Fecha 15 septiembre 2023

Informe de proceso de análisis aplicado sobre el dataset de Googgle Maps en relación al sector de restaurantes de comida “Seafood” en los estados “Carolina del Norte”, “Carolina del Sur”, “Maryland”, “Georgia”, “Virginia” y “Florida”.

Introducción al contexto técnico del análisis.

A partir de la solicitud del cliente interesado en invertir por primera vez en EEUU en el sector comercial de restaurantes y bajo la premisa de “selección” del estado más propicio para hacerlo, se emprende a desarrollar un análisis sobre los datos disponibles de una de las bases de datos más amplias y confiables del mercado, la de Google maps.

Dado el requerimiento del cliente se extrajo solo la data concerniente a los estados mencionados al inicio de este documento, mediante procedimientos de ingeniería de datos que no serán descritos en este documento. El actual informe solo se enfocara a exponer los criterios y secuencias de calculo que se consideraron prudentes y aplicables a efectos de obtener la información más precisa y amplia posible.

A continuación, la estructura del presente documento:

1. Descripción de estructura y datos disponibles.
2. Estrategia de análisis.
3. Implementación de estrategia.

Descripción de estructura y datos disponibles.

Luego del ETL, se nos dispuso la siguiente estructura:

Se nos dispuso un resumen estructurado en tres archivos, los cuales se relacionan por el "id_estado" y el "gmap_id". Conjunto de datos base con los cuales era posible extraer información de valor.

Entre todos los atributos identificados, procedemos a describir los más importantes:

Del archivo "estados":

- "id_estado", "estado". Son la base del manejo de la información puesto que también es la base del proyecto, la selección de un estado.
- "superficie_km2". Se asoció data externa a la base de datos. Área de cada estado en análisis.
- "población". Data externa inherente a la población censada al año 2021 por estado.

Del archivo "restaurante":

- "gmap_id". Identificador único por comercio, en este caso por restaurante del tipo en estudio.
- "cant_reviews". Almacena la cantidad de veces que un restaurante ha recibido algún tipo de comentario por parte de sus consumidores.
- "valoración promedio". Almacena la apreciación del consumidor hacia un comercio en base a una escala del 1 al 5.

Del archivo "revisiones":

- "fecha". Echa de acciones que nos interesan analizar.
- "valoración". Almacena la valoración que un consumidor coloco en la fecha expuesta.

Estrategia de análisis.

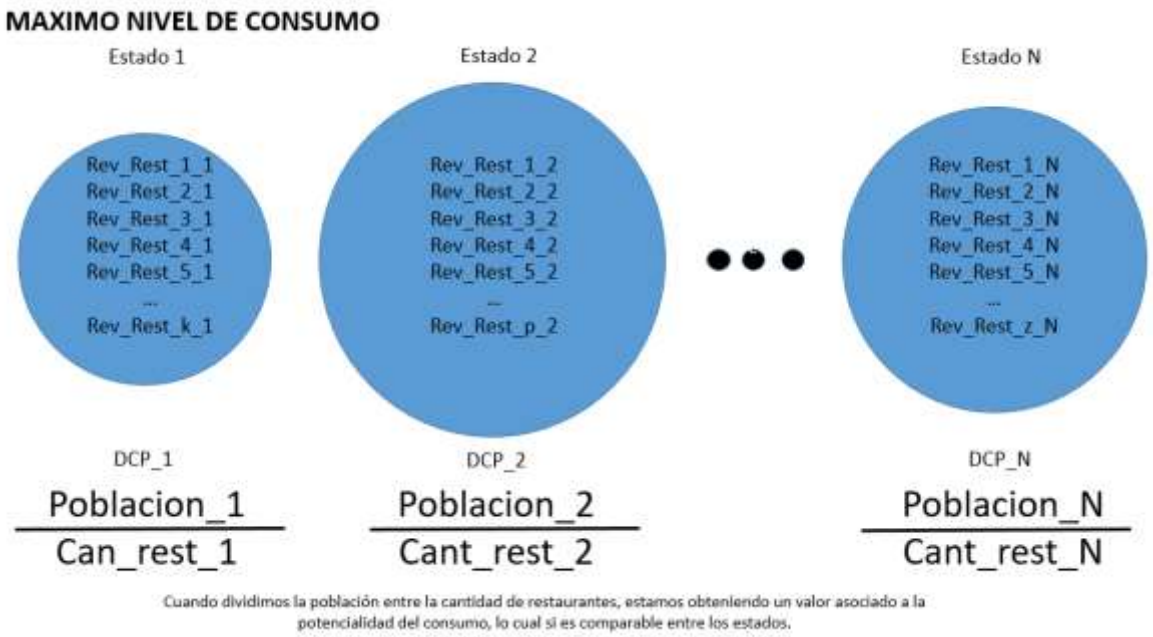
1. Con la data disponible identificar las suposiciones de logro de información y proceder a diseñar una secuencia “teórica” de análisis a fin de luego implementarlo.
2. Implementar el diseño preliminar en una sección reducida de datos.
3. Identificar dificultades y alteraciones o distorsiones posibles.

