Calidad de Datos
Universidad de Buenos Aires
Primer Cuatrimestre de 2020
Trabajo Práctico Final

Informe Tecnico

Diego Santos, Alvaro Machicado & Maximiliano Cabezón Alvarez

1 Dataset

Uitilizamos el set de datos `Obras Registradas Junio Septiembre de 2019` disponible en la página del Gobierno de la Ciudad de Buenos Aires

https://data.buenosaires.gob.ar/dataset/obras-registradas

En el cual se lleva un registro catastral de las obras realizadas en la Ciudad de Buenos Aires durante el período.

El dataset no tiene un diccionario de datos por lo cual la información que contiene fue inferida e interpretada de dataset similares

Columna	Tipo de Dato	Significado
long	Numérico	Coordenada de longitud
lat	Numérico	Coordenada de latitud
nro_exp	String	Número de expediente
nomenclacion_par	String	Union de Seccion - Parcela - Manzana
fecha_registro_p	Fecha	Fecha de registro
direccion	String	Calle abreviada y número de parcelas afectadas
direccion_normalizada	String	Calle y número de la obra
tipo_obra	String	Tipo de obra
metros_cuadrados_obra	Numérico	Cantidad de metros cuadrados
calle_nombre	String	Calle de la obra
calle_altura	Numérico	Número de la obra
barrio	String	Barrio
comuna	Numérico	Número de comuna
comuna	String	Comuna
codigo_postal	Numérico	Código postal CABA
codigo_postal_argentino	String	Código postal argentino

1.1 Análisis de Diagnóstico

Encontramos varios puntos que permiten decir que la calidad del dataset no es buena.

Hay problemas para la interpretación debido a:

1. No existe un diccionario de datos.
2. La nomenclatura usada en las columnas es poco declarativa
3. El orden de las columnas no ayuda a comprender rapidamente de que se trata.

Y problemas con la calidad de los atributos como:

1. Las columnas dirección, dirección_normalizada y calle_nombre se refieren a la calle donde se realiza la obra y se nota de forma distinta en todas.
2. Las columnas dirección, dirección_normalizada y calle_altura se refieren a la altura donde se realiza.
3. Hay dos columnas que se llaman comuna, tienen la misma información solo que una tiene el agregado de la palabra comuna
4. Hay dos columnas que se refieren al código postal.
5. Hay datos missing en la columna barrio
6. Hay datos missing en las dos columnas referidas al código postal
7. Hay missing en la columa metros_cuadrados
8. Hay missing en una de las columnas de comuna, pero no en la otra.
9. Hay errores de escritura en la columna metros_cuadrados
10. La columna nomenclacion_par es el resultado de la concatenación de 4 atributos distintos (zona - sección - parcela - manzana)

La cantidad de columnas con datos missing y su porcentaje se descubrirán con el análisis univariado

2 Limpieza del Dataset

Para la limpieza del set de datos tomamos las siguientes medidas:

1. Creamos la columna `Calle` a partir de la columna `calle_nombre`
2. Creamos la columna `Altura` a partir de la columna `calle_altura`
3. Creamos la columna `AlturaObra` tomando la altura de las parcelas de la columna `direccion` descartando la calle
4. Separamos la columna `nomenclacion_p` en los valores que representa creando las columnas `Zona`, `Sector`, `Manzana` y `Parcela`
5. Eliminamos la columna `codigo_postal` dejando `codigo_postal_argentino`
6. Eliminamos la columna `comuna` que tenía la palabra Comuna antes del número
7. Eliminamos la columna `calle_nombre`
8. Eliminamos la columna `calle_altura`

9. Eliminamos la columna `direccion`
10. Eliminamos la columna `direccion_normalizada`
11. Renombramos la columna `comuna` por `Comuna`
12. Renombramos la columna `lat` por `Latitud`
13. Renombramos la columna `long` por `Longitud`
14. Renombramos la columna `nro_exp` por `Expediente`
15. Renombramos la columna `fecha_registro_p` por `FechaRegistro`
16. Renombramos la columna `tipo_obra` por `TipoObra`
17. Renombramos la columna `barrio` por `Barrio`
18. Renombramos la columna `metros_cuadrados_obra` por `MetrosCuadrados`
19. Renombramos la columna `codigo_postal_argentino` por `CodigoPostal`

Luego reordenamos las columnas con el fin de que sea más declarativo visualmente de la siguiente forma:

Expediente	FechaRegistro	Calle	Numero	AlturaObra	Zona	Seccion	Manzana

Parcela	Barrio	Comuna	CodigoPostal	Tipo	MetrosCuadrados	Longitud	Latitud

De esta forma eliminamos la rebundacia de datos que generaban ambiguedades dejando solo una columna correspondiente a la calle, altura, comuna y código postal de la obra. Separando la nomenclatura en lo que representa permite una mayor amplitura de los datos facilitando su explotación. Además asigamos nombres declarativos a las columnas.

Utilizando la herramienta `open refine` corregimos manualmente los errores de carga en `MetrosCuadrados`.

El dataset obtenido lo presentamos en el archivo `obras_con_missing.csv` . El tratamiento de los datos faltantes queda detallado más adelante en el informe.

2.1 Diccionario de Datos

Columna	Tipo de Dato	Significado
Expediente	Texto (string)	Número de expediente de la obra
FechaRegistro	Fecha ISO-8601 (date)	Fecha de inicio de expediente
Calle	Texto (string)	Calle a la que pertenece la obra
Numero	Número entero (number)	Número de calle a la que pertenece la obra
AlturaObra	Texto (string)	Numeraciones de la parcela de la obra
Zona	Número entero (number)	Número de zona a la que pertenece la obra
Seccion	Número entero (number)	Número de sección a la que pertenece la obra
Manzana	Número entero (number)	Número de manzana en la sección de la obra
Parcela	Texto (string)	Número de parcela en la manzana de la obra
Barrio	Texto (string)	Barrio al que pertenece la obra
Comuna	Número entero (number	Número de Comuna a la que pertenece la obra
CodigoPostal	Texto (string	Código Postal Argentino de la obra
Tipo	Texto (string)	Tipo de obra
MetrosCuadrados	Número decimal (number)	Metros cuadrados de la obra
Latitud	Número decimal (number)	Latitud de la obra
Longitud	Número decimal (number)	Longitud de la obra

2.2 Mecanismos de Remediación Automático

Para la automatización de las transformaciones realizadas al dataset proveemos un script desarrollado en python adjunto en:

`scripts/limpiar.py`

Antes de su uso recomendamos verificar la sección de `Requerimientos`

3 Análisis Univariado

Para la confección de la tabla de frecuencias reemplazamos los valores nulos en las columnas categoricas por `Desconocido` y en las númericas por `-9999` , a fin de poder identificar los missing de forma más rapida.

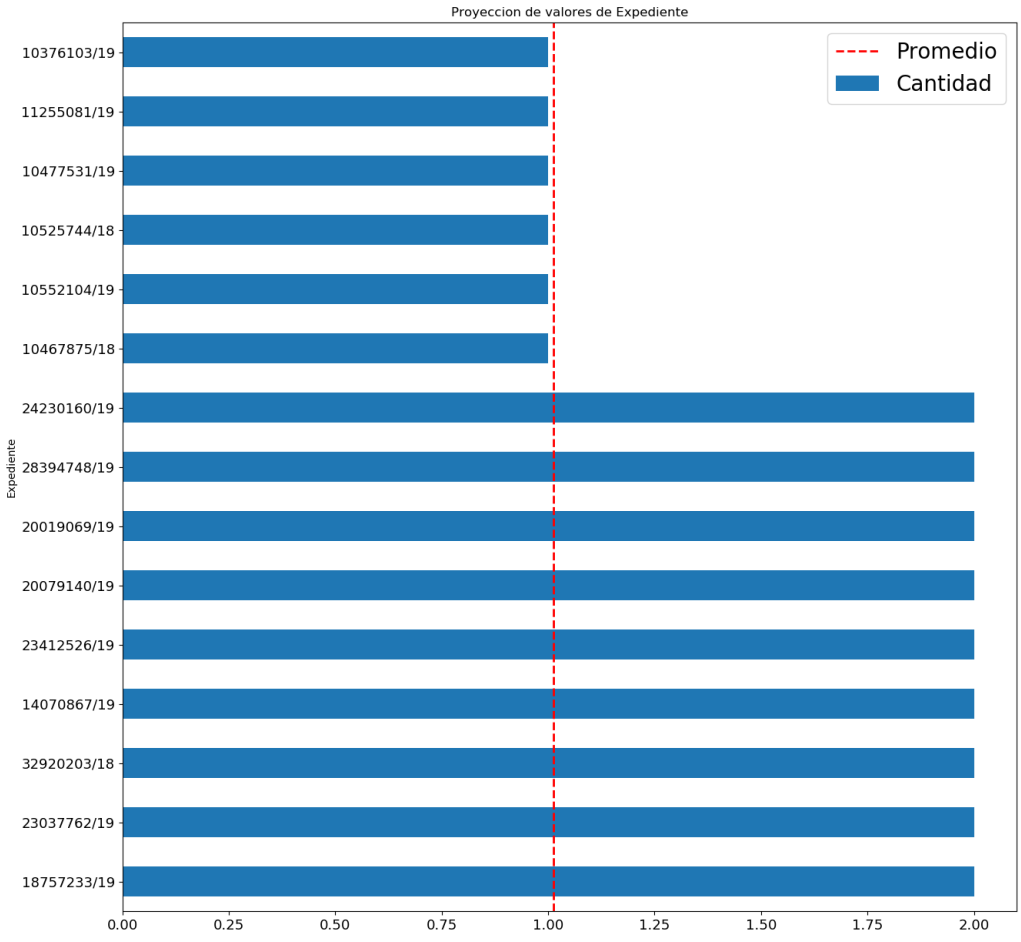
El dataset consta de 692 registros, en algunos atributos no fue posible representar la tabla de frecuencia completa, por lo cual se adjunta en un csv dentro de la carpeta `analisisUnivariado` identificado por nombre de columna.

3.1 Expediente

La tabla de frecuencias para los primeros 15 atributos es

	Expediente	Cantidad	Porcentaje
0	18757233/19	2	0.289
1	23037762/19	2	0.289
2	32920203/18	2	0.289
3	14070867/19	2	0.289
4	23412526/19	2	0.289
5	20079140/19	2	0.289
6	20019069/19	2	0.289
7	28394748/19	2	0.289
8	24230160/19	2	0.289
9	10467875/18	1	0.145
10	10552104/19	1	0.145
11	10525744/18	1	0.145
12	10477531/19	1	0.145
13	11255081/19	1	0.145
14	10376103/19	1	0.145

Esta columna no posee missing. Graficamente su distribución puede verse como:



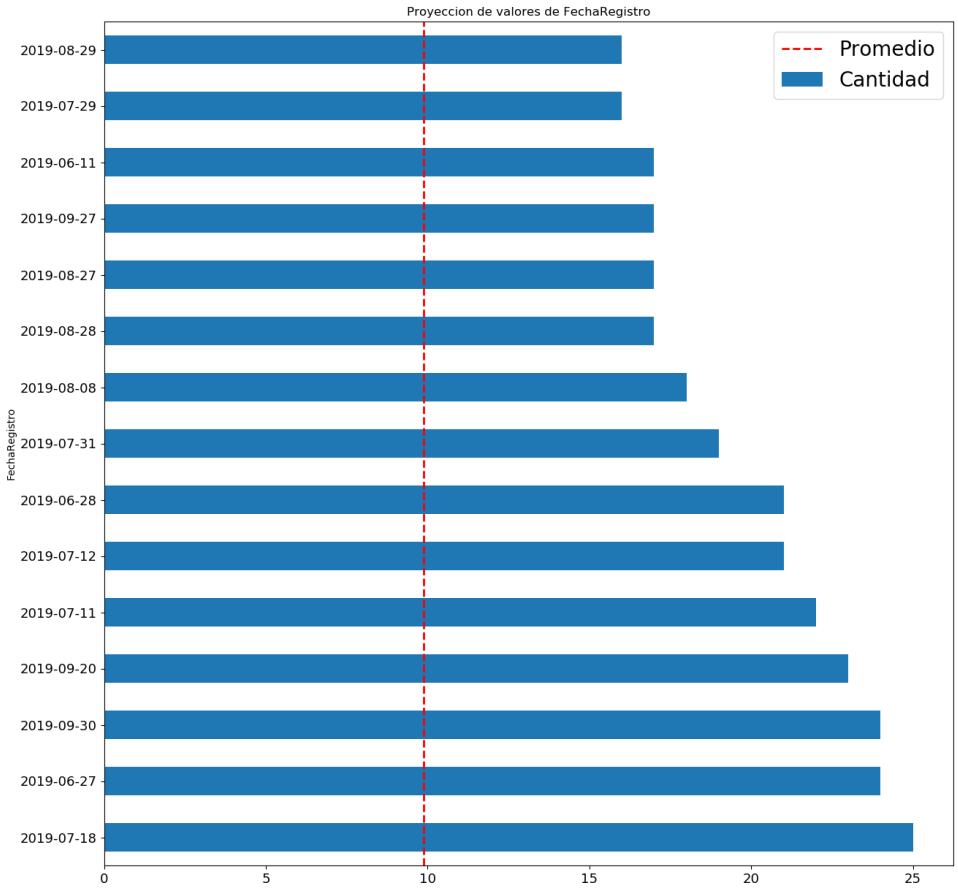
La moda es :14070867/19
La mediana es :25747086/19

