Proyecto 1.
Obtención y Limpieza de los datos

INTRODUCCIÓN:

Los tipos y las fuentes de datos como se encuentran en las organizaciones son realmente diversos. El primer trabajo de un científico de datos es acceder a las fuentes y preparar los datos para que puedan ser analizados. Usualmente esto lleva un gran trabajo, pero sin hacerlo el riesgo de llegar a resultados erróneos es demasiado alto. Con la realización de este proyecto, no solo aprenderá a utilizar las herramientas que le permitan acceder a los datos, sino a limpiarlos, haciéndolo de la forma más transparente posible.

Competencias:

- Utiliza las herramientas que tiene a su disposición para obtener los datos de la fuente especificada.
- Modifica los datos que obtuvo realizando procesos de limpieza que allanen el camino del analista de datos.
- Hace el proceso de limpieza transparente y reproducible para cualquiera que lo desee verificar.
- Elabora un “Code Book” detallado que contenga todos los metadatos que sean necesarios para analizar los datos.

ACTIVIDADES

1. Descargue los datos de los establecimientos educativos de todo el país que lleguen hasta el nivel de diversificado (NIVEL ESCOLAR: DIVERSIFICADO). Los puede encontrar en el siguiente vínculo: http://www.mineduc.gob.gt/BUSCAESTABLECIMIENTO_GE/
2. Guarde los datos crudos en archivos .csv.
3. Describa el estado de los datos y las operaciones de limpieza que considera que hará.
4. Haga los procesos de limpieza que considere necesarios para tener un conjunto de datos listo para el análisis. Debe dejar constancia de cada una de las acciones que ejecutó. Debe explicar la razón por la cual dio cada paso. Todo debe ser reproducible. No debe hacer análisis sobre el conjunto de datos, tampoco tomar decisiones sobre eliminación de filas y columnas (a menos que no haya datos), solo limpiar.
5. Se le solicita que el conjunto de datos resultante quede lo más consistente posible y que se preste especial atención a los campos que identifican el establecimiento como nombre, dirección y teléfono. Tenga en cuenta que los textos pueden tener errores tipográficos. Debe revisar que un mismo establecimiento no quede varias veces escrito de formas diferentes. Hay establecimientos que funcionan en varios horarios.
6. Genere **un** conjunto con la unión de los datos de todos los departamentos totalmente limpio. Si crea variables, estas deben especificarse en el libro de códigos con sus respectivos metadatos.

7. Elabore un Libro de códigos, donde describa el significado de cada variable, los valores posibles que puede tomar. Incluya la descripción general del conjunto de datos.

EVALUACIÓN

- **(5 puntos)** Carga de los conjuntos de datos.
- **(15 puntos)** Análisis del estado de los datos crudos: Se hace un análisis detallado del estado de los datos antes de comenzar las operaciones de limpieza.
- **(45 puntos)** Operaciones de limpieza y explicación de las decisiones tomadas. Se hacen operaciones de limpieza en todas las variables, especialmente en el nombre, dirección y teléfono de los establecimientos. Se verifica que no estén duplicados, o con variaciones en el nombre.
- **(25 puntos)** Libro de códigos. El libro de código debe ser comprensible, estar bien detallado y organizado. Contiene los metadatos necesarios para comprender el conjunto de datos, incluidas las fechas en las que se extrajeron y la fuente.
- **(10 puntos)** Generación del conjunto de datos Limpios. Se genera un solo conjunto de datos con la información de todos los departamentos. Los datos generados están limpios y pueden utilizarse para hacer análisis.

NOTA: Todos los integrantes del grupo deben contribuir al repositorio y code book. No se evaluará a quien no tenga contribuciones significativas.

MATERIAL A ENTREGAR

- Archivo .r, .rmd, .ipynb, o .py, con el código de las acciones tomadas desde que se carga el conjunto de datos hasta que se termina de limpiar.
- Link del repositorio con el código.
- Documento del Google docs donde se trabajó el libro de códigos
- Archivo pdf, con el libro de códigos
- Archivo csv con los datos limpios