SEM309 – 10160 SEMINARIO DE PRÁCTICA EN CIENCIA DE DATOS

TRABAJO PRÁCTICO N° 4

RESEÑA DE VINOS – VISUALIZACIÓN E INTERPRETACIÓN DE DATOS

MARIANO NATIELLO (VLDC000632)

NOVIEMBEE 2024

ÍNDICE

1. CONTEXTO

El mundo del vino, como otras industrias, está cambiando de una forma más rápida de la esperada. Podemos mencionar a Big Data como uno de los impulsores de este cambio, apoyado por ejemplo, en el uso de herramientas de análisis de datos. Cada día, más empresas vinícolas de gran envergadura se apoyan en la recopilación y el análisis de grandes cantidades de información para la toma de decisiones (en tiempo real).

En su oportunidad, se realizó un estudio que contiene 130.000 reseñas de vinos, incluyendo información sobre el país origen del vino, una descripción, bodega que produjo el vino, viñedo dentro de la misma, variedad de uva, catador que realizó la reseña, puntuación que recibió el vino, entre otros datos.

2. OBJETIVO

Considerando lo elaborado a nivel del trabajo práctico N° 2, el presente trabajo práctico tiene como objetivo principal explorar y aplicar técnicas avanzadas de visualización de datos utilizando herramientas como Python o R, la interpretación de dichas visualizaciones cuyas conclusiones actúen como guía en la toma de decisiones tal como fue señalado al inicio del presente documento.

3. ENTORNO Y HERRAMIENTAS UTILIZADAS

En la realización del presente trabajo se utilizó Google Colab como entorno de desarrollo, Python como lenguaje principal y Google Drive como repositorio para almacenar los datos necesarios del proyecto.

4. ACTIVIDADES

Para poder dar cumplimiento al objetivo establecido, se realizaron una serie de operaciones que permitieron contar con datos depurados, identificar patrones y relaciones significativas y crear visualizaciones cuya interpretación refleje información útil para los productores y comerciantes de vino.

En primera instancia, se revisaron las principales características del dataset utilizado (Entregable-4-DocExtra.csv), fue posible ver que existen columnas relevantes como por ejemplo: country, province, region_1, variety, winery, price, points y description.

Dentro de los casi 130.000 filas que cuenta el dataset, se identificaron: i) variables numéricas (price, points); ii) variables categóricas (country, province) y variables de texto (description).

A continuación, detallamos las actividades y su correspondiente desarrollo:

1) Preparación de los datos

En este punto se realizaron los siguientes pasos:

a) Carga de datos

i) En Python, se ha especificado la ruta en Google Drive donde guardamos los datos a utilizar en el presente trabajo practico:

```
file_path = "/content/drive/MyDrive/SEMCDS21/Entregable-4-DocExtra.csv"
```

ii) Se leyó el archivo considerando comillas dobles para su correcto manejo, junto con la coma como delimitador.

```
df = pd.read_csv(file_path, delimiter=',', quotechar='"', header=None)
```

iii) Se dividió el archivo leído (Entregable-4-DocExtra.csv) en columnas de la siguiente manera:

En este caso, se utilizó una expresión regular para dividir por comas fuera de las comillas dobles. Esta expresión regular, buscó en particular “comas” que no se encuentren dentro de las comillas dobles.

```
df_dividido = df.iloc[:, 0].apply(lambda x: re.split(r',(?=(?:[^"]*"[^"]*")*[^"]*$)', x))
```

iv) Se presentó el número de columnas antes de realizar la división de columnas:

```
print(f"Cantidad de columnas antes de dividir: {df.shape[1]}")
```

v) Se convirtió la lista resultante en un DataFrame

```
df_dividido = pd.DataFrame(df_dividido.tolist())
```

vi) Se validó la cantidad de columnas luego de realizar la división de las mismas:

```
print(f"Cantidad de columnas después de dividir: {df_dividido.shape[1]}")
```

vii) Como resultado de lo anterior, se obtuvieron 16 columnas. Dado que las últimas dos columnas no presentaron datos, se han eliminado las mismas. Para esto, se seleccionaron todas las filas y todas las columnas excepto las dos últimas columnas:

```
df_dividido = df_dividido.iloc[:, :-2]
```

viii) Se guardó el DataFrame, con las columnas divididas, en un nuevo archivo que será nuestro insumo para las visualizaciones de datos:

```
output_path='/content/drive/MyDrive/SEMCDS21/Entregable-4-DSFinal.csv'
df_dividido.to_csv(output_path, index=False)
```

b) Revisión inicial de la estructura

i) Como parte de la misma, luego de haber guardado el dataset final (Entregable-4-DSFinal.csv) en el paso anterior, se actualizó la primera columna con el valor “ID” y se guardaron los cambios.

