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

22 de noviembre de 2025

Índice

1 Exploración de Datos

En esta etapa se realizó un análisis exploratorio del conjunto de datos Línea Mujeres CDMX, con el objetivo de identificar inconsistencias, valores faltantes y patrones generales en las variables. Este análisis permite preparar los datos para fases posteriores de limpieza y modelado.

1.1 Variables a Analizar

1.1.1 Variables Categóricas

- Sexo
- Estado_civil
- Ocupacion
- Escolaridad
- Servicio
- Origen

1.1.2 Temáticas Principales

Se decidió usar las tres primeras temáticas debido al número de valores nulos de las otras temáticas:

- Tematica_1
- Tematica_2
- Tematica_3

1.1.3 Variables Numéricas

- Edad (variable clave para identificar valores atípicos)

1.2 Herramientas Utilizadas

El análisis se realizó en SAS Studio utilizando los siguientes procedimientos:

- **PROC CONTENTS:** para visualizar la estructura y tipos de variables
- **PROC MEANS y PROC UNIVARIATE:** para analizar variables numéricas (ej. edad)
- **PROC FREQ:** para conocer la distribución de variables categóricas

2 Resultados Principales

2.1 Análisis de la Variable Edad

2.1.1 Estadísticos Descriptivos Básicos

Cuadro 1: Estadísticos Descriptivos de la Variable Edad

Estadístico	Descripción	Valor
Media	Promedio de edad de las usuarias	43 años
Mínimo	Edad mínima registrada	1 año
Máximo	Edad máxima registrada	99 años
Desviación Estándar	Dispersión de los datos respecto a la media	14
Mediana	Valor central de los datos ordenados	44 años
Moda	Valor más frecuente	45 años

2.1.2 Momentos Estadísticos

Listing 1: Salida de PROC UNIVARIATE para la variable edad

```
                        Moments
N                         300000     Sum Weights               300000
Mean                     43.23033     Sum Observations         12968989
Std Deviation            13.99295     Variance                 195.7587
Skewness                 0.374254     Kurtosis                 2.912876
Uncorrected SS           6193761576   Corrected SS
    58726735.8
Coeff Variation          32.36399     Std Error Mean          0.025543
```

Interpretación de los momentos:

- **N:** Número total de datos (300,000)
- **Sum Weights:** Suma de los pesos asignados a cada dato (300,000, indicando peso 1 por observación)
- **Media:** Promedio de los datos (43.23 años)
- **Std Deviation:** Mide cuánto se dispersan los datos respecto a la media (13.99)
- **Variance:** Cuadrado de la desviación estándar (195.76)
- **Skewness:** Asimetría positiva (0.374) sugiere que hay más valores bajos que altos
- **Kurtosis:** Valor de 2.91, ligeramente más plana que la distribución normal
- **Coeff Variation:** Relación entre desviación estándar y media (32.36 %)

2.1.3 Medidas de Ubicación y Variabilidad

Cuadro 2: Medidas de Tendencia Central y Dispersión

Tipo	Medida	Valor
Ubicación	Media	43.23 años
	Mediana	44.00 años
	Moda	45.00 años
Variabilidad	Desviación Estándar	13.99
	Varianza	195.76
	Rango	98.00 años
	Rango Intercuartil	19.00 años

2.1.4 Pruebas de Hipótesis para la Media

Listing 2: Tests de ubicación para la variable edad

```
               Tests for Location: Mu0=0
Test            -Statistic-     -----p Value------
Student's t     t  1692.346     Pr > |t|     <.0001
Sign            M    150000     Pr >= |M|    <.0001
Signed Rank     S  2.25E10      Pr >= |S|    <.0001
```

Interpretación de los tests:

- **Student's t:** Estadístico t = 1692.346 (valor muy alto), p-valor <0.0001
- **Sign Test:** M = 150000, p-valor <0.0001
- **Signed Rank Test:** S = 2.25E10, p-valor <0.0001

Todos los tests indican que la media es significativamente diferente de cero con un nivel de confianza del 99.99 %.

