Informe de Calidad de Datos

Para solucionar esta prueba, se realiza una importación de datos para entenderlos de una manera adecuada por medio de Python.

Realizamos la importación de la tabla en Python llamada DataFrame:

```
[2] df_persona= pd.read_csv('/content/drive/MyDrive/Persona2.csv', sep=';')
    df_persona
```

	NumIdPersona	id_empresa	Genero	FechaNacimiento	Edad	Salario	Categoria	Segmento_poblacional	segmento_grupo_familiar	DepartamentoPersona	MunicipioPersona	EstratoPersona
0	1174205.0	56353	F	1/01/1968	52	1800000	b	Medio	NaN	NaN	NaN	NaN
1	800759.0	13625	M	10/11/1930	88	4139944	C	Medio	PAREJA CONYUGAL	DISTRITO CAPITAL	BOGOTA D.C.	5.0
2	809497.0	13922	M	24/10/1929	89	2052757	b	Medio	FAMILIA NUCLEAR INTEGRAL	DISTRITO CAPITAL	BOGOTA D.C.	5.0
3	849342.0	15448	M	5/11/1930	88	2331227	b	Medio	PAREJA CONYUGAL	NaN	NaN	NaN
4	896732.0	17649	M	11/11/1928	90	2372000	b	Medio	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
594186	1430358.0	59001	F	20/12/1993	26	877803	A	Joven	AFILIADO SIN GRUPO FAMILIAR	NaN	NaN	NaN
594187	1430632.0	59092	M	4/03/1989	30	828116	A	Basico	AFILIADO SIN GRUPO FAMILIAR	NaN	NaN	NaN
594188	1431245.0	42960	M	6/11/1991	28	829000	A	Joven	AFILIADO SIN GRUPO FAMILIAR	DISTRITO CAPITAL	BOGOTA D.C.	2.0
594189	1431565.0	41484	F	28/07/1995	24	828116	A	Basico	AFILIADO SIN GRUPO FAMILIAR	DISTRITO CAPITAL	BOGOTA D.C.	NaN
594190	1437508.0	45740	M	3/10/1974	45	877803	A	Basico	FAMILIA MONOPARENTERAL	NaN	NaN	NaN

594191 rows × 12 columns

1. Perfilamiento de la Base de Datos: Tipo de datos, mayor y mínimo valor, longitud del campo:

Tipo de datos: Se determina el tipo de datos de cada columna (float, int y object)

```
df_persona.dtypes
```

```
NumIdPersona              float64
id_empresa                  int64
Genero                     object
FechaNacimiento            object
Edad                        int64
Salario                     int64
Categoria                  object
Segmento_poblacional       object
segmento_grupo_familiar    object
DepartamentoPersona        object
MunicipioPersona           object
EstratoPersona            float64
dtype: object
```

Valores máximos y mínimos: Determinamos los valores máximos y mínimos de cada variable

```
Valores máximos:
NumIdPersona              1437512.0
id_empresa                    74045
Genero                     MASCULINO
FechaNacimiento            9/12/2000
Edad                             120
Salario                   1550000000
Categoria                         b
Segmento_poblacional          Medio
EstratoPersona                  6.0
dtype: object
----------------------------
Valores mínimos:
NumIdPersona                    4.0
id_empresa                        1
Genero                            1
FechaNacimiento           1/01/1927
Edad                              0
Salario                           0
Categoria                         A
Segmento_poblacional          Alto
EstratoPersona                  0.0
```

Longitud del campo: Determinamos la longitud de cada una de las variables en sus respectivas filas

	NumIdPersona	id_empresa	Genero	FechaNacimiento	Edad	Salario	Categoria	Segmento_poblacional	segmento_grupo_familiar	DepartamentoPersona	MunicipioPersona	EstratoPersona
0	9	5	1	9	2	7	1	5	3	3	3	3
1	8	5	1	10	2	7	1	5	15	16	11	3
2	8	5	1	10	2	7	1	5	24	16	11	3
3	8	5	1	9	2	7	1	5	15	3	3	3
4	8	5	1	10	2	7	1	5	3	3	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...
594186	9	5	1	10	2	6	1	5	27	3	3	3
594187	9	5	1	9	2	6	1	6	27	3	3	3
594188	9	5	1	9	2	6	1	5	27	16	11	3
594189	9	5	1	10	2	6	1	6	27	16	11	3
594190	9	5	1	9	2	6	1	6	22	3	3	3

