AirBnb Data enunciado práctica Análisis de Datos 21_22. GMAT2

Grupo y nombre de cada práctica

2021-2022

Contenidos

Título:

Autores:

1. Apellidos, Nombre
2. Apellidos, Nombre
3. Apellidos, Nombre
4. Apellidos, Nombre
5. Apellidos, Nombre

1 Taller evaluable en grupos datos AirBnb

- Aquí tenéis el enlace a estos datos de AirBnb
- Generad un proyecto nuevo.
- Bajad lo datos de AirBnb a un carpeta/directorio que se llame `AirBnb` y dentro de `AirBnb` crear una carpeta/directorio que se llame `data`.
- Podéis (tenéis) que utilizar las ayudas del taller de estos datos.

1.1 Instrucciones

- Entregad en grupos de prácticas.
- Se puede hacer con R o Python.
- Hay que entregar el Rmd/Notebook junto con su salida en html/pdf
- Máxima longitud: equivalente a 10 páginas en pdf.
- Hay que cuidar la presentación, ortografía y redacción.
- Fecha límite entrega y lugar de entrega consultad el espacio moodle de la asignatura.

1.2 Contexto de los datos

La página Inside Airbnb http://insideairbnb.com/ contiene información sobre los datos de los apartamentos o residencias vacacionales en alquiler en diversas localizaciones del mundo.

Los datos recogidos están repartidos por diversas regiones, provincias, departamentos, condados... del mundo. Los datos son Open Source y los podemos usar (ver licencia About Inside Airbnb http://insideairbnb.com/about.html.)

En resumen el acceso y los diccionarios de datos y otras utilidades son accesibles desde la página principal o en los siguientes enlaces:

Data Resources

- Get the data
- View Data Dictionary
- Read Data Policies including aligning data availability to the mission, Community Guidelines and policies on Archived and New Data
- Make a Data Request for Archived Data or Data for a new región

¡¡Atención!! El último servicio es de pago para datos de más de un año de antigüedad.

1.2.1 Acceso a los datos

En el enlace Get the data podemos descargar para cada ciudad los ficheros siguientes:

File Name	Description
listings.csv.gz	Detailed Listings data for Name City.
calendar.csv.gz	Detailed Calendar Data for listings in Name City.
reviews.csv.gz	Detailed Review Data for listings in Name City.
listings.csv	Summary information and metrics for listings Name City (good for visualisations).
reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing) N/A Name City.
neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files N/A Name City.
neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

Definid una carpeta `data` y dentro una carpeta por zona `Mallorca`, `Valencia`,`Barcelona`, etc. Bajad los datos, podéis utilizar el el código del script `download_city_inside_airbnb.R` que se encuentra la raíz del github de la práctica.

1.2.2 Especificación de las tablas de datos

Para entender cada tabla de datos tenemos que acceder al Diccionario de datos.

Tenemos que comprender qué variables vamos a cargar y el tipo de datos. Como hay datos de todo tipo tenemos que ir con especial atención:

- A los id de enteros largos que se puedan confundir con variables numéricas: Hay que leerlos como cadenas de caracteres.
- A las variables numéricas que puedan contener caracteres especiales: símbolo de $, símbolo de €, el símbolo de %, separadores de miles...
- Variables que sean listas; por ejemplos extras de la vivienda: wifi,TV, piscina,... .
- Otros tipos especiales de variables: latitud, longitud, texto, etc.

Como el problema es de datos sin una estructura clara cada grupo tendrá que estudiar las zonas:

- Mallorca
- Valencia
- Barcelona

- Varias ciudad más hasta completar (junto con las tres anteriores) el número de miembros del grupo.

1.2.3 Bibliografía y software adicional

- Gráficos dinámicos con plotly: https://plotly.com/r/animations/
- MAPAS de España: https://www.cienciadedatos.net/documentos/58_mapas_con_r.html fijos

1.3 Cuestión 1: Contexto del problema y modelo de datos (25%)

1. Cargar fichero `listing.csv`, `calendar.csv` y `reviews.csv` de cada ciudad. Tenéis que estudiarlas y decidid qué tipo de dato y qué variables cargáis. Hay que el explicar las transformaciones que realicéis para manipular los datos; por ejemplo 50$ lo transformo a 50, "2020-01-30" lo leo en tipo `date`... .
2. Definid un **modelo de datos** con todas las tablas. Por ejemplo unid todos los listings de vuestras ciudades en una sola tabla, añadiendo una variable que especifique la zona: Mallorca, Valencia, Barcelona, CiudadX, CiudadY....
3. Guardar el modelo de datos en ficheros `.csv` o `Robj` para la segunda parte de la práctica.
4. Redactar un informe explicando los tres apartados anteriores.

1.4 Cuestión 2: Análisis exploratorio de datos (EDA). (50%)

En las siguientes preguntas aplica todo lo que hemos visto acerca de la documentación en el EDA: Título de gráficos, etiquetas de los ejes, coloreado con información, leyendas, tablas bien presentadas (knitr)...

1. Calcula la frecuencia del número de reviews por apartamento. Es decir cuántos apartamentos tienen 1 review, 2 reviews, 3 reviews y así sucesivamente. ¿Sigue la frecuencia del número de reviews por apartamentos vacacionales y el rango de reviews una "power law" (relación potencial)?
2. Del número de reviews por zona, barrio, día de las semana y por meses realizar un amplio análisis descriptivo y gráfico. Se valorará positivamente la correcta elección de estadísticos y gráficos correspondientes.
3. De cada ciudad selecciona los 5 zonas/barrios con más apartamentos vacacionales. De estas zonas y para cada ciudad compara los precios medios (de todo el periodo), el número de habitaciones y el número de camas para cada apartamento.
4. De cada ciudad calcula y compara (analítica y gráficamente) la serie temporal de los precios medios, máximos y mínimos por día, semana y mes (de todo el periodo).
5. De cada ciudad representa gráficamente un mapa donde se aprecien el número de apartamentos y su precio medio durante el periodo de navidad (23 de diciembre al 2 de enero), por latitud y longitud o agrupando por cuadrados. Selecciona la representación gráfica más informativa.

1.5 Cuestión 3: Presentación final. (25%)

Presentación final:

- Presentación de 15 minutos tipos transparencias: ioslide, powerpoint... (recordar que las ioslides con Rmd tienen salida a pdf, html y ppt).
- Todos los miembros del grupo deben presentar y se sorteará el orden de presentación.
- Se podrán dos notas una a la presentación global y otra a la individual.
- Cuestión extra: panel interactivo +20%

Podréis conseguir puntos adicionales haciendo un panel interactivo con alguna de las librerías o programas vistos en el la asignatura: shiny, graphana, Power BI, tableau.