

## **ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL INTRODUCCIÓN A CIENCIA DE DATOS**

2021 -1

Prof. Carmen Vaca, PhD

### **FECHAS DE ENTREGA**

Tarea : Lunes 19 de julio, 22 horas

Avance 1: Lunes 26 de julio, 22 horas

Avance 2 : Domingo 15 de agosto, 22 horas

Presentación proyecto final: Jornada de presentaciones– Lunes 23, Miercoles 25 de agosto de 2021

### **Introducción**

El proyecto de curso busca entregar un producto de analítica con una estructura en la cual se hace uso adecuado de técnicas diversas para implementar las diferentes partes de un pipeline de ciencia de datos. Cada una de las etapas se documentará adecuadamente: recolección de datos, propuestas de preguntas de negocio y análisis exploratorio de datos, comunicación de insights a través de visualizaciones.

### **TAREA : Lunes 19 de julio, 22 horas**

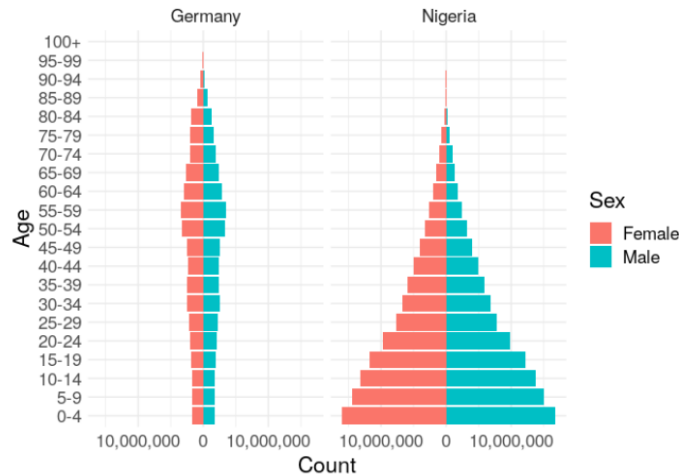
- Descripción del dataset , análisis exploratorio de datos con respuestas de al menos 3 preguntas de negocios
- Entregables: Jupyter notebooks y slides

### **AVANCE 1 COMPLETO: Lunes 26 de julio, 22 horas**

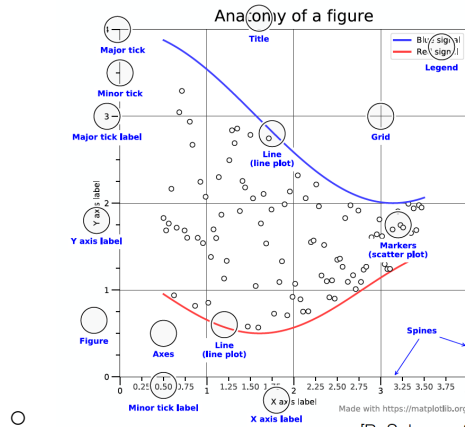
#### **Requerimientos**

- Colecte y describa su dataset. En la descripción del dataset debe incluir un overview del mismo: 80% de los registros cumplen esta condición, etc.
- Maneje datos faltantes si los hubiera.
- Definir 6 preguntas de investigación [incluyendo las tres del avance anterior] a responder en su proyecto. Si las preguntas a explorar son muy básicas y no requieren de ninguna técnica de análisis, no se contarán dentro de las 6 preguntas requeridas. El viernes

- Crear métricas o variables categóricas que ayuden a encontrar grupos en sus datos. Respalde esta métrica en literatura:
  - Usar apply / filter / transform para hacer cálculos de índices a partir de los datos originales.
- Combine datos de varios datasets, uno de los cuales debe haber sido recolectado por usted.
- Haga análisis de outliers
- Incluir visualizaciones para análisis univariado y multivariado. Sus visualizaciones deben incluir: (varias de sus visualizaciones podrían incluir algunas de estas sugerencias, en ciertos casos como el de ejes etiquetados, todas las visualizaciones deben implementarlo)
  - Faceted plots
  - Multivariables. Use visualizaciones avanzadas que muestren la información de manera ordenada. La visualización siguiente por ejemplo muestra 4 variables sin sobrecargar el gráfico:



- Visualizaciones geográficas (choropleth). Recuerde normalizar la información si es necesario.
- Visualizaciones que combinan variables categóricas y numéricas
- Comparación de distribuciones
- Ejes etiquetados
- Título para el gráfico
- Rotación de eje de etiquetas X de ser necesario (que no haya overlap en las etiquetas)
- Visualizaciones comparativas (side by side) en las que se muestran por ejemplo: distribuciones de dos grupos distintos en sus datos.
- Personalizar los ticks (es decir las marcas del eje x o eje y)