A continuación, un resumen del diseño y suposiciones preliminares antes de iniciar la corrida preliminar.

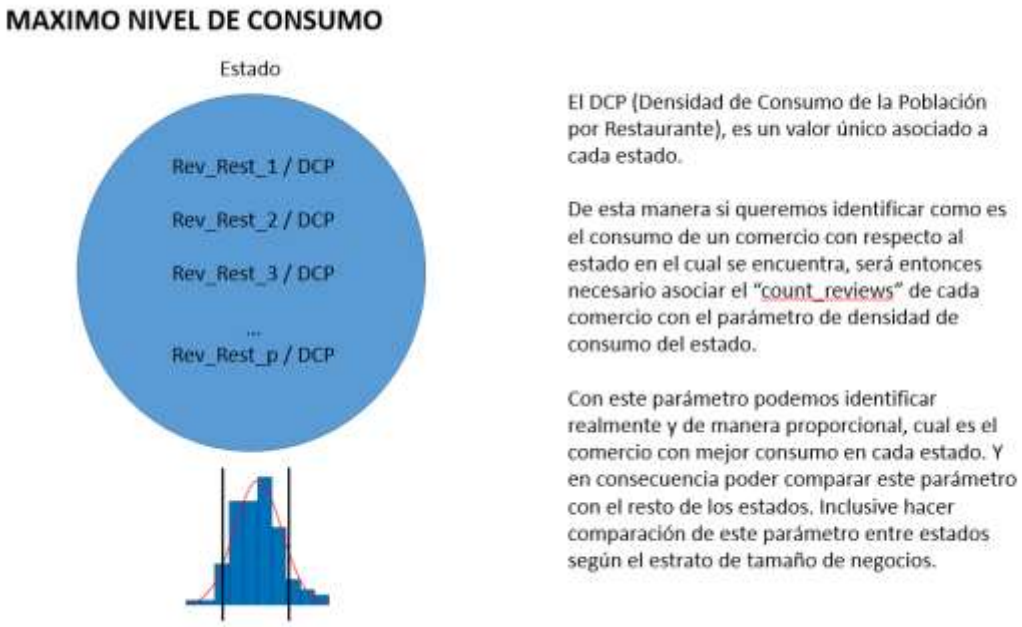
La primera métrica trazada y objetivo con la data disponible seria, una métrica de concentración de restaurantes relacionada al tamaño del territorio.

La segunda métrica a conseguir seria otra de tipo “densidad” dado a que la relacionaríamos esta vez a la población.

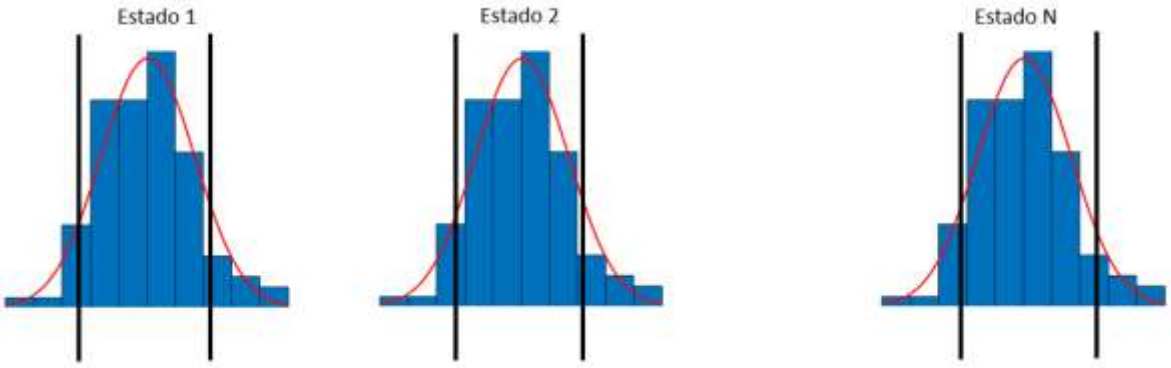
La importancia de estas primeras dos métricas radican en que están asociadas exclusivamente al estado en su generalidad, no tienen relación a sus interacciones internas, en este caso, no tienen que ver con el consumo actualmente “operativo” sino más bien permite entender la “potencialidad” de una nueva participación en estos mercados.