594191 rows × 12 columns

2. Fallas por Variable:

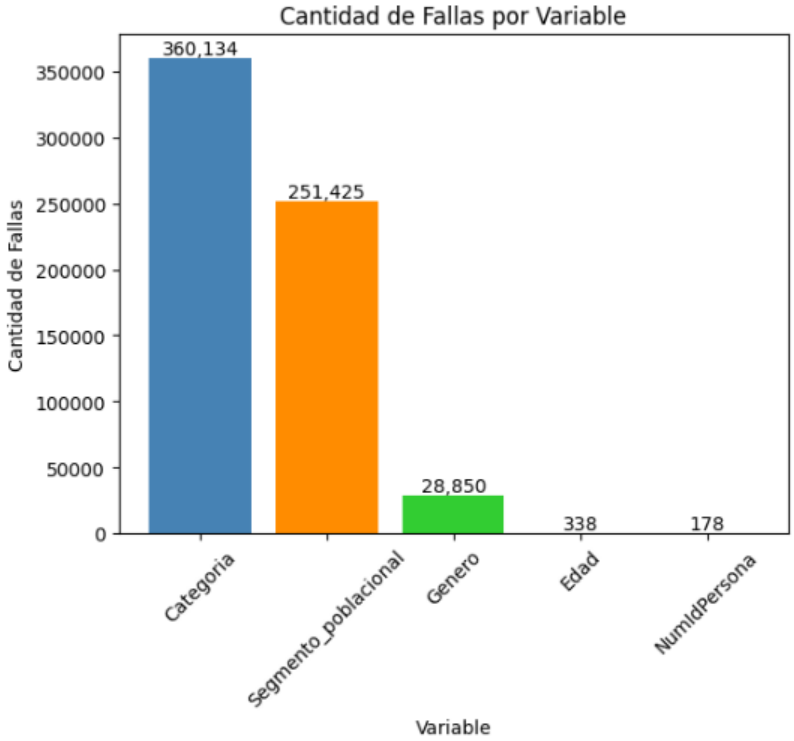
- NumIdPersona: 178 valores vacíos

- Genero: 28.850 valores que no tienen la denominación ‘F’ o ‘M’

- Segmento: 251.425 cantidad de valores que no están entre las opciones Alto, Básico, Joven y Medio (Los errores presentados fueron porque en la base de datos entregada el campo ‘Básico’ esta sin la tilde y esto causo todos los errores para esta variable)

- **Edad:** 338 cantidad de valores que no tienen 0 (30) o que su longitud en este campo es de más de dos cifras (308)

- **Categoria:** 360.134 cantidad de valores que no cumplen con las opciones del campo segmento_poblacional



Indicadores de Calidad:

1. Indicador de Calidad General:

- **Registros sin errores**: 322.596 registros

- **Total de registros:** 594.191 registros

- **Indicador de calidad general:** 54.3% de los datos se encuentran bien

2. Indicador de Calidad por Variable:

- NumIdPersona:

```
Número de campos llenos en 'NumIdPersona': 594,013
Porcentaje de campos diligenciados en 'NumIdPersona': 99.97 %
```

- Genero:

```
Número de campos validos en la columna 'Genero': 565,341
Porcentaje de campos validos en la columna 'Genero': 95.14 %
```

-**Segmento_poblacional**:

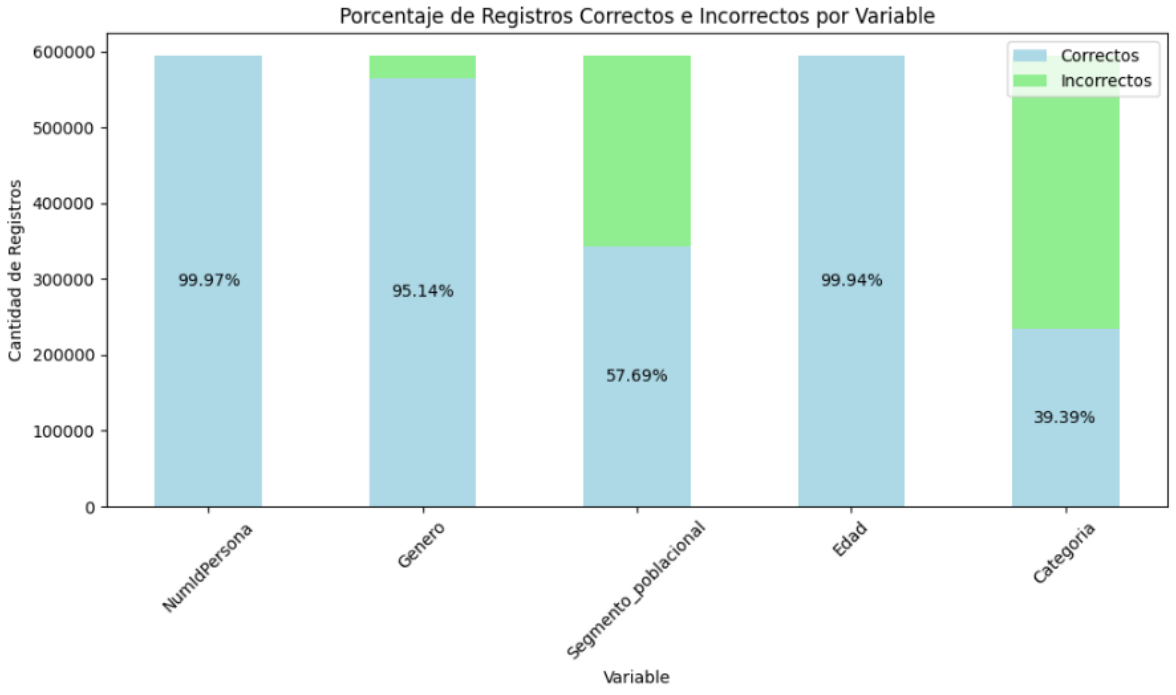
```
Número de campos validos en la columna 'Segmento poblacional': 342,766
Porcentaje de campos validos en la columna 'Segmento poblacional': 57.69 %
```