3.2 FechaRegistro

La tabla de frecuencias para los primeros 15 atributos es

	FechaRegistro	Cantidad	Porcentaje
0	2019-07-18	25	3.613
1	2019-06-27	24	3.468
2	2019-09-30	24	3.468
3	2019-09-20	23	3.324
4	2019-07-11	22	3.179
5	2019-07-12	21	3.035
6	2019-06-28	21	3.035
7	2019-07-31	19	2.746
8	2019-08-08	18	2.601
9	2019-08-28	17	2.457
10	2019-08-27	17	2.457
11	2019-09-27	17	2.457
12	2019-06-11	17	2.457
13	2019-07-29	16	2.312
14	2019-08-29	16	2.312

Esta columna no posee missing. Graficamente su distribución puede verse como:



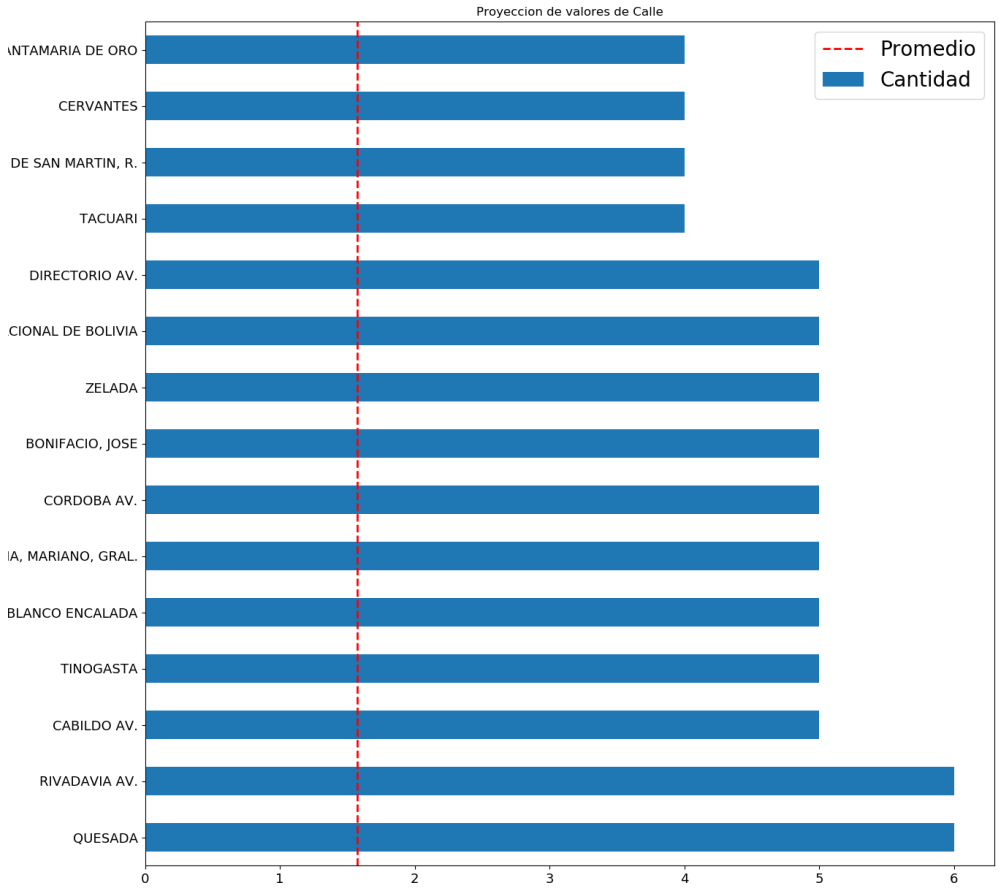
La moda es :2019-07-18
La mediana es :2019-07-29

3.3 Calle

La tabla de frecuencias para los primeros 15 atributos es

	Calle	Cantidad	Porcentaje
0	QUESADA	6	0.867
1	RIVADAVIA AV.	6	0.867
2	CABILDO AV.	5	0.723
3	TINOGASTA	5	0.723
4	BLANCO ENCALADA	5	0.723
5	ACHA, MARIANO, GRAL.	5	0.723
6	CORDOBA AV.	5	0.723
7	BONIFACIO, JOSE	5	0.723
8	ZELADA	5	0.723
9	ESTADO PLURINACIONAL DE BOLIVIA	5	0.723
10	DIRECTORIO AV.	5	0.723
11	TACUARI	4	0.578
12	ESCALADA DE SAN MARTIN, R.	4	0.578
13	CERVANTES	4	0.578
14	FRAY JUSTO SANTAMARIA DE ORO	4	0.578

Esta columna no posee missing. Graficamente su distribución puede verse como:



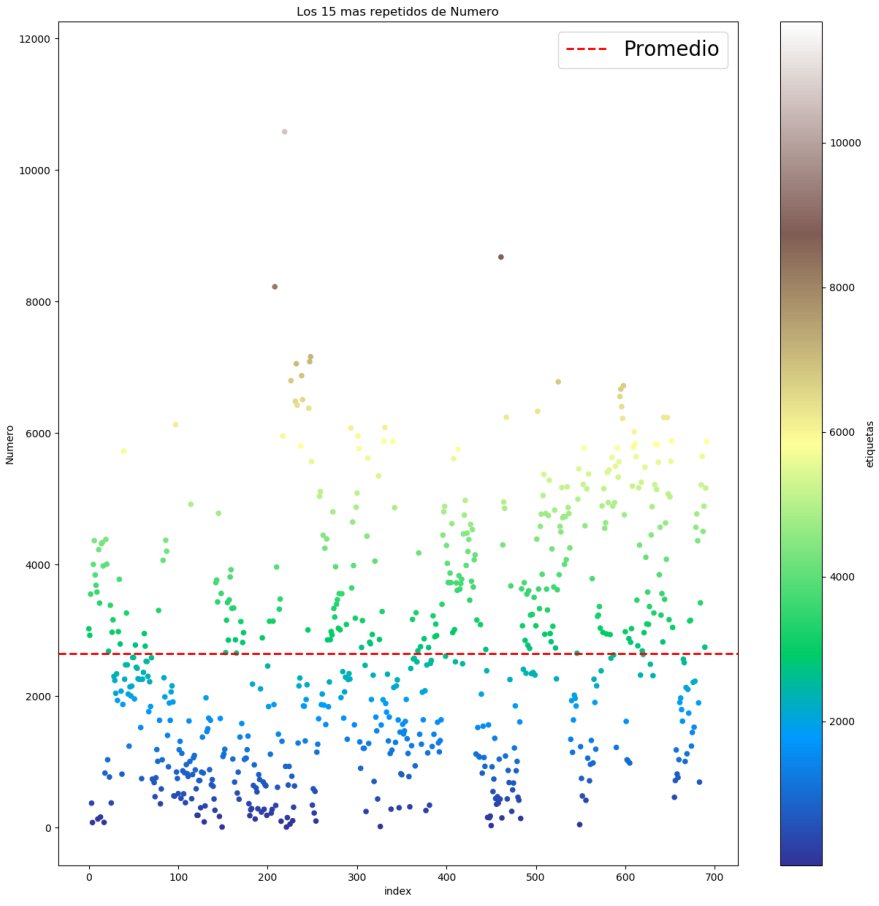
La moda es :QUESADA
La mediana es :HUGO, VICTOR

3.4 Numero

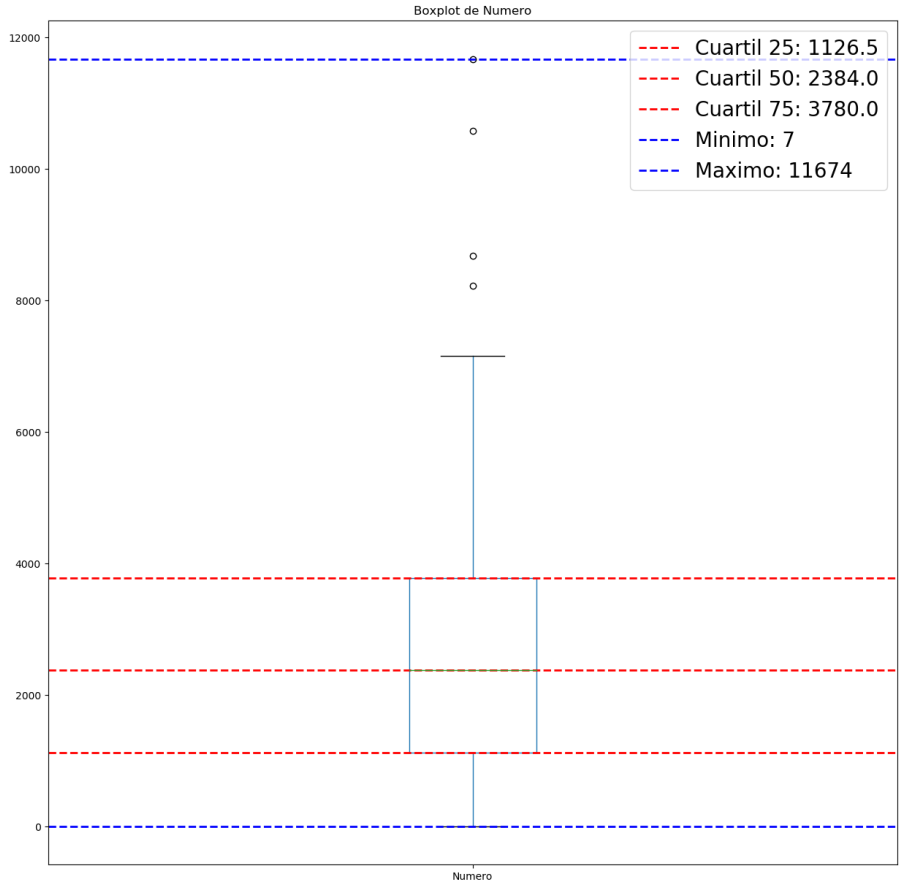
La tabla de frecuencias para los primeros 15 atributos es

	Numero	Cantidad	Porcentaje
0	715	3	0.434
1	2256	3	0.434
2	2155	3	0.434
3	130	2	0.289
4	684	2	0.289
5	3658	2	0.289
6	3726	2	0.289
7	1267	2	0.289
8	1947	2	0.289
9	1902	2	0.289
10	1317	2	0.289
11	3845	2	0.289
12	2586	2	0.289
13	1205	2	0.289
14	4942	2	0.289

Esta columna no posee missing. Graficamente su distribución puede verse como:



La moda es :715
La mediana es :2383

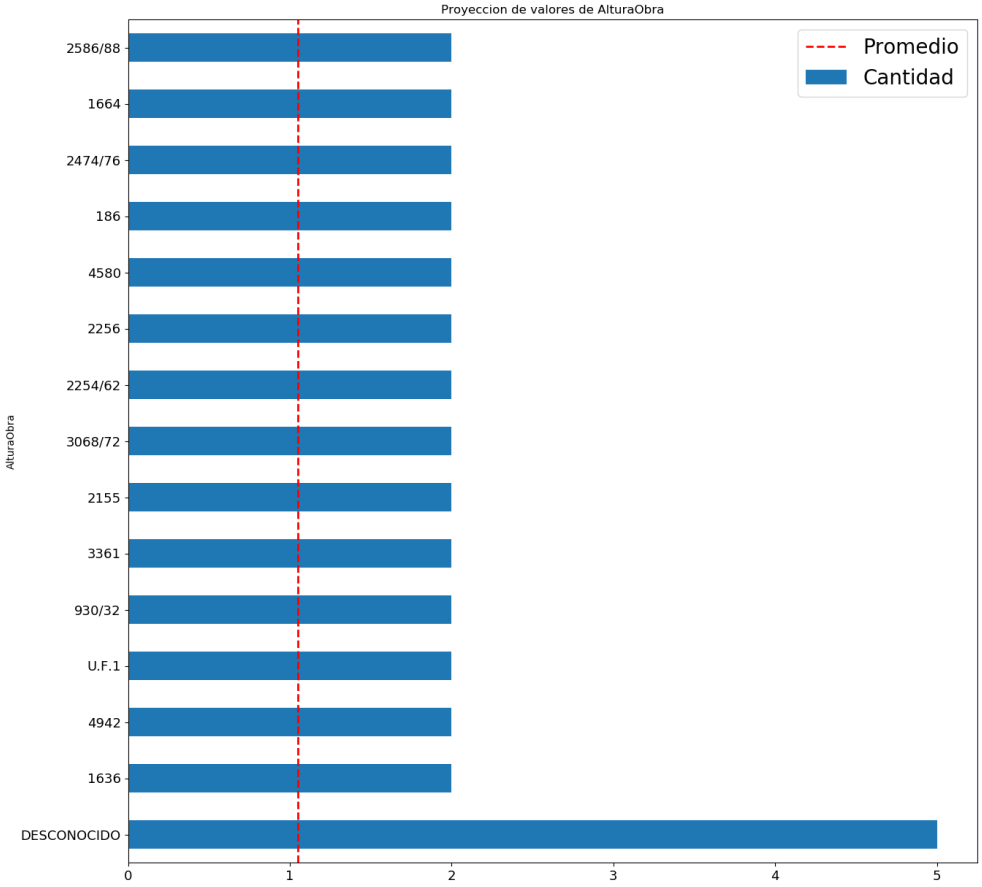


3.5 AlturaObra

La tabla de frecuencias para los primeros 15 atributos es

	AlturaObra	Cantidad	Porcentaje
0	DESCONOCIDO	5	0.723
1	1636	2	0.289
2	4942	2	0.289
3	U.F.1	2	0.289
4	930/32	2	0.289
5	3361	2	0.289
6	2155	2	0.289
7	3068/72	2	0.289
8	2254/62	2	0.289
9	2256	2	0.289
10	4580	2	0.289
11	186	2	0.289
12	2474/76	2	0.289
13	1664	2	0.289
14	2586/88	2	0.289

