Laboratorio 1
Estadística Computacional

Universidad Técnica Federico Santa María
Departamento de Informática

José García
<jigarcia@alumnos.inf.utfsm.cl>

Sebastián Bórquez
<sborquez@alumnos.inf.utfsm.cl>

Héctor Allende
<hallende@inf.utfsm.cl>

Rodrigo Naranjo
<rodrigo.naranjo@alumnos.usm.cl>

5 de abril de 2019

Análisis Exploratorio de Datos

El objetivo de esta experiencia será realizar un análisis exploratorio de un conjunto de datos, por lo que será necesario usar herramientas de visualización incluidas en R o librerías como ggplot2, o en las librerias de ggplot y Seaborn en en caso de Python. Cada respuesta **debe** incluir un gráfico, y este debe ser claro y entendible (título, unidades, legible, etc). Recuerde también escribir las interpretaciones de cada gráfico.

1. Contexto

El dataset a trabajar se encuentra en Moodle en la sección Entregas. Este consiste de datos históricos de los juegos olímpicos. El trabajo de este laboratorio consiste en lograr extraer información a partir de los datos usando visualizaciones interesantes y simples. Las tareas más comunes son las de carácter descriptivo, por lo que será necesario usar distintas categorías y variables para explorar el dataset.

2. Preguntas

1.- Defina 3 requerimientos de tipo descriptivo y respóndalos con un gráfico.

2.- Construya un scatterplot con al menos 3 variables en consideración.

3.- Construya un gráfico de boxplot para las variables edad y peso para los países involucrados.

4.- Construya un gráfico temporal en el que se pueda observar como varia la cantidad de atletas mujeres y hombres.

5.- Construya un dataframe derivado de los datos compuestos por las siguientes variables:

- País.

- Fecha de la Olimpiada.
- Número de atletas enviados a competir.
- Número de atletas hombres.
- Número de atletas mujeres.
- Número de medallas de bronce obtenidas.
- Número de medallas de plata obtenidas.
- Número de medallas de oro obtenidas.

Luego, obtenga un correlograma de las variables numéricas del nuevo dataframe.

3. Conclusiones

Mencione las conclusiones más relevantes e interesantes que ha encontrado en el análisis. Aclaración:

- Sí es conclusión relevante: *El país con más medallas de oro en los últimos 10 años es ...*
- No es conclusión relevante: *Rstudio permite trabajar una gran cantidad de datos ...*

La conclusión también lleva puntaje, tome su tiempo para encontrar información útil.

4. Sobre el desarrollo

Las sesiones y material usados serán hechas en R y Python. El desarrollo puede ser realizado con R o Python utilizando las herramientas presentadas en las sesiones. Las herramientas para el desarrollo son R Markdown y Jupyter Notebooks, respectivamente. Para usar R se recomienda trabajar en RStudio, y para Python usar Jupyter Notebooks junto con Spyder, recomendado trabajar con Anaconda.

5. Sobre la Entrega

El informe puede realizarse en parejas o tríos. El informe **debe incluir el código** que usó en la ejecución, por lo que es necesario que use notebooks en el trabajo. Se aplicarán **descuentos** por código desordenado, ilegible o no modularizado. Se recomienda leer las siguientes convenciones de código: https://github.com/google/styleguide. La fecha de entrega es **Jueves 18 de Abril**. El archivo a subir **debe ser el notebook** con el que trabajaron con los scripts ejecutados en formato HTML (o .ipynb en caso de usar Jupyter Notebooks) con nombre "Nombre1Apellido1-Nombre2Apellido2" a la sección de entregas de Moodle. En caso de atrasos, si el atraso es de 1 día, la nota máxima será 80. 2 o más días tendrán nota 0.