DATA WOMEN'S COMPANY

WHERE DATA MAKES SENSE

DATA WOMEN'S COMPANY

WHERE DATA MAKES SENSE

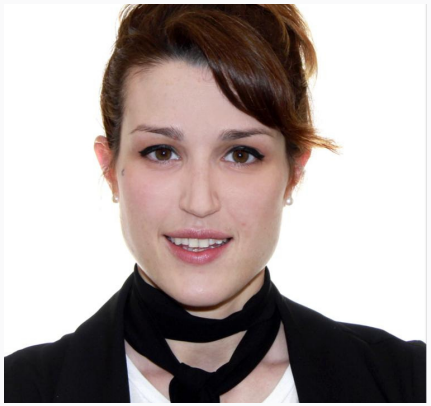
Diana Maria Toro López

María de Lluch Gual Pérez

Sanja Aleksova

Sandra Julieth Castaño Reina

Elsa Cembrero Bonet



Mireia Hernández Lozano

ÍNDICE

1. INTRODUCCIÓN Y OBJETIVOS

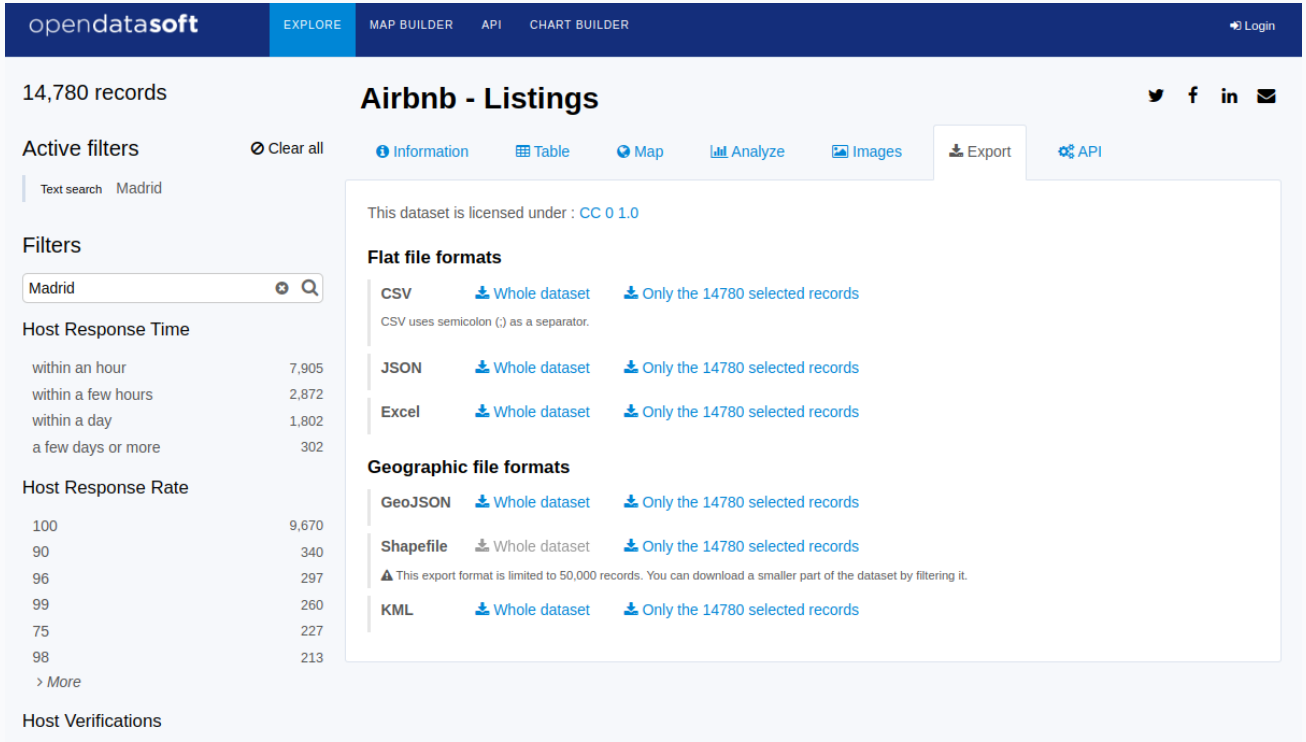
python™

PostgreSQL

R

+ableau

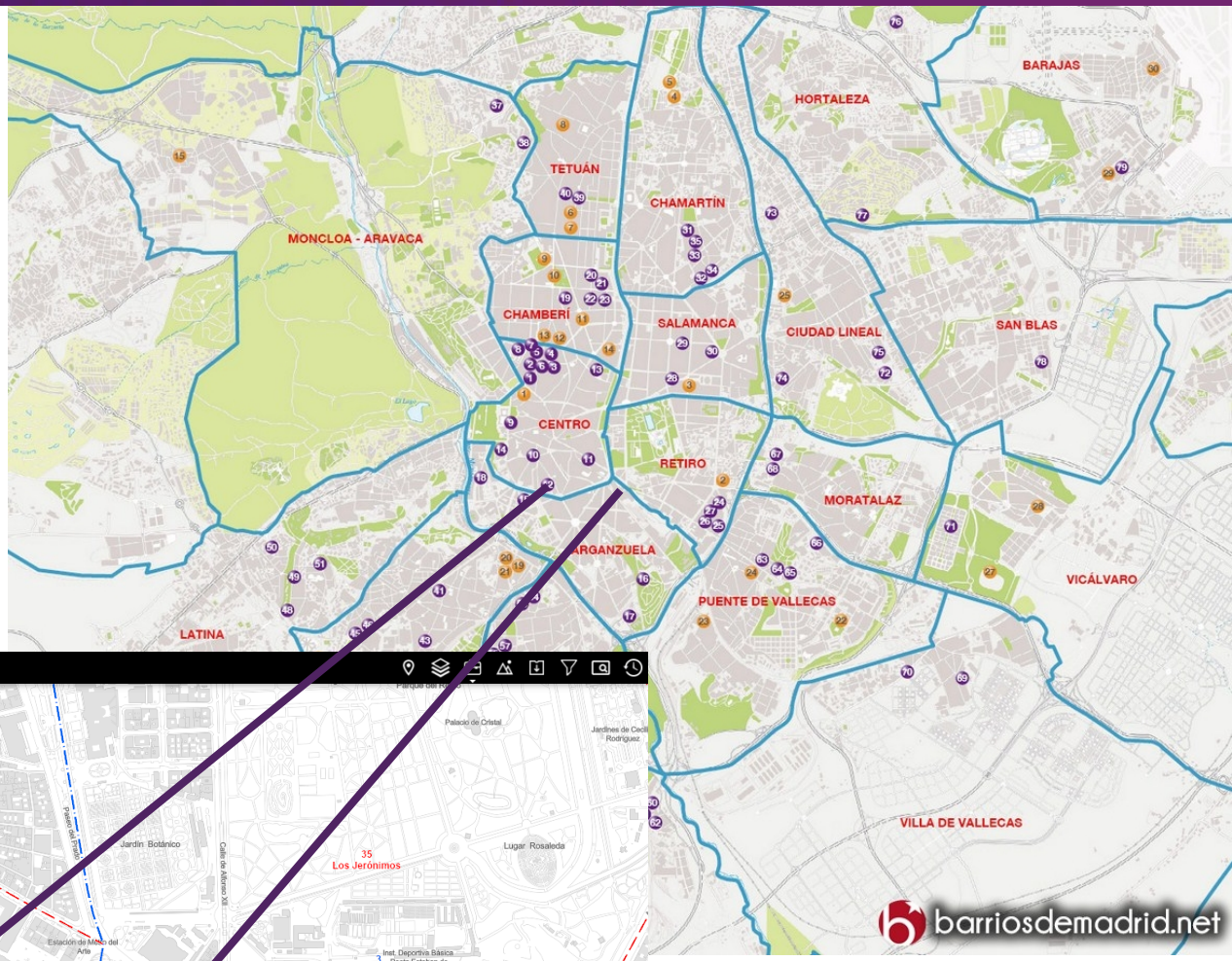
2. DEFINIENDO EL DATASET

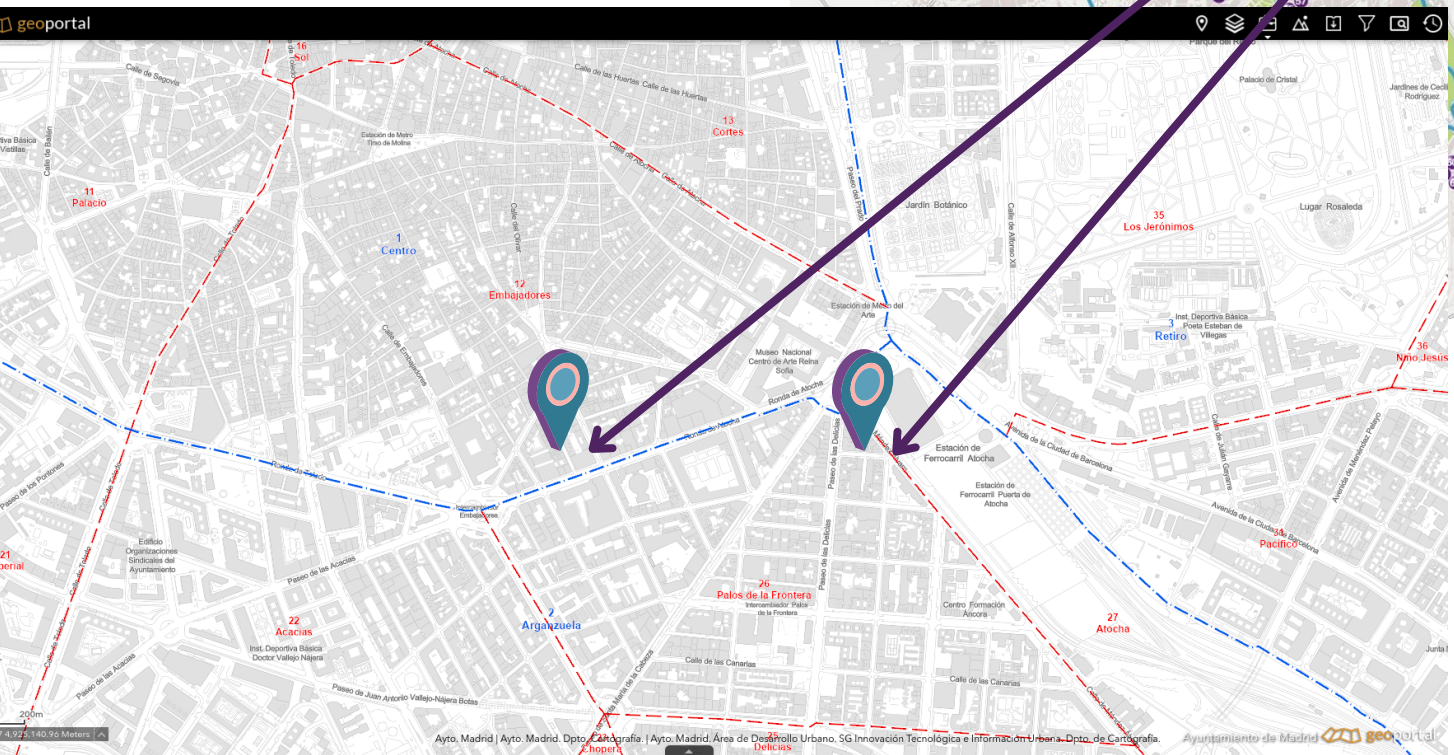



```
Data columns (total 89 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   ID                            14780 non-null  int64
 1   Listing Url                   14780 non-null  object
 2   Scrape ID                     14780 non-null  int64
 3   Last Scraped                  14780 non-null  object