En este caso se abrió el dataset omitiendo la primera fila en la lectura mediante skiprows:

```
df2 = pd.read_csv(output_path,skiprows=1)
```

Actualizamos la columna con el nombre “ID”:

```
df2.rename(columns={df2.columns[0]: "ID"}, inplace=True)
```

Guardamos los cambios:

```
df2.to_csv(output_path, index=False)
```

ii) Se eliminaron las filas 18884, 18885, 21521 y 51400 ya que no presentaron un valor numérico en la primera columna (ID). Al realizar dicha operación se tuvo en cuenta que la cantidad de filas a eliminar (4) es poco significativa en relación al total de filas del conjunto de datos (129.976), representando solo el 0,31%.

Para lo anterior, se resto uno a la fila tal como aparece en Microsoft Excel con el objetivo de que corresponda a “skip_rows”:

```
skip_rows = [18884 - 1, 18885 - 1, 21521 - 1, 51400 - 1]
```

Se leyó el archivo .csv omitiendo las filas especificas y se guardaron los cambios:

```
df2 = pd.read_csv(output_path, skiprows=skip_rows)
df2.to_csv(output_path, index=False)
```

Previo a realizar esta operación y luego de realizar la misma, se verificaron la cantidades de registros para control utilizando len(df2)

Al ejecutar lo anterior, se obtienen los siguientes resultados:

- Cantidad de registros previo a quitar las 4 filas: 129975
- Cantidad de registros actuales: 129971

iii) Detectamos si en la columna “price” existe algún valor atípico

```
print("Valores atípicos en la variable 'price':")
df2['price'] = pd.to_numeric(df2['price'], errors='coerce')
print(df2['price'].max())
```

```
El valor máximo de 'price' es: 3300.0
El valor del campo 'ID' asociado al valor máximo es: 80290
Media: 35.272434270632424
Mediana: 25.0
Valor mínimo: 4.0
Valor máximo: 3300.0
```

Del resultado obtenido, se confirmó que el ID 80290 presenta el valor 3300 en la columna “price”. Considerando que la media es de 35.27 y la mediana de 25, sin duda que el hallazgo anterior es un valor atípico de la columna “price”.

c) Limpieza de datos

i) Se identificaron valores nulos en nuestra DataFrame actual:

```
Los valores nulos previo a la actualización son:
ID                          0
country                   481
description               948
designation             38224
points                    2065
price                     11281
province                  2704
region_1                  23445
region_2                  80416
taster_name               28714
taster_twitter_handle33691
title                     2896
variety                   2956
winery                    2956
```

ii) Reemplazamos los valores nulos por la mediana de los puntos (points):

```
df2['points'] = pd.to_numeric(df2['points'], errors='coerce')
```

iii) Reemplazamos los valores nulos por la mediana de los precios (price):

```
df2['price'] = pd.to_numeric(df2['price'], errors='coerce')
```

iv) Reemplazamos los valores nulos en las columnas country, description, designation, province, region_1, region_2, taster_name, taster_twitter_handle, title, variety y winery, con un valor utilizando la nomenclatura "Desconocido+nombre columna". Ejemplo "Desconocido_country" para el caso de la columna country.

```
df2.loc[:, 'country'] = df2['country'].fillna('Desconocido_country')
df2.loc[:, 'description'] = df2['country'].fillna('Desconocido_description')
df2.loc[:, 'designation'] = df2['country'].fillna('Desconocido_designation')
df2.loc[:, 'province'] = df2['country'].fillna('Desconocido_province')
df2.loc[:, 'region_1'] = df2['country'].fillna('Desconocido_region_1')
df2.loc[:, 'region_2'] = df2['country'].fillna('Desconocido_region_2')
df2.loc[:, 'taster_name'] = df2['country'].fillna('Desconocido_taster_name')
df2.loc[:, 'taster_twitter_handle'] =
df2['country'].fillna('Desconocido_taster_twitter_handle')
df2.loc[:, 'title'] = df2['country'].fillna('Desconocido_title')
df2.loc[:, 'variety'] = df2['country'].fillna('Desconocido_variety'
```