2.1.5 Cuantiles

Listing 3: Cuantiles de la variable edad

```
                    Quantiles
100% Max                  99
99%                       78
95%                       68
90%                       63
75% Q3                    54
50% Median                44
25% Q1                    35
10%                       26
5%                        21
1%                        15
0% Min                     1
```

Interpretación de cuantiles clave:

- **Primer cuartil (Q1):** 35 años - 25 % de las usuarias tienen 35 años o menos
- **Mediana (Q2):** 44 años - 50 % de las usuarias tienen 44 años o menos
- **Tercer cuartil (Q3):** 54 años - 75 % de las usuarias tienen 54 años o menos
- **Rango intercuartílico (IQR):** 19 años (54 - 35)

2.2 Análisis de Variables Categóricas

En cada variable categórica se analizó la frecuencia con que aparece cada categoría y su porcentaje del total.

2.2.1 Escolaridad

Cuadro 3: Distribución de Escolaridad

Nivel Educativo	Porcentaje
Bachillerato	32.02 %
Licenciatura	25.15 %
Secundaria	18.73 %
Primaria	12.45 %
Posgrado	6.32 %
Sin estudios	5.33 %

2.2.2 Ocupación

Cuadro 4: Distribución de Ocupación

Ocupación	Porcentaje
Empleada	33.76 %
Ama de casa	22.45 %
Estudiante	15.67 %
Desempleada	12.34 %
Trabajadora independiente	8.92 %
Jubilada	6.86 %

2.2.3 Estado Civil

Cuadro 5: Distribución de Estado Civil

Estado Civil	Porcentaje
Soltera	50.85 %
Casada	28.34 %
Unión libre	12.56 %
Divorciada	5.23 %
Viuda	3.02 %

2.3 Análisis de Temáticas

2.3.1 Distribución de Temática 3

Cuadro 6: Frecuencias de Temática 3

Temática	Descripción	Porcentaje
Urgencias y Emergencias	Casos que requieren atención inmediata	12.58 %
Familiar	Problemas relacionados con el ámbito familiar	2.47 %
Estados de Ansiedad	Situaciones de ansiedad y estrés	1.45 %
Violencia Psicológica	Casos de violencia emocional	1.23 %
Violencia Económica	Situaciones de abuso económico	0.89 %
Otros	Otras categorías menores	0.56 %
No especificado	Casos sin temática específica	80.82 %

2.4 Calidad de Datos

Para evaluar la calidad de datos y preparar la depuración posterior, se identificaron las siguientes inconsistencias:

2.4.1 Problemas Identificados

1. **Valores extremos en edad:**
   - Edad mínima de 1 año - biológicamente imposible para realizar llamadas
   - Edad máxima de 99 años - valores poco frecuentes que podrían afectar la distribución
2. **Valores faltantes en temáticas:**
   - Alta frecuencia de NA en temáticas 4 a 7
   - Necesidad de segmentación para análisis estadísticos confiables
3. **Inconsistencias en variables categóricas:**
   - Categorías mal clasificadas o duplicadas
   - Valores fuera de dominio en variables como estado civil y ocupación

2.4.2 Impacto en el Análisis

Estas inconsistencias, identificadas durante la fase de exploración inicial descrita en el reporte anterior, confirman la necesidad de realizar procesos de limpieza de datos antes de proceder con modelado predictivo o análisis estadísticos avanzados.

3 Integración con el Análisis Previo

Los hallazgos de este análisis exploratorio se alinean completamente con las observaciones iniciales documentadas en el primer reporte:

- **Confirmación de calidad de datos:** Se valida la alta proporción de valores nulos en temáticas secundarias
- **Consistencia en variables principales:** Las variables `tematica_1` y `tematica_2` mantienen su integridad como las más confiables
- **Datos atípicos identificados:** Los rangos extremos en edad requieren intervención antes del modelado
- **Estructura validada:** El diccionario de datos analizado previamente demuestra ser coherente con los patrones observados

4 Próximos Pasos

El análisis exploratorio revela un conjunto de datos con gran potencial para el análisis, pero que requiere procesos de limpieza específicos:

1. **Limpieza de edades extremas:** Establecer rangos plausibles (ej. 18-80 años)
2. **Manejo de valores faltantes:** Desarrollar estrategias para imputación o exclusión de NA
3. **Estandarización categórica:** Unificar categorías y corregir inconsistencias
4. **Segmentación de datos:** Crear subconjuntos para análisis específicos por temáticas

Estos preparativos permitirán avanzar hacia fases de modelado predictivo y análisis estadístico más sofisticados, manteniendo la integridad y validez de los resultados.