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. Es por esto que, 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 libro de variables detallado que contenga los significados de los valores posibles y significados de las mismas.

ACTIVIDADES

1. Descargue los datos de “Fallecidos y Lesionados”, “Hechos de tránsito” y “Vehículos involucrados” de todos los años disponibles (2009 - 2017) del Instituto Nacional de Estadísticas de Guatemala (INE). Los puede encontrar en el siguiente enlace: https://www.ine.gob.gt/index.php/estadisticas-continuas/accidentes-de-transito
2. Guarde los datos crudos en archivos csv.
3. Describa el estado de los datos y las operaciones de limpieza que considera que hara.
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.
5. Genere **un** conjunto con la unión de los datos de todos los años (2009-2017) totalmente limpio.
6. 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

- **(10 puntos)** Carga de los conjuntos de datos.
- **(15 puntos)** Análisis del estado de los datos crudos
- **(40 puntos)** Operaciones de limpieza y explicación de las decisiones tomadas.
- **(25 puntos)** Libro de códigos.
- **(10 puntos)** Generación del conjunto de datos Limpios

MATERIAL A ENTREGAR

- Archivo .r, .rmd o .py o flujo de trabajo en knime, con el código y/o los pasos en caso de que se use Knime de las acciones tomadas desde que se carga el conjunto de datos hasta que se termina de limpiar.
- Archivo pdf, con el libro de códigos
- Archivo csv con los datos limpios