Proyecto asignatura Estadística (MAT3) GIN2 Grupos 1 y 3: PARTE 1 y PARTE 2. CUrso 2020-2021

Poned los nombres de los autores

Contenidos

1 Parte 1: Estadística Descriptiva (1 punto sobre 10 de la nota final. Entregad por Aula Digital antes del 21 de abril) **1**

2 Parte 2: Estadística Inferencial (1 punto sobre 10 de la nota final. Entregad por Aula Digital antes del 15 de junio) **2**

1 Parte 1: Estadística Descriptiva (1 punto sobre 10 de la nota final. Entregad por Aula Digital antes del 21 de abril)

La proyecto se entrega en **grupos de 4 persones**. Cada grupo tendrá asignado un **nombre de ciudad** que podréis consultar en la página de la asignatura en Aula Digital. La práctica consiste en obtener los datos de la ciudad asignada, correspondientes a **febrero de 2020**, de la web de Airbnb y redactar un informe utilizando Rmarkdown respondiendo a las siguientes cuestiones

1. Cargad en un `dataframe` los datos del fichero `listings.csv` (descomprimido a partir de `listings.csv.gz`). Esta tabla de datos contiene datos sobre muchas variables:
   - `neighbourhood property_type`
   - `accommodates`
   - `beds`
   - `bedrooms`
   - `bathrooms`
   - `price`
   - `security_deposit`
   - `minimum_nights`
   - `review_scores_rating`
   - `number_of_reviews`
   - `host_response_time`
   - `requires_license`
   - `review_scores_cleanliness`
   - `cleaning_fee`
   - `etc.`

   Construid un nuevo `data frame` que contenga solo la información relativa a como mínimo 3 variables no numéricas y 7 variables numéricas. Hay que **escoger forma obligatoria las variables** `neighbourhood` y `price`. Las variables relativas a precioso (por ejemplo, `price` y `security_deposit`) se consideran numéricas, pero primero deben convertirse a valores numéricos eliminando los símbolos `$` o `€`; por ejemplo utilizando la función `gsub(pattern="[\\$]|[,]",replacement="",x="$10,000")`; haced (`help(gsub)`)
2. Renombrar las columnas del `dataframe` con nombres en castellano o catalán.

3. Para las variables numéricas, calcular los siguientes estadísticos y mostrar en una tabla los siguiente valores: cantidad de valores no válidos, mínimo, máximo, media, varianza, cuartiles y mediana.
4. Para las variables no numéricas, generad las tablas de frecuencias absolutas de cada uno de sus valores.
5. Dibujar diagramas de cajas(**boxplots**) e histogramas de todas las variables numéricas, mostrando un mínimo de 2 **boxplot** por fila. Ajustad la altura de los gráficos para que no queden demasiado pequeños.
6. Dibuja un diagrama de tarta para una de las variables no numéricas agrupando como "Otros" los valores que representen un porcentaje inferior al 1% del total.
7. Calcular el valor medio de alguna de las variables numéricas según el barrio, de menor a mayor y sin tener en cuenta nombres de Barrio incorrectas (`""` o `"N/A"`).
8. Dibuja un **boxplot** de la variable precio, para los 5 barrios más caros (precio medio más alto) y los 5 barrios más baratos (precio medio más bajo), en un mismo diagrama. Los **boxplots** deben indicar también el valor medio de los datos.
9. Para 4 de las variables numéricas dibujar los diagramas de dispersión dos a dos, con colores diferentes para cada barrio ("neighbourhood").
10. Por las mismas variables elegidas en el apartado anterior calcular los coeficientes de correlación dos a dos de las variables, sin tener en cuenta los valores `NA`.
11. Selecciona dos variables numéricas, y para cada una de ellas organiza sus valores en un máximo de 5 intervalos con la función `cut`.
12. A partir de los datos en intervalos obtenidos en el apartado anterior construid una tabla de contingencia de las dos variables y dibuja el diagrama de mosaico asociado a la mesa.

