Tarea 2 Curso Análisis y Limpieza de datos.

Posgrado en Data Science UDD

Profesor responsable: Jorge Alexis Castillo Sepúlveda

El objetivo de esta tarea es evaluar el conocimiento práctico en lo que respecta al análisis descriptivo de datos. Se espera además que los conocimientos sean acumulativos, así que también está considerado evaluar destrezas respecto a la carga y limpieza de datos, además de leer archivos de distintos tipos y fuentes, para armar un solo archivo de análisis. Se también se evaluará el buen uso de markdown para la generación del ipynb, por lo cual se espera que el estudiante presente un trabajo bien redactado y con buena estética ofimática.

Instrucción principal: Generar un archivo .ipynb donde redacte de manera clara y ordenada, todos los desafíos que están enumerados en esta tarea.

Para esta tarea, usted dispone de los datos de nacimientos en Chile desde 1990 a 2017 en esta dirección: https://bit.ly/35N4GD6

Estos datos fueron obtenidos por su profesor hace mucho tiempo, pero que, por temas de contingencia, no es posible obtenerlos de nuevo (la página del DEIS de origen de esos datos ya no es pública), por lo que no se cuenta con la documentación completa. Sin embargo, los nombres de las columnas se pueden deducir por lo general, por ejemplo, día_nac refiere al día en que nació, y mes_nac al mes correspondiente, edad_p corresponde a la edad del padre, y edad_m a la edad de la madre, etc.

Los datos entre 1990 y 1995 están guardados en formato mdb, que refiere a Microsoft Access Database, que era bien usado por los estadísticos de esa época para gestionar bases de datos. Se pueden guardar en formato Excel (xls, xlsx) su usted los abre y los guarda en ese formato. Si no usa SO Windows, puede cargar una consola virtual. Vamos a suponer que esta base de datos de 28 años cerrados es una muestra suficiente para hacer inferencias sobre los nacimientos de la población en Chile.

1. Juntar todos los archivos de todos los años en un solo dataframe global para efectuar los análisis posteriores.
2. ¿Cuál es el mes más frecuente de nacimientos en Chile? Comentar al respecto.
3. ¿Cuál es el día del año más común en el que la gente en Chile está de cumpleaños?
4. Calcular covarianza y correlación entre peso y talla a nivel general (tomando toda la base). Luego hacerlo por año. ¿Cambia con el paso de los años?
5. Calcular covarianza y correlación entre la edad del padre y la edad de la madre, a nivel general (tomando toda la base). Luego hacerlo por año. ¿Cambia con el paso de los años?
6. Investigue las condiciones para que un bebé cuando nazca se considere “prematuro”, “a término” y “postérmino”. Hacer diagramas de caja para el peso y la talla para estas 3 categorías. Comentar al respecto.
7. Crear una columna llamada “indicador” que valga “1” si el bebé nació en una ambulancia y que valga “2” si el bebé nació en el trayecto (para los datos desde 1996). Caracterice los datos atípicos (outliers) usando el IQR y el primer y tercer cuartil para cada una de estas variables peso, talla, edad del padre y madre, en cada caso con indicador 1 o 2. Comentar al respecto.

8. Hacer un diagrama de distribución por tipo de establecimiento donde nacen los bebés desde 1996. Notar que los distintos hospitales deben agruparse en una sola categoría (lo mismo corre para las otras categorías). Comentar al respecto.
9. Suponer que los datos de nacimiento (variables continuas) provienen de una distribución normal (desde 1990). Un intervalo de confianza, al 95% de confianza, asumiendo distribución normal, se calcula como [promedio-1.96$\frac{desviación\ estándar}{\sqrt{n}}$, promedio+1.96$\frac{desviación\ estándar}{\sqrt{n}}$], donde n indica el número de datos de la muestra. Calcular intervalos de confianza para la talla y el peso de los bebés nacidos en Chile. Adicionalmente, calcular el intervalo [Q_1-1.5IQR, Q_3+1.5IQR] para ambas variables talla y peso. Comparar ambos intervalos y comentar.
10. Comentar las características de los bebés nacidos cuando la madre tiene más de 40 años. Hacer lo mismo para cuando la madre tiene menos de 18 años.

La entrega es individual. Debe ser enviada a j.castillo@udd.cl antes del jueves 1 de octubre a las 23:59 horas.