 4   Name                          14779 non-null  object
 5   Summary                       14189 non-null  object
 6   Space                         10888 non-null  object
 7   Description                   14774 non-null  object
 8   Experiences Offered           14780 non-null  object
 9   Neighborhood Overview         9134 non-null   object
 10  Notes                         5644 non-null   object
 11  Transit                       9066 non-null   object
 12  Access                        8318 non-null   object
 13  Interaction                   8228 non-null   object
 14  House Rules                   9619 non-null   object
 15  Thumbnail Url                 11960 non-null  object
 16  Medium Url                    11960 non-null  object
 17  Picture Url                   14751 non-null  object
 18  XL Picture Url                11960 non-null  object
 19  Host ID                       14780 non-null  int64
...
 87  Geolocation                   14780 non-null  object
 88  Features                      14779 non-null  object
dtypes: float64(23), int64(13), object(53)
memory usage: 10.0+ MB
```

3. NUESTRO SERVICIO DE CONSULTORÍA

- ¿A qué **precio base** debería publicar los apartamentos?
- ¿Cómo puede variar el **precio base** según la **cantidad de reviews**?
- Grado de implicación: ¿Es rentable ser **superanfitrión**?





- Apartamentos
- Completos
- 2 habitaciones
- 2 camas
- 1 baño
- Hasta 6 huéspedes
- 1 incluido en precio

4. ARQUITECTURA Y VALIDACIÓN DE LOS DATOS

Procesos ETL: Extracción

FUENTE DE DATOS

- Analizar
- Interpretar
- Convertir
- Cuidado y cautela
- Horarios para realizar el proceso