Esta columna posee 5 missing que representan un 0.7 % del total . Graficamente su distribución puede verse como:



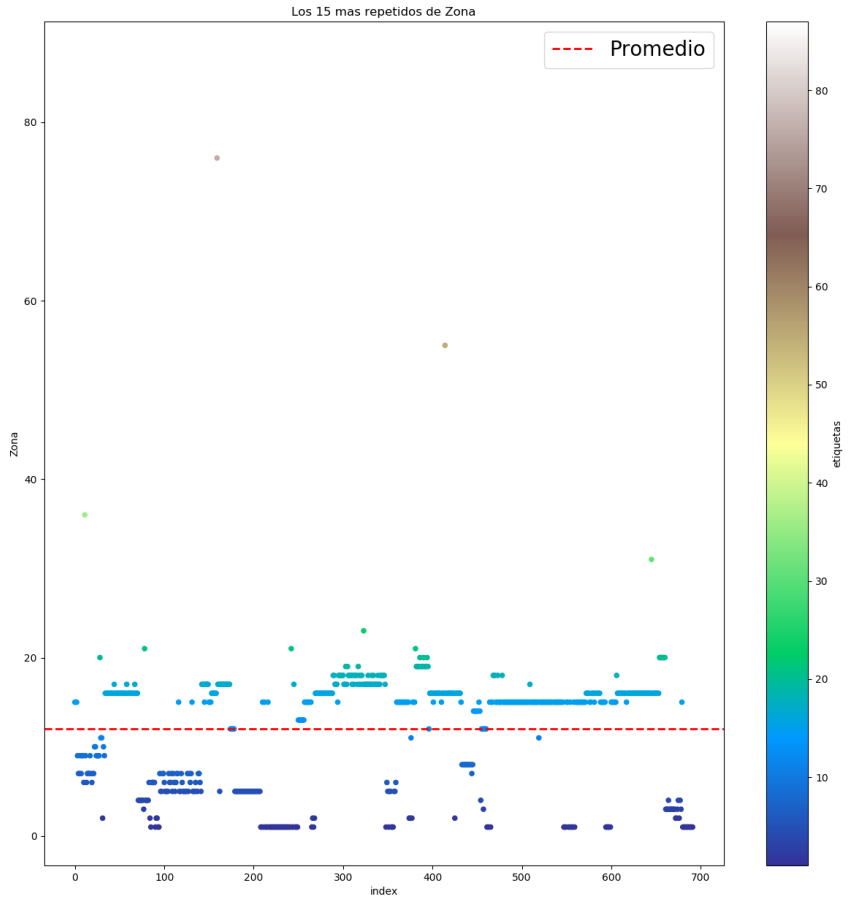
La moda es :1205/09
La mediana es :3220/50

3.6 Zona

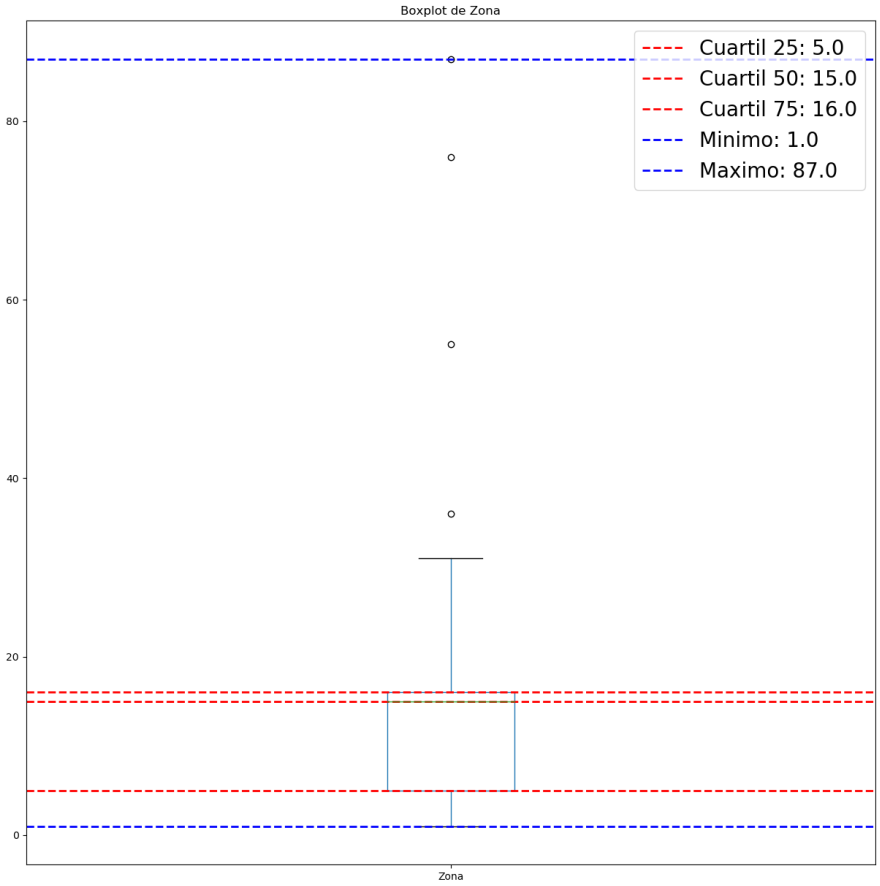
La tabla de frecuencias para los primeros 15 atributos es

	Zona	Cantidad	Porcentaje
0	15	149	21.532
1	16	148	21.387
2	1	79	11.416
3	5	57	8.237
4	17	51	7.370
5	18	34	4.913
6	7	24	3.468
7	6	19	2.746
8	4	14	2.023
9	19	14	2.023
10	3	14	2.023
11	2	13	1.879
12	8	12	1.734
13	20	11	1.590
14	9	11	1.590

Esta columna no posee missing. Graficamente su distribución puede verse como:



La moda es :15.0
La mediana es :15.0

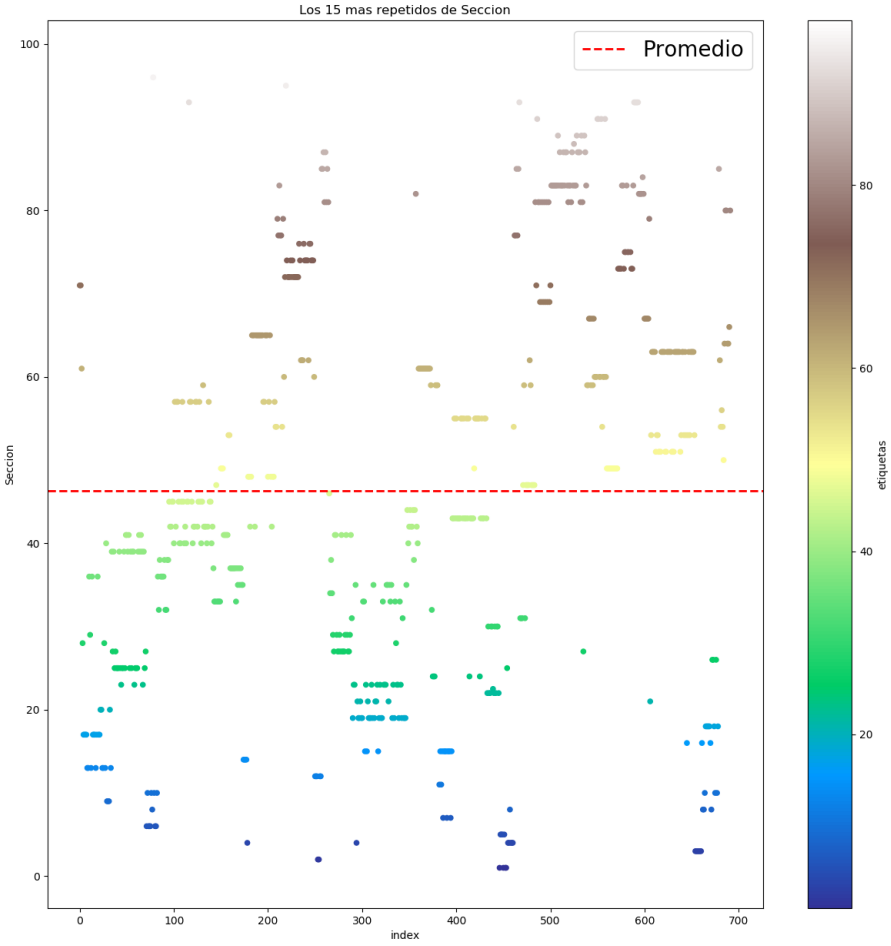


3.7 Seccion

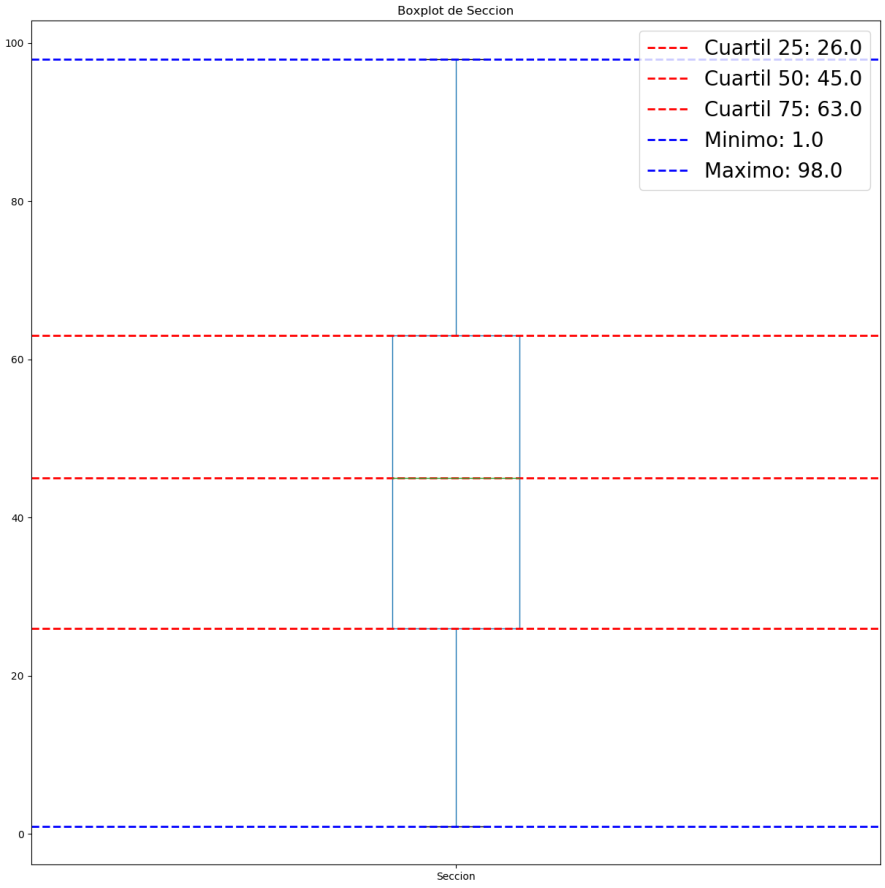
La tabla de frecuencias para los primeros 15 atributos es

	Seccion	Cantidad	Porcentaje
0	63.0	27	3.902
1	83.0	24	3.468
2	19.0	20	2.890
3	43.0	19	2.746
4	42.0	19	2.746
5	23.0	16	2.312
6	49.0	16	2.312
7	55.0	15	2.168
8	25.0	15	2.168
9	65.0	15	2.168
10	41.0	15	2.168
11	81.0	15	2.168
12	45.0	14	2.023
13	57.0	14	2.023
14	61.0	14	2.023

Esta columna no posee missing. Graficamente su distribución puede verse como:



La moda es :63.0
La mediana es :44.0

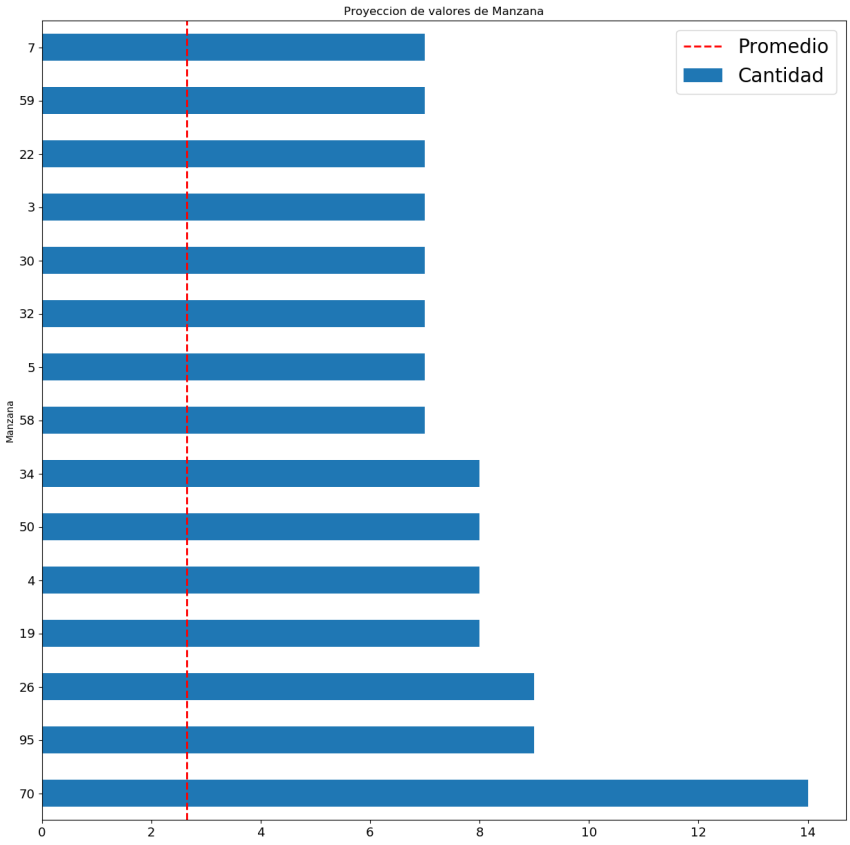


3.8 Manzana

La tabla de frecuencias para los primeros 15 atributos es

	Manzana	Cantidad	Porcentaje
0	70	14	2.023
1	95	9	1.301
2	26	9	1.301
3	19	8	1.156
4	4	8	1.156
5	50	8	1.156
6	34	8	1.156
7	58	7	1.012
8	5	7	1.012
9	32	7	1.012
10	30	7	1.012
11	3	7	1.012
12	22	7	1.012
13	59	7	1.012
14	7	7	1.012

