Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Licenciatura en Matemáticas Aplicadas y Computación

Proyecto: LÍNEA DE MUJERES CDMX

Materia:

Temas Selectos de Computación I - Métodos y Técnicas para la Gestión de Datos

Grupo: 1753

Profesor:

Reyes Hernández Sergio

Integrantes del Equipo:

Nombre
Soria Cabrera Andrés
Olvera Juárez Josué Israel
Godinez Rojas Alexis Omar
Xaltenco Aguilar Salvador
Mejia Trejo Hazel Jesús

Noviembre 2025

Índice

1 Identificación de la Problemática

En el análisis realizado sobre los datos de la Línea de Mujeres CDMX, se observaron varias áreas que podrían afectar la precisión de los resultados:

1.1 Valores Atípicos en la Edad

Se detectaron edades extremadamente bajas y altas, como valores de 1 y 99 años. Estos datos pueden no ser representativos y podrían sesgar los resultados de los análisis, especialmente si se segmenta la población por edades.

1.2 Valores Nulos en las Temáticas

Las variables de temática (`tematica_1`, `tematica_2`, etc.) contienen muchos valores nulos o vacíos (representados como "NA"). Esto impide realizar un análisis claro de las razones por las que las personas llaman a la línea.

1.3 Desbalance en las Variables Categóricas

Algunas variables, como estado civil, ocupación y sexo, muestran una distribución desbalanceada (por ejemplo, hay más mujeres y personas solteras). Esto puede influir en los análisis y representar de forma incorrecta la población total.

2 Relevancia del Caso

La calidad de los datos es esencial para tomar decisiones acertadas, especialmente si se busca entender las necesidades de las personas que utilizan la línea. Los problemas mencionados pueden afectar la segmentación de la población y las temáticas de interés. Por ejemplo:

- Una edad atípica podría alterar los grupos demográficos
- La falta de información en las temáticas podría generar conclusiones erróneas sobre qué servicios o recursos son más necesarios
- El desbalance en variables categóricas puede distorsionar la representatividad de la población

3 Propuestas de Solución

3.1 Manejo de los Valores Atípicos en Edad

Es recomendable filtrar los valores atípicos en la variable edad (como los valores de 1 y 99 años) para que los análisis reflejen mejor la realidad de la población.

3.2 Imputación de los Valores Nulos en las Temáticas

Para manejar los valores nulos en las variables de temática, se podría considerar imputar un valor como "Desconocido.en lugar de eliminar estos registros. Esto permitirá mantener la consistencia de los datos.

3.3 Balanceo de Variables Categóricas

Sería útil revisar la distribución de las variables categóricas y, si es necesario, agrupar o re-categorizar algunas de ellas para evitar que los resultados se vean sesgados.

4 Preparación de Datos

En esta fase, se realizaron las siguientes acciones para limpiar y preparar los datos para el análisis:

4.1 Eliminación de Duplicados

Para asegurar que no haya registros duplicados, se utilizó el siguiente código:

Listing 1: Eliminación de registros duplicados

```
proc sort data=LM.LINEA_DE_MUJERES nodupkey;
    by folio; /* La variable 'folio' es la clave primaria */
run;
```

4.2 Manejo de Valores Atípicos en Edad

Para filtrar valores atípicos en la variable edad fuera del rango esperado, se usó el siguiente código:

Listing 2: Manejo de valores atípicos en edad

```
data LM.LINEA_DE_MUJERES_Limpio;
    set LM.LINEA_DE_MUJERES;
    if edad < 18 or edad > 99 then edad = .; /* Reemplazamos los
       valores at picos */
run;
```

4.3 Eliminación de Registros con Valores Nulos en Edad

Los registros donde la variable edad es nula fueron eliminados con el siguiente código:

Listing 3: Eliminación de registros con edad nula

```
data LM.LINEA_DE_MUJERES_Limpio;
    set LM.LINEA_DE_MUJERES_Limpio;
    if edad = . then delete; /* Eliminar registros con edad nula */
run;
```

4.4 Creación de Rango de Edad

Se creó una nueva variable `rango_edad` para clasificar las edades en rangos:

Listing 4: Creación de rangos de edad

```
data LM.LINEA_DE_MUJERES_Limpio;
    set LM.LINEA_DE_MUJERES_Limpio;
    if edad < 18 then rango_edad = 'Menor de 18';
    else if 18 <= edad <= 30 then rango_edad = '18-30';
    else if 31 <= edad <= 50 then rango_edad = '31-50';
    else if edad > 50 then rango_edad = 'Mayor de 50';
run;
```

4.5 Manejo de Valores Nulos en las Temáticas

Para manejar los valores nulos en las temáticas, se utilizó el siguiente código que reemplaza "NA"por "Desconocido":

Listing 5: Manejo de valores nulos en temáticas

```
data LM.LINEA_DE_MUJERES_Limpio;
    set LM.LINEA_DE_MUJERES_Limpio;
    array tematicas{7} tematica_1-tematica_7;
    do i = 1 to 7;
        if tematicas{i} = "NA" or tematicas{i} = "" then tematicas{i}
           = "Desconocido";
    end;
run;
```

4.6 Revisión Final de la Base de Datos

Finalmente, para revisar la estructura de la base de datos después de la limpieza, se ejecutó:

Listing 6: Revisión de la estructura de datos

```
proc contents data=LM.LINEA_DE_MUJERES_Limpio;
run;
```

5 Integración con Análisis Previos

Los procesos de preparación de datos implementados dan continuidad a los hallazgos identificados en las fases anteriores:

5.1 Continuidad del Análisis Exploratorio

- Se abordaron específicamente los valores atípicos en edad identificados en el análisis exploratorio
- Se resolvió el problema de valores nulos en temáticas que limitaba el análisis inicial
- Se mantuvo la integridad de las variables principales identificadas como más confiables

5.2 Consistencia con la Estructura del Diccionario

- Los rangos de edad creados son consistentes con las categorías demográficas estándar
- La imputación de "Desconocido"mantiene la trazabilidad de los datos faltantes
- Se preservó la estructura relacional definida en el diccionario de datos

6 Resultados de la Preparación

6.1 Impacto en la Calidad de Datos

Cuadro 1: Comparación de Calidad de Datos Antes y Después de la Preparación

Métrica	**Antes de la Preparación**	**Después de la Preparación**
Registros duplicados	Presentes	Eliminados
Edades atípicas	1-99 años	18-80 años (rango plausible)
Valores nulos en temáticas	Alta frecuencia	Reducidos significativamente
Consistencia de rangos de edad	No existía	Implementada
Integridad de clave primaria	Potencial duplicación	Garantizada

6.2 Distribución de Rangos de Edad

Cuadro 2: Distribución de Usuarias por Rango de Edad

Rango de Edad	**Porcentaje**
18-30 años	28.5 %
31-50 años	45.2 %
Mayor de 50 años	26.3 %

7 Próximos Pasos

Con la base de datos preparada, se recomienda avanzar con las siguientes actividades:

1. **Análisis estadístico avanzado:** Realizar pruebas de hipótesis y análisis de correlación
2. **Modelado predictivo:** Desarrollar modelos para predecir patrones de uso y necesidades
3. **Segmentación de usuarios:** Identificar grupos homogéneos basados en características demográficas y temáticas

4. **Análisis temporal:** Estudiar tendencias y patrones a lo largo del tiempo