**DIPLOMADO EN DATA SCIENCE**

Módulo 7: Fundamentos de Visualización de Datos | 2021

Profesores: Gonzalo Donoso | Patricio Torres



**Trabajo Práctico Finalización Módulo.**

1. **Objetivo**

Demostrar los conocimientos y habilidades adquiridas en el Módulo, mediante la representación visual de datos.

1. **Instrucciones**

Realización de un proceso de producción de analítica mediante el apoyo de visualizaciones.

1. **Planteamiento y antecedentes:** Se le ha otorgado la tarea de analizar una base de datos de una tienda retail de aproximadamente 7 meses. La cual busca incorporar análisis de la ciencia de datos para mejorar sus operaciones y fidelización de clientes. Los directivos que deben tomar las decisiones le han pedido su apoyo para este trabajo, sin embargo, no cuentan con conocimiento para interpretar datos ni estadística, es por esto, que han solicitado representaciones visuales de cada hallazgo para facilitar el entendimiento.

**Además buscan responder las siguientes preguntas:**

* 1. ¿Existen productos favoritos que compran los clientes?
  2. ¿Existe algún horario especifico en el que los clientes compran? ¿ En cual horario sugeriría incorporar apoyo a los servidores de la tienda para evitar un colapso?
  3. ¿Qué frecuencia de compra tiene un cliente de alta frecuencia? ¿ Cuanto debería gastar un cliente para considerarlo sobre la media?
  4. ¿Cómo se distribuye las antigüedades de los clientes?
  5. Si clusterizamos en 3 grupos a los clientes en base a su frecuencia de compra ¿Qué diferencias existen entre cada clúster? (Utilice el algoritmo de clusterización Kmeans)

Tenga en cuenta además que el dataset entregado , puede no estar depurado, es decir, podrían existir devoluciones, ingresos manuales de caja , que podrían ensuciar su análisis.

1. **Fuente de Datos:**
   1. Descripción y contexto del data set.
   2. Resumen de atributos iniciales del data set: número y clase de campos, número de registros.
   3. Descripción de, al menos, 2 (dos) atributos (campos / variables) presentes en el data set, y que se utilizarán en el análisis.
2. **Preprocesamiento:**
   1. Principales problemas y estrategias de solución
   2. Procedimientos de limpieza (si aplica)
   3. Decisiones tomadas respecto de los atributos que se utilizarán en el análisis.
3. **Resultados:** Deberán presentarse resultados apoyados por visualizaciones de datos, los cuales deberán ser explicados teniendo en consideración que el destinatario será un público no entendido en la materia del reporte. Concretamente, deberá contener, al menos:
   * 1. **Visualización de datos:**
        1. 1 Gráficos en la etapa de pre procesamiento que evidencien problemáticas con los datos y que correspondan a lo que se busca representar.
        2. 1 Gráfico de cada pregunta solicitada y que represente la información buscada por los directivos. Entre ellos deben estar:
           1. Grafico de cajas
           2. Dispersión o Burbujas
           3. Histograma
           4. Gráfico de distribución
           5. 1 Grafico adicional a elección del estudiante.

Cada gráfico debe estar construido de forma tal que permita diferenciar apropiadamente grupos , categorías o información relevante, mediante , por ejemplo colores o formas distintos si se requiere.

* + 1. **Textos:**
       1. El documento deberá plantearse siguiendo lineamentos de jerarquía textual.
       2. Cada visualización deberá ser referenciada en el documento y comentada brevemente indicando insights del dataset analizado.

1. **Conclusiones:** Principales conclusiones respecto de cómo el planteamiento escogido permitió cumplir los objetivos y responder las preguntas de investigación.
2. **Consideraciones**
   1. Los estudiantes deberán trabajar con los datos provistos por el equipo docente.
   2. El trabajo deberá realizarse de forma individual.
   3. El formato de entrega deberá ser PPT (Power Point) con las representaciones visuales junto al desarrollo del encargo.
   4. Se deberá entregar el notebook (colab o jupyter notebook, formato ipynb) de trabajo con los códigos respectivos.
   5. Las expresiones visuales y textuales deberán ceñirse a los contenidos abordados por el curso.
3. **Entregables**
   1. Presentación de no más de 30 láminas (slides de PPT) en formato PDF.
   2. Jupyter o Colab notebook (formato ipynb)
4. **Fecha de Entrega**

- Miércoles 28 de septiembre, hasta las 23:59 hrs., al correo del profesor Gonzalo Donoso gonzalo.donoso.p@usach.cl