, dataset, data, properties, value, commercial, residential, pedestrian, traffic

Acknowledgements.

This project could not have taken place without the work done by:

- The developers of Sqlite3[6.1.].

License and info for contributors.

Please read the following files included with this project:

- README.md: contains info on setting up your system and installing the files related to this module.

- CONTRIBUTING.md: if you want to contribute to this project.

- COPYING: for license information.

Table of Contents

1. Introduction.

This is a dataset presented in Sqlite3 database format containing data on real estate values for the downtown area of the city of Jose C. Paz, Buenos Aires, Argentina, gathered in 2019.

It is presented in Sqlite3 database format and contains values in US$ per square meter averaged (built and not built), plus data on the number of pedestrians strolling thorough the streets of the commercial district of the city.

SQL queries useful for processing the data contained in the data set are also provided.

2. Conventions, caveats and base conditions for development.

Take into consideration these factors.

2.1. Terms and abbreviations used.

AB: Address block.

C: Commercial zone.

CO: Commercial zone, outside the commercial district.

R: Residential zone.

RO: Residential zone, outside the commercial district.

SQL: Structured query language.

TB: Time block.

2.2. Caveats.

You should be aware of:

2.3.1. Samples per AB.

The number of samples per address block might be not be the same in all cases, but in all cases the number of samples performed is high enough to be statistically significant.

2.3.2. Samples per TB.

The number of samples per time block might not be the same in all cases. Since this data set is geared among other things, towards gather information on pedestrian traffic in relation to the commercial value of different address blocks samples were mostly obtained between 7:00 and 19:00 hours.

However, in some cases night hours might make a difference, so a few samples were taken during night hours as well, when they imply business activities.

2.3.3. Properties on sale per AB.

The number of sampled properties on sale per address block might not be the same in all cases. This is due to market reasons and depends on the decision of owners, which its outside the control possibilities of the authors of this work.

Simply put, it might be possible to find one or more properties per address block being sold over a sufficiently long period of time, but for the purposes of this dataset and the time alloted to building it, such a homogeneity of data is not achievable.

2.3.4. Urban codes.

Urban codes establish that middle to intensive commercial activities shall only take place on blocks of streets that have been declared commercial zones. There are streets that can be partially commercial and partially residential, or being commercial altogether, display a very different pedestrian density depending on factors like the presence of schools or bus stops.

Urban codes guarantee that there is no middle to heavy commercial, or industrial development in residential areas. In this part of the city, there is no allowance for industrial activities so they are not being considered in this dataset.

Thew price of properties per square meter is influenced by urban codes: indeed, residential areas have lower prices since there are less options to build and develop. Commercial areas, on the other side, allow for both commercial as well as residential construction. Of course, few people tend to prefer living in a commercial zone but they can.

This is the reason why in areas that have been declared commercial only recently and there are still private residents, some people might want to use as a reference the residential square meter price instead of a commercial value.

But since this policy does not make well with common sense, once home owners realize the difference between commercial square meter prices and those for homes, they start selling their properties at the commercial market value.

2.3.5. The effect of public activities.

In some cases, commercial prices per square meter have significant variations between fairly close or even contiguous properties. This is mostly due to certain activities like.

2.3.5.1. Banking. The presence of banks tend to markedly increase the number of pedestrian in a given street block and those found in their way from the train station or bus stops. Normal banking hours go from 10:00 to 16:00.

However, the presence of people using an ATM or waiting in queues at the banks – particularly pensioneers and people receiving unemployment subsidies – is significant almost around the clock.

2.3.5.2. Schools and a local university. They have a similar effect as banks, but with reduced hours. Many schools run from 7:00 to 23:00.

Such a schedule includes lessons for primary and secondary school children, up to 17:00, and from then on courses for adults, night lessons for working people, vocational courses, etc.

2.3.5.3. Trains and buses. By far, the train station in the city produces the heaviest pedestrian density in nearby areas, but bus stops also produce significant pedestrian numbers. Moreover: legislation established that within cities there must be a bus stop every 200 m along the route of any bus line.

In commercial areas this guarantees a significant and homogeneous pedestrian density, especially if banks or schools are found nearby.

2.3.5.4. Large commercial establishments: a shopping center, a couple of discos, etc. are also responsible for a relatively large number of regular, visiting pedestrians.