De esta manera, se buscó identificar de una forma simple aquellos datos con valores nulos.

v) Se guardaron los datos luego de los reemplazos realizados:

```
df2.to_csv(output_path, index=False)
```

vi) Listamos los valores nulos de cada columna luego de las actualizaciones:

```
print("Los valores nulos luego de actualizar son:")
for column in df2.columns:
        print(f"{column:<20}{df2[column].isnull().sum()}")
```

```
Los valores nulos luego de actualizar son:
ID                    0
country               0
description           0
designation           0
points                0
price                 0
province              0
region_1              0
region_2              0
taster_name           0
taster_twitter_handle0
title                 0
variety               0
winery                0
```

2) Contextualización de los datos y formulación de preguntas claves

Previo a contextualizar los datos y poder formular preguntas claves que guíen el análisis, ha sido fundamental enfocarse en el propósito para el cual los datos fueron generados y los objetivos que se desea alcanzar. Dependiendo del caso, dichos objetivos pueden incluir la realización de análisis descriptivos para identificar patrones y tendencias generales, la exploración de relaciones significativas entre variables, hasta inclusive la construcción de modelos predictivos que aporte un valor estratégico.

Como alternativas de análisis, se encuentra al análisis descriptivo y el automático. Recordemos que el primero corresponde al proceso de resumir y visualizar datos utilizando tablas y gráficos, mientras que el análisis automático está más orientado al uso de algoritmos y modelos para identificar patrones de forma automática, utilizándose herramientas como Machine Learning, IA, entre otros.

De estas alternativas, se decidió avanzar con el análisis descriptivo, utilizando al análisis estratégico como complemento para la toma de decisiones. A continuación, se detallaron una serie de preguntas con el fin de entender los datos y contextualizar los objetivos del análisis:

- **Pregunta 1:** ¿Cuál es la relación entre precio y puntuación del vino?

  Justificación: permite a los comerciantes, desde el lugar de consumidores, determinar si vale la pena pagar más por un vino de mayor puntuación, encontrando de esta manera la mejor relación precio-calidad. A nivel de los comerciantes de vino, esta pregunta puede identificar tendencias en las potenciales preferencias de los consumidores.

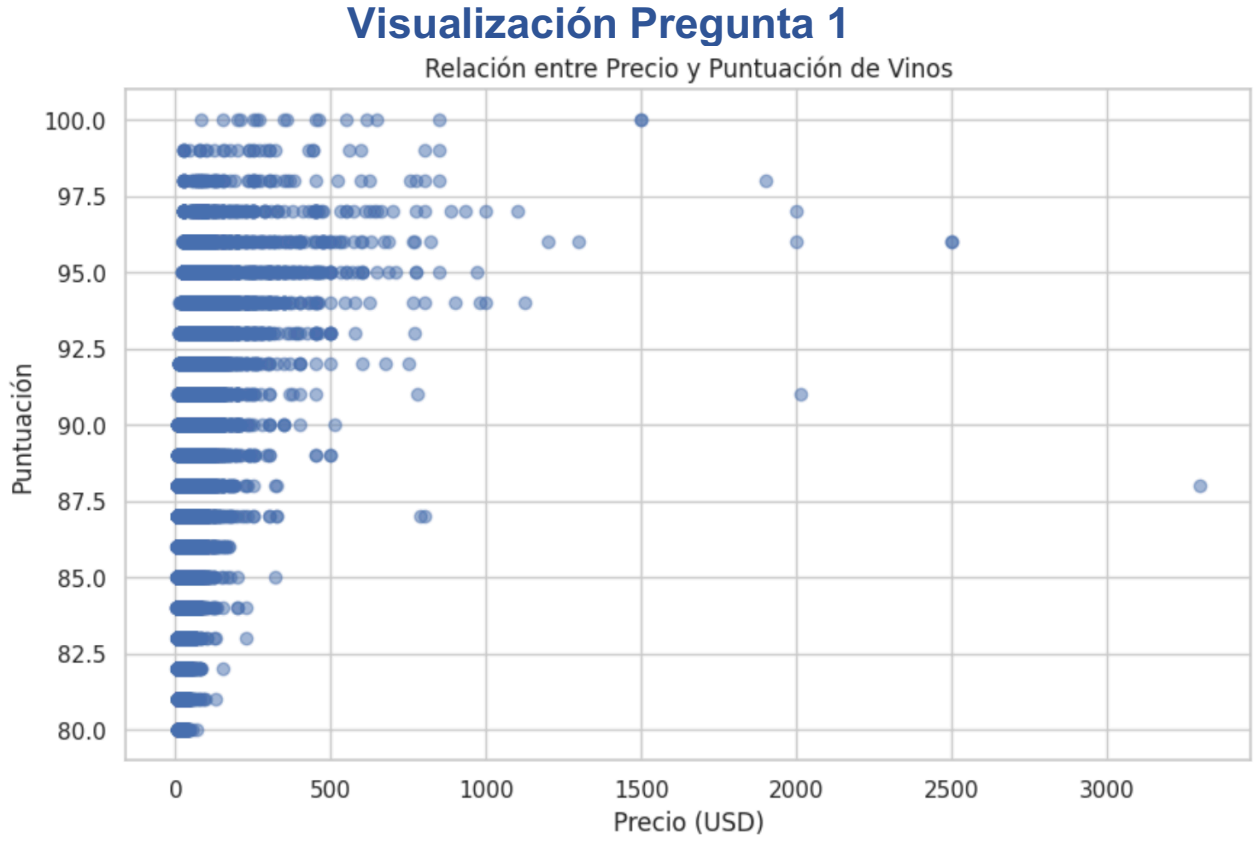
- **Pregunta 2:** ¿Cuál es la relación entre el puntaje y las variedades principales de vino?