Esta columna no posee missing. Graficamente su distribución puede verse como:



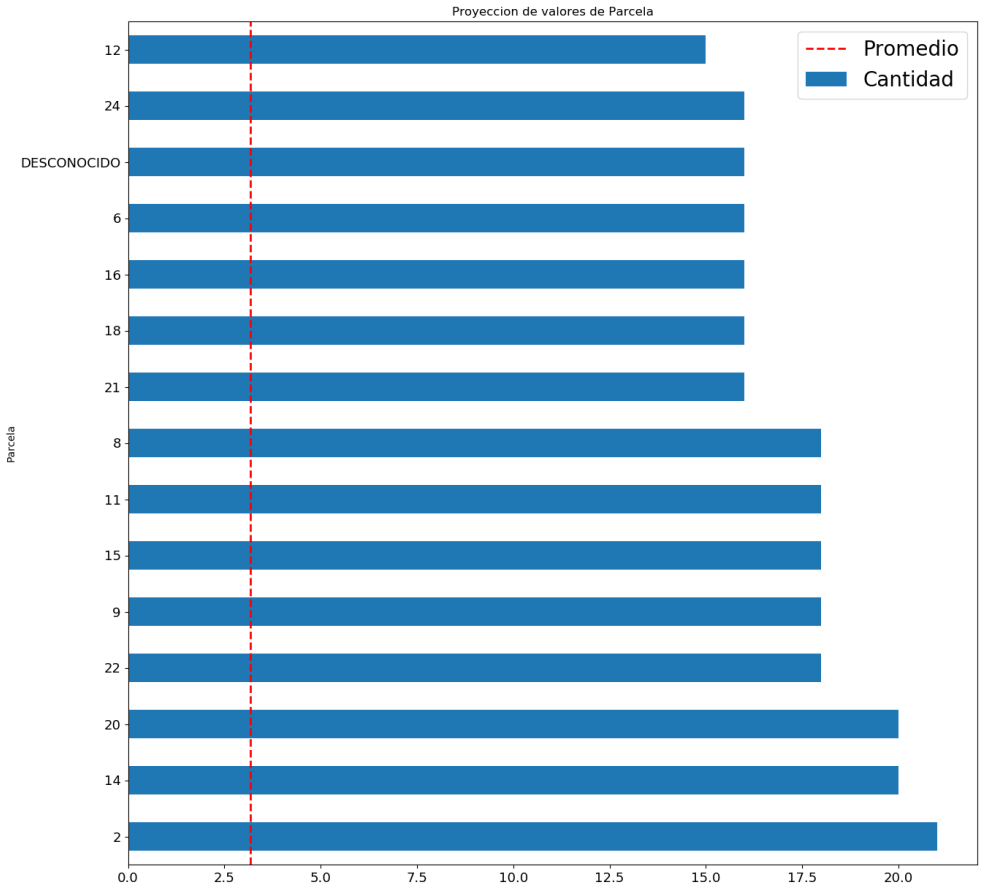
La moda es :70
La mediana es :33

3.9 Parcela

La tabla de frecuencias para los primeros 15 atributos es

	Parcela	Cantidad	Porcentaje
0	2	21	3.035
1	14	20	2.890
2	20	20	2.890
3	22	18	2.601
4	9	18	2.601
5	15	18	2.601
6	11	18	2.601
7	8	18	2.601
8	21	16	2.312
9	18	16	2.312
10	16	16	2.312
11	6	16	2.312
12	DESCONOCIDO	16	2.312
13	24	16	2.312
14	12	15	2.168

Esta columna posee 16 missing que representan un 2.31% . Graficamente su distribución puede verse como:



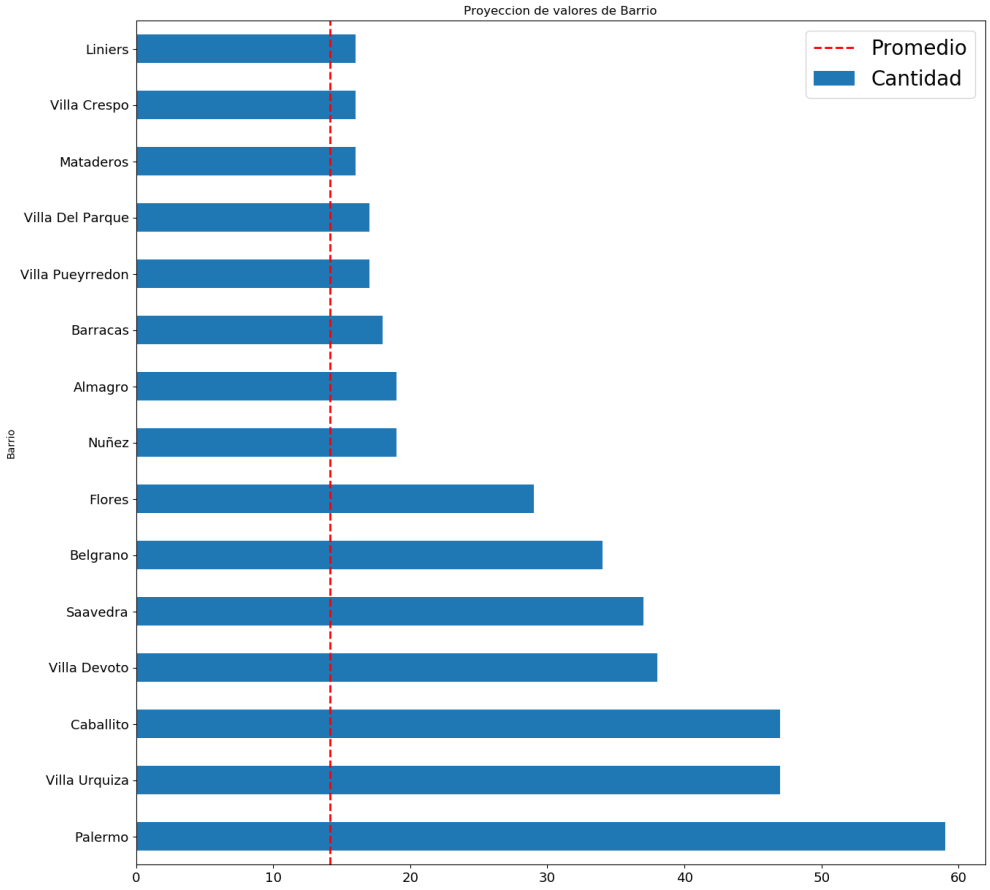
La moda es :2
La mediana es :22

3.10 Barrio

La tabla de frecuencias para los primeros 15 atributos es

	Barrio	Cantidad	Porcentaje
0	Palermo	59	8.526
1	Villa Urquiza	47	6.792
2	Caballito	47	6.792
3	Villa Devoto	38	5.491
4	Saavedra	37	5.347
5	Belgrano	34	4.913
6	Flores	29	4.191
7	Nuñez	19	2.746
8	Almagro	19	2.746
9	Barracas	18	2.601

Esta columna posee 13 missing que representan un 1.87% . Graficamente su distribución puede verse como:



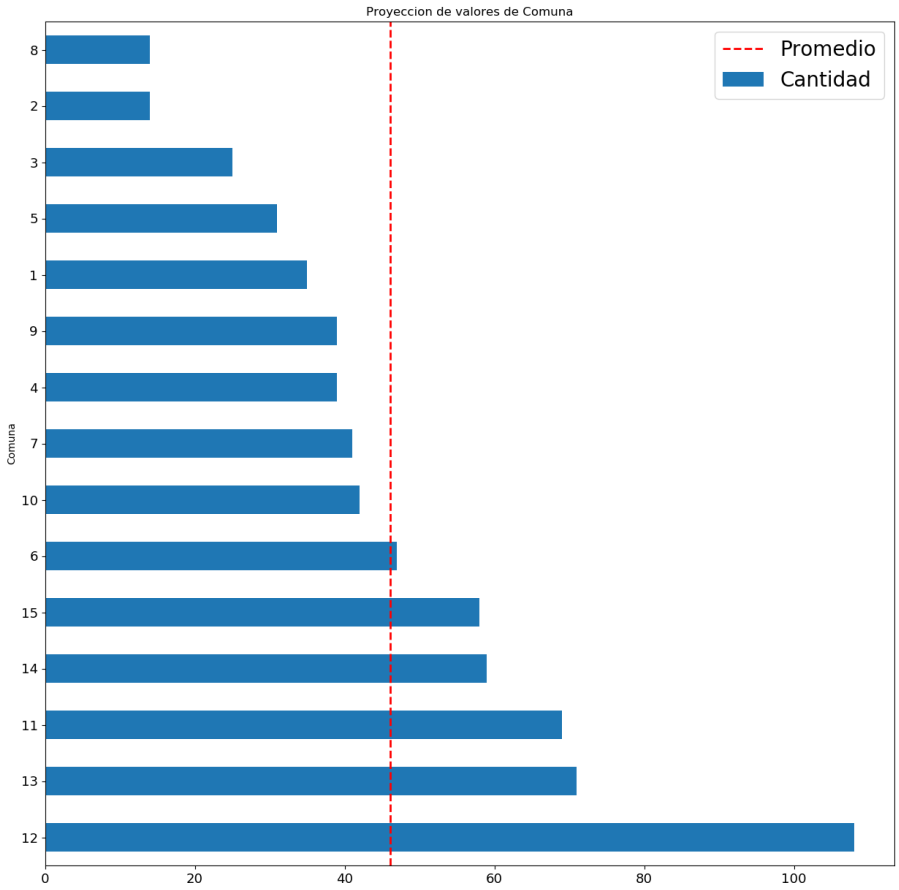
La moda es :Palermo
La mediana es :Palermo

3.11 Comuna

La tabla de frecuencias para los primeros 15 atributos es

	Comuna	Cantidad	Porcentaje
0	12	108	15.607
1	13	71	10.260
2	11	69	9.971
3	14	59	8.526
4	15	58	8.382
5	6	47	6.792
6	10	42	6.069
7	7	41	5.925
8	4	39	5.636
9	9	39	5.636
10	1	35	5.058
11	5	31	4.480
12	3	25	3.613
13	2	14	2.023
14	8	14	2.023

Esta columna no posee missing . Graficamente su distribución puede verse como:



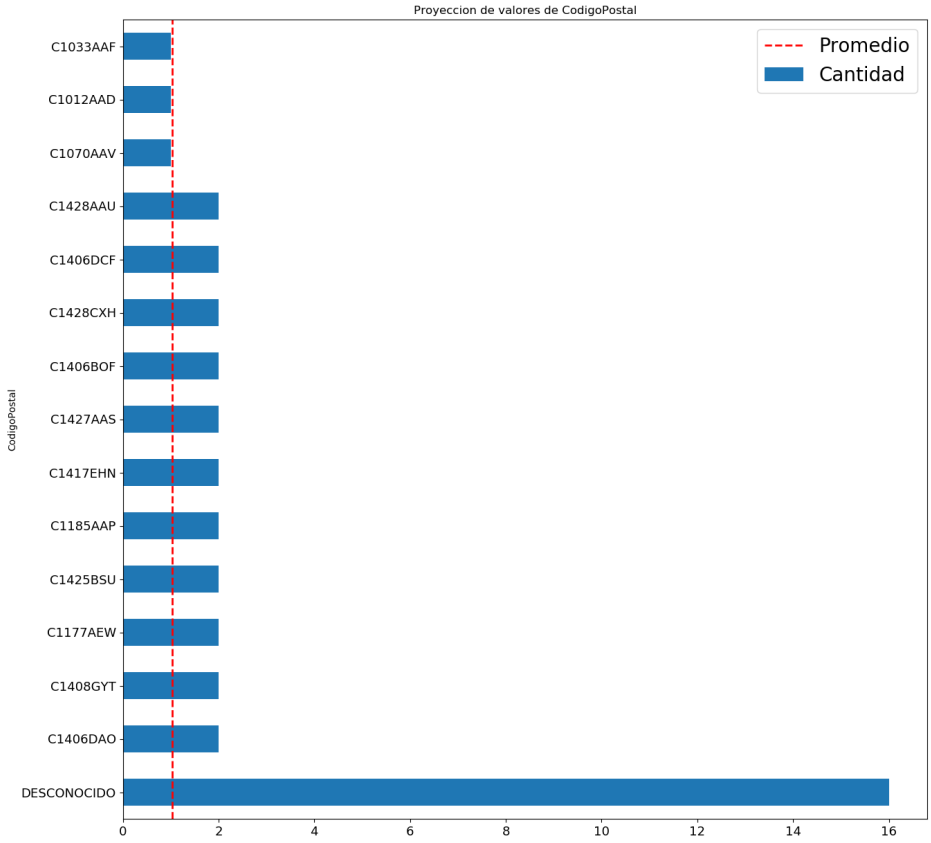
La moda es :12
La mediana es :11

3.12 CodigoPostal

La tabla de frecuencias para los primeros 15 atributos es

	CodigoPostal	Cantidad	Porcentaje
0	DESCONOCIDO	16	2.312
1	C1406DAO	2	0.289
2	C1408GYT	2	0.289
3	C1177AEW	2	0.289
4	C1425BSU	2	0.289
5	C1185AAP	2	0.289
6	C1417EHN	2	0.289
7	C1427AAS	2	0.289
8	C1406BOF	2	0.289
9	C1428CXH	2	0.289
10	C1406DCF	2	0.289
11	C1428AAU	2	0.289
12	C1070AAV	1	0.145
13	C1012AAD	1	0.145
14	C1033AAF	1	0.145