Procesos ETL: Transformación

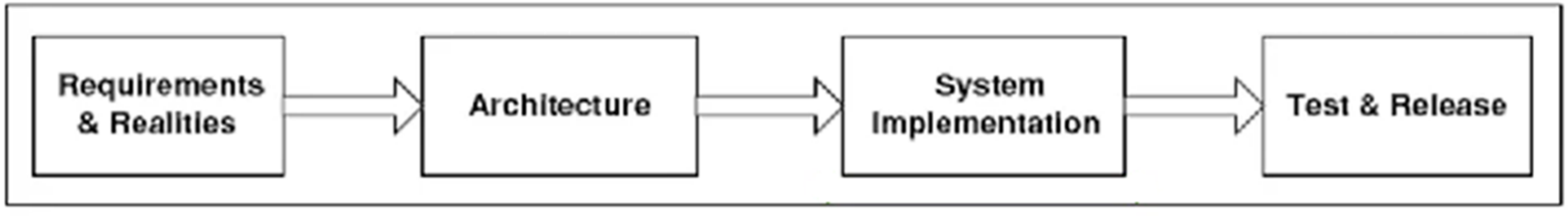
Transformación

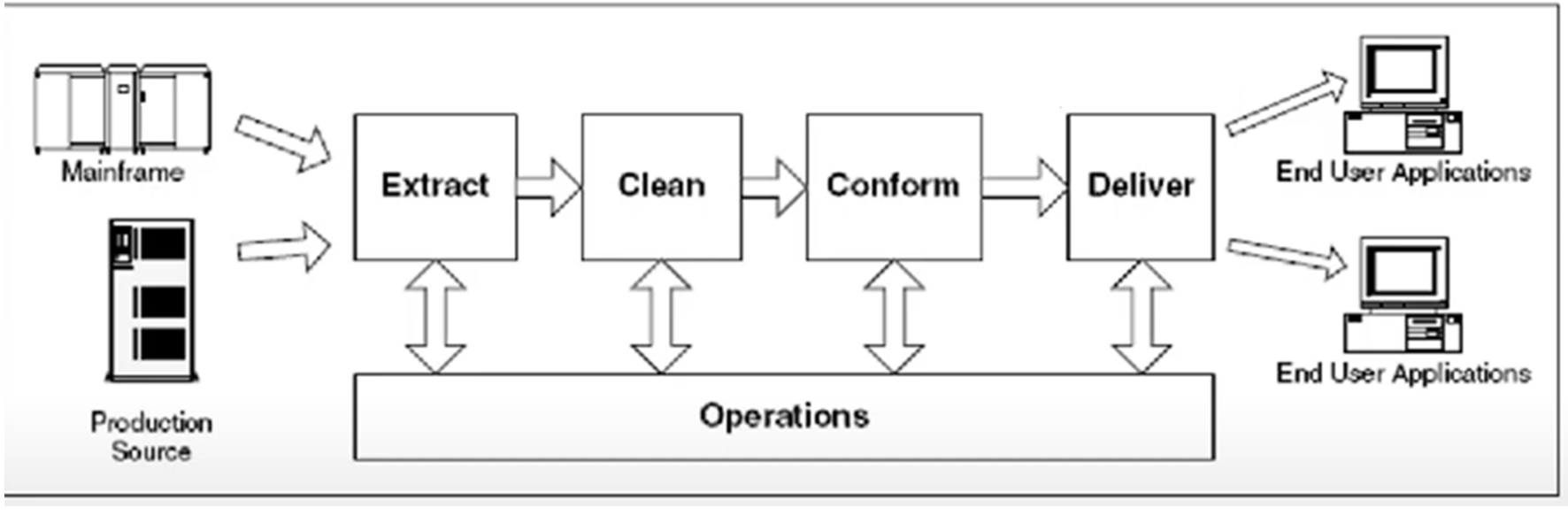
- Aplicar regla de negocio
  - Declarativas
  - Independientes
  - Inteligibles
  - Útiles
- Practico al momento de transformar datos
  - Seleccionar columnas
  - Traducir códigos
  - Codificar valores libres
  - Valores calculados
  - Unir datos de múltiples fuentes
  - Generar campos claves
  - Transponer o pivotar
  - Dividir columnas en varias

DATOS DE DESTINO

Procesos ETL: Carga

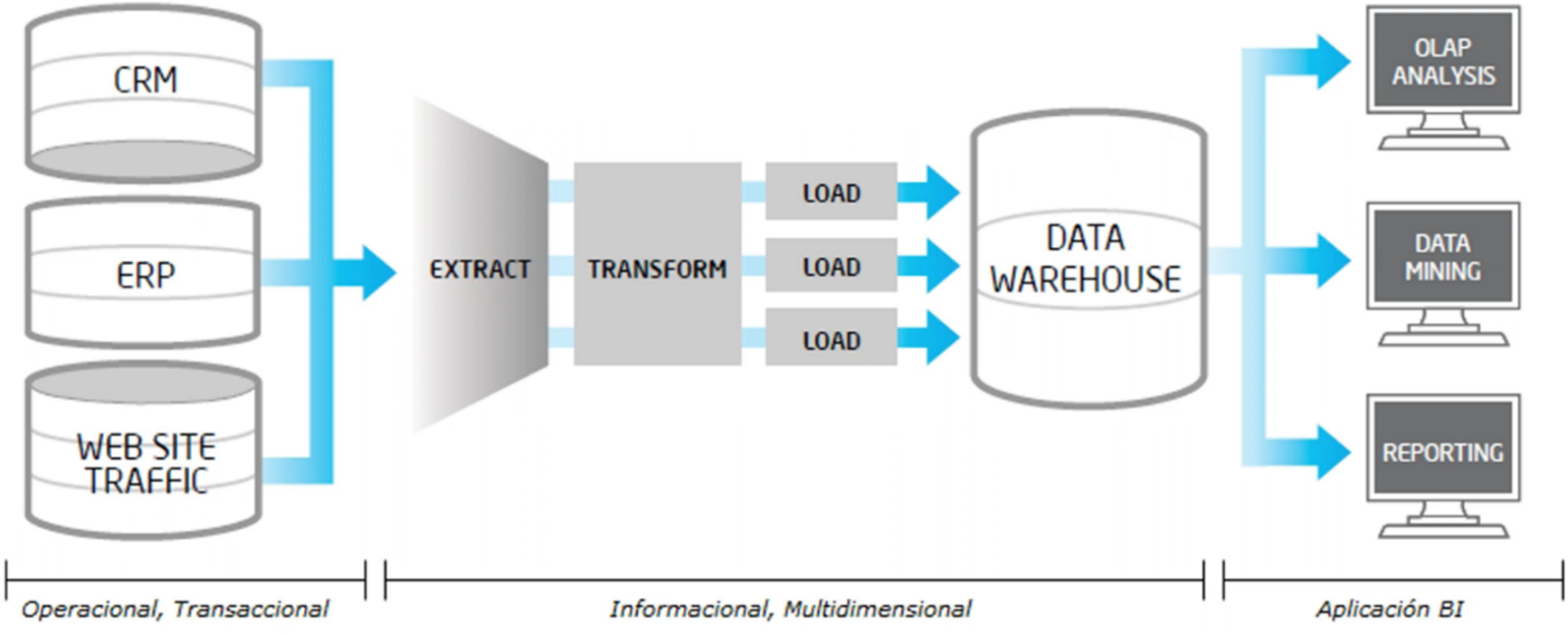
- Carga de datos al sistema de destino
- Políticas de la organización al cargar datos
- Procesos de carga
  - Acumulación simple
  - Rolling


