Descorchando Datos: Exploración de Reseñas de Vinos con Python para un Análisis Avanzado [TP2]

AUTOR: GERARDO JUAN MARTIN SERRANO GALLEGO

FECHA: 06/10/2024

ÍNDICE

Análisis Exploratorio de Datos - Estructura General de los Datos

El conjunto de datos contiene varias columnas que describen diferentes atributos de los vinos. Entre ellas se destacan las siguientes:

- **country**: País de origen del vino.
- **description**: Descripción textual del vino.
- **designation**: Nombre específico del vino en algunos casos.
- **points**: Puntuación otorgada al vino (rango de 80 a 100).
- **price**: Precio del vino.
- **province**: Provincia o región dentro del país.
- **region_1** y **region_2**: Subregiones dentro de la provincia.
- **variety**: Tipo de uva utilizada.
- **winery**: Nombre de la bodega.
- **taster_name**: Nombre del crítico que evaluó el vino.
- **taster_twitter_handle**: Identificador de Twitter del crítico.

Este conjunto de datos incluye variables categóricas y numéricas. Las variables categóricas, como country (país de origen) y variety (variedad de uva), proporcionan información cualitativa, mientras que variables numéricas como points (puntuación) y price (precio) permiten un análisis cuantitativo de las propiedades de los vinos.

1.2. Resumen Estadístico Inicial

Un resumen estadístico de las variables numéricas principales, como points y price, revela lo siguiente:

Puntos (points):

- Rango de valores: entre 80 y 100.
- Media: alrededor de 88 puntos.
- La mayoría de los vinos tiene puntuaciones altas, lo cual es característico de los conjuntos de datos de reseñas, donde los productos mal evaluados tienden a estar subrepresentados.

Precio (price):

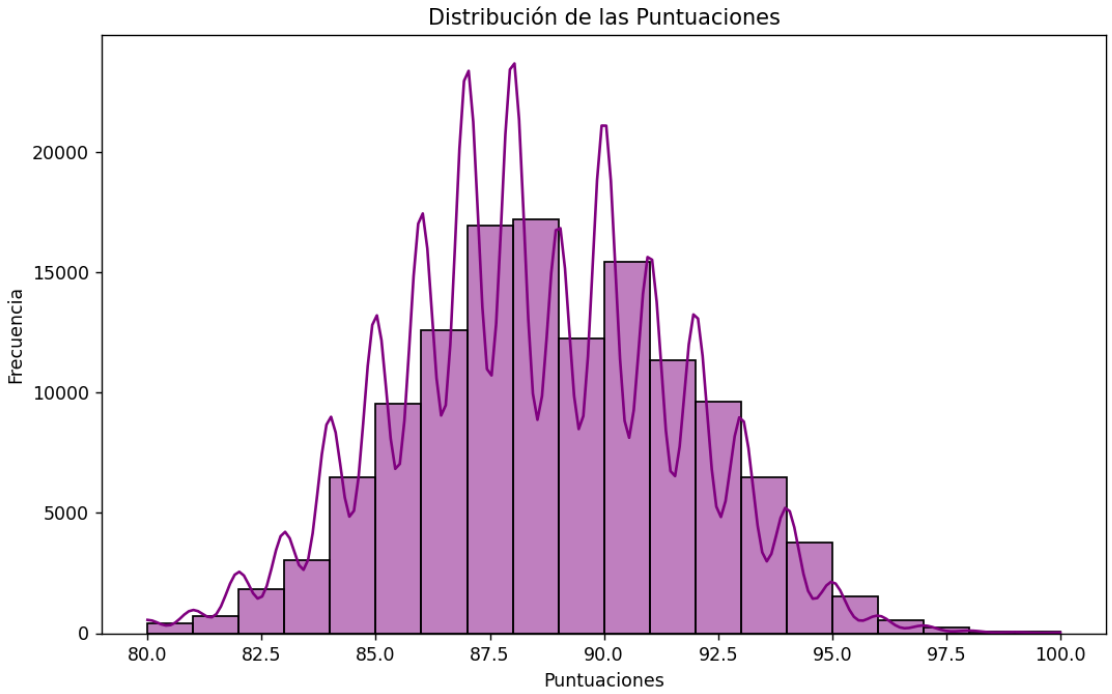
Existe una gran variabilidad en los precios de los vinos, pero también, se observan vinos con precios extremadamente altos, que podrían considerarse como valores atípicos.

1.3. Distribución de Variables Numéricas y Categóricas

Distribución de las puntuaciones:

La puntuación de los vinos sigue una distribución asimétrica hacia la derecha, concentrándose principalmente entre 85 y 92.5 puntos. La mayoría de los vinos obtiene puntuaciones elevadas, lo que indica que las evaluaciones son en su mayoría positivas.