Esta columna posee 16 missing que representan un 2.31% . Graficamente su distribución puede verse como:



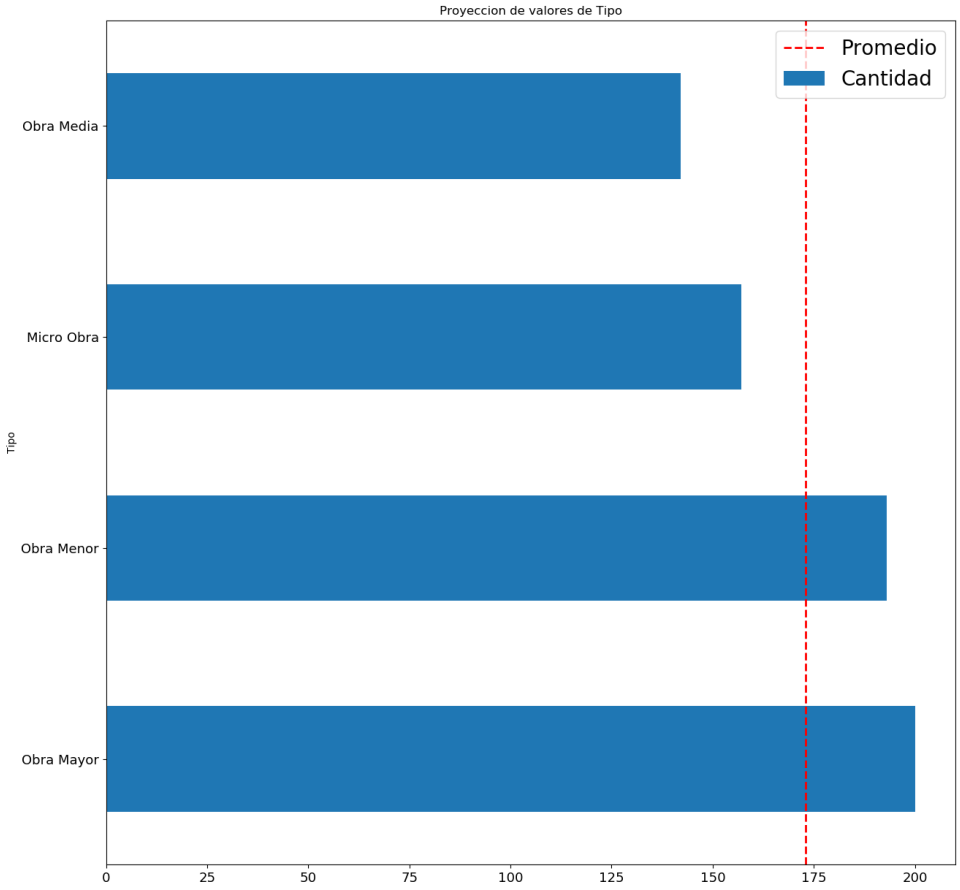
La moda es :C1177AEW
La mediana es :C1417FNZ

3.13 Tipo

La tabla de frecuencias para los primeros 15 atributos es

	Tipo	Cantidad	Porcentaje
0	Obra Mayor	200	28.902
1	Obra Menor	193	27.890
2	Micro Obra	157	22.688
3	Obra Media	142	20.520

Esta columna no posee missing . Graficamente su distribución puede verse como:



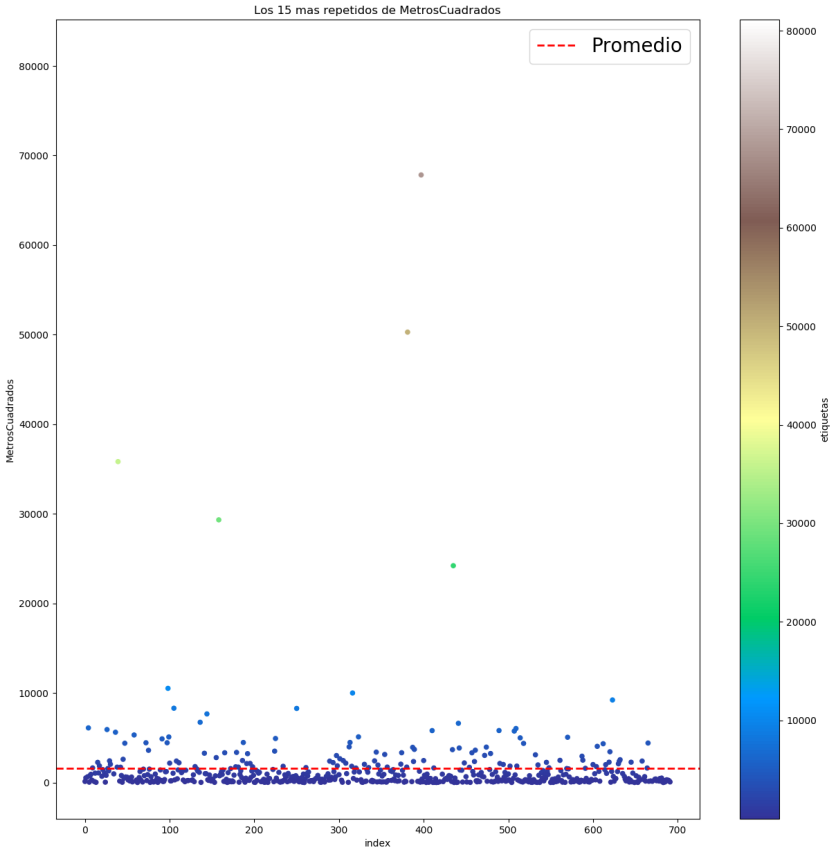
La moda es :Obra Mayor
La mediana es :Obra Mayor

3.14 MetrosCuadrados

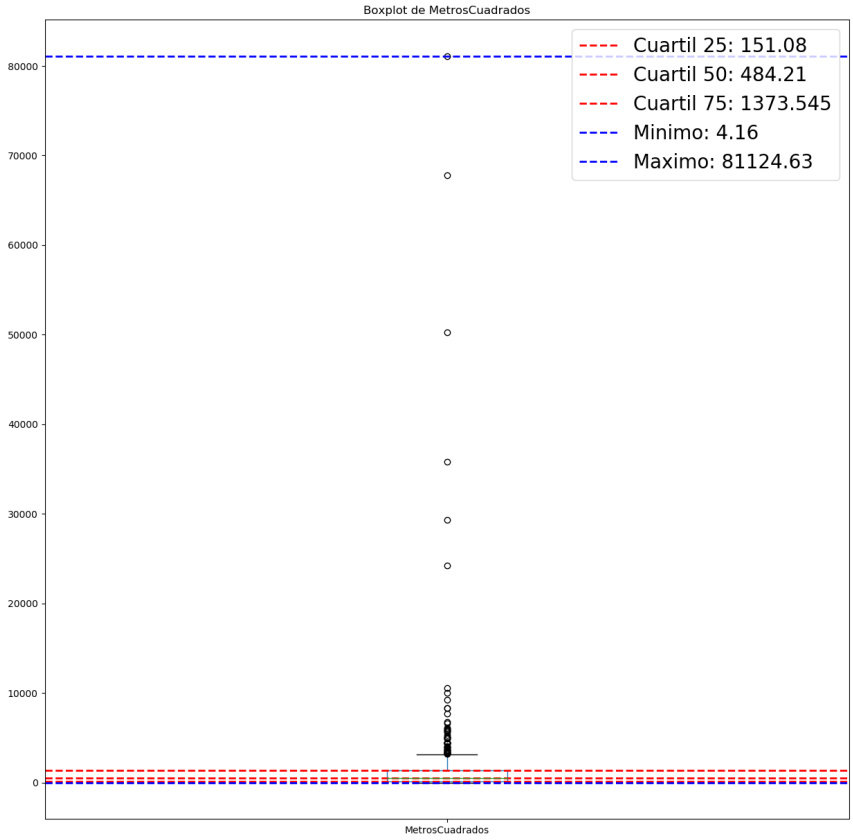
La tabla de frecuencias para los primeros 15 atributos es

	CodigoPostal	Cantidad	Porcentaje
0	DESCONOCIDO	16	2.312
1	C1406DAO	2	0.289
2	C1408GYT	2	0.289
3	C1177AEW	2	0.289
4	C1425BSU	2	0.289
5	C1185AAP	2	0.289
6	C1417EHN	2	0.289
7	C1427AAS	2	0.289
8	C1406BOF	2	0.289
9	C1428CXH	2	0.289
10	C1406DCF	2	0.289
11	C1428AAU	2	0.289
12	C1070AAV	1	0.145
13	C1012AAD	1	0.145
14	C1033AAF	1	0.145

Esta columna posee 129 missing que representan un 18.64% . Graficamente su distribución puede verse como:



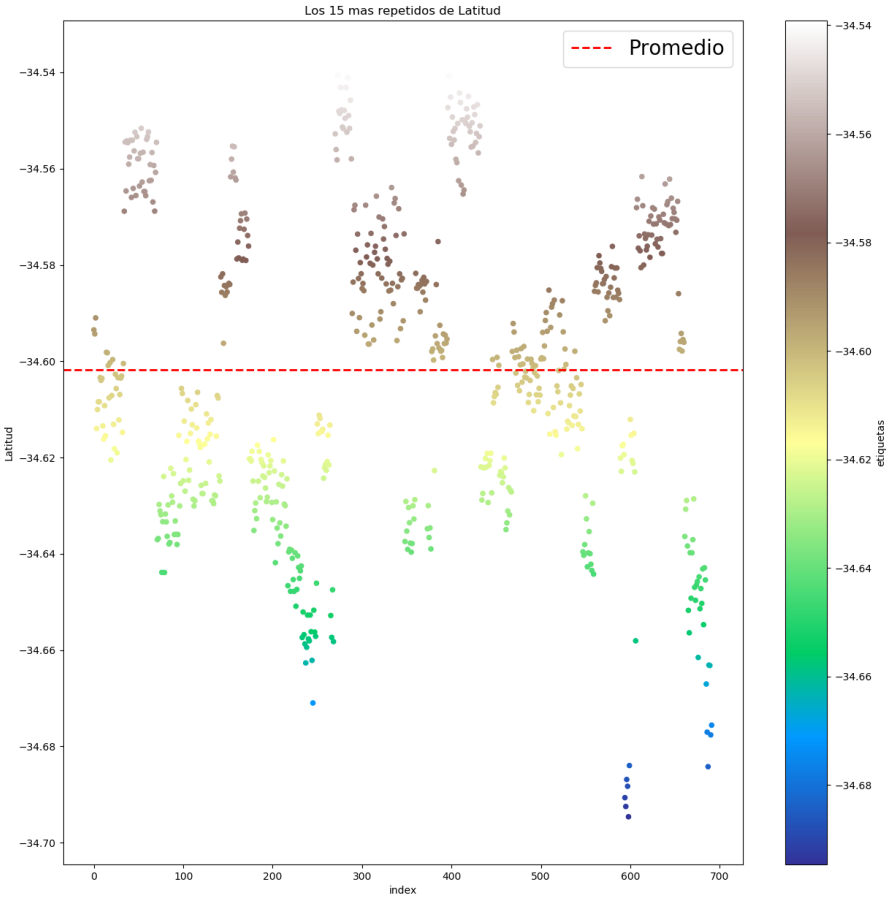
La moda es :221.93
La mediana es :476.37