Requirements & Realities
Architecture
System Implementation
Test & Release


Mainframe
Production Source
Extract
Clean
Conform
Deliver
Operations
End User Applications
End User Applications

PostgreSQL

python



5. VISUALIZACIÓN DE LAS MÉTRICAS

- Selección de 4 constantes:
  - Apartamento
  - Precio por día
  - 2 Habitaciones
  - Distritos

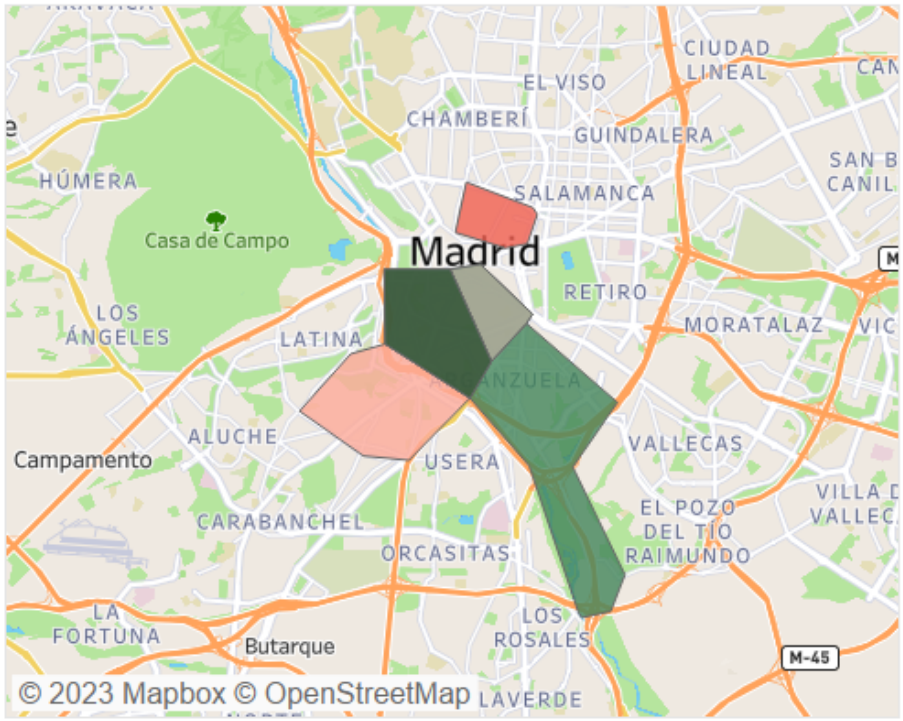
- Visualización de un mapa con los dos distritos y la mediana del precio por día

- 3 gráficos de barras con variables que afectan al precio
  - Afecta positivamente:
    - Ser superhost
    - El barrio
  - Afecta negativamente:
    - Número de reviews