Para ilustrar esta distribución, se utilizó un **histograma** con suavizado de densidad (kde), que mostró que el valor más común ronda los 87-88 puntos.



Análisis de la Distribución

Forma general de la distribución:

La distribución de las puntuaciones tiene una forma aproximada de campana, lo que indica que la mayoría de los vinos reciben una calificación en un rango intermedio. Esto sugiere que los críticos tienden a evitar puntuar vinos con valores extremadamente bajos o extremadamente altos, y la mayoría se concentra en un rango medio-alto.

Aunque la distribución parece cercana a una distribución normal, tiene ligeras irregularidades, con varios picos locales alrededor de las puntuaciones 87.5, 90 y 92.5. Estos picos sugieren que los críticos pueden estar utilizando escalas de calificación en estos puntos específicos con más frecuencia.

- **Pico principal**:

  El pico principal de la distribución se encuentra entre los 88 y 90 puntos, lo que indica que esta es la puntuación más frecuente otorgada a los vinos. Esta concentración en puntuaciones altas es un fenómeno común en los conjuntos de datos de reseñas, donde las malas evaluaciones suelen ser subrepresentadas.

- **Asimetría y cola hacia la derecha**:

  La distribución muestra una ligera asimetría con una cola más larga hacia la derecha, hacia las puntuaciones más altas (por encima de 90). Esto sugiere que algunos vinos excepcionales reciben puntuaciones cercanas a 100, aunque son significativamente menos frecuentes. Pero todo pareciera indicar que su presencia afecta la simetría de la distribución.

Interpretación y Utilidad

Este histograma es útil por varias razones:

1. **Comprensión de la calidad percibida de los vinos**:

   Al observar la concentración de puntuaciones entre 85 y 95, podemos deducir que la mayoría de los vinos son evaluados favorablemente, con pocos vinos recibiendo puntuaciones muy bajas. Esto también refleja la tendencia de los críticos a evaluar principalmente vinos de calidad media-alta.

2. **Picos locales**:

   Los picos en puntuaciones específicas como 87.5, 90 y 92.5 sugieren que ciertos puntos de referencia son comunes entre los críticos. Esto podría ayudar a entender cómo los críticos perciben la calidad de los vinos y cómo asignan calificaciones consistentes, como si hablar de variables cualitativas se tratase tales como “Decente, Muy Bueno, Excelente”.

3. **Análisis de patrones en la puntuación**:

   La asimetría y la cola hacia la derecha revelan que existen algunos vinos excepcionales que rompen con la tendencia general, recibiendo puntuaciones excepcionalmente altas. Estos vinos son menos frecuentes, pero podrían representar productos de nicho.

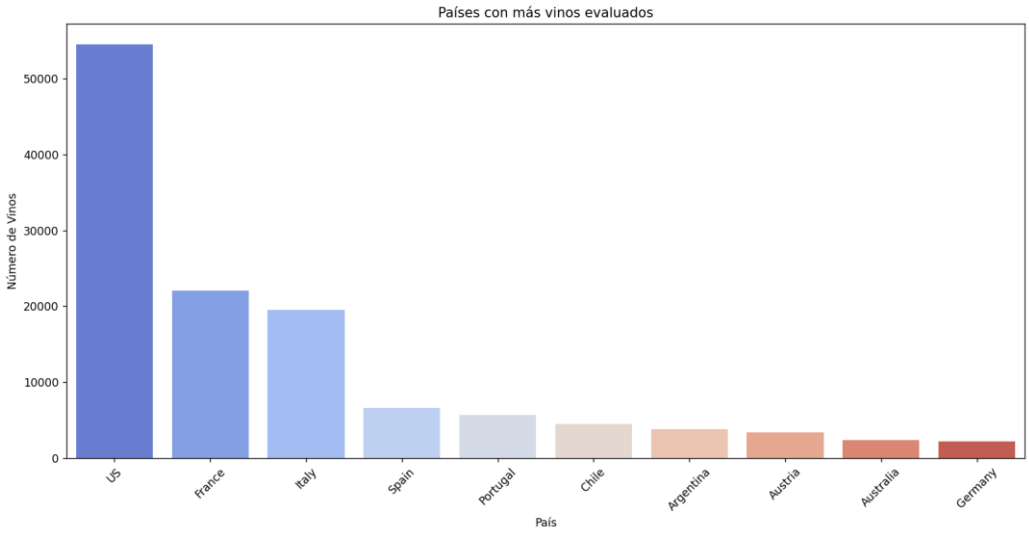
4. **Distribución para modelado**:

   Si se planea usar las puntuaciones como variable objetivo en futuros modelos de predicción (como para predecir la relación entre precio y puntuación), la distribución sesgada hacia puntuaciones altas puede requerir transformaciones para asegurar un mejor ajuste en los modelos.