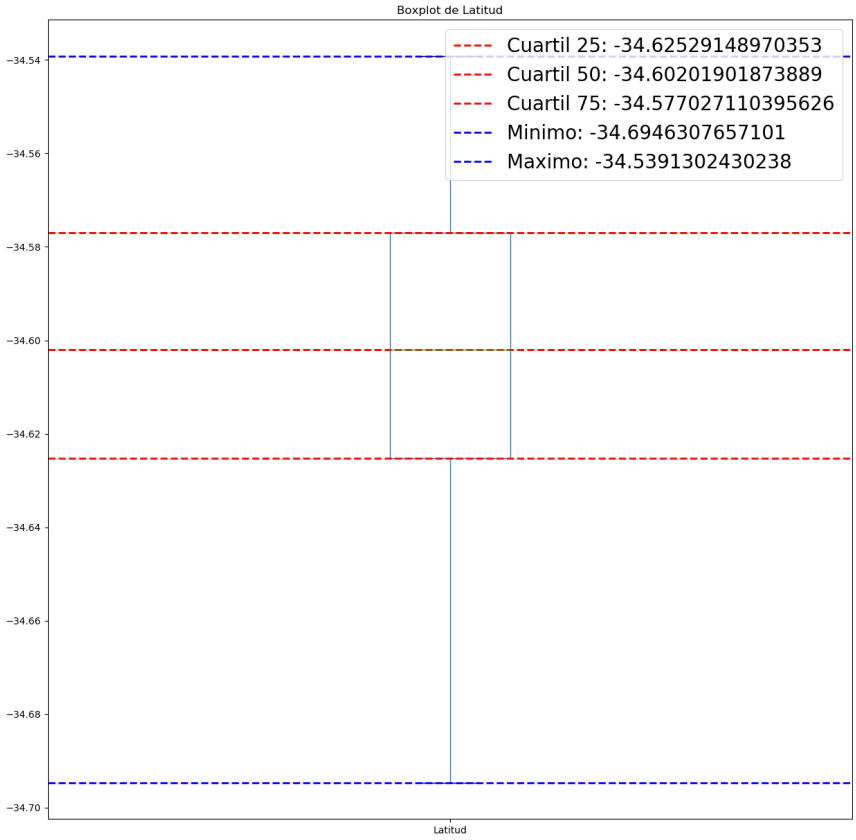
3.15 Latitud

	Latitud	Cantidad	Porcentaje
0	-34.600	2	0.289
1	-34.648	2	0.289
2	-34.554	2	0.289
3	-34.565	2	0.289
4	-34.617	2	0.289
5	-34.596	2	0.289
6	-34.627	2	0.289
7	-34.558	2	0.289
8	-34.687	1	0.145
9	-34.678	1	0.145
10	-34.684	1	0.145
11	-34.684	1	0.145
12	-34.658	1	0.145
13	-34.688	1	0.145
14	-34.691	1	0.145

Esta columna no posee missing . Graficamente su distribución puede verse como:



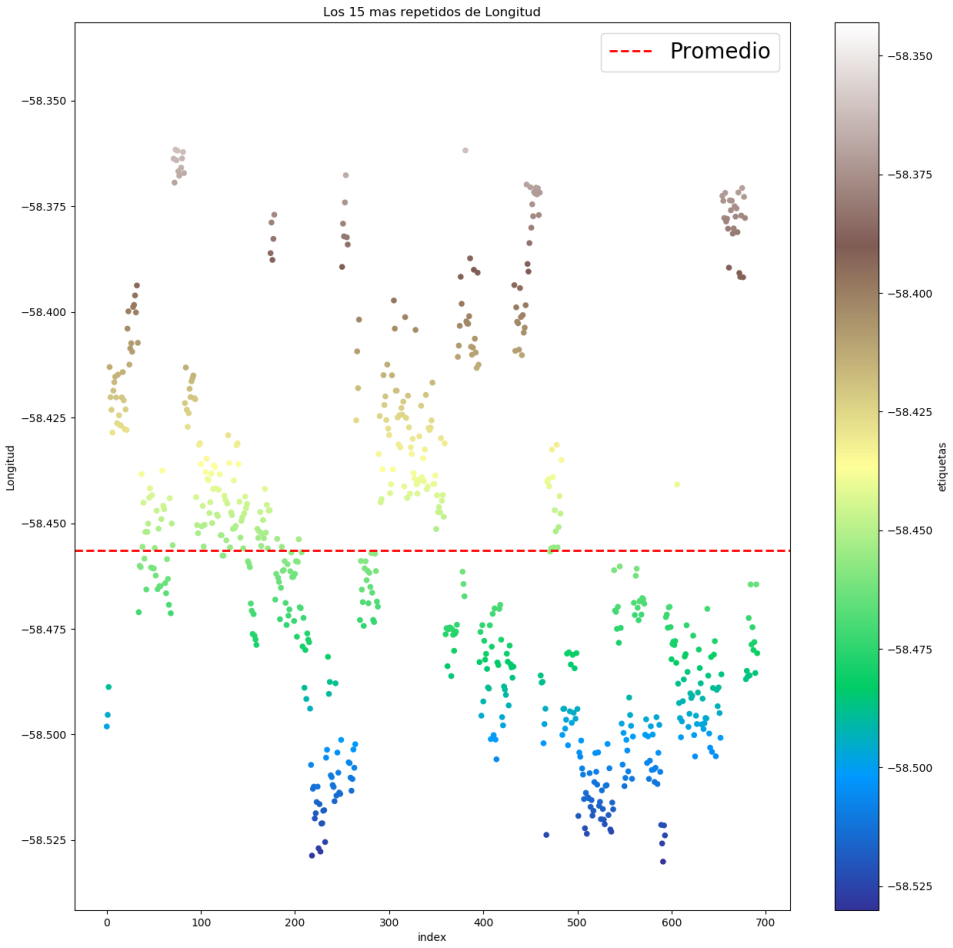
La moda es :-34.647807201376
La mediana es :-34.6021021876619



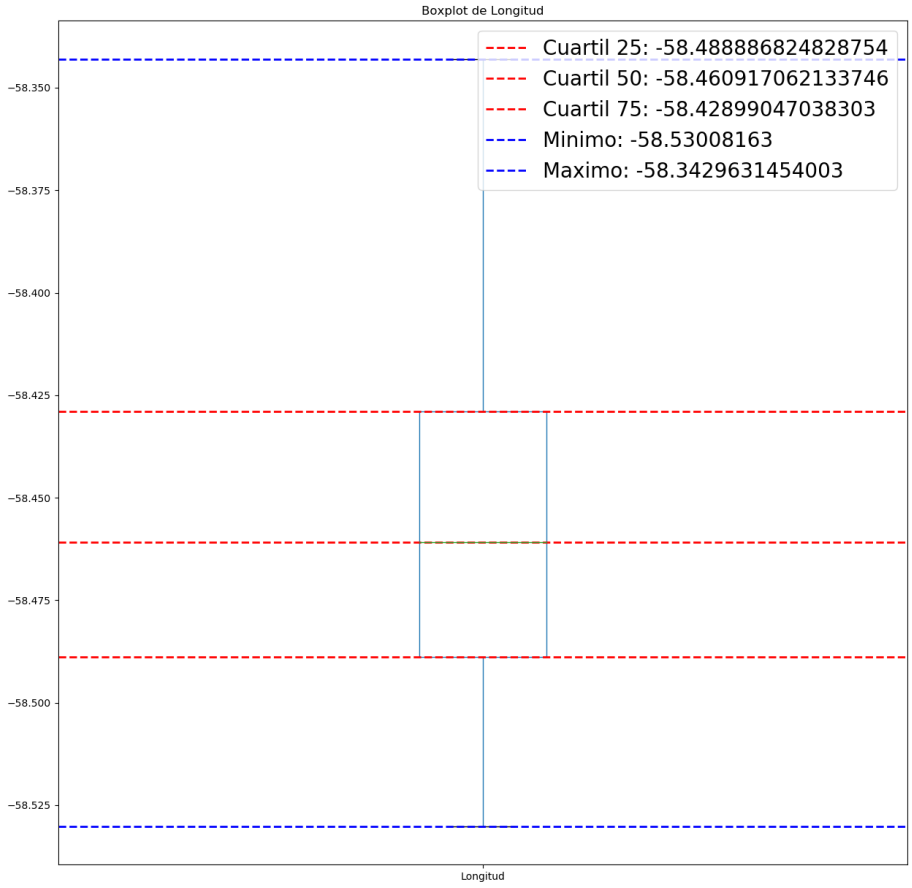
3.16 Longitud

	Longitud	Cantidad	Porcentaje
0	-58.494	2	0.289
1	-58.452	2	0.289
2	-58.512	2	0.289
3	-58.465	2	0.289
4	-58.419	2	0.289
5	-58.453	2	0.289
6	-58.458	2	0.289
7	-58.461	2	0.289
8	-58.526	1	0.145
9	-58.524	1	0.145
10	-58.524	1	0.145
11	-58.525	1	0.145
12	-58.520	1	0.145
13	-58.527	1	0.145
14	-58.528	1	0.145

Esta columna no posee missing . Graficamente su distribución puede verse como:



La moda es :-58.5123330398802
La mediana es :-58.461075629955594



3.17 Tratamiento de Missing

Efectuado el análisis univariado detectamos el porcentaje de missing por columna y estos podemos clasificarlos en:

1. Remediables
2. No Remediables

Los `No Remediables` son los de la columna `MetrosCuadrados`, `Parcela` ya que para obtenerlos deberiamos tener acceso al expediente y constatar si existe el valor.

Los `Remediables` son los correspondientes a las columnas `Barrio`, `CodigoPostal` y `AlturaObra` .

Para completar el código postal utilizamos la página web provista por el Correo Argentino a partir de la calle y altura.

https://www.correoargentino.com.ar/formularios/cpa

Y para completar el barrio realizamos la consulta a partir de la calle y altura en

https://mapa.buenosaires.gob.ar

Es posible automatizar esta acción utilizando APIs pero por falta de tiempo tuvimos que hacerlo manualmente.

Y para la `AlturaObra` tomamos el valor de `Altura`.