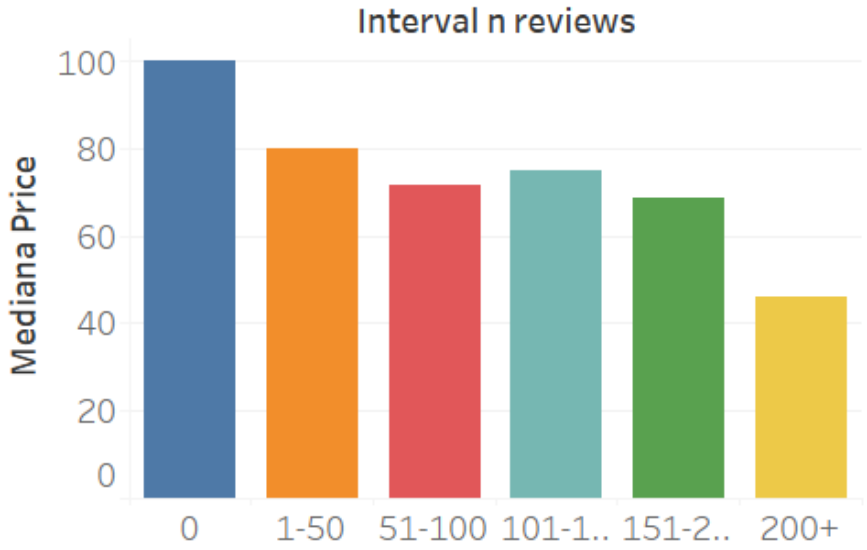
Property Type
apartment

Median Price
60,00 100,00

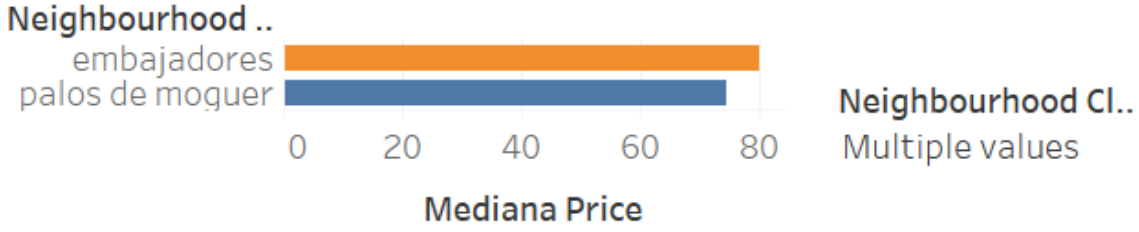
Mapa precio por distrito



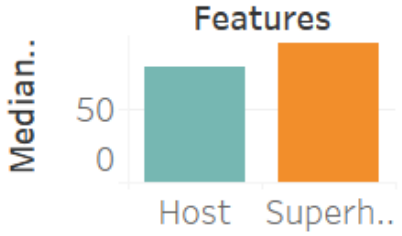
Precio por número de reviews



Precio base por barrio



Precio super host vs. host por barrio



6. PRE-PROCESAMIENTO Y MODELADO

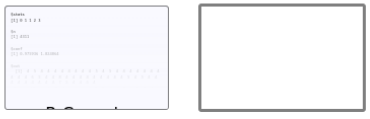
- Selección -> caracterísitcas de los alojamientos

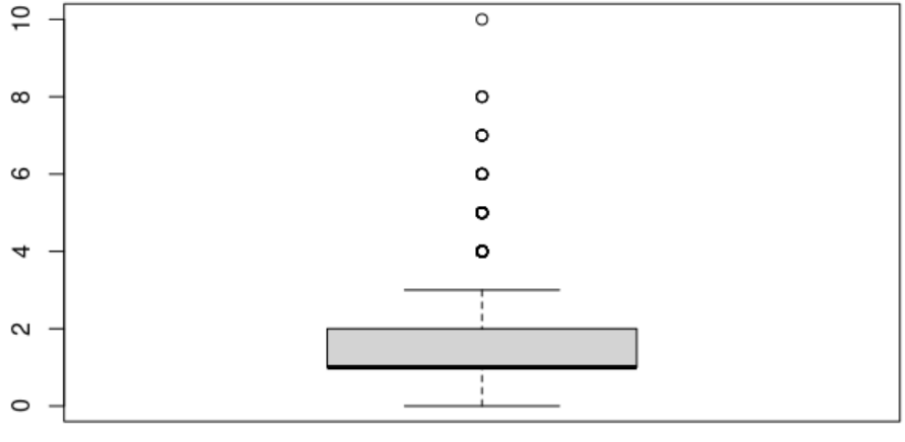
```r
df_amparito <- dplyr:: filter(df, Neighbourhood.Cleansed == "embajadores", Property.Type == "apartment", Room.Type == "entire home/apt")
df_amparito
```

```r
df_alvarito <- dplyr:: filter(df, Neighbourhood.Cleansed == "palos de moguer", Property.Type == "apartment", Room.Type == "entire home/apt")
df_alvarito
```

- Detección de outliers - características

```r
boxplot.stats(df_amparo$Bedrooms)
boxplot(df_amparo$Bedrooms)
```





- Nueva selección con rangos de outliers

```r
amparito_s <- dplyr:: filter(df_amparito, Bedrooms <=3, Beds <= 5, Number.of.Reviews <= 121, Price <=140, Accommodates <=7)
amparito_s
```

```r
alvarito_s <- dplyr:: filter(df_alvarito, Bedrooms <=3, Beds <= 4, Guests.Included <=3, Number.of.Reviews <= 49, Price <=130, Accommodates <=8)
alvarito_s
```

- Correlacionar -> relaciones entre variables
- Gráfica y análisis de correlación
- Generación de modelo con todas las variables observando las que no son útiles
- Comprobación mediante la función **step ()** que selecciona las variables previstas
- Obtención de fórmulas