2.3.5.5. Visitors from nearby towns. Since some towns in the region do not offer certain services, people from those usually visit the city to shop, run errands, etc.

2.3.6. Area of sampling blocks.

In Argentina, square blocks tend to measure 100 x 100 meters and streets are laid in a square grid pattern for the most part. In this case, since the urban area of the city includes railway tracks, there are some diagonal streets but since the data set includes data from all the streets in the commercial zone of the town, the numbers represent well the presence and distribution thorough the day and days of the week of pedestrians (customers) in the whole area.

This means that samples were obtained in segments of 100 m on each street. The total number of pedestrians per minute per block indicate the number of pedestrian per each segment of a hundred meters on each street, regardless of the urban characteristics of each segment. That is, if in a commercial street one block instead of being a built up area is a park, pedestrians were accounted for on that segment too.

In some cases, streets have been partially zonified as commercial areas, while in others the same streets remains being residential.

Blocks located within the commercial area of the city were sampled regardless of whether they were defined as commercial or residential blocks based on the fact that such characteristics of a street do influence the value of properties and the number of pedestrians passing thorough.

2.3.7. Currency used for appraisal.

In Argentina, the currency used for real estate appraisals has been traditionally the US dollar. This is mainly a collateral effect of the chronic high inflation of the Argentine peso. Dollars provide a better framework for buyers, investors and real estate agents.

In more recent times, the Bitcoin has been acquiring some relevance in real estate operations in the country. However, the volatility and speculative values surrounding this crypto currency has put a limit to this growth, at least for now.

Hence, for the purposes if this work we will stick with the US dollar.

2.3.8. RO and CO properties.

As expressed in **[2.1.]**, there are some properties classified as RO and CO. These lie outside the commercial district studied but were taken as external control references for the purposes of this study. Their values and characteristics are not part of the statistic itself.

2.3.9. The effect of urban design.

The way in which accesses for pedestrians, bus stops, etc. are distributed also seems to have an effect of pedestrian transit, affecting the number of people that:

2.3.9.1. Visit each street. Undeniably, the town's architecture, good or bad, has an effect on how many people walk on different areas. Some places are better suited for pedestrians than others on the basis of their characteristics such as vehicle traffic, street lights for crossing, viaducts, etc.

This study does not involve itself into the details of urban design except for the fact that the count of pedestrians on each sampling block might vary, for this, among other reasons.

2.3.9.2. Walk on each side of some streets. Aside from what is stated in **[2.3.9.2.]**, the characteristics of urban design, development, etc. influence even how people walk on one side or the other of some streets.

This study takes into account this fact but is not designed to make such differentiation. Statistics on each block include the sum of pedestrians from both sides of each involved street.

3. Development.

Regarding the numbers representing customers in the area, data was gathered in each street, from each block, by counting the pedestrians – people on foot, not vehicles – passing on a period of approximately one minute. People inside shops or buildings were not counted.

Data on commercial and residential values in the area was obtained from the properties being offered publicly for sale at the time.

In order to better pinpoint the date and time in which data was added to the database, each record was provided with a time stamp field.

4. Structure and data.

This section should be considered as the data dictionary of the database.

4.1 Data tables in the dataset.

4.1.1. *trf_pers*

This table contains data regarding the number of persons measured on each street block.

4.1.1.1. *Id*: Primary key. Unique integer that identifies each record in the table.

4.1.1.2. *Context*: describes the area or district within the town.

4.1.1.3. *Status*: *'enabled'* for records to be considered, *'disabled'* for records on hold.

4.1.1.4. *Abr*: A two to four letter code that identifies each street. Correspondence between *Abr* and the actual name of the street is contained in table *'trf_streets'* between fields *Item* and *Abr*.

4.1.1.5. *Alt*: Street coordinate - i.e. street number. In Argentina normally each street block carries numbers by a century or hundred meters counted from what is considered to be the starting point of the street, for a given town. For example, being at the corner of Lavalle 1700 means that this is the 17ht block of that street.

Houses are identified by meters. So, Lavalle 1723 is the address that stands at 1723 meters from the beginning of the street. For the purposes of this study, addresses represent the starting point of each street block from which a count of pedestrians is made.

