**Actividad 3:**

**Análisis de datos en menú interactivo**

|  |  |
| --- | --- |
| Asignatura: | Big Data |
| Integrantes: | Ignacio Arancibia  Eduardo Silva |
| Profesor:  Fecha de entrega: | Edio Mardones  22/08/2024 |
|  |  |

**Índice General**

[1. Introducción 2](#_Toc175179608)

[2. Desarrollo del contenido 3](#_Toc175179609)

[3. Conclusiones 10](#_Toc175179610)

[4. Referencias 11](#_Toc175179611)

# Introducción

Una habilidad crucial para el análisis y manejo de grandes cantidades de datos es la toma y manipulación efectiva de datos, a continuación se mostrará el desarrollo de un menú interactivo diseñado para facilitar estas prácticas enfocadas a datos sacados desde un archivo Excel que muestra diferentes tipos de datos sobre empleados contratados. Conteniendo datos como el ID de cada persona, su nombre, su salario, el departamento en el que desempeñan sus funciones, la fecha de ingreso a la empresa y la locación en que residen.

El objetivo de este menú es principalmente ofrecer una solución intuitiva para permitir a usuarios explorar, analizar y modificar datos de una manera activa, utilizando python y librerías como Pandas,Numpy, Streamlit, Matplotlib, Seaborn, etc.

Más adelante se enseñarán las fases del desarrollo del menú y datos, las funciones y usos del menú, el impacto de la implementación de este archivo. Este código intenta buscar la forma de demostrar que las herramientas interactivas pueden transformar la manera en que se manejan, analizan y modifican datos en un entorno profesional relacionado a empresas o derivados.

# Desarrollo del contenido

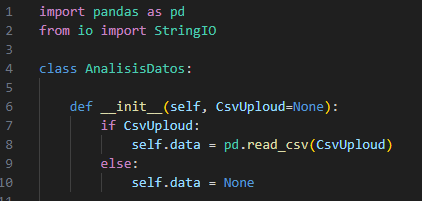
Marco teórico:

* Librería pandas: Librería principal que gestiona la carga y el análisis de datos.
* Librería numpy: Librería que proporciona la gestión de arrays.
* Librería streamlit: Librería que permite la creación de dashboards interactivos en una página web.
* Librería matplotlib: Librería que crea estadísticas interactivas, animadas o estáticas. Genera histogramas, gráficos de barras, etc
* Librería seaborn: Librería que crea gráficos estadísticos usando parte de la librería de matplotlib.
* Limpieza de datos: Paso necesario para realizar cualquier tarea importante en el ámbito de análisis de datos. Proceso que incluye el detectar, eliminar, corregir o transformar datos anómalos, incorrectos o irrelevantes. Este método ayuda a obtener resultados confiables y a que los datos sean fidedignos.
* Remover outliers: Outlier se refiere a cada dato que difiere de manera significativa con el resto de datos. Un ejemplo de esto es encontrar un dato en la variable “edad” con una cantidad mayor a 200, 300, 400, etc.
* Trabajo con datos nulos: Los datos nulos se refieren a información no introducida en las variables. Esto debido a falta de información o un error de entrada. Es necesario extrapolar estos datos efectuando transformaciones.

Metodología:

La principal fuente de los datos viene de Kaggle y tiene como nombre “Synthetic data”, esta fuente nos proporciona datos como el nombre de un empleado, el departamento en el cual trabaja, la fecha de ingreso a la empresa y su ubicación.

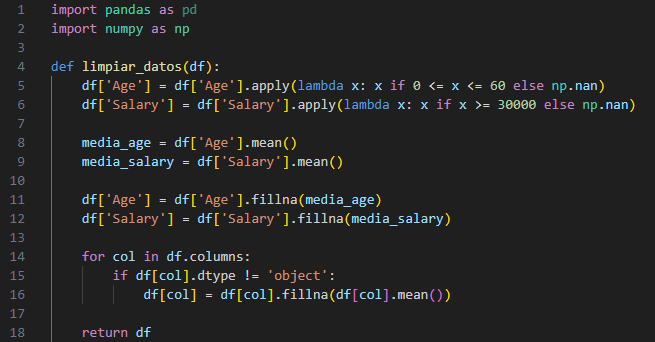
A continuación se muestra el código de la implementación de los datos en Visual Studio Code utilizando las librerías y lenguaje ya mencionados. Este código permite leer cualquier tipo de archivo “.csv” dentro de una página web.