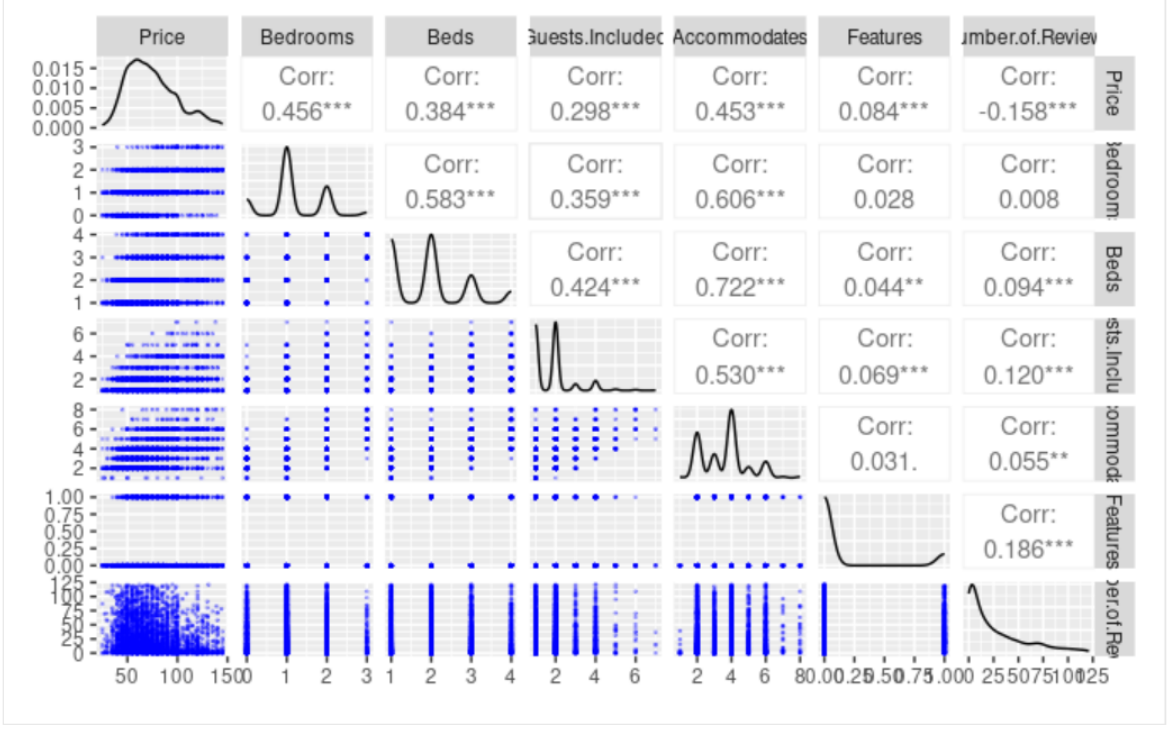
```r
{r}
round(cor(x = amparito_s[,c("Price","Bedrooms", "Beds", "Guests.Included", "Accommodates","Features",
"Number.of.Reviews")], method = "pearson"), 3)
```

```r
{r}
round(cor(x = amparito_s[,c("Price","Bedrooms", "Beds", "Guests.Included", "Accommodates","Features",
"Number.of.Reviews")], method = "spearman"), 3)
```

```r
{r}
ggpairs(amparo_s[,c("Price","Bedrooms", "Beds", "Guests.Included", "Accommodates","Features", "Number
.of.Reviews")], lower = list(continuous = wrap("points", alpha = 0.3,size=0.3,color='blue')))
```

```r
{r}
model_amparito_s<- lm(data=amparito_s, formula = Price~Bedrooms+Beds+Guests.Included+Accommodates
+Features+Number.of.Reviews)
summary(model_amparito_s)
```

```r
{r}
step(object = model_amparito_s, direction = "both")
```



Fórmula

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_kX_k$$

	NOMBRE / CANTIDAD DE LISTINGS		REGRESIÓN LINEAL				
	Base		Predictores aplicados	Fórmula $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_kX_{ki}+\epsilon_i$	Multiple R^{2*} - p-value	Precio apartamento	
Selección	Nombre	Total				Host	Superhost
Dataset	*airbnb-listings.csv*	14780	-	-	-	-	
Madrid	*df_madrid*	6970	Bedrooms + Beds + Accommodates + Features + Number.of.Reviews	*Y=30.59119+17.32385*Br+3.96598*B+6.30426*A+6.22138*F-0.15766*NR*	**0.3389** / < 2.2e-16	$111,00	$117,22
Host	*df_host*	6068	Bedrooms + Beds + Guests.Included + Accommodates + Number.of.Reviews	*Y=31.55054+17.24586*Br+4.24305*B-1.18408*GI +6.58241*A-0.17421*NR*	**0.3306** / < 2.2e-16	$111,65	-
Superhost	*df_superhost*	902	Bedrooms + Guests.Included + Accommodates + Number.of.Reviews	*Y=30.67782+17.98233*Br+5.59964*GI+6.02448*A-0.07431*NR*	**0.4383** / < 2.2e-16	-	$108,39
Dº Centro (c/ de Amparo)	*df_amparo*	**4311**	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=32.3670+17.8795*Br +3.3104*B+1.3557*GI+6.1332*A+6.2986*F-0.1518*NR*	**0.4078** / < 2.2e-16	**$112,90**	**$119,20**
Bº Embajadores (c/ de Amparo)	*df_amparito*	**1099**	Bedrooms + Beds + Accommodates + Features + Number.of.Reviews	*Y=29.1843+14.1807*Br +2.5654*B+6.0246*A +10.7265*F-0.1191*NR*	0.4636 / < 2.2e-16	**$98,82**	**$109,55**
Dº Arganzuela (c/ de Méndez Álvaro)	*df_alvaro*	303	Bedrooms + Guests.Included + Accommodates + Number.of.Reviews	*Y=35.2937+9.3598*Br-2.9502*GI+6.9069*A-0.1186*NR*	0.2925 / < 2.2e-16	**$92,50**	**$92,50**
Bº Palos de Moguer (c/ de Méndez Álvaro)	*df_alvarito*	**114**	Bedrooms + Guests.Included + Accommodates + Number.of.Reviews	*Y=35.8010+10.6005*Br-8.5652*GI+7.8809*A-0.2023*NR*	0.3888 / 4.947e-11	**$95,75**	**$95,75**

**En regresiones lineales múltiples, suele ser difícil conseguir un coeficiente de determinación múltiple mayor de un 30%.*