4.1.1.6. *Value*: contains the number of pedestrians counted in the *[Alt, Alt + 100)* interval for that particular street. That is, the values from the whole block, for a given street as expressed in *Abr*.

4.1.1.7. *Hr:* hour interval at which the sample is taken.

4.1.1.8. *Dayn:* number corresponding to the day of the month.

4.1.1.9. *Dayw:* day of the week, as a three letter abbreviation.

4.1.1.10. *Monthn:* number corresponding to the month.

4.1.1.11. *Year:* self explanatory.

4.1.1.12. *Tstamp:* time stamp, provided by the RDBMS itself.

4.1.2. *trf_ppcm*

Thus table contains data regarding known prices asked for actual properties in the area of interest.

4.1.2.1. *Id*: see [**4.1.1.1.**].

4.1.2.2. *Context*: see [**4.1.1.2.**].

4.1.2.3. *Status*: see [**4.1.1.3.**].

4.1.2.4. *Abr*: see [**4.1.1.4.**].

4.1.2.5. *Alt*: see [**4.1.1.5.**].

4.1.2.6. *Value*: see [**4.1.1.6.**].

4.1.2.7. *Zone*: a one or more letter long string that identifies commercial and residential areas relevant to the study. “C” identifies an active commercial zone, “R” identifies an active residential zone.

4.1.2.8. *Monthn*: see [**4.1.1.10.**].

4.1.2.9. *Tstamp*: see [**4.1.1.12.**].

4.1.3. *trf_streets*

4.1.3.1. *Id*: see [**4.1.1.10.**].

4.1.3.2. *Abr*: see [**4.1.1.2.**].

4.1.3.3. *Item*: official, current street name.

4.1.3.4. *Alias*: unofficial or not current names or aliases by which the street might also be identified.

4.1.3.5. *Avg_pers*: average number of people per block, per sample instance.

4.1.3.6. *Samples_total*: total number of samples per street.

4.1.3.7. *Avg_ppmc*: average price per square meter, commercial.

4.1.3.8. *Avg_ppmr*: average price per square meter, residential.

4.1.3.9. *Value_ppers*: value per person.

4.1.3.10. *Est_final_ppmc:*estimated final price per square meter.

This table contains data considered to be part of the conclusion of the study.

5. Use.

SQLite3 databases are fully self-contained. Users need to have Sqlite3 installed on their system, and preferably a visual database editor. Using SQL queries it is possible to obtain any section of the gathered data, and it can be converted to any data format used by libraries written in R, Python, C++ etc.

If you plan to operate on this dataset from an environment written in a specific language, you will likely have to install some libraries. However, don’t worry since Sqlite3 is widely used and there are many such libraries available.

You might ask yourself why is this dataset presented in an Sqlite3 format. The main reason is simple: it is far easier to operate on a SQL – compatible dataset than in any other case. SQL systems have limitations, but for the most part or at least, on average, relational databases are better than any other format.

It is recommendable for users to review the structure of the data tables as well as the data itself in order to understand the whole package and what can be done with it. In general, tables contain data either as strings, integers or real numbers.

5.1. Installation

See the file *README.md* included in this package.

5.2. Software

This dataset comes with and also requires some software in order to be useful. Regarding the libraries and software required to use this dataset, see *README.md*.

5.2.1. *jcp_inm.scm*

This is a program written in GNU Guile, which is the GNU flavor of the Scheme language. In order to use it you need to have the required version of GNU Guile installed on your system[6.3.].

Also you will need a program called *sqlp*[6.2.]. This is a program I wrote in C++, and serves as an interface between high level languages and Sqlite3 and HDF5 databases. GNU Guile and sqlp are open source, free software.

You will find more information on how to get and install them on *README.md*. So I will concentrate now on the characteristics of *jcp_inm.scm*, assuming that you have all the required software or libraries installed on your system.

In order to start the program, cd to the /src/scheme folder and then write

guile jcp_inm.scm

The first time you run this program, Guile will compile it first. After a few seconds, the main menu of the program will appear on your terminal.

5.2.1.1. Main menu

On the main menu of *jcp_inm.scm* you will find the following options:

0 - Exit.

1 - Input data.

2 - Process.

3 – Report.

Entering the respective number and pressing <ENT> (enter), will take you to different sections of the program, mostly presented as further menus.

