Proyecto vivienda

Gonzalo Terry

2022-12-21

Primera fase: Planificación

La misión de este proyecto es estudiar la relación que hay entre el precio medio de la vivienda y la renta per cápita en las diferentes provinvias españolas, así como estudiar el tipo de vivienda más frecuente en cada una de las distintas provincias. También se pretende estudiar la relación entre el tipo de vivienda y el precio de la misma.

Limpieza de Datos con SQL y Google Sheets

Al empezar este proyecto se cogió un conjunto de datos públicos de Kaggle, sobre las carácterísticas de la vivienda ofertada en las distintas provincias de España.

La limpieza consisitió en las siguientes fases:

- Eliminación de las viviendas repetidas.
- Eliminación para el análisis del precio de la vivienda de las viviendas con precio erróneo (como por ejemplo un chalet que costaba 15€).Esto se realizó creando una tabla con precios de vivienda menores al 10% del de todas las viviendas y viendo tambien el 10% mas caro, para encontrar posibles anomalías en los precios.
- Añadir una columna de renta per cápita por provincia actualizada a la tabla original. Esta tabla se obtuvo de Wikipedia.
- Cambiar formatos, ya que todo estaba en "string", hubo que cambiar las variables numéricas a "float" o "int".
- Eliminación de columnas que no nos eran útiles como cuantos baños tenía cada vivienda o la fecha de construcción.
- Redducción de la columna "tipo de vivienda", ya que había distintas etiquetas que hacían referencia la mismo tipo de vivienda. Reduje las etiquetas la mitad.

Análisis de Datos con R

Comenzando instalando los paquetes que se van a utilizar, en este proyecto se va a utilizar el paquete "tidyverse".

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Se procede también a leer el archivo ".csv" de los datos

```
ventas <- read_csv("spain_v_ventas_america.csv")
```

```
## Rows: 90302 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): population_prov, energetic_certif, house_type, loc_zone
## dbl (12): house_id, air_conditioner, bath_num, chimeney, construct_date, gar...
## num  (1): renta_media_prov
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

A continuación se va a estudiar la relación entre varias variables de un conjunto de datos público sobre la venta de vivienda en España

Relación entre el precio de la vivienda por comunidades y la renta per cápita

Primero de todo se va a crear una nueva tabla con los datos que nos interesan, que en este caso son las provincias, su precio medio de vivienda, y su renta per cápita.

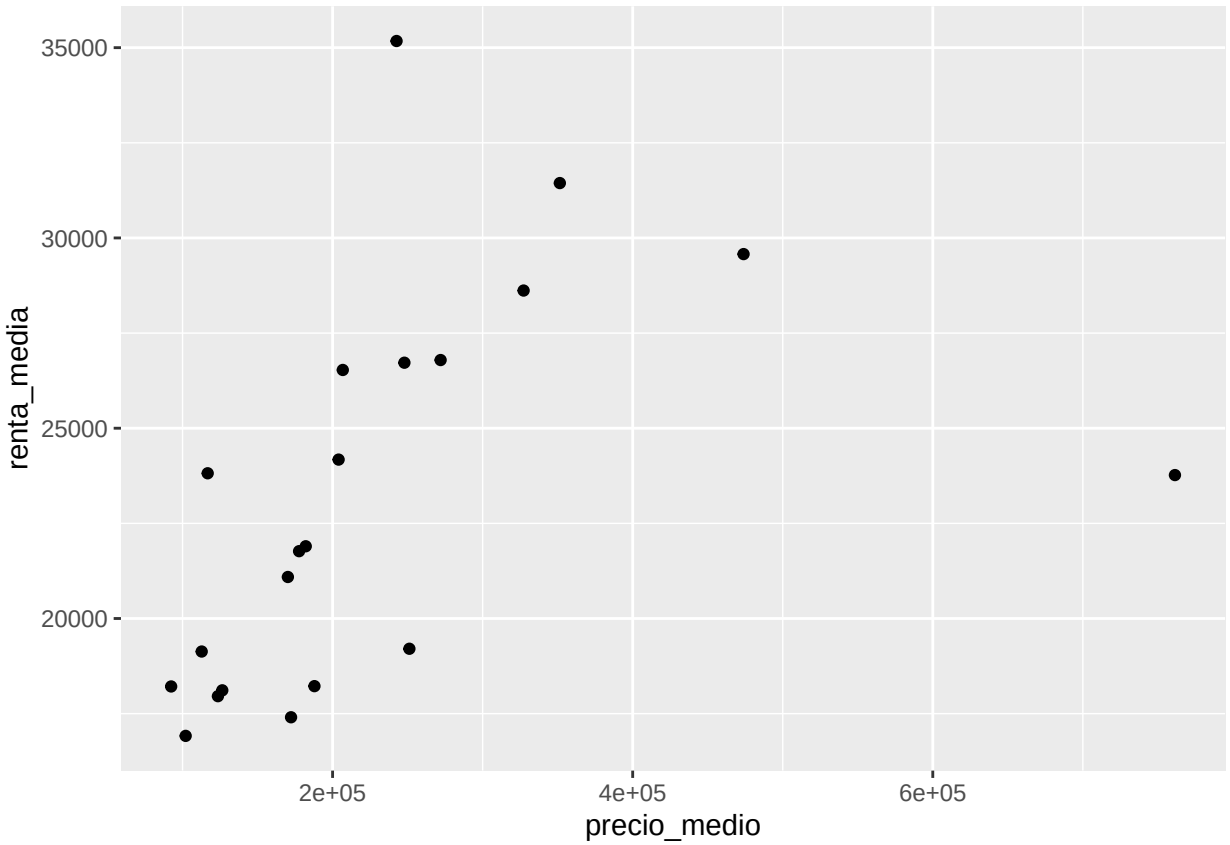
```
tabla1 <- ventas %>%
  group_by(loc_zone) %>%
  summarize(renta_media = mean(renta_media_prov), precio_medio = mean(price))