	NOMBRE / CANTIDAD DE LISTINGS		REGRESIÓN LINEAL - SIN OUTLIERS				
	Limpio / Sin outliers		Predictores aplicados	Fórmula $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_kX_{ki}+\epsilon_i$	Multiple R^2 - p-value	Precio apartamento	
Selección	Nombre	Total				Host	Superhost
Dataset	*airbnb-listings_clean_R.csv*	13264	-	-	-	-	
Madrid	*madrid_s*	5229	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=49.34626+7.18268*Br+1.77273*B-1.73606Gl+4.91038*A+8.93240*F-0.23564*NR*	0.2012 < 2.2e-16	$94,98	$103,92
Host	*host_s*	4497	Bedrooms + Beds + Guests.Included + Accommodates + Number.of.Reviews	*Y=48.66326+7.25955*Br+2.06097*B-1.11485*Gl+4.85718*A-0.29265*NR*	0.2058 < 2.2e-16	$95,33	
Superhost	*super_s*	644	Bedrooms + Guests.Included + Accommodates	*Y=57.976+8.573*Br+-4.891*Gl+4.984*A*	0.1562 < 2.2e-16		$100,14
D° Centro (c/ de Amparo)	*amparo_s*	**3495**	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=45.2834+9.4199*Br+1.4329*B+2.2162*Gl+4.1568*A+7.0679*F-0.1714*NR*	**0.305** < 2.2e-16	$96,15	$101,21
B° Embajadores (c/ de Amparo)	*amparito_s*	**922**	Bedrooms + Beds + Guests.Included + Accommodates + Features + Number.of.Reviews	*Y=38.4211+9.0340*Br+2.5862*B+1.2239*Gl+3.6121*A+9.7052*F-0.1477*NR*	0.3544 < 2.2e-16	**$84,56**	**$94,26**
D° Arganzuela (c/ de Méndez Álvaro)	*alvaro_s*	219	Bedrooms + Beds + Features + Number.of.Reviews	*Y=46.1461+4.1607*Br+4.6348*B+6.3421*F-0.1417*NR*	0.1387 1.666e-05	**$63,74**	$70,08
B° Palos de Moguer (c/ de Méndez Álvaro)	*alvarito_s*	**86**	Bedrooms + Beds+ Guests.Included + Accommodates	*Y=37.900+7.713*Br+5.142*B-7.518*Gl+4.947*A*	0.3454 5.275e-07	**$85,77**	**$85,77**

**En regresiones lineales múltiples, suele ser difícil conseguir un coeficiente de determinación múltiple mayor de un 30%.*

7. RESULTADOS

	REGRESIÓN LINEAL		REGRESIÓN LINEAL - SIN OUTLIERS		TABLEAU	
	Precio apartamento		Precio apartamento		Precio apartamento	
Selección	Host	Superhost	Host	Superhost	Host	Superhost
Bº Embajadores (c/ de Amparo)	$98,82	$109,55	**$84,56**	**$94,26**	**$80**	**$95,00**
Bº Palos de Moguer (c/ de Méndez Álvaro)	$95,75	**$95,75**	$85,77	**$85,77**	**$74,50**	**$95,00**

REVIEWS:

- Barrio Embajadores:

$$Y=29.1843+14.1807*BR+2.5654*B+6.0246*A+10.7265*F-0.1191*NR$$

- Barrio Palos de Moguer:

$$Y=35.8010+10.6005*BR-8.5652*GI+7.8809*A-0.2023*NR$$

8. CONCLUSIONES

Precios base

- Barrio Embajadores:
  - Host: **$80,00 - 84,16**
  - Superhost: **$94,26 - 95,00**

- Barrio Palos de Moguer:
  - Host: **$74,90 - 85,77**
  - Superhost: **$95,00 - $95,75**

(sin regresión lineal)

- El precio es similar en ambos barrios.
- La regresión lineal de ambos y Tableau indica:

(-) precio~nº de reviews

- Ser superhost implica

(+) ~ **$15,00**

¿QUÉ TE HA APORTADO EL DESARROLLAR ESTE PROYECTO?

- Satisfacción de superar retos y ser resolutivas
- Consolidación de conocimientos mediante práctica
- Trabajo en equipo implementando SCRUM
- Hacernos amigas de la frustración,

¿QUÉ HAS APRENDIDO?

- Saber cómo y en qué momento aplicar las herramientas
- A trabajar en un proyecto pseudorreal
- A conocer los límites personales
- A implementar nuevas funciones de herramientas

¿QUÉ ES LO QUE NO VOLVERÍAS A HACER DE LA MISMA MANERA?

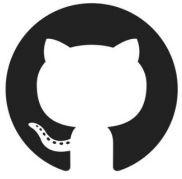
- Big Data es aprendizaje diario, tenemos mucho que descubrir y aprender
  HARÍAMOS IGUAL PERO CON MÁS TIEMPO

¿QUÉ COSAS SEGUIRÍAS HACIENDO EN EL FUTURO PARA MEJORAR EL PROYECTO?

- Continuar con el modelado realizando regresiones logísticas.
- Conseguir más métricas continuas o métricas temporales para mejorar los resultados

¡SEGUIR **TRABAJANDO** Y **MEJORANDO** NUESTRAS **SKILLS**!

Enlaces de interés:



Fórmula

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_k X$$

DATA WOMEN'S COMPANY

WHERE DATA MAKES SENSE

Nos gustaría dar las gracias a todo el equipo de KeepCoding, Glovo y a todas las compañeras que nos han acompañado en este increíble viaje.

CON CARIÑO

DATA WOMEN'S COMPANY

WHERE DATA MAKES SENSE