

## Laboratorio 27

### Sesión #27 Limpieza, estadística descriptiva

**Título del Laboratorio:** Limpieza y estadística descriptiva de los datos en los productos de licor registrados en Risaralda para la distribución.

**Duración:** 2 horas

**Objetivos del Laboratorio:**

1. *Llevar a cabo un proceso completo de limpieza de datos y aplicar técnicas de estadística descriptiva al dataset de productos de licor, con el fin de mejorar la calidad de los datos y proporcionar insights significativos que puedan ser utilizados para la toma de decisiones estratégicas.*
2. *Esto incluye identificar y resolver problemas de datos como valores nulos, duplicados e inconsistencias, y resumir las principales características del dataset mediante medidas estadísticas que faciliten la comprensión y análisis de la información*

**Materiales Necesarios:**

1. *Computador con conexión a internet.*
2. *Herramientas de limpieza de los datos y aplicación de estadística descriptiva.*

**Estructura del Laboratorio:**

1. Entregar un informe escrito que detalle el paso a paso, debe incluir las primeras observaciones sobre la estructura de los datos, los hallazgos, conclusiones y recomendaciones, se debe entregar en PDF o Word.
2. Responder las preguntas planteadas.
  - a) ¿Qué información se debe analizar en el Dataset para garantizar el cumplimiento de las normativas sanitarias y optimizar la distribución de los productos?
    - ❖ **Fecha de caducidad y condiciones de almacenamiento:** Asegurar que los productos cumplan con las normativas de caducidad y se mantengan en condiciones adecuadas durante su almacenamiento y distribución.
    - ❖ **Registros de calidad e inspección:** Verificar los controles de calidad y las inspecciones sanitarias realizadas a los productos para asegurar que cumplen con las regulaciones.
    - ❖ **Distribución de productos:** Analizar la distribución de los productos para garantizar que se cumplan las normativas sanitarias locales o regionales.
    - ❖ **Inventario y movimientos de productos:** Mantener el inventario actualizado para evitar productos caducados o deteriorados en el proceso de distribución.
    - ❖ **Datos de transporte:** Asegurar que el transporte de los productos cumpla con las normativas sanitarias, como los requisitos de temperatura o higiene.

- b) ¿Cómo pueden los valores nulos y las inconsistencias en el texto afectar la precisión y calidad del análisis de los datos?
- ❖ **Valores nulos:** Los valores nulos pueden causar vacíos de información crucial, como fechas de caducidad o condiciones de almacenamiento, lo que puede afectar el cumplimiento normativo y la distribución adecuada de productos. Es necesario aplicar técnicas de limpieza de datos para manejar estos valores, ya sea imputando o eliminando registros incompletos.
  - ❖ **Inconsistencias en el texto:** Las inconsistencias en el registro de datos (como nombres de productos o fechas incorrectas) pueden afectar la precisión del análisis, creando confusión o errores en el control de inventarios y el análisis de tendencias. Normalizar los datos y asegurar un formato consistente ayuda a mitigar estos problemas.
- c) ¿Qué estrategias se pueden implementar para asegurar que el análisis de datos sea preciso y útil para la toma de decisiones estratégicas?
- ❖ **Limpieza de datos:** Implementar un proceso robusto de limpieza para corregir valores nulos, inconsistencias y errores en los datos.
  - ❖ **Validación de datos:** Realizar validaciones durante la captura de datos para garantizar su precisión y consistencia desde el inicio.
  - ❖ **Análisis predictivo:** Usar herramientas analíticas avanzadas para detectar patrones y prever problemas, como una baja en la demanda o un incumplimiento de las normativas.
  - ❖ **Alertas automáticas:** Establecer alertas que notifiquen sobre posibles incumplimientos de normas o problemas en la cadena de distribución.
  - ❖ **Visualización de datos:** Utilizar gráficos y dashboards interactivos para facilitar la toma de decisiones, presentando la información crítica de manera clara y accesible.
  - ❖ **Evaluación continua del cumplimiento:** Realizar auditorías periódicas para garantizar que los procesos se mantengan en cumplimiento con las normativas y detectar áreas de mejora.