Se valorará la claridad y los comentarios de los resultados obtenidos. Si se detectan **trabajos copiados** quedarán **suspendidos todos los alumnos implicados**. Todas las preguntas se pueden contestar a partir de la información de los manuales de AprendeR1, AprendeR2 y concretamente AprendeR2 tema 8, a parte de los enlaces proporcionados en este documento, y, en general, haciendo búsquedas en internet. También puede basar su informe en los ejemplos publicados en la web de la asignatura, Además, puede consultar dudas con el resto de los compañeros de curso a través del Foro de la asignatura a Aula Digital o en discord.

El documento en formato .Rmd y el informe en .html o .pdf se debe en la actividad correspondiente del **Aula Digital antes del 21 de abril**.

2 Parte 2: Estadística Inferencial (1 punto sobre 10 de la nota final. Entregad por Aula Digital antes del 15 de junio)

Supongamos que los datos de la ciudad que ha sido asignada a cada grupo corresponden a una muestra aleatoria simple de todas las viviendas que se podrían alquilar en la ciudad. Utilizando esta muestra se pide:

1. Calcular una estimación puntual de la media para la variable `price` y el error estándar del estimador.
2. Calcular un intervalo de confianza, al nivel de confianza del 95%, para la variable `price`.
3. Calcular un intervalo de confianza, al nivel de confianza del 99%, para la proporción de viviendas que tienen un `review_scores_rating` inferior a 95%.
4. Supongamos que un responsable de Airbnb asegura que la media de los valores de `review_scores_rating` de las viviendas de su portal es superior a 95. Contrastad esta hipótesis.
5. Calcular el intervalo de confianza, con un nivel de confianza del 95%, asociado al contraste del ejercicio anterior.

6. Considera ahora los datos de `price` para la ciudad de New York del mes de febrero de 2020 (están en http://insideairbnb.com/get-the-data.html, y debe pulsar en 'show archived fecha). Compararemos los valores de esta variable con los correspondientes a la ciudad que tiene asignada. Haga un contraste de hipótesis para decidir si las desviaciones típicas de los precios de las dos ciudades son iguales o diferentes. Considera que las distribuciones de los valores de precio en las poblaciones son normales.

7. A partir de los resultados del apartado anterior contratad la hipótesis de que los precios medios en las dos ciudades son iguales.

8. Utilice el test de Kolmogorov-Smirnov-Lilliefors para confirmar o rechazar la hipótesis de que la distribución de los valores de la variable `price` es normal, decidid el resultado del contraste con el p-valor.

9. Analizad la dependencia entre las variables `Price` y `review_scores_rating` de la ciudad que tiene asignada. Seguid los siguientes pasos:

    - a) Seleccione del data frame las muestras que tienen valores diferentes de NA por las dos variables.
    - b) A continuación agrupau los valores de cada variable utilizando los intervalos siguientes: $[\min, Q_1), [Q_1, Q_2), [Q_2, Q_3)$ y $[Q_3, \max]$. Los valores min y max son el mínimo y el máximo de la variable, respectivamente. Mientras que Q_1, Q_2 y Q_3 · representan los cuartiles primero, segundo (mediana) y tercero. Si los valor mínimo y máximo de algún intervalo son iguales elimine este intervalo.
    - c) Organizad los datos agrupados en intervalos en una tabla de contingencia `Price` versus `review_scores_rating` .
    - d) A partir de esta tabla haced un test χ^2 de independencia para determinar si las dos variables son independientes, con un nivel de significación del 0.05.

Comentarios:

- Para hacer los cálculos solicitados en los apartados anterior se deben eliminar los valores no disponibles (`NA`) de las variables.
- Siempre que sea posible se deben utilizar las funciones de R explicadas en clase para resolver los ejercicios.
- Debe redactar un documento utilizando Rmarkdown con las respuestas a estas preguntas y que incluya el código R utilizado. También debe generar (Knit) una versión HTML del documento.

El documento, en formato .Rmd y .html o .pdf , se debe **entregar a Aula Digital antes del 15 de junio**.