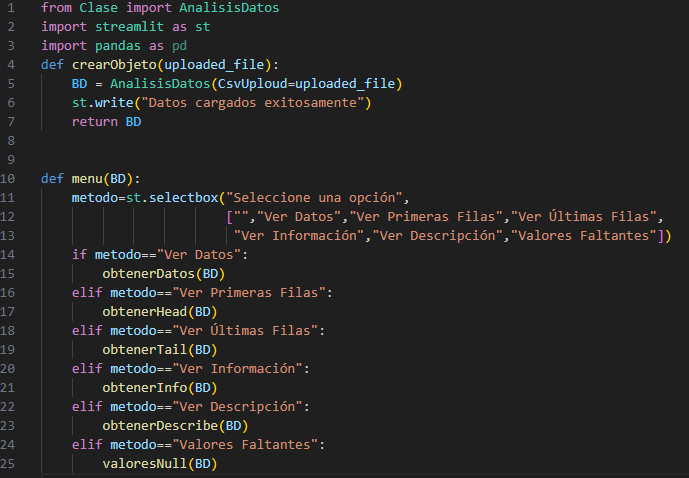
Para la limpieza de datos e imputación de valores nulos, se utilizó la media de cada variable para indicar un valor realista, como se muestra en el siguiente código:

En la edad y en el salario se puede ver la aplicación de este código, las restricciones para mostrar la edad y salario de cada empleado es que debe ser mayor a cero y menor a sesenta (edad) y mayor o igual a 30000 (salario).

Este proceso fue automatizado utilizando las librerías de “pandas” y “numpy”, que permitieron una manipulación eficiente de grandes volúmenes de datos.

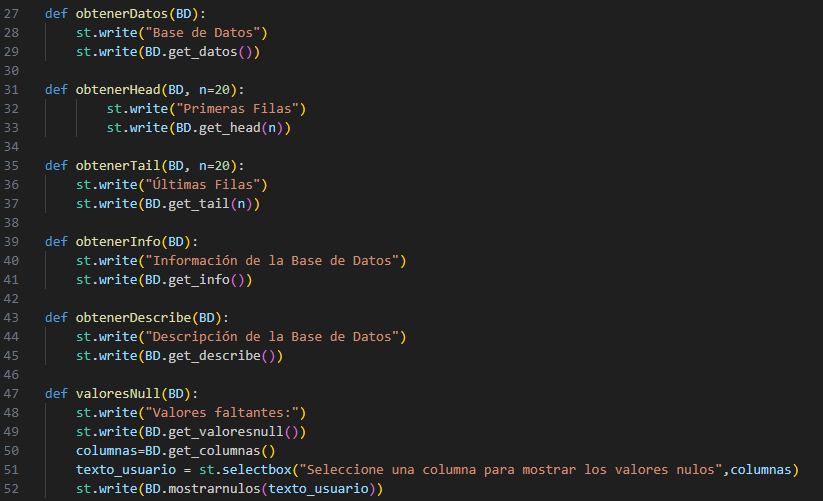


Al momento de crear el menú se implementó el siguiente código que primero crea la base de datos para luego analizarla si es que cumple con el formato correcto. Este menú permite ver los datos de cada tabla, sus primeras filas, las últimas filas, información de las tablas, su descripción y los valores faltantes

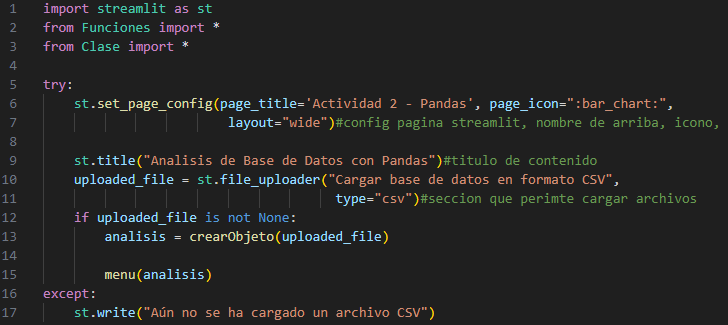


A continuación está el código que permite cumplir las funciones de cada una de las opciones del menú:

Obtener datos, saca los datos desde la base de datos. Obtener head, obtiene las primeras 20 filas de la base de datos. Obtener tail, obtiene los últimos 20 datos. Obtener info, obtiene la información de cada dato. Obtener descibre, muestra la descripción de la base de datos. Valores null, muestra los valores faltantes y su columna.



Creación de página con streamlit:



# Conclusiones

Se desarrolló una página con menú interactivo para facilitar la limpieza y análisis de datos. El menú permitió la ejecución de tareas automatizando procesos críticos y garantizando resultados consistentes.

Los objetivos que se lograron fueron la limpieza de datos de forma correcta, así permitiendo una corrección de datos inconsistentes y valores nulos o fuera de rango.

También se logró un satisfactorio análisis descriptivo que permitió visualizar estadísticas para una mejor y mayor comprensión de los datos.

Estos hallazgos impactan de manera significativa para campos donde es de vital importancia la rápida toma de decisiones, la investigación académica o estudios de mercado.

La capacidad de manejar grandes volúmenes de datos y realizar análisis de los mismos de forma detallada implica una alta aplicabilidad y utilidad en la disciplina que se requiera.

# Referencias

Torres, A. (2021, 9 de diciembre). freeCodeCamp. <https://www.freecodecamp.org/espanol/news/limpieza-de-datos-en-pandas-explicado-con-ejemplos/>

Burrueco, D. (2020, 12 de abril). InteractiveChaos. <https://interactivechaos.com/es/manual/tutorial-de-pandas/creacion-de-un-dataframe-partir-de-un-diccionario-de-listas>

*Dare Planet*. (s.f.). Dare Planet. <https://dareplanet.tech/insights/analisis-de-datos-libreria-pandas-python/>