**VARIABLES BASE DE DATOS**

* **ID:** número asociado a cada casa (sobra seguramente)
* **Listing URL:** link (sobra)
* **Scrape Id:** otro número asociado (también sobra?)
* **Last Scrapped:** qué significa, última vez que se ha buscado?
* **Name:** nombre del anuncio (sobra)
* **Summary:** cada resumen es diferente (sobra)
* **Space:** igual (sobra)
* **Description:** igual (sobra)
* **Experiences Offered: REVISAR (pone todo “none”)**
* **Neighborhood Overview:** párrafo largo (sobra)
* **Notes:** párrafo largo (sobra)
* **Transit:** párrafo largo (sobra)
* **Access:** párrafo largo (sobra)
* **Interaction:** párrafo largo (sobra)
* **House Rules:** párrafo largo (sobra)
* **Thumbnail URL:** link (sobra)
* **Medium URL:** link (sobra)
* **Piccture URL:** link (sobra)
* **XL Picture URL:** link (sobra)
* **Host ID:** numero único (sobra)
* **Host URL:** link (sobra)
* **Host Name:** único, aunque haya varios apartamentos con el mismo dueño no veo sentido a quedarnos con ella.
* **Host Since:** fecha desde la cual es host, seguro que influye algo.
* **Host Location:** donde vive el host, en algunas aparece texto largo, por lo que ya no podemos utilizarla
* **Host About:** párrafo largo (sobra)
* **Host Response Time:** pone el tiempo en plan texto (por ej “within an hour”). Mirar si se puede transformar bien. Aun así, no creo que tenga influencia sobre el precio de la casa. (Probablemente sobra)
* **Host Response Rate:** porcentaje de respuestas del host, numérica por lo que a priori la podemos utilizar, aunque no se si tendrá que ver con el precio final de la casa
* **Host Acceptance Rate:** puede ser el porcentaje de solicitudes que acepta? Puede ser que cuanto más cara menos acepte. De todas formas, la muestra que he cogido de las primeras son todo datos missing, pendiente de revisar el resto para confirmar si se puede hacer algo con ella.
* **Host Thumbnail URL:** link(sobra)
* **Host Picture URL:** link(sobra)
* **Host Neighbourhood:** el barrio donde vive el casero. No creo que influya. A lo mejor si tiene casa mas caras vivirá en un barrio más caro. Revisar.
* **Host Listings Count:** que es? No creo que sea el número de casas xq tienen valores de 162, 207, etc. Puede ser el numer total de huéspedes acogidos. En este caso también puede ser útil.
* **Host Total Listings Count:** los valores de esta variable son exactamente los mismos que los de arriba (por lo menos los de la muestra que he cogido). Confirmarlo, Revisar.
* **Host Verifications:** métodos de verificación de identidad del host. A mas números de métodos cuadra que sea mayor precio. De todas formas hay que transformarla, está en modo lista (por ej “email,phone,reviews,kba”)
* **Street: texto (sobra)**
* **Neighbourhood:**
* **Neighbourhood Cleansed:** la primera observación es un numero y la segunda un barrio, hay algo raro.
* **Neighbourhood Group Cleansed:**
* **City:** ciudad de la casa, convertirla en categorías.
* **State:** lo de estado tiene sentido para EEUU, pero por ejemplo para España no. Hay un caso en el que pone Barcelona en ciudad y Barcelona en estado, y la siguiente pone Barcelona en ciudad y Cataluña en estado. Si no tiene siempre el mismo criterio la quitamos y nos dejamos de problemas. Probablemente este correlacionada con ciudad
* **Zipcode:** estará correlacionada con ciudad, pero puede que aporte más información. Convertirla en categorías
* **Market:** sale ciudad, igual que city, quitarla.
* **Smart Location:** igual que la de arriba, quitarla
* **Country Code:** nos la quedamos, convertirla en categorías (country code es “ES”, “US”)
* **Country:** nos vale con el country code, quitarla.
* **Latitude:** ya tenemos suficiente info de localización, quitarla
* **Longitude:** igual.
* **Property Type:** nos sirve (loft, apartamento, villa, etc). Convertirla en categorías con numeros en función del tipo
* **Room Type:** nos sirve igual que la de arriba.
* **Accomodates:** numero de gente que cabe. Nos la quedamos, muy importante a la hora de establecer el precio. REVISAR CON GUESTS INCLUDED
* **Bathrooms:** nos la quedamos obvio
* **Bedrooms:** igual
* **Beds:** igual
* **Bed type:** tipos de cama, habrá que convertirla en categorías, nos la quedamos. No se si hay alguna observación donde aparecen varias tipos dentro de una misma casilla (xq hay varias camas en la casa), eso nos puede complicar
* **Amenities:** servicios que tiene la casa: tele, internet, etc. Aporta información relevante al precio, revisar como transformarla xq no lo veo fácil.
* **Square feet:** metros cuadrados, nos la quedamos
* **Price:** VARIABLE OBJETIVO
* **Weekly Price:** Revisar que hacemos con esta y la siguiente. Va a ser parecido a la variable objetivo, no la podemos utilizar como dato de entrada xq va a acaparar la estimación xq estará muy correlacionada. Igual con la siguiente.
* **Monthly Price:** igual
* **Security Deposit:** Numérica fácil de meter.
* **Cleaning fee:** Numerica fácil de meter
* **Guests included:** parece lo mismo que otra variable. REVISAR CON ACCOMODATES
* Extra people: valores de 0,44,39, … no se que siginifica
* **Minimum nights:** numérica y fácil de meter.
* **Maximum nights:** igual
* **Calendar updated:** fecha en la que se ha actualizado el calendario. Se puede convertir en numérica, no se hasta que punto será útil.
* **Has availability:** entiendo que será YES o NO, la muestra que he cogido son todo missings. Puede que influya en el precio de la noche en el alojamineto: si no tiene disponibilidad será que esta muy demandada y más cara por ello**.**
* **Availability 30:** días disponibles de los próximos 30.
* **Availability 60:** días disponibles de los próximos 60.
* **Availability 90:** días disponibles de los próximos 90.
* **Availability 365:** días disponibles de los próximos 365.
* **Calendar Last Scrapped:** ni idea. Fecha, SOBRA
* **Number of reviews:** numérica fácil de usar
* **First Review:** Fecha, yo la quitaría, SOBRA
* **Last Review:** Fecha, yo la quitaría, SOBRA
* **Review Scores Rating:**
* **Review Scores Accuracy:**
* **Review Scores Cleanliness:**
* **Review Scores Checkin:**
* **Review Scores Communication:**
* **Review Scores Location:**
* **Review Scores Value:**
* **License:** fuera, numero de licencia
* **Jurisdiction Names:** fuera
* **Cancellation policy:** categorizarla, se puede utilizar.
* **Calculated Host Listings count:** numérica, me lio con otras que hablan del numero de hosts y no se que significa exactamente.
* **Reviews per month:** numérica, fácil de usar, no se si influye en el precio.
* **Geolocation:** fuera, ya tenemos suficiente info de la localización
* **Features:** ejemplo: “Host Is Superhost,Host Has Profile Pic,Host Identity Verified,Is Location Exact”. Contiene información importante, ver como transformarla.