Take into consideration that this is a CLI program, with no graphical user interface, so you will find no fancy graphs, but the menus are organized in a similar way.

5.2.1.1.1. *Exit* takes you out of the program with no further ado.

5.2.1.1.2. *Input data* does what the words say. You can add data to the *jcp_inm.db* database. By selecting this option you will be take to another menu with the following options:

0 - Back to main.

1 - Update prep.sql.

2 - Enter block data.

5.2.1.1.2.1. *Back to main* takes you to the main menu.

5.2.1.1.2.2. *Update prep.sql* lets you generate a new *prep.sql* file. This is a .sql file that contains UPDATE queries relative to data that will be common to all components to one data block (see **[5.2.1.1.2.3.]**) and that would be needlessly repetitive to enter each time that you input block components.

For example, if you are going to enter data concerning one batch of samples, each taken within the same day and hour, instead of having to write those data fields in the case of each record, you enter them just one for the whole batch.

Once you are finished with each batch, *prep.sql* will be executed and the UPDATE sql statements will be passed to the database as a composite query. If you have a doubt about what a composite query is, please see *sqlp.pdf* contained in the sqlp distribution file.

5.2.1.1.2.3. Enter block data lets you pass records or registers related to new samples to be added to the *jsp_inm.db* database. Once you select this option, you will be presented with the following menu:

0 - No more records.

1 - Enter a record.

5.2.1.1.2.3.1. *No more records* will finish the data entry operation by executing *prep.sql* and then taking you back to the upper level menu described in **[5.2.1.1.2.]**.

5.2.1.1.2.3.2. *Enter a record* will take you to a series by simple questions and menu options regarding the data to be passed for each record, designed to minimize data entry errors. Once you are finished with one record, you will be taken back to menu **[5.2.1.1.2.3.]** in order to add new more records or execute *prep.sql* once the whole batch of records has been entered.

5.2.1.1.3. *Process* executes a batch sql program called *calc.sql*, which is contained in the src/sql folder. This program performs a number of calculations on the dataset. You can see the code of *calc.sql* by opening it with any text editor.

These calculations involve finding the number of pedestrians per street and street block, property average prices per section, etc.

5.2.1.1.4. *Report* presents some results of the calculations described in **[5.2.1.1.3.]**. This option is yet a bit underdeveloped and needs more work, since it only presents information in the sqlp query data output format, which is contained in a file called *sqlp_results.txt* and intended to be used by other programs.

For more information on *sqlp_results.txt*, see *sqlp.pdf*.

5.2.1.2. Limitations.

As you might have noticed, this program still lacks in some aspects:

5.2.1.2.1. There are no delete or edit facilities. Currently, in those cases, which are rare, I use a database editor. However, t might be useful to develop such a thig, but first it would require working on **[5.2.1.2.3.]**.

5.2.1.2.2. It has no GUI. Granted, a terminal is not not the nicest interface, but the good part of this is that it was never meant to have a nice one. Writing GUI programs requires a lot of spaghetti code, and that conspires against the main goals of *jcp_inm.scm*.

The program is a basic tool to enter data and serve as proof of concept for sqlp. No more no less, and a writing it to work with a GUI, while it might be interesting in the future, would unnecessarily obfuscate its code right now.

5.2.1.2.3. Reporting is very basic. As expressed in **[5.2.1.1.4.]**, things could and should be better in this case because while the kind of output interface this program has can be used by other programs, it is of little use for the human user as it is.

This is one of the main areas of further development for this program.

6. Sources.

6.1. Sqlite.org. (2000). SQLite Home Page. [online] Available at: https://www.sqlite.org/index.html [Accessed 26 Aug. 2019].

6.2. Edronkin, P. (2019). sqlp - Simple terminal query and .sql file processing for Sqlite3 and HDF5 databases. [online] Available at: https://peschoenberg.github.io/sqlp/ [Accessed 8 Nov. 2019].

6.3. GNU contributors (2019). GNU's programming and extension language - GNU Guile. [online] Gnu.org. Available at: https://www.gnu.org/software/guile/ [Accessed 2 Sep. 2019].

Alphabetical Index

A

B

C

D

E

F

H

I

J

L

M

N

O

P

Q

R

S

T

U

V

Y

Z