  Justificación: conocer si las principales variedades de vino tienen una puntuación que supera los 80 puntos, permitiendo establecer una relación entre top 5 de las variedades y el puntaje recibido en la reseña.

- **Pregunta 3:** ¿Cuál es la distribución de precios de vino?

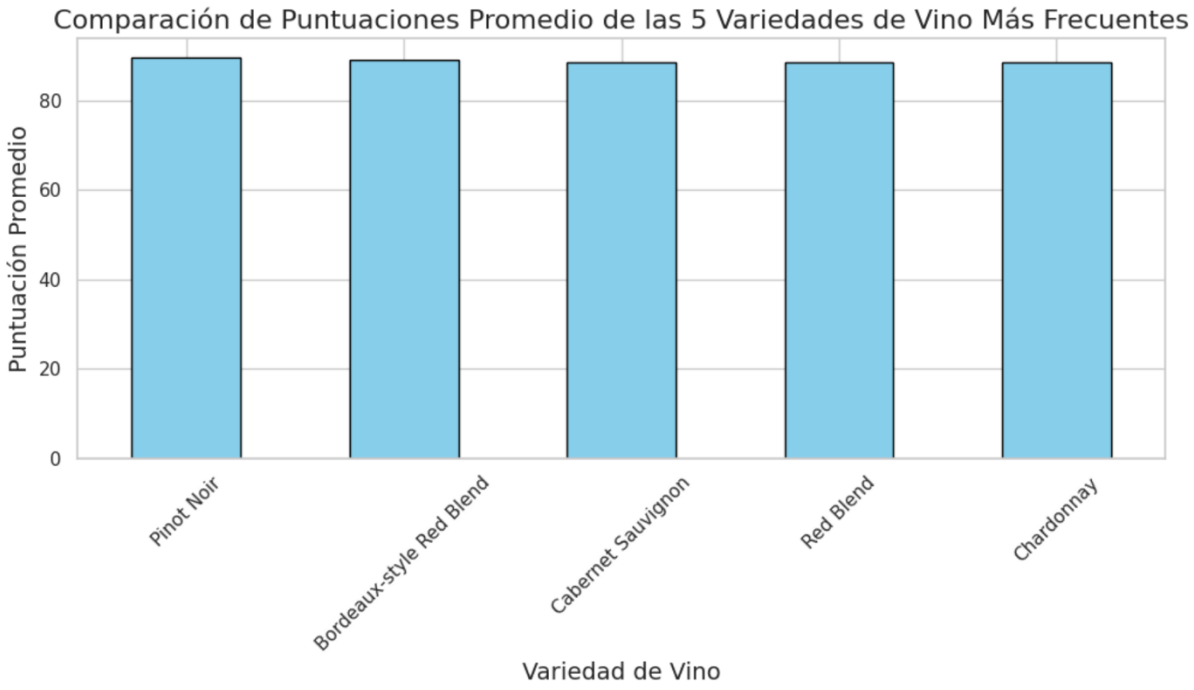
  Justificación: analizar la distribución de los precios de los vinos para identificar las categorías de precios más comunes y ayudar en la fijación de precios o segmentación de mercado.