tabla1
```

```
## # A tibble: 21 x 3
##    loc_zone    renta_media precio_medio
##    <chr>             <dbl>        <dbl>
##  1 A Coruña          21898      182104.
##  2 Álava             35175      242635.
##  3 Albacete          18113      126484.
##  4 Alicante          17405      172310.
##  5 Barcelona         26531      206795.
##  6 Cádiz             16916      102070.
##  7 Ciudad Real       18214       92415.
##  8 Girona            26722      247847.
##  9 Guipúzcoa         31442      351448.
## 10 Huelva            17959      123584.
## # ... with 11 more rows
```

A continuación vamos a realizar la representación gráfica. Vamos a realizar un gráfico de dispersión colocando en el eje x al precio medio de la vivienda por provincia, y en el eje y la renta per cápita de la misma provincia.

```
ggplot(data = tabla1, aes(x = precio_medio, y = renta_media)) + geom_point()
```



A simple vista podemos intuir que los puntos siguen una posible tendencia lineal, vamos a calcular la correlación estadística.

```
tabla1 %>%
  summarize(correlacion = cor(precio_medio, renta_media))
```

```
## # A tibble: 1 x 1
##   correlacion
##         <dbl>
## 1       0.465
```

Obtenemos una correlación de 0.465, no es extremadamente fuerte, pero en la tabla vemos que hay un precio medio de vivienda que se aleja de la tendencia, este es el punto del precio medio de vivienda en las Islas Baleares, ya que hay en proporción mas casa en venta de alta gama y lujosas comparado con el resto de las provincias.

Vamos a quitar este punto y realizar los mismo pasos que antes.

Empezamos por definir una nueva tabla, quitando la fila de las Islas Baleares.

```
tabla2 <- tabla1[-11,]
tabla2
```

```
## # A tibble: 20 x 3
##    loc_zone               renta_media precio_medio
##    <chr>                        <dbl>        <dbl>
##  1 A Coruña                     21898      182104.
##  2 Álava                        35175      242635.
##  3 Albacete                     18113      126484.
##  4 Alicante                     17405      172310.
```

```
##  5 Barcelona                   26531      206795.
##  6 Cádiz                       16916      102070.
##  7 Ciudad Real                 18214       92415.
##  8 Girona                      26722      247847.
##  9 Guipúzcoa                   31442      351448.
## 10 Huelva                      17959      123584.
## 11 Madrid                      29576      473896.
## 12 Santa Cruz de Tenerife      19205      251180.
## 13 Segovia                     21769      177672.
## 14 Sevilla                     18223      187903.
## 15 Soria                       23816      116754.
## 16 Tarragona                   26792      271986.
## 17 Valencia                    21091      170199.
## 18 Valladolid                  24176      204000.
## 19 Vizcaya                     28618      327294.
## 20 Zamora                      19132      112762.
```

Ahora calculamos de nuevo la correlacion.

