LAB RETO 2

```
Datos 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  \
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  \
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
0  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
1  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
2  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
3  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
4  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
```

```
Valores nulos por columna antes de la limpieza:
ORIGEN                                0
PRODUCTO                              0
GRADO DE ALCOHOL                      0
REGISTRO SANITARIO                  134
VIGENCIA DE REGISTRO SANITARIO        0
RESOLUCIÓN DE REGISTRO SANITARIO   6306
PRODUCTOR                           168
NOMBRE EMPRESA DISTRIBUIDORA          0
NIT                                   0
dtype: int64

Número de filas duplicadas antes de la limpieza:
2574
```

```
Datos después 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
4      N      Ap. Crema Sab. A Ron Harvey Mackys              14.0
6      N       Ap. Crema Sabor A Whisky Brissart              14.0

     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
4  INVIMA 2014L-0007076                     12/05/2024
6  INVIMA 2019L-0009865                      8/03/2029

   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.
4                        14009708.0  CANDIOTA DE VINOS Y LICORES S.A.
6                        19005235.0           LICORES BRISSART S.A.S

              NOMBRE EMPRESA DISTRIBUIDORA        NIT
0  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
1  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
2  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
4  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640
6  RODRIGUEZ Q. ALEXANDER/J&M DIST BRISSART  100278640

--- Resumen de la Limpieza ---
Total de valores nulos antes de la limpieza: 6608
Total de valores nulos después de la limpieza: 0
Total de filas duplicadas antes de la limpieza: 2574
Total de filas duplicadas después de la limpieza: 0
Total de filas eliminadas: 0

El dataset limpio ha sido guardado como 'Productos_licores_limpiados.csv'.
```

Dataset: Productos Licores.csv

Informe de Análisis de Datos: Limpieza y Estadística Descriptiva del Dataset de Productos de Licor

El proyecto se enfoca en limpiar y analizar estadísticamente el conjunto de datos **"Productos_licores.csv"** que incluye información importante sobre licores distribuidos en el Departamento de Risaralda. El objetivo es garantizar que los datos sean de alta calidad y ofrecer información importante para tomar decisiones sobre el cumplimiento de normas y mejorar la distribución

Paso a Paso del Análisis

Cargar y Explorar el Dataset

Se utilizó Python (pandas) para cargar el archivo Productos_licores.csv y se visualizaron las primeras filas del conjunto de datos. Durante la exploración inicial se analizó la estructura de los datos y se identificaron las variables principales: origen, producto, grado de alcohol, registro sanitario, vigencia de registro sanitario, resolución de registro sanitario, productor, nombre de la empresa distribuidora y NIT. Se detectaron y se registraron los valores en blanco en cada columna y se comprobó cuántas filas estaban repetidas. Se encontraron columnas críticas con información faltante y se identificaron 2,574 filas duplicadas.

Proceso de Limpieza de Datos

- **Eliminar valores nulos en columnas críticas:** Es importante asegurarse de que las columnas críticas como ORIGEN, PRODUCTO, GRADO DE ALCOHOL y REGISTRO SANITARIO no tengan valores nulos. Se eliminaron las filas con valores nulos en estas columnas para asegurar la exactitud del análisis.
- **Rellenar Valores Nulos en Columnas No Críticas:** En casos donde era necesario, se decidió reemplazar los valores nulos en columnas como RESOLUCIÓN DE REGISTRO SANITARIO y PRODUCTOR por "Desconocido", con el fin de garantizar que los datos siguieran siendo útiles y no se perdiera información relevante.
- **Eliminar Filas Duplicadas:** Se eliminaron las filas repetidas para garantizar la precisión de los datos y evitar posibles errores en los resultados.
- **Conversión de Tipos de Datos:** Se convirtió la columna "GRADO DE ALCOHOL" a un tipo numérico, corrigiendo cualquier error para garantizar que los datos se interpretaran de forma precisa.
- **Revisión Final de Valores Nulos:** Se llevó a cabo una revisión final para verificar que no hubiera valores nulos después de las conversiones y que los datos estuvieran preparados para su análisis.

Estadística Descriptiva y Resumen de la Limpieza

- **Generar Resumen de la Limpieza:** Se indicaron cuántos valores estaban vacíos y cuántas filas estaban duplicadas, tanto antes como después de

realizar la limpieza de los datos. Esto ayudó a entender cómo la limpieza afectó al conjunto de datos.

- **Total de Filas Eliminadas:** Se determinó y se exhibió la cantidad total de filas eliminadas, lo que demostró el nivel de depuración y cómo esta contribuyó a mejorar la calidad del conjunto de datos.

Hallazgos y Conclusiones

Hallazgos

- Las columnas ORIGEN y PRODUCTO son esenciales para separar y estudiar los productos de licor de forma eficiente.
- Se eliminaron muchos valores faltantes y duplicados, lo que mejoró la calidad de los datos de forma notable.
- Es necesario desarrollar estrategias específicas basadas en el contenido de alcohol para optimizar la distribución, dado que la variabilidad en el grado de alcohol es importante.

Conclusiones:

Limpiar los datos es fundamental para obtener información precisa y confiable. Se realizó un proceso para asegurar que el conjunto de datos esté libre de errores importantes, lo que permitirá llevar a cabo un análisis estadístico descriptivo preciso y útil para tomar decisiones estratégicas.

Recomendaciones

Mejorando constantemente el proceso de limpieza: Se recomienda utilizar procesos automatizados para detectar y manejar valores nulos y duplicados en próximas recolecciones de datos.

Normalización de datos de texto: Garantizar que la entrada de datos sea consistente para evitar errores de capitalización y tipográficos.

Seguimiento de la Vigencia del Registro Sanitario: Crear un sistema de alerta para mantener actualizados los registros sanitarios de los productos, garantizando el cumplimiento de las normativas vigentes.

Uso de Herramientas Complementarias:

Excel: Para una limpieza rápida y tareas manuales.

Python en Google Colab: Para análisis avanzados y automatización de procesos.

Formular preguntas iniciales para guiar el análisis.

1. **¿Qué información se debe analizar en el dataset para garantizar el cumplimiento de las normativas sanitarias y optimizar la distribución de los productos?**
   Es importante verificar que el registro sanitario de cada producto esté al día y cumpla con las leyes vigentes. Es fundamental verificar de dónde proviene el producto y cuánto alcohol contiene, ya que estos aspectos influyen en cómo se comercializa y distribuye.

2. **¿Cómo pueden los valores nulos y las inconsistencias en el texto afectar la precisión y calidad del análisis de los datos?**
   Los valores faltantes pueden causar errores en las conclusiones si no se gestionan correctamente, especialmente en columnas importantes como "RESOLUCIÓN DE REGISTRO SANITARIO" y "PRODUCTOR". Los errores en el texto dificultan organizar y analizar los datos, lo que dificulta la identificación de patrones y la realización de comparaciones precisas.

3. **¿Qué estrategias se pueden implementar para asegurar que el análisis de datos sea preciso y útil para la toma de decisiones estratégicas?**
   Es necesario realizar una limpieza completa de los datos, eliminando filas repetidas, completando los valores faltantes y estandarizando el formato del texto para garantizar la coherencia. También se pueden utilizar métodos de estadística descriptiva para resumir y entender las principales características del conjunto de datos, lo que ayudará a tomar decisiones informadas.