- Edad:

```
Número de campos validos en la columna 'edad': 593,853
Porcentaje de campos validos en la columna 'edad': 99.94 %
```

-Categoria:

```
Número de campos con conbinaciones validas: 234,057
Porcentaje de campos con combinaciones validas categoria y segmento_poblacional: 39.39 %
```



Proceso Desarrollado:

1. Perfilamiento de la Base de Datos:

- Se realizó un análisis de los tipos de datos presentes en cada variable, así como el mayor y mínimo valor y la longitud del campo.

2. Identificación de Fallas:

- Se detectaron y contabilizaron las fallas presentadas por cada variable, indicando la cantidad de fallas encontradas.

3. Cálculo de Indicadores de Calidad:

- Se determinó el número de registros sin errores y el total de registros, obteniendo el indicador de calidad general como el porcentaje de registros sin errores sobre el total.

4. Indicadores de Calidad por Variable:

- Se calculó el indicador de calidad por cada variable como el porcentaje de registros sin errores sobre el total de registros para esa variable.

ANALISIS DATOS OBTENIDOS

Basándonos en los datos proporcionados y los resultados adicionales, podemos profundizar en los análisis de calidad por variable de la siguiente manera:

1. NumIdPersona:

Indicador de calidad: El 99.97% de los campos en la variable NumIdPersona están llenos, lo que indica que casi todos los registros tienen valores en esta columna. Esto muestra un alto nivel de integridad en esta variable, con una mínima cantidad de valores faltantes.

2. Genero:

Indicador de calidad: El 94.14% de los valores en la variable Genero cumplen con la denominación correcta, es decir, son 'F' o 'M'. Esto indica que la mayoría de los registros tienen la información de género correctamente registrada. Sin embargo, existe un porcentaje pequeño de registros que no cumplen con esta regla de calidad.

3. Segmento_poblacional:

Indicador de calidad: El 57.69% de los datos en la variable Segmento_poblacional se encuentran correctamente diligenciados. Esta baja proporción indica que una gran cantidad de registros en esta variable no cumplen con las opciones permitidas (Alto, Básico, Joven y Medio). Además, se mencionó previamente que existieron errores relacionados con la falta de tilde en la opción 'Básico', lo que puede haber contribuido a esta baja calidad.

4. Edad:

Indicador de calidad: El 99.94% de los valores en la variable Edad se encuentran correctamente diligenciados. Esto significa que la mayoría de los registros tienen un valor válido para la edad, ya sea un número distinto de 0 o un número de dos cifras. La baja proporción de errores en esta variable indica un buen nivel de calidad en términos de integridad y formato de los datos.

5. Categoria:

Indicador de calidad: El 39.39% de los datos en la variable Categoria cumplen con la regla de negocio establecida. Esto indica que una gran proporción de registros en esta variable no cumplen con las opciones permitidas en el campo segmento_poblacional. Es importante destacar que este bajo indicador de calidad puede tener un impacto significativo en la integridad y coherencia de los datos en relación con las categorías y segmentos poblacionales asociados.

En resumen, estos análisis de calidad nos brindan una visión detallada de las fallas presentadas por variable y el indicador general de calidad de los datos. Estos resultados resaltan las áreas específicas que requieren atención y corrección para mejorar la integridad y confiabilidad de la base de datos. Se recomienda realizar acciones correctivas y de limpieza de datos en las variables identificadas con un bajo nivel de calidad, garantizando así la consistencia y confiabilidad de los datos utilizados para el análisis y toma de decisiones.