Universidad Francisco Marroquín

Data Wrangling - 2019

# Proyecto Final: Resultado del Censo 2018

El Instituto Nacional de Estadística realizó el XII Censo Nacional de Población y el VII Censo Nacional de Vivienda en el año 2018. Los datos recolectados contiene información e indicadores a nivel nacional, departamental y municipal sobre características de la población guatemalteca, sexo, edad, educación, pueblo e idioma, tecnologías de la comunicación, acceso a servicio, composición de hogares, estructura de vivienda entre otras.

Una delegación del instituto se encargó de tabular los datos recolectados en todo el territorio nacional. Estos han sido debidamente tabulados por municipio y departamento. El INE obtuvo conocimiento de su capacidad de manejo, extracción y limpieza de datos por debido a las consultorías que ha trabajado en el año y han decidido contratarlo para el análisis de datos del estado actual del país.

Se le ha proveído un total de 16 archivos, los cuales cada uno contiene información recopilada de los temas previamente mencionados.

**Requisitos**

* Utilizar los lenguajes de programación vistos en clase. Python o R a su discreción. **No utilizar Excel.**
* Notebooks o Rmarkdowns subidos a cuenta de Github.Tomar en cuenta que esto puede formar parte de su currículum al momento de aplicar a algún puesto de análisis o científico de datos.
* Trabajo individual o en parejas.

**Entregables**

1. Base de datos
   1. Tablas (15) en formato Tidy.
   2. Código con el cual transformaron los datos.
2. Reporte escrito
   1. Introducción y objetivos
   2. Metodología
   3. Exploración (EDA)
   4. Feature engineering (Si aplica. Incluye también cualquier transformación de variables y/o expresiones regulares)
   5. Gráficas (histogramas, boxplots, scatterplots, etc)
   6. Correlación de variables (Sugerencia: unir las tablas)
   7. Hallazgos
   8. Conclusiones
   9. Recomendaciones y siguientes pasos
   10. Anexos: Diccionario de variables, definiciones, etc.
3. Presentación final: Tendrán 15 minutos para presentar y se calificarán los siguientes elementos:
   1. Comprensión del problema (20%)
   2. Objetivo y descripción teórica (20%)
   3. Claridad y Limpieza (10%)
   4. Análisis y Resultados (30%)
   5. Conclusiones, recomendaciones y siguientes pasos (20%)

*Nota: La presentacion sera realizada en el salón de clase correspondiente (a confirmar) donde posiblemente asistirán entre 1-3 jueces adicionales para evaluar su trabajo.*

Ponderación:

1. Base de datos: 30%.
2. Reporte escrito: 40%.
3. Presentación final. 30%.

Recomendaciones:

* No presentar líneas de código en su presentación. Al momento de mostrar resultados, el foro no está interesado en las tecnicidades que realizaron para llegar sus conclusiones. El proceso será evaluado por medio de su repositorio de GitHub.
* De igual forma, el reporte escrito no debe de incluir líneas de código.
* El reporte escrito debe de ser conciso al igual que la presentación.