UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Guatemala 25 de Julio de 2024

Jimena Hernandez - 21199
Mark Albrand - 21004
Alejandro Azurdia - 21242
Diego Morales - 21146
Angel Castellanos - 21700

Proyecto 1- Avances

---

1. **Describa el estado de los datos y las operaciones de limpieza que considera que hará.**
   a. Describa el set de datos, cuantas filas tiene inicialmente con los datos crudos, y cuantas variables.

      El conjunto de datos cuenta con 22 bases de datos, correspondientes a cada uno de los departamentos de Guatemala. Cada una de estas bases (en formato csv) cuentan con 17 variables y datos crudos. A continuación se describen las variables encontradas en cada uno de los archivos.

      1. ‘Código'
      2. 'Distrito'
      3. 'Departamento',
      4. 'Municipio'
      5. 'Establecimiento'
      6. 'Dirección'
      7. 'Teléfono'
      8. 'Supervisor'
      9. 'Director'
      10. 'Nivel'
      11. 'Sector'
      12. 'Área'
      13. 'Status'
      14. 'Modalidad'
      15. 'Jornada'
      16. 'Plan'
      17. 'Departamental'

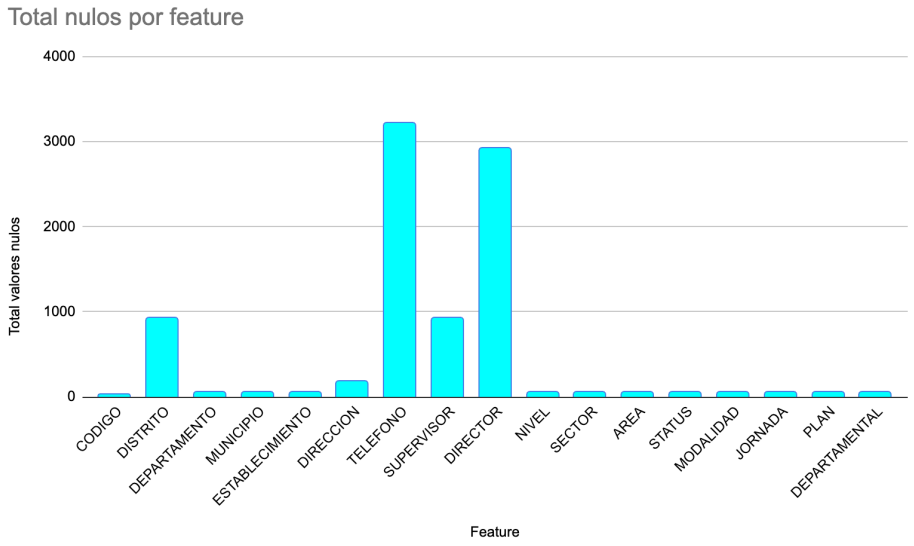
A continuación se puede observar la distribución de filas por cada departamento.

Departamento	Filas	Columnas
Sacatepéquez	321	17
Capital	5529	17
Suchitepéquez	387	17
Santa Rosa	159	17
San Marcos	576	17
Baja Verapaz	116	17
Petén	368	17
Alta Verapaz	377	17
Chiquimula	2511	17
Izabal	370	17

Sololá	140	17
Retalhuleu	318	17
Zacapa	96	17
Chimaltenango	2458	17
Quetzaltenango	493	17
Quiche	246	17
Guatemala	1481	17
Escuintla	2932	17
Huehuetenango	5160	17
Jalapa	153	17
Totonicapán	92	17
Jutiapa	312	17

b. Liste las variables que más operaciones de limpieza necesitarán.

Después de realizar un análisis exhaustivo de las variables, fue posible observar que existe una cantidad significativa de valores "nulos" para las siguientes variables: teléfono, director, distrito supervisor. Por esto, serán las variables que más operaciones de limpieza necesitarán.



Para lidiar con estos valores nulos y considerando que son variables categóricas, se intentará eliminar las filas con valores faltantes si la columna (variable) tiene al menos un 70% de datos no nulos. En caso de que las celdas con valores sean menores al 70%, entonces se dejarán los valores nulos.

El formato en mayúsculas del texto se mantendrá. Se eliminará cualquier carácter especial para conservar la uniformidad de los datos. Por último, es importante eliminar cualquier repetición. Esto es posible gracias al código único de la institución.