5. VISUALIZACIONES

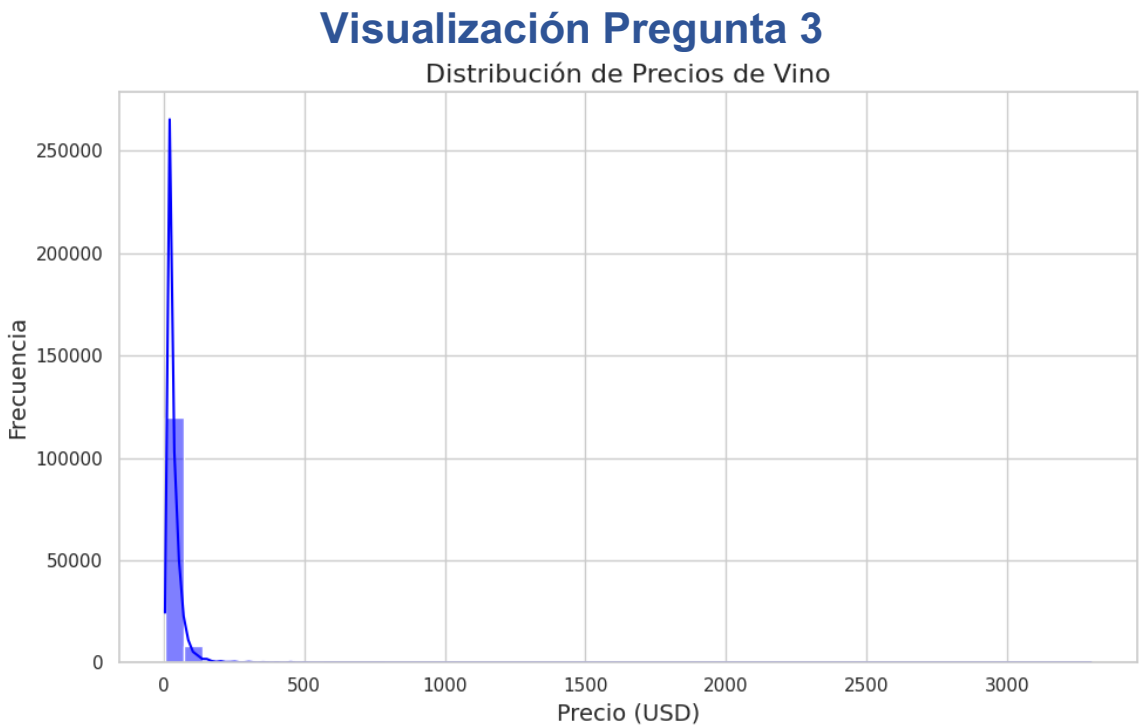


Interpretación: este gráfico de dispersión permite visualizar si existe una correlación entre el precio de los vinos y las puntuaciones que recibieron en la reseñas. Esto permite determinar si los vinos más caros tienden a obtener mejores calificaciones y la relación que puede existir entre precio-calidad.

Visualización Pregunta 2



Interpretación: este gráfico de línea permite establecer que existe una relación entre las 5 variedades mas frecuentes y el puntaje de la reseña, en donde se cumple que las 5 variedades principales de vino tienen un puntaje sobre 80 puntos.



Interpretación: en este histograma, es posible considerar que el pico en la frecuencia podría sugerir que el rango de precios es un estándar para una vino en particular.

6. CONCLUSIONES

Este análisis basado en las reseñas de los vinos, podría contribuir en la toma de decisiones principalmente a los comerciantes, pudiendo ayudar a comprender las tendencias y preferencias de los consumidores y de esa manera, optimizar su estrategia de marketing para aumentar las ventas y posicionamiento.

7. REFERENCIA BIBLIOGRÁFICA

Material de Lectura N° 4 de la materia Seminario de práctica en Ciencia de Datos de la carrera Licenciatura en Ciencias de Datos.

Enunciado del Trabajo Práctico N° 4 de la materia Seminario de práctica en Ciencia de Datos de la carrera Licenciatura en Ciencias de Datos.

Utilización de Google Colab.

Utilización de Microsoft Excel.