INFORME LIMPIEZA DE DATOS

DESARROLLADO POR:

ING. JULIAN ANDRES QUIMBAYO CASTRO

ENTREGADO AL MAESTRO:

MARLON CARDENAS BONETT

UNIVERSIDAD INTERNACIONAL DE LA RIOJA UNIR

ASIGNATURA BASE DE DATOS PARA EL BIG DATA

2022

1. DESARROLLO DE LA ACTIVIDAD

Como primera medida el dataset consta de una información sobre el estado de san francisco en estados unidos en cuanto a tipologías de asaltos y robos efectuados, con sus causas, tiempos, momento de la ofensiva y lugares. Son alrededor de 10051 filas por 12 columnas.

En total los datos NAN (Not a number) son 10375 entendiendo que la columna range no posee información y en city existen 321 datos inconsistentes. Existen 10 columnas de tipo Object y solo dos numéricas como crimeID y Range.

Herramienta usadas:

- Excel
- Python 3.10
- Pandas y numpy
- Manejo de funciones
- Visual studio code

1.1. descripción de problemas encontrados

Descripción del Problema 1: Columna OriginalCrimeTypeName

Esta columna posee inconsistencias en la información ya que la tipología de origen del crimen no está bien escrita, posiblemente por errores de digitación, inconsistencias en las bases de datos o perdida de información. Esto se detecto utilizando la librería pandas con su método value_counts() que permite realizar frecuencias absolutas o relativas contando por cada categoría encontrada en la columna.

```
## Frecuencias absolutas para cualquier columna
def revCategoricos(df, nomCol):
   rev = df[nomCol].value_counts()
   return rev

## Frecuencias relativas para cualquier columna
def revCategoricosFreq(df, nomCol, dec):
   return round(df[nomCol].value_counts()/np.float64(len(df)),dec)*100
```

Estas dos funciones permiten evidenciar estas anomalías, pero existía otro inconveniente era la incapacidad de poder leer más de 575 opciones diferentes por tanto se utilizó un filtro que permitía en orden alfabético ir sacando porciones de la data e ir corrigiendo una a una.

`Forma de solucionar problema 1:`

El primer paso encontrar las inconsistencias apoyado en palabras iniciadas por ‘A’, ‘B’, así sucesivamente. Usando la siguiente función:

```
dataA = data[data['OriginalCrimeTypeName'].str.startswith("A")]
revCategoricos(dataA, 'OriginalCrimeTypeName')
```

Seguidamente se encontraban las inconsistencias y con la teoría del modelo de negocio de la data se pasaba a corregir usando la siguiente función.

```
# Función para datos categóricos modificaciones
def imputacionCat(df, nomCol, busqueda, reempl):
   df[nomCol] = np.where(df[nomCol] == busqueda, reempl, df[nomCol])
   return df[nomCol]
```

Esta función permite encontrar el parámetro y cambiarlo por la data que corresponda.

```
data['OriginalCrimeTypeName']= imputacionCat(data, 'OriginalCrimeTypeName', 'Assault / Battery Dv','Assault / Battery')
data['OriginalCrimeTypeName']= imputacionCat(data, 'OriginalCrimeTypeName', 'Agg Assault / Adw','Assault / Battery')
data['OriginalCrimeTypeName']= imputacionCat(data, 'OriginalCrimeTypeName', 'Att','Attempt')
```

Finalmente, si el porcentaje de las opciones restantes es menor al 5% del total de la data se eliminaban.

```
data = data.drop(data[(data.OriginalCrimeTypeName=='Awol') | (data.OriginalCrimeTypeName == 'Adv') |
                                (data.OriginalCrimeTypeName == 'Adv To 0123')|
(data.OriginalCrimeTypeName == 'A')
                                | (data.OriginalCrimeTypeName == 'Atc')|
(data.OriginalCrimeTypeName == 'Ams')
                                | (data.OriginalCrimeTypeName == 'Areport')|
(data.OriginalCrimeTypeName == 'Amplified')
                                | (data.OriginalCrimeTypeName == 'Adv To Co A')].index)
```

Este proceso se realize hasta la letra N y finalmente se extrajo el dataframe para terminar por medio de filtros en Excel.

Descripción del Problema 2: Columna City

Existián mas de 300 datos NAN de dicha columna y ciudades que no corresponden a san francisco, como daly city, yerba buena y Brisbane. Esto es un problema de inconsistencia en la data.

Forma de solucionar problema 2:

Inicialmente se revisan los datos categóricos encontrando que el dato que mas se repite es san francisco utilizando la moda en su posición cero se reemplaza los datos NAN y se ajusta un SAN FRANCISCO en mayúscula a minúscula. Usando las siguientes funciones:

```
# Función para datos categóricos modificaciones
def imputacionCat(df, nomCol, busqueda, reempl):
   df[nomCol] = np.where(df[nomCol] == busqueda, reempl, df[nomCol])
   return df[nomCol]
```

```
dataFinal['City']= imputacionCat(dataFinal, 'City', 'SAN FRANCISCO','San Francisco')
dataFinal['City']= imputacionCat(dataFinal, 'City', 'Treasure Isla','Treasure Island')

#Función para imputación de datos NA
def imputacionCatNa(df, nomCol):
  df[nomCol] = df[nomCol].fillna(df[nomCol].mode()[0])
  return df[nomCol]
```

Descripción del Problema 3: Columna Disposition

Se presenta el conocimiento de unos registros con la palabra Not Recorded y el valor 22, se reemplazan por los valores más repetidos como es HAN. Usando las mismas funciones anteriores.

Descripción del Problema 4: Columna Agendy ID

Se presenta el conocimiento de registros con valores truncados donde es ID = 1 tenía en cambio CA se cambia directamente por 1 en este caso por medio de Excel.

Descripción del Problema 5: Columna State

Se presenta el conocimiento de datos truncados donde se colocaba 1 en vez de CA se procede a realizar el cambio en Excel.

Descripción del Problema 6: Columna Range

Mas del 25% de datos NA, no se sabe a que se refiere la columna y se procede a eliminar.

Ubicación del dataframe limpio en formato .csv y Json con formato Orient Index en el siguiente GitHub:
https://github.com/jaquimbayoc7/MaestriaAnalitica/tree/main/Bases%20de%20Datos%20para%20el%20Big%20Data/Actividad_1_limpieza

{"0":{"Name":"Jay","Age":16,"Course":"BBA"},

"1":{"Name":"Jack","Age":19,"Course":"BTech"},

"2":{"Name":"Mark","Age":18,"Course":"BSc"}}

2. Formato JSON

El formato JSON propuesto es de tipo index tal como se comentó en la sección anterior se utilizó pandas y Python para exportarlo en dicho formato.

3. Metodología sugerida

La metodología sugerida consta de 6 pasos:

1. Revisar la idea del dataset y su modelo de negocio que vamos a resolver con estos datos.
2. Identificar el contexto de cada variable y su tipología
3. Limpiar datos categóricos por medio de revisiones de frecuencia o filtros.
4. Limpiar datos NAN con ayuda de la moda o una técnica como MICE
5. Revisar incoherencias

6. Perfilamiento para machine learning.

4. Mejoras al conjunto de datos

Las mejoras sugeridas son las formas de captura de la data deben corregirse si es por medio de un software se deben parametrizar.

Evitar permitir al usuario que coloque lo que le convenga y este abierto seleccionar o dar opciones es más adecuado.

Evitar la columna range si no se va a utilizar.