**Examen Final Exploración de Datos Multivariados**

El conjunto de datos asignado corresponde a salarios de trabajadores de análisis de datos, usted deberá tratar de obtener información descriptiva sobre los salarios y encontrar características de los salarios mejores pagos.

A continuación encontrará en el enlace de descarga del dataset asignado para defender en el examen final.

**Descarga del dataset** [**https://drive.google.com/file/d/1NnOvk3UtVGYxelG\_qlK\_TkaYu47sYthx/view?usp=sharing**](https://drive.google.com/file/d/1NnOvk3UtVGYxelG_qlK_TkaYu47sYthx/view?usp=sharing)

| Columna | Descripción |
| --- | --- |
| work\_year | El año en que se pagó el salario. |
| experience\_level | El nivel de experiencia en el trabajo durante el año con los siguientes valores posibles: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director |
| employment\_type | El tipo de empleo para el puesto: PT A tiempo parcial FT A tiempo completo CT Contrato FL Freelance |
| job\_title | El rol funcionó durante el año. |
| sueldo | El monto total del salario bruto pagado. |
| salary\_currency | La moneda del salario pagado como un código de moneda ISO 4217. |
| salarioenusd | El salario en USD (tasa de cambio dividida por la tasa promedio de USD para el año respectivo a través de fxdata.foorilla.com). |
| employee\_residence | El país de residencia principal del empleado durante el año laboral como un código de país ISO 3166. |
| remote\_ratio | La cantidad total de trabajo realizado de forma remota, los valores posibles son los siguientes: 0 Sin trabajo remoto (menos del 20%) 50 Parcialmente remoto 100 Totalmente remoto (más del 80%) |
| company\_location | El país de la oficina principal o sucursal contratante del empleador como un código de país ISO 3166. |
| company\_size | El número promedio de personas que trabajaron para la empresa durante el año: S menos de 50 empleados (pequeños) M 50 a 250 empleados (medianos) L más de 250 empleados (grandes) |

**Indicaciones**

Para el examen final deberá presentar dos archivos: el primero es un colab con todo el análisis realizado y las conclusiones pertinentes, donde el código debe poder ejecutarse. Y el segundo archivo consiste en un documento Word donde se entregará el informe final seleccionando la información que surgió como importante del análisis realizado. En base a este informe deberá realizar una presentación para exponer el día del examen, haciendo hincapié en las cuestiones más relevantes, y proporcionando información (tablas, gráficos, medidas) que permitan apoyar la información que se presente.

La defensa debe tener una duración aproximada de 25 minutos. Luego de la exposición el jurado realizará preguntas que podrán ser sobre el código implementado, o sobre lo presentado y las cuestiones que validan lo realizado, o sobre la no aplicación de alguna herramienta estudiada en el cursado.

**Algunas cuestiones a tener en cuenta:**

* Un cliente que llega con una base de datos muchas veces no sabe sobre la calidad de sus datos y la información que contiene su dataset, por lo cual siempre es imprescindible describir la base datos recibida.
* Siempre hay alguna pregunta o prioridad a tratar de contestar, no obstante si bien el análisis debe poner el foco ahí, siempre es posible extraer información adicional que puede resultar de utilidad para el cliente, el análisis debe ser exhaustivo.
* No todas las variables son útiles, hay información que tal vez no tiene sentido considerar, y que el analista debe saber reconocer. Por ejemplo el nombre de una persona en un análisis de venta.
* Por otra parte siempre es posible generar nuevas variables que tal vez resulten de utilidad para el análisis, por ejemplo si se conoce la dirección de una persona, se puede generar una variable con la localidad o el barrio, y esto puede servir para segmentar los clientes de algún determinado rubro.
* La calidad de un informe no se mide por su extensión, puede haber un informe de cinco páginas claro y conciso que contenga toda la información relevante, como puede haber un informe de 50 páginas que redunde sobre lo mismo o muestre información irrelevante para el problema.
* El informe no debe contener ni todas las pruebas que se realizaron, ni todas las tablas, ni todos los gráficos, debe seleccionar o confeccionar solo aquellos que sirvan para comprender los datos y que sean de valor para el cliente. Es decir si encontró que hay una correlación entre la altura y el peso, pero lo que está estudiando es el rendimiento académico de un estudiante, dicha información resultará irrelevante.

**Criterios de evaluación:**

* Descripción del dataset: cantidad de registros, cantidad de registros mal cargados, cantidad de variables, cantidad de valores nulos, y cualquier otra información que dé cuenta del estado en el que recibe el dataset para el análisis.
* Tareas de limpieza realizadas: cambio del nombre de variables para una lectura de datos más fluida, creación de nuevas variables que aporten al problema, eliminación o arreglo de valores mal cargados o nulos, etc.
* Análisis univariado: información relevante obtenida de este análisis, lectura de tablas de frecuencias, interpretación de medidas de resumen, selección de gráficos apropiados.
* Análisis bivariado: selección de test a realizar según las variables que se cruzan, análisis con variables cuantitativas, análisis con variables cualitativas, uso correcto de test teniendo en cuenta los supuestos de cada test, lectura de la información obtenida, selección de gráficos apropiados. Interpretación de resultados.
* Análisis multivariado: selección de gráficos que permitan aportar mayor información al problema, uso de test si correspondiese.

*Esperamos que disfrute la ardua tarea de poner en juego lo aprendido.*

*Equipo de Cátedra*