```
tabla2 %>%
  summarize(correlacion = cor(precio_medio, renta_media))
```

```
## # A tibble: 1 x 1
##   correlacion
##         <dbl>
## 1       0.726
```

Ahora se obtiene una correlación estadística mucho mas alta, de aproximadamente 0.726.

Con estos resultamos no podemos afirmar que haya causalidad entre la renta media de una provincia y el precio medio de la vivienda en dicha provincia.

Lo que podemos afirmar con esta moderada correlación es que si que hay una tendencia entre estas dos variables, por lo menos, en el mercado español, ya que esta tendencia podría variar dependiendo del modelo económico de cada país, a una corralación más floja o más fuerte.

Relación entre la provincia y le tipo de vivienda

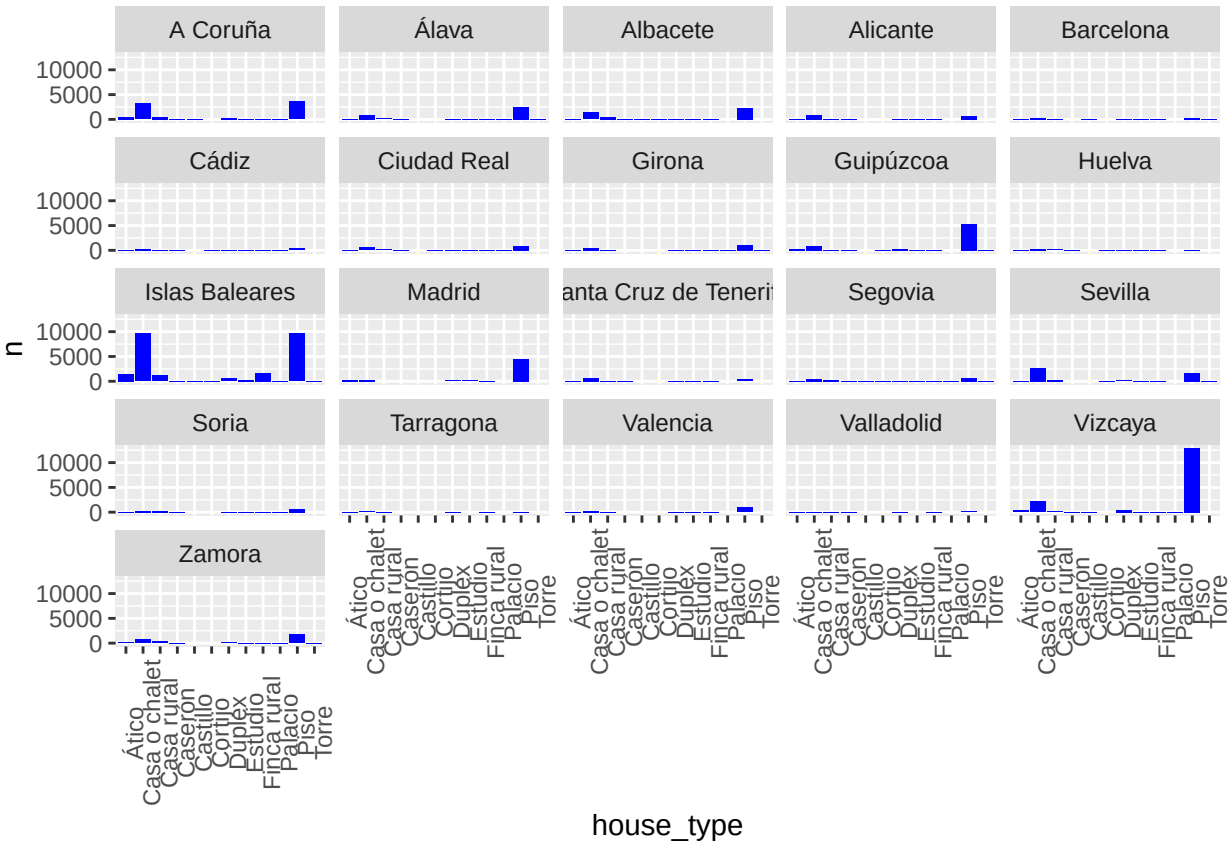
Ahora podemos analizar la tendencia entre las distintas provincias de España y que tipo de vivienda se prefiere en cada una de las provincias. Con esto podríamos intuir cuan interesante es construir un tipo de vivienda en cada una de las provincias.

Vamos a comenzar por hacer una tabla auxiliar con los datos que nos interesan para abordar este problema, que es el número de cada tipo de vivienda que hay en cada provincia.

```
tabla2 <- ventas %>%
  select(house_type, loc_zone) %>%
  group_by(loc_zone)%>%
  count(house_type)
```

Ahora vamos a representar estos datos en un gráfico del tipo “ggplot2”.

```
ggplot(data=tabla2, aes(y = n, x=house_type)) + geom_bar(stat = 'identity', fill = 'blue') + facet_wrap
```



Donde podemos ver que Baleares afecta a la correlación entre la renta per cápita y el precio medio de la vivienda, ya que en este set de datos la mayor parte de vivienda ofertada en Baleares es vivienda de lujo.

En el resto de provincias las viviendas más ofertadas son casas o pisos, aunque en las provivncias más rurales también hay significativa presencia de casas de tipo rústico.

Fase de Visualización

La visualización de los datos se ha realizado con la herramineta Power BI

.

Conclusión

La conclusión a la que podemos llegar es que si hay una correlación entre la renta per cápita y el precio de la vivienda en cada una de las regiones, con la excepción de las Islas Balerares, donde la vivienda ofertada es mayormente de lujo comparado con el resto de las provincias.

El mayor tipo de vivienda ofertada es “Piso” (50,5%), seguido de “Casa o Chalet” (26,6%), y seguido de “Casa Rural” (4,3%).

Las tres provincias con el precio más alto de vivienda son: Islas Baleares, Madrid y Gipúzcua. En ese orden.

Y los tres tipos de vivienda más caros son: Castillos, Palacios y Fincas Rurales. En ese orden.

Advertencia

Esto es un estudio con carácter académico con un conjunto de datos en el que no aparecen representadas todas las provincias españolas y que puede no ser correcto o no representar fielmente la tendencia de la vivienda en España.