El dataset obtenido lo presentamos en el archivo `obras.csv`

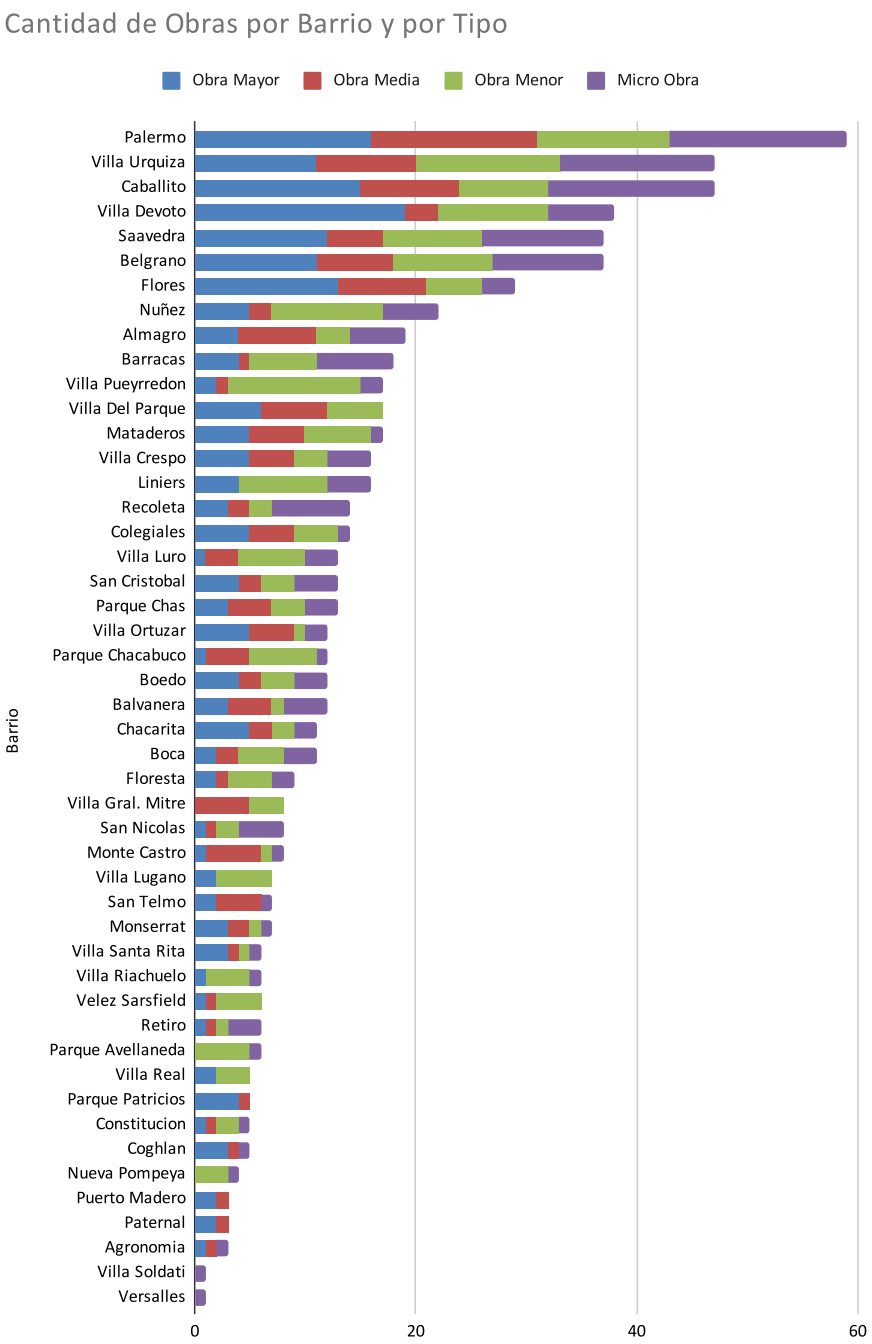
4 Análisis Bivariado

4.1 Barrio y Tipo

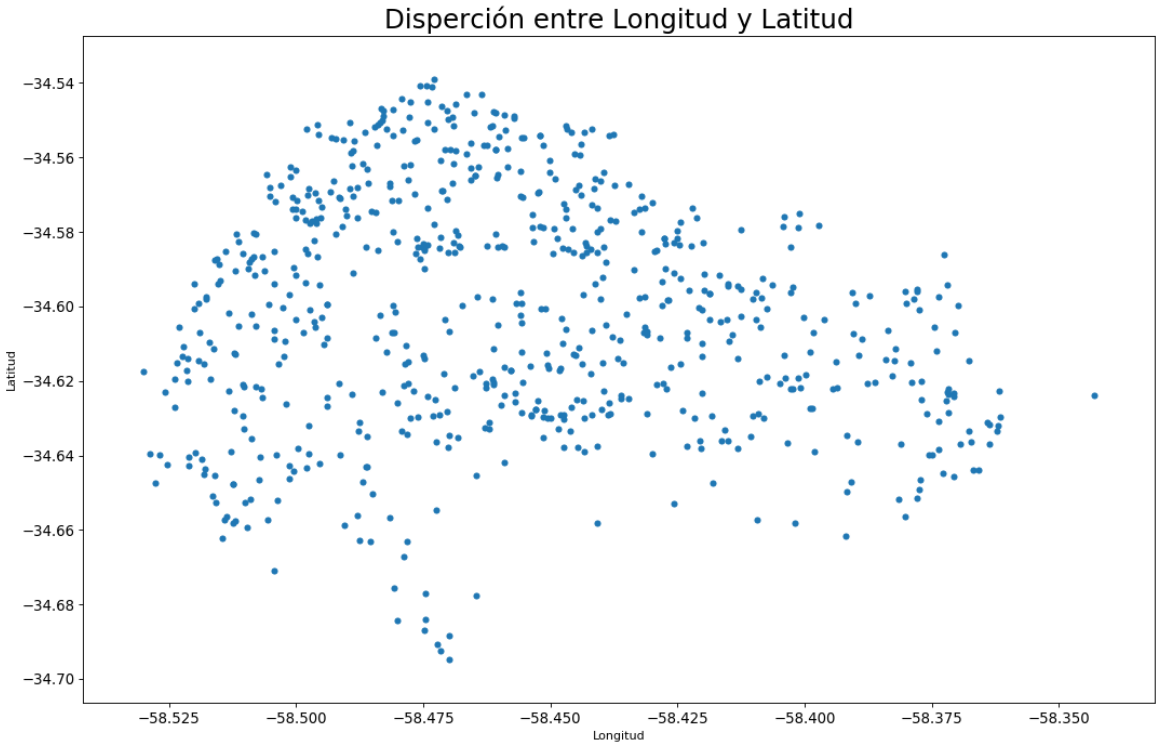
4.1.1 Tabla de contingencia

Barrio	Micro Obra	Obra Mayor	Obra Media	Obra Menor	Grand Total
Palermo	16	16	15	12	59
Villa Urquiza	14	11	9	13	47
Caballito	15	15	9	8	47
Villa Devoto	6	19	3	10	38
Saavedra	11	12	5	9	37
Belgrano	10	11	7	9	37
Flores	3	13	8	5	29
Nuñez	5	5	2	10	22
Almagro	5	4	7	3	19
Barracas	7	4	1	6	18
Villa Pueyrredon	2	2	1	12	17
Villa Del Parque		6	6	5	17
Mataderos	1	5	5	6	17
Villa Crespo	4	5	4	3	16
Liniers	4	4		8	16
Recoleta	7	3	2	2	14
Colegiales	1	5	4	4	14
Villa Luro	3	1	3	6	13
San Cristobal	4	4	2	3	13
Parque Chas	3	3	4	3	13
Villa Ortuzar	2	5	4	1	12
Parque Chacabu	1	1	4	6	12
Boedo	3	4	2	3	12
Balvanera	4	3	4	1	12
Chacarita	2	5	2	2	11
Boca	3	2	2	4	11
Floresta	2	2	1	4	9
Villa Gral. Mitre			5	3	8
San Nicolas	4	1	1	2	8
Monte Castro	1	1	5	1	8
Villa Lugano		2		5	7
San Telmo	1	2	4		7
Monserrat	1	3	2	1	7
Villa Santa Rita	1	3	1	1	6
Villa Riachuelo	1	1		4	6
Velez Sarsfield		1	1	4	6
Retiro	3	1	1	1	6
Parque Avellane	1			5	6
Villa Real		2		3	5
Parque Patricios		4	1		5
Constitucion	1	1	1	2	5
Coghlan	1	3	1		5
Nueva Pompeya	1			3	4
Puerto Madero		2	1		3
Paternal		2	1		3
Agronomia	1	1	1		3
Villa Soldati	1				1
Versalles	1				1
Grand Total	**157**	**200**	**142**	**193**	**692**

4.1.2 Barras apiladas



4.2 Latitud y Longitud



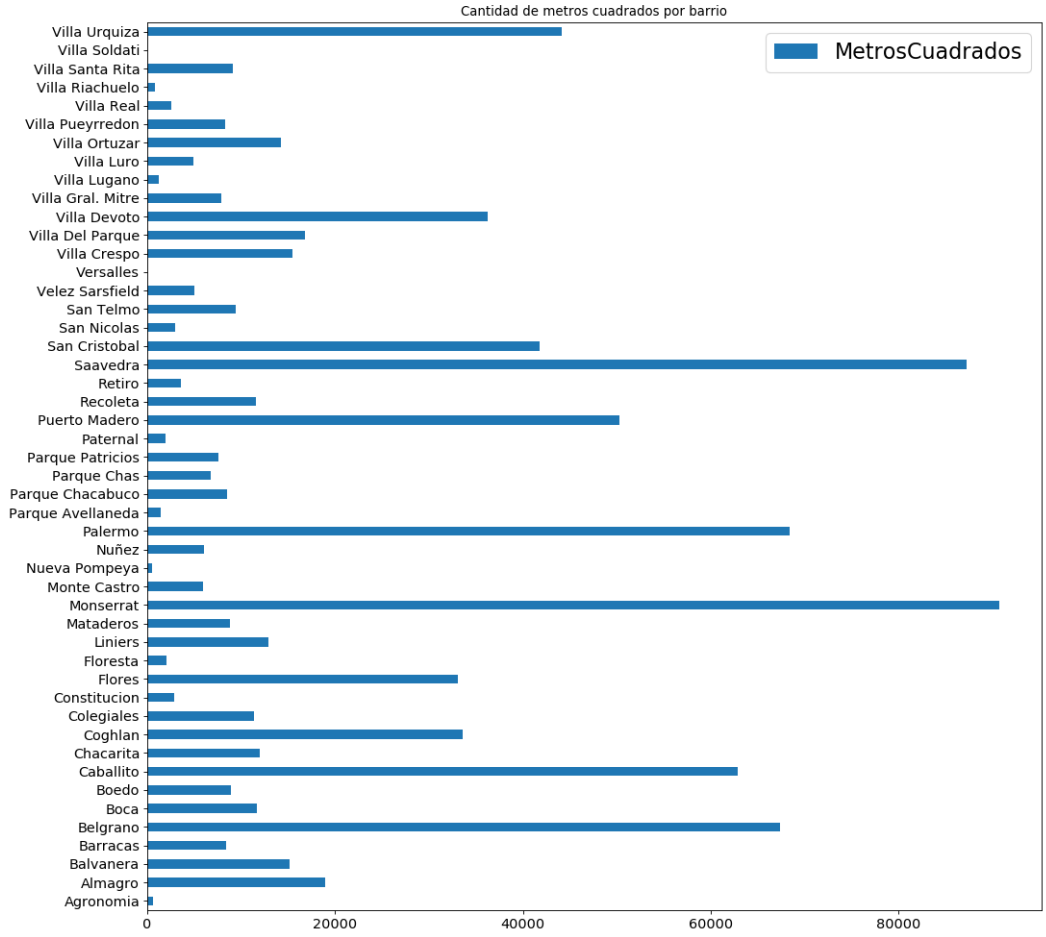
La media de Latitud es -34.60
La media de Longitud es -58.45
La diferencia de medias es -23.85

4.3 Barrio y MetrosCuadrados

4.3.1 Tabla de contingencia

La tabla de contingencias para estas variables es:

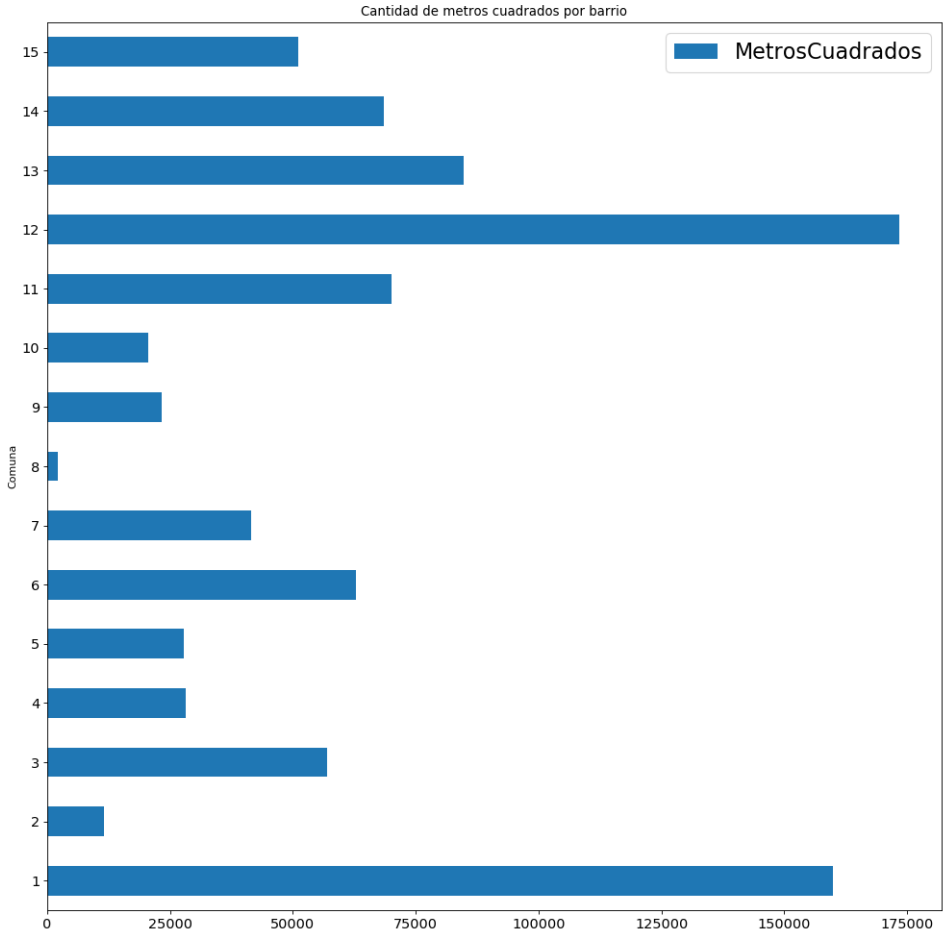
Barrio	MetrosCuadrados
Agronomia	642.18
Almagro	18910.10
Balvanera	15192.31
Barracas	8439.32
Belgrano	67396.58
Boca	11654.15
Boedo	8908.91
Caballito	62859.68
Chacarita	11999.75
Coghlan	33629.64
Colegiales	11363.49
Constitucion	2857.84
Flores	33023.90
Floresta	2066.62
Liniers	12938.66
Mataderos	8804.20
Monserrat	90694.67
Monte Castro	5968.68
Nueva Pompeya	506.68
Nuñez	6051.50
Palermo	68448.90
Parque Avellaneda	1463.14
Parque Chacabuco	8475.17
Parque Chas	6792.87
Parque Patricios	7608.46
Paternal	1992.08
Puerto Madero	50284.82
Recoleta	11616.41
Retiro	3582.95
Saavedra	87285.63
San Cristobal	41790.08
San Nicolas	2952.05
San Telmo	9458.31
Velez Sarsfield	5041.77
Versalles	0.00
Villa Crespo	15427.29
Villa Del Parque	16826.06
Villa Devoto	36205.19
Villa Gral. Mitre	7894.03
Villa Lugano	1271.52
Villa Luro	4866.83
Villa Ortuzar	14240.19
Villa Pueyrredon	8305.21
Villa Real	2573.07
Villa Riachuelo	848.04
Villa Santa Rita	9107.03
Villa Soldati	0.00
Villa Urquiza	44118.41



4.4 Comuna y MetrosCuadrados

4.4.1 Tabla de contingencia

Comuna	MetrosCuadrados
1	159830.64
2	11616.41
3	56982.39
4	28208.61
5	27819.01
6	62859.68
7	41499.07
8	2119.56
9	23206.00
10	20516.97
11	70032.31
12	173338.89
13	84811.57
14	68448.90
15	51094.36