Distribución de los países:

Se realizó un análisis de los países con más vinos evaluados. Los 10 países principales en términos de vinos evaluados incluyen grandes productores de vino como Estados Unidos, Francia, Italia, y España.

Un gráfico de barras muestra claramente que estos países dominan el mercado del vino en términos de cantidad de vinos evaluados.



Análisis de la Distribución

1. **Estados Unidos (US) como líder**:

   Estados Unidos lidera de forma clara con más de 50,000 vinos evaluados. Esta cifra es más del doble de la del segundo país en la lista, Francia, lo que indica una predominancia significativa de Estados Unidos en la base de datos. Esta representación puede estar relacionada con el hecho de que la base de datos puede tener una orientación hacia mercados estadounidenses o que la industria del vino en los EE.UU. es una de las más dinámicas en términos de volumen y diversidad de vinos evaluados.

2. **Francia e Italia**:

   Francia e Italia, que ocupan el segundo y tercer lugar, respectivamente, tienen cada uno alrededor de 20,000 vinos evaluados. Estos dos países son tradicionalmente conocidos por su legado histórico en la producción de vinos, y su presencia prominente en el gráfico no es sorprendente. Francia e Italia son potencias vinícolas globales, con regiones famosas como Burdeos, Champagne, Toscana y Piamonte, que contribuyen significativamente a la calidad y diversidad de vinos evaluados

3. **Diferencias significativas**:

   El gráfico revela una brecha significativa entre los tres primeros países (Estados Unidos, Francia e Italia) y los demás países, lo que refleja una mayor producción vinícola o, alternativamente, una mayor representación de estos países en la base de datos. Esta diferencia en volumen puede estar relacionada tanto con la cantidad de vino producido como con el interés comercial y de mercado en estas regiones.

Interpretación y Utilidad del Gráfico

Este gráfico de barras proporciona información valiosa en varios contextos:

1. **Diversidad y volumen de producción**:

   La cantidad de vinos evaluados por país puede ser un reflejo de la diversidad y volumen de la producción vinícola en cada uno de estos países. Estados Unidos, con más de 50,000 vinos evaluados, podría estar indicando un mercado vinícola muy activo y diverso, mientras que Francia e Italia mantienen su liderazgo mundial por su larga tradición vinícola.

2. **Prominencia en el mercado global**:

   La prominencia de ciertos países en este gráfico puede también señalar su relevancia en el mercado global del vino, al igual que el enfoque de los críticos en estas regiones. Estados Unidos, por ejemplo, tiene una gran relevancia en el mercado global y un enfoque fuerte en la evaluación de vinos locales. Esto es particularmente importante para comerciantes y productores que desean posicionar sus productos en mercados internacionales.

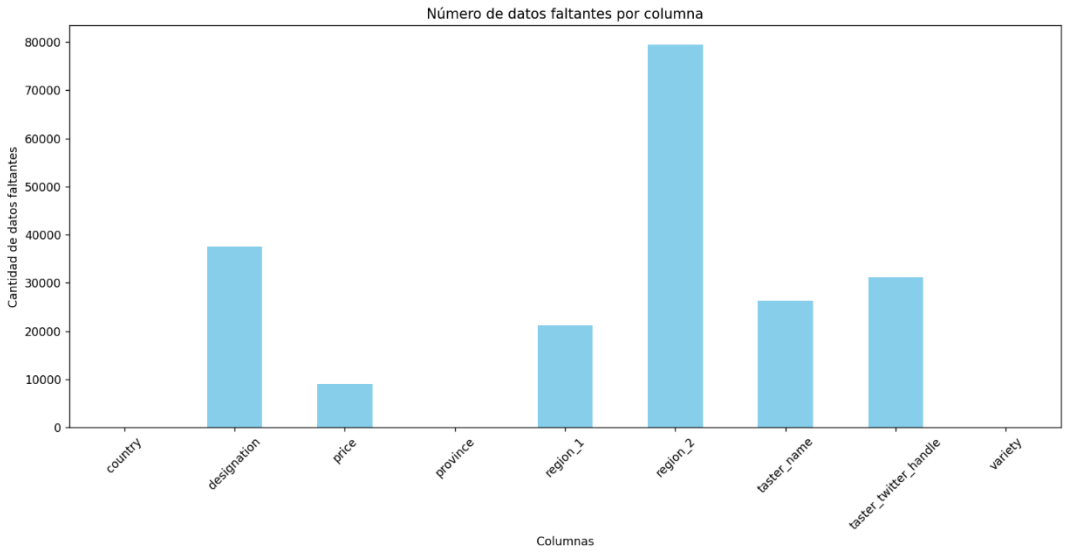
3. **Estrategias comerciales**:

   - Para comerciantes de vino, este gráfico es útil para comprender en qué países se están concentrando las evaluaciones. Países como Estados Unidos, Francia e Italia pueden ser mercados clave para productos de alta gama o de exportación, mientras que Chile y Argentina podrían representar oportunidades para vinos de calidad-precio que atraen a consumidores internacionales.