- Entregables:
  - Artículo: Introducción, dataset, métodos
  - Jupyter notebooks
  - Slides

AVANCE 2 COMPLETO: Domingo 15 de agosto, 22 horas  
(se proveerá mayor detalle luego)

- Visualizaciones temporales
  - Compare trends, seasonality.
- Aplicar las técnicas revisadas en clases:
  - Clustering
  - .... (Esto se completará en la publicación del siguiente avance.)
- Mejorar el análisis exploratorio realizado en entregable anterior agregando 4 Business Questions a las ya presentadas
- A las visualizaciones mostradas antes, agregue , **una visualización dinámica usando Dash, Bokeh o Altair.**
- Asegúrese de interpretar los resultados de los clusters y caracterizarlos:
  - Presentar estadísticas por cluster. Por ejemplo, qué diferencia hay en una métrica dada para el grupo con valor más alto y el de valor más bajo.
- Mida y explique el efecto de normalizar (o no) la data en su problema de análisis.
- Escoja casos de estudio. Por ejemplo: los barrios de Santiago con mayor/menor cantidad de gente con acceso a educación y compárelos.
- Entregables:
  - Artículo completo
  - Jupyter notebooks
  - Slides

## NOTAS SOBRE LOS ENTREGABLES

Para la presentación de su proyecto/avances, su trabajo debe incluir tres entregables.

### 1. JUPYTER NOTEBOOK

Dentro de un **archivo .ipynb** deben incluirse TODOS los análisis y experimentos realizados durante su investigación.

Cada sección de análisis debe estar comentada y explicada según el análisis correspondiente.

Debe usar celdas **markdown** para describir los métodos empleados y separar los análisis hechos en diferentes secciones.

### 2. SLIDES

Desarrolle una presentación .pptx (o Keynote) de su trabajo de analítica. Esto debe incluir al menos: introducción donde justifique la motivación del trabajo (1 diapositiva), pregunta(s) de investigación, metodología utilizada y las razones de su elección, resultados positivos más relevantes, conclusiones y discusión. No olvide que siempre los resultados y las conclusiones deben intentar responder la pregunta de negocio. Considere lo siguiente:

- En la presentación evite frases como “Existen bastantes estudiantes con promedio bajo”.

Lo apropiado es usar descripciones cuantitativos como: “Cerca del 40% (37.8) de estudiantes tienen un promedio menor a \_\_\_\_\_. Esto considerando los tres últimos años de data estaría tres puntos por debajo del promedio.”

- Utilice fondo blanco para las presentaciones
- Numere los slides
- Si compara dos gráficos, colóquelos en la misma diapositiva
- Enfatique el/los findings más importantes de su análisis.

Al final, incluya una sección ANEXOS que resuma los experimentos fallidos que haya tenido. Ej: Al aplicar SNA no obtuvo ningún hallazgo. Este análisis debe ir en anexos.

### 3. ARTÍCULO CIENTÍFICO

Consiste en un **archivo PDF** escrito a modo de artículo científico formato IEEE (<https://www.ieee.org/conferences/publishing/templates.html>) . Las secciones que se deben incluir son:

- a. Introducción: Problema, y motivaciones para resolverlo.
- b.. Dataset: Campos, cantidad de registros, métodos de extracción y limpieza.

- c. Metodología: Debe incluir información extraída de las distribuciones de sus variables, Por ejemplo:
  - i. El 60% de transacciones tiene un valor menor a 4 dólares.
  - ii. El 10% de clientes que facturan más de 300 dólares representa el 90% de ingresos en dólares.
  - iii. El 60% de los medidores no tienen variaciones por el clima.
- d. Métodos: Métricas y técnicas utilizadas. NO EXPLICAR QUE ES K-MEANS O LA TÉCNICA QUE UTILIZAN sino, la aplicación de esa técnica a su problema.
- e. Resultados: Explique cómo el resultado de aplicar una técnica ayuda a responder un Business Question. En el caso particular de clustering, debe caracterizar los clusters y describir lo que se observa.
- f. Discusión: Reflexionar sobre los hallazgos y las posibilidades de aplicar su análisis en otro contexto o para nuevos análisis.

No olvide incluir tanto en el artículo como en las diapositivas, gráficos relevantes a su investigación sean visualizaciones geográficas, de grafos o estadísticos. VENDA SU IDEA

Cada entregable debe tener un avance significativo ( 80% ? ) para ser evaluado.