4.5 Comuna y Barrio

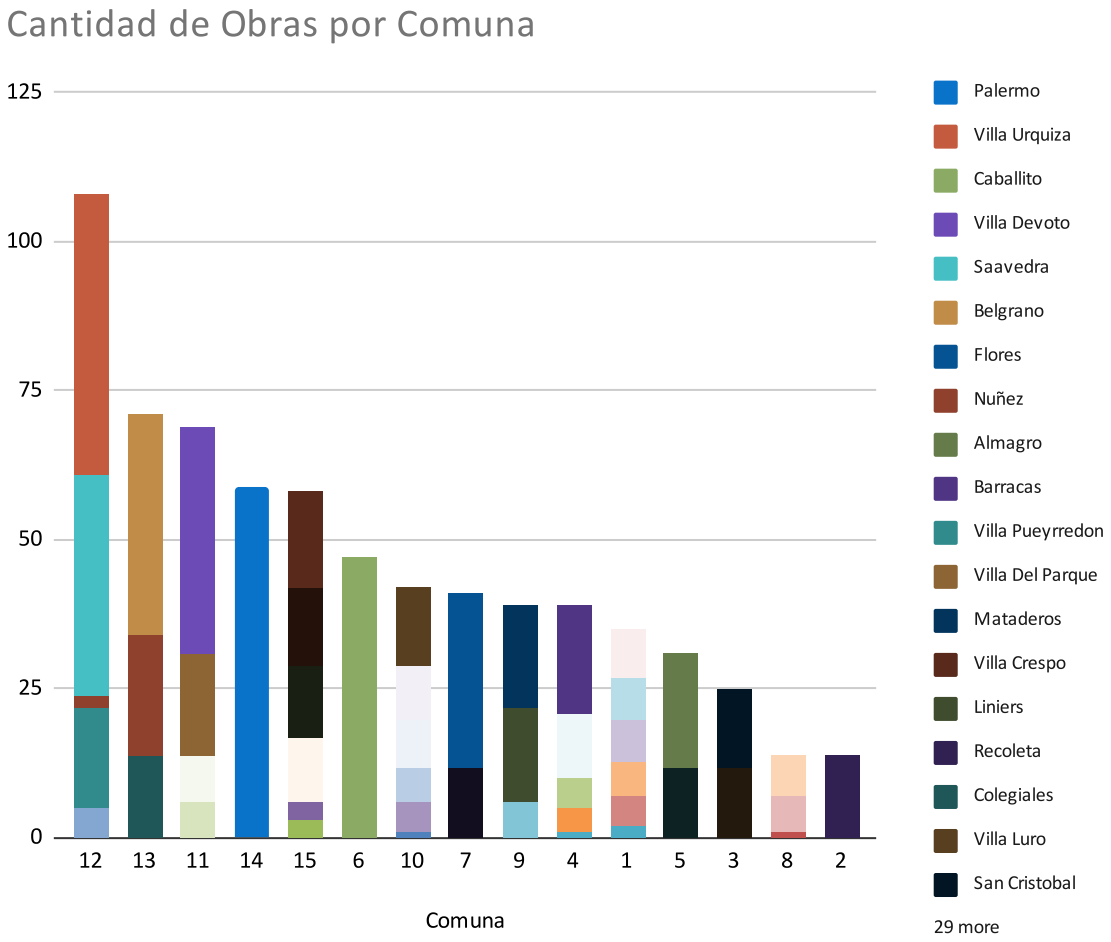
4.5.1 Tabla de contingencia

Debido a la esparcidad de la tabla de contingencia original decidimos representarla de la siguiente forma.

Comuna	*Barrio*	Cantidad
12	Coghlan	5
12	Nuñez	2
12	Saavedra	37
12	Villa Pueyrredon	17
12	Villa Urquiza	47
12 Total		108
13	Belgrano	37
13	Colegiales	14
13	Nuñez	20
13 Total		71
11	Villa Del Parque	17
11	Villa Devoto	38
11	Villa Gral. Mitre	8
11	Villa Santa Rita	6
11 Total		69
14	Palermo	59
14 Total		59
15	Agronomia	3
15	Chacarita	11
15	Parque Chas	13
15	Paternal	3
15	Villa Crespo	16
15	Villa Ortuzar	12
15 Total		58
6	Caballito	47
6 Total		47
10	Floresta	9
10	Monte Castro	8
10	Velez Sarsfield	6
10	Versalles	1
10	Villa Luro	13
10	Villa Real	5
10 Total		42

7	Flores	29
7	Parque Chacabuc	12
7 Total		41
9	Liniers	16
9	Mataderos	17
9	Parque Avellanec	6
9 Total		39
4	Barracas	18
4	Boca	11
4	Nueva Pompeya	4
4	Parque Patricios	5
4	Puerto Madero	1
4 Total		39
1	Constitucion	5
1	Monserrat	7
1	Puerto Madero	2
1	Retiro	6
1	San Nicolas	8
1	San Telmo	7
1 Total		35
5	Almagro	19
5	Boedo	12
5 Total		31
3	Balvanera	12
3	San Cristobal	13
3 Total		25
8	Villa Lugano	7
8	Villa Riachuelo	6
8	Villa Soldati	1
8 Total		14
2	Recoleta	14
2 Total		14
Grand Total		**692**

4.5.2 Barras apiladas


Cantidad de Obras por Comuna
125
100
75
50
25
0
12
13
11
14
15
6
10
7
9
4
1
5
3
8
2
Comuna
Palermo
Villa Urquiza
Caballito
Villa Devoto
Saavedra
Belgrano
Flores
Nuñez
Almagro
Barracas
Villa Pueyrredon
Villa Del Parque
Mataderos
Villa Crespo
Liniers
Recoleta
Colegiales
Villa Luro
San Cristobal
29 more

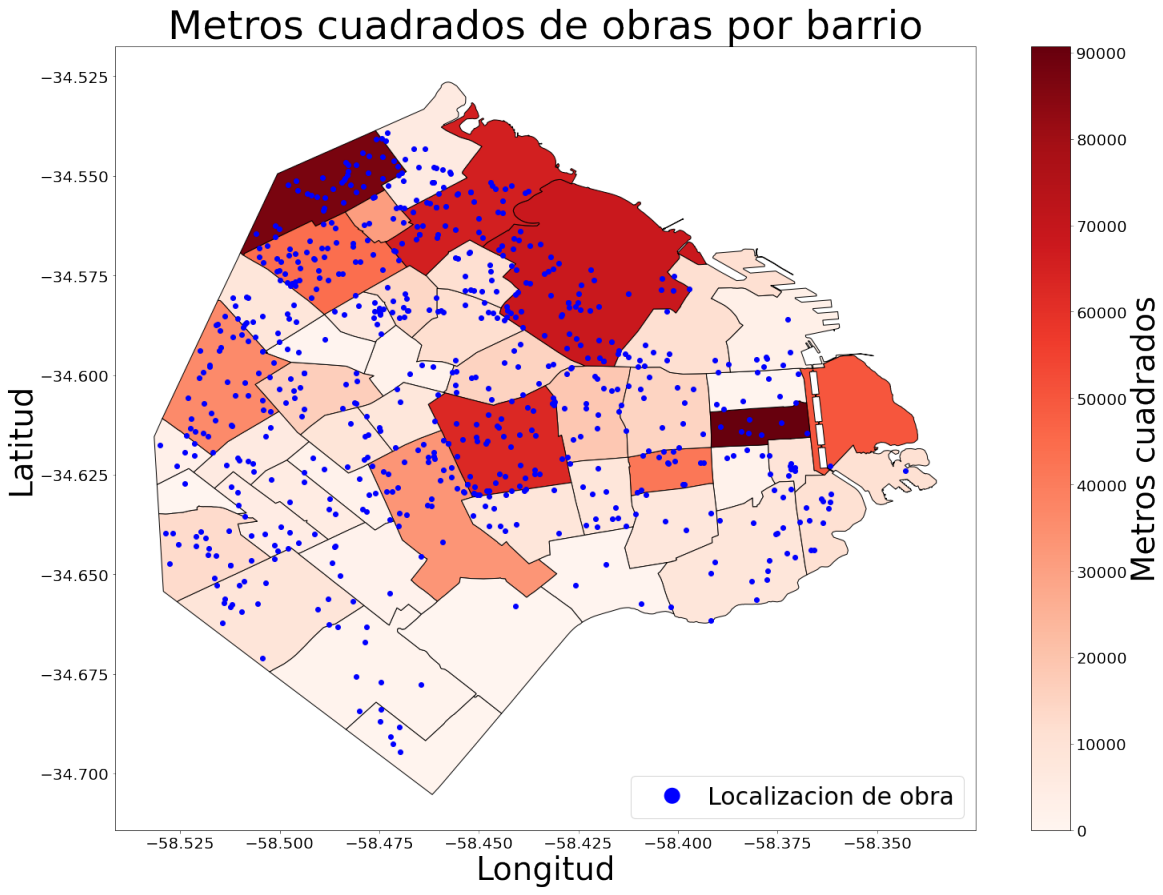
5 Calidad Obtenida y Reglas de Negocio

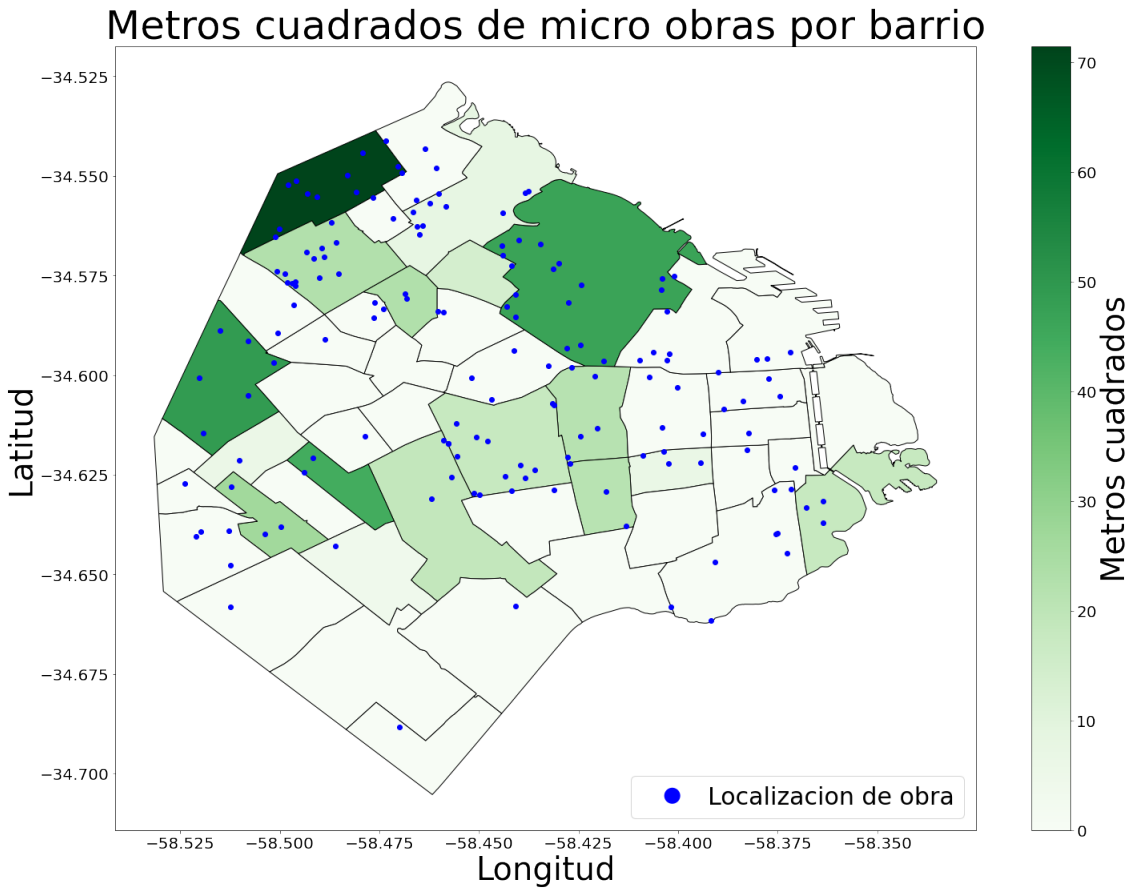
A partir de las transformaciones realizadas al set de datos inicial obtuvimos un dataset que muestra de forma clara y precisa las obras que se realizaron en la Ciudad de Buenos Aires durante el período de Junio a Septiembre de 2019.

Permitiendo establecer reglas de negocio e insights como:

1. Geolocalización de las obras a partir de los campos Latitud y Longitud
2. Ubicación exacta de la obra a partir de Calle, Numero, Zona, Sector, Manzana y Parcela
3. Cantidad de metros cuadrados en obra por Comuna y Barrio. Permitiendo desglozarse por Calle y Altura.
4. Identificar la cantidad de obras de cada tipo por Comuna, Barrio y Calle
5. Identificar las parcelas con más obras.

Algunos de los usos posibles para este set de datos pueden ser :





6 Adjuntos

- Informe Gerencial: InformeGerencial.pdf
- Informe Técnico: InformeTecnico.pdf
- Carpeta `bases` con:
  - obras.csv: Base de datos final con todas las transformacciones y correcciones.
  - obras_con_missing.csv: Base de datos obtenida luego de aplicar el script limpiar.py.
  - obras-registradas-junsep2019.csv: Base de datos inicial descargada del sitio del Gobierno de la Ciudad.
- Carpeta `analisisUnivariado` con:
  - archivo .csv con tabla de frecuencias de cada columna.
  - gráficos de la sección Analisis Univariado.
- Carpeta `mapas` con:
  - dataset de geolocalización usado para graficar los mapas
  - mapas de la sección Calidad Obtenida y Reglas de Negocio
- Carpeta `scripts` con:
  - limpiar.py: script usado para la creación del dataset final.

  - analisisUnivariado.py: script que ejecuta las funciones y graficos del análisis.
  - analisisBivariado.py: script que ejecuta las funciones y graficos del análisis.
  - mapadecalor.ipynb: script usado para generar el mapa de calor por metros cuadrados.
  - mapadecalor_micro_obra.ipynb: script usado para generar el mapa de calor por obra.

- Carpeta `analisisBivariado` con:
  - archivo .csv con tabla de frecuencias de barrio y comuna.
  - gráficos de la sección Analisis Bivariado.

6.1 Requerimientos

Para ejecutar los scripts de Python es necesario tener instalado:

- Python 3.6 ó superior
- Pandas version 1.04 ó superior
- GeoPandas para graficar los mapas con localizacion de obras