

Curso: Análisis de datos

Práctico 1

Obtención, exploración y pretratamiento de datos

A continuación se presentarán distintos ejercicios a realizar utilizando el dataset *permisos_construccion.csv*, el cual trata sobre permisos de construcción emitidos en la ciudad de San Francisco, Estados Unidos. El mismo, y su metadata, se encuentran disponibles en la pestaña “Pretratamiento de datos” del curso.

Para estos ejercicios es necesaria la utilización de python y la librería pandas. También pueden utilizar matplotlib y seaborn para graficar resultados.

Se debe entregar un informe con justificaciones de lo realizado y de los resultados obtenidos en un archivo .ipynb (colab o jupyter notebook). Indicar integrantes del equipo de trabajo.

1. Cargar los datos a estudiar.
2. Hacer un resumen con la explicación de las variables del juego de datos.
3. Realizar un análisis exploratorio de los datos (incluir gráficas y métricas de lo que consideren necesario).
 - a. Calcular y mostrar la cantidad de filas y columnas.
 - b. Observar y mostrar las primeras 5 filas.
 - c. Evaluar la existencia de datos faltantes y duplicados. Cuantificarlos y calcular el porcentaje sobre el total de filas.
 - d. Para los datos faltantes, evaluar posibles motivos de esto en cada caso.
 - e. Para variables discretas, evaluar los posibles valores de cada variable (valores únicos).
 - f. Cuantificar los valores únicos del punto e) y realizar histogramas.
 - g. Evaluar la existencia de datos inconsistentes.
4. Limpieza de datos:
 - a. Tomar una decisión respecto a los datos faltantes, duplicados, inconsistencias y cualquier inconveniente que encuentren con los datos.
 - b. Realizar las modificaciones necesarias para cumplir con el punto a.

5. Normalización de datos
 - a. Normalizar las variables continuas (numéricas).
 - b. Investigar ventajas y posibles usos de esta tarea.
6. Utilizar la librería de Data Profiling vista en el curso y comparar las inconsistencias encontradas con lo que muestra el informe.