### 3. Adjunta el Dataset, los script o archivo de la exploración.

```
[ ] # Valores antes de la limpieza

print("Valores antes de la limpieza: ")
print(df.head())
```

Valores antes de la limpieza:

	ORIGEN	PRODUCTO	GRADO DE ALCOHOL
0	N	Ap. Brissart Sabor A Cafe	10.0
1	N	Ap. Crema Con Sabor A Chocolate M Harv	14.0
2	N	Ap. Crema Sab Whisky Harvey Mackys 14	14.0
3	N	Ap. Crema Sab Whisky Harvey Mackys 14	14.0
4	N	Ap. Crema Sab. A Ron Harvey Mackys	14.0

REGISTRO SANITARIO VIGENCIA DE REGISTRO SANITARIO

	REGISTRO SANITARIO	VIGENCIA DE REGISTRO SANITARIO
0	INVIMA 2019L-0010049	19/06/2029
1	INVIMA 2016L-0008145	14/04/2026
2	INVIMA 2017L-0008812	17/07/2027
3	INVIMA 2017L-0008812	17/07/2027
4	INVIMA 2014L-0007076	12/05/2024

RESOLUCIÓN DE REGISTRO SANITARIO PRODUCTOR

	RESOLUCIÓN DE REGISTRO SANITARIO	PRODUCTOR
0	19022399.0	LICORES BRISSART S.A.S
1	16009721.0	CANDIOTA DE VINOS Y LICORES S.A.
2	17025467.0	CANDIOTA DE VINOS Y LICORES S.A.
3	17025467.0	CANDIOTA DE VINOS Y LICORES S.A.
4	14009708.0	CANDIOTA DE VINOS Y LICORES S.A.

NOMBRE EMPRESA DISTRIBUIDORA NIT

	NOMBRE EMPRESA DISTRIBUIDORA	NIT
0	RODRIGUEZ Q. ALEXANDER/38M DIST BRISSART	100278640
1	RODRIGUEZ Q. ALEXANDER/38M DIST BRISSART	100278640
2	RODRIGUEZ Q. ALEXANDER/38M DIST BRISSART	100278640
3	RODRIGUEZ Q. ALEXANDER/38M DIST BRISSART	100278640
4	RODRIGUEZ Q. ALEXANDER/38M DIST BRISSART	100278640

```
[ ] nulos_antes = df.isnull().sum()
duplicados_antes = df.duplicated().sum()

print("Duplicados y valores nulos antes de la limpieza: ")
print(f"Duplicados: {duplicados_antes}")
print(f"Valores nulos: {nulos_antes}")
```

Duplicados y valores nulos antes de la limpieza:

Duplicados: 2574

Valores nulos:

	ORIGEN	PRODUCTO	GRADO DE ALCOHOL	REGISTRO SANITARIO	VIGENCIA DE REGISTRO SANITARIO	RESOLUCIÓN DE REGISTRO SANITARIO	PRODUCTOR	NOMBRE EMPRESA DISTRIBUIDORA	NIT
0	0	0	0	134	0	6306	168	0	0

dtype: int64

```
[30] # Empezamos con la limpieza

# a. Eliminar filas con valores nulos en columnas críticas

columnas_criticas = ['ORIGEN', 'PRODUCTO', 'GRADO DE ALCOHOL']
df = df.dropna(subset=columnas_criticas)
```

[ ] nulos\_antes = df.isnull().sum()

print(nulos\_antes)

ORIGEN 0

PRODUCTO 0

GRADO DE ALCOHOL 0

REGISTRO SANITARIO 0

VIGENCIA DE REGISTRO SANITARIO 0

RESOLUCIÓN DE REGISTRO SANITARIO 6270

PRODUCTOR 168

NOMBRE EMPRESA DISTRIBUIDORA 0

NIT 0

dtype: int64

```
[33] # Rellenar valores nulos en caso de ser necesario
```

```
df['NOMBRE EMPRESA DISTRIBUIDORA'].fillna('Desconocido', inplace=True)
```

```
print(df.isnull().sum())
```

```
ORIGEN      0
PRODUCTO    0
GRADO DE ALCOHOL  0
REGISTRO SANITARIO  0
VIGENCIA DE REGISTRO SANITARIO  0
RESOLUCIÓN DE REGISTRO SANITARIO  6270
PRODUCTOR    168
NOMBRE EMPRESA DISTRIBUIDORA  0
NIT          0
dtype: int64
```

<ipython-input-33-046489e7c87b>:3: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series resulting in inplace modification. The behavior will change in pandas 3.0. This inplace method will never work

For example, when doing 'df[col].method(value, inplace=True)', try using 'd

```
df['NOMBRE EMPRESA DISTRIBUIDORA'].fillna('Desconocido', inplace=True)
```

```
df=df.dropna()
nulos_despues = df.isnull().sum()
print(nulos_despues)
```

```
ORIGEN      0
PRODUCTO    0
GRADO DE ALCOHOL  0
REGISTRO SANITARIO  0
VIGENCIA DE REGISTRO SANITARIO  0
RESOLUCIÓN DE REGISTRO SANITARIO  0
PRODUCTOR    0
NOMBRE EMPRESA DISTRIBUIDORA  0
NIT          0
dtype: int64
```