2. Tratamiento de Datos Faltantes

Un aspecto crítico del análisis de datos es el manejo de los valores faltantes. En este conjunto de datos, algunas variables presentan datos faltantes, lo que podría afectar el análisis si no se trata adecuadamente.

Columnas con Más Datos Faltantes



region_2: Tiene la mayor cantidad de datos faltantes (cerca de 80,000).

designation: Segundo mayor con aproximadamente 37,000 datos faltantes.

taster_twitter_handle: Tercer lugar con alrededor de 31,000 datos faltantes.

Columnas con Menos Datos Faltantes

country, province, variety: No muestran barras visibles, lo que sugiere pocos o ningún dato faltante.

price: Tiene una cantidad relativamente baja de datos faltantes (menos de 10,000).

Recomendaciones para el Tratamiento

a) Para region_2:

Dado que tiene muchos datos faltantes, considerar si esta columna es crucial para el análisis.

Si es importante, se podría usar una categoría "Desconocido" o combinar con region_1.

Alternativamente, se podría eliminar si no es esencial para el análisis principal.

b) Para designation y taster_twitter_handle:

Evaluar la importancia de estas columnas para el análisis general.

Para designation, considerar usar "Sin designación especial" para los valores faltantes.

Para taster_twitter_handle, se podría usar "No disponible" o eliminar si no es crucial.

c) Para price:

Dado que tiene relativamente pocos datos faltantes, se podría imputar usando la mediana o media del precio por categoría de vino o región.

d) Para columnas con pocos datos faltantes (country, province, variety):

Mantener estas columnas sin cambios, ya que parecen estar casi completas.

3. Análisis de Datos Atípicos (Outliers)

Detección de Valores Atípicos

Se empleó el método del rango intercuartílico (IQR) para detectar valores atípicos en la variable price. Se calcularon los cuartiles y se establecieron los límites superior e inferior.

Número de outliers en el precio:

Se encontraron varios valores atípicos en la variable price, que podrían influir en el análisis.

Evaluación de la Naturaleza de los Valores Atípicos

Los valores atípicos pueden deberse a vinos excepcionales o inusuales en términos de precio. Sin embargo, es importante evaluar cómo estos outliers pueden afectar las decisiones basadas en los datos.

Impacto en el Análisis

Se comparó el puntaje promedio con y sin los outliers para entender su efecto. La diferencia entre ambas medias puede ser insignificante, pero los outliers pueden sesgar algunos análisis de correlación, especialmente en el análisis de la relación entre precio y puntuación.

4. Comentarios y Documentación

Se agregaron otros análisis extra dentro del mismo archivo .py para experimentar con el dataset tales como

País más productivo: England con un puntaje promedio de 91.58

Marca más productiva: Araujo con un puntaje promedio de 98.00

que se podrían utilizar como potenciales insights en un futuro desarrollo con preguntas más puntuales.

INSTRUCCIONES

#0. Descargar winemag-data-130k-v2.xlsx de mi repositorio
https://github.com/serranogallegogerardo/S21-WIne-Analysis

#1. winemag-data-130k-v2.xlsx debe estar en la misma ruta que este archivo .py

#2. USAR PYTHON 3.10.10

#2. Instalar las dependencias del requirements usando pip install -r requirements.txt

Referencias

- **Wine Reviews Dataset**: Dataset utilizado en el análisis, disponible en https://github.com/serranogallegogerardo/S21-WIne-Analysis

  Artículos consultados sobre tratamiento de datos faltantes: **Little, R. J. A., & Rubin, D. B. (2002)**. *Statistical Analysis with Missing Data*. Wiley-Interscience.
  Este libro proporciona un análisis profundo de las técnicas de imputación y el impacto de los datos faltantes en el análisis estadístico.

- **Van Buuren, S., & Groothuis-Oudshoorn, K. (2011)**. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software, 45*(3), 1-67.
  Este artículo detalla el paquete R "mice", que permite la imputación de datos faltantes usando ecuaciones encadenadas.

  Publicaciones consultadas sobre análisis de outliers: Estudio de métodos para detectar y manejar valores atípicos en análisis de datos.

  Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outliers. Sage Publications.

- Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.