Luego de estas dos métricas base, nos interesaría comenzar a entender primero, como se comporta el consumo en la actualidad dentro de cada estado. Para ello, aunque no tendríamos datos explícitos de consumo, nos sirve a efectos relativos el dato “cant_riviews” para entender el comportamiento del consumo. Es decir, tendríamos que inferir que el comportamiento de la cantidad de opiniones está relacionado o correlacionado con los niveles de consumo o con la calidad del consumidor, lo que a su vez estaría relacionado con la calidad o impacto del comercio. En tal sentido se hace uso de dicho parámetro para generar un “índice” que nos permita estratificar a los restaurantes en tipos según su “tamaño”. Segmento 1, segmento 2 y segmento 3.

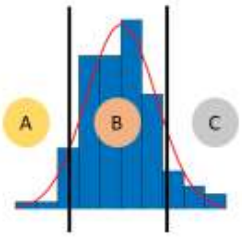


Por tanto, la idea sería estratificar a cada estado en estos tres segmentos. Antes de implementar esta aspiración presumimos una distribución normal, pero será en dicho momento de implementación que identificaríamos las cualidades de la distribución de frecuencia del parámetro “cant_reviews” y en consecuencia como sería su segmentación.



Bajo esta premisa de estratificación, podríamos tener entonces un índice que describa cada segmento en función no solo del consumo, sino que luego utilizaríamos dicha estratificación para entender el comportamiento de la valoración del consumidor por cada uno de dichos segmentos.

RESUMEN PARCIAL (Índice de consumo estratificado por tamaño del comercio)



C — Comercios sobre los cuales se infiere alta afluencia de clientes (ventas, consumo). 15% en una distribución teórica normal.

B — Comercios sobre los cuales se infiere una afluencia de rango promedio de clientes (ventas, consumo). 70% en una distribución teórica normal.

A — Comercios sobre los cuales se infiere una baja afluencia de clientes (ventas, consumo). 15% en una distribución teórica normal.

Densidad de Consumo del comercio (Según el estrato de tamaño), por estado (DCC).
Tenemos un conjunto de KPIs que proveen información relacionada a como es el nivel de consumo según la estratificación de comercios por “tamaño”. El índice construido permite comparar como es el consumo máximo y promedio en cada estrato de comercios según su tamaño, entre estados.

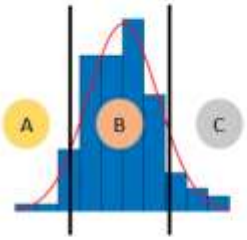
Sea Dre, la Densidad de revisiones en el restaurante “r” dentro del estado “e”.

$$Dre = \frac{Rev\_Rest\_r}{DCP\_e}$$

Máximo del índice de consumo por comercio $MAX\{D1a, D2a, D3a, \ldots, Dna\}$

Promedio del índice de consumo por comercio $PROM\{D1a, D2a, D3a, \ldots, Dna\}$

RESUMEN PARCIAL (Índice de percepción por estrato de tamaño del comercio)



C — Comercios sobre los cuales se infiere alta afluencia de clientes (ventas, consumo). 15% en una distribución teórica normal.

B — Comercios sobre los cuales se infiere una afluencia de rango promedio de clientes (ventas, consumo). 70% en una distribución teórica normal.

A — Comercios sobre los cuales se infiere una baja afluencia de clientes (ventas, consumo). 15% en una distribución teórica normal.

Percepción (Según el estrato de tamaño), por estado (**PEE**).
Tenemos un conjunto de KPIs que proveen información relacionada a como es percibida la calidad de servicio de los comercios según el estrato de tamaño al cual ha sido clasificado.

Estado 1

MAX(PA1)	P1A , P2A , P3A , … , PnA	PROM(PA1)	P1A , P2A , P3A , … , PnA
MAX(PB1)	P1B , P2B , P3B , … , PnB	PROM(PB1)	P1B , P2B , P3B , … , PnB
MAX(PC1)	P1C , P2C , P3C , … , PnC	PROM(PC1)	P1C , P2C , P3C , … , PnC

Finalmente, todo el conjunto de métricas, tanto externas (que describen al estado en su generalidad) como internas (que describen a los segmentos generados) conforman un conjunto de métricas a las cuales es posible asignar o identificar objetivos referenciales. Razón por la cual las terminamos denominando “KPI”.

Compendio total de KPIs disponibles

Los KPIs descriptores de cada estado CRE y DCP dependen de información adicional, Tamaño de territorio y Población, respectivamente.
Los KPIs internos de cada estado están diseños para poder ser comparados con los mismos parámetros de otros estados.

Implementación de la estrategia.

A continuación, estaremos entonces exponiendo de manera resumida la secuencia de actividades inherentes a la generación de métricas, análisis y valores que implicaron algunas decisiones y criterios técnicos adicionales.

1. Calculo de métrica (KPI) relacionara a la superficie del territorio denominada "Concentración de Restaurantes por Estado (CRE).
2. Calculo de métrica (KPI) relacionada a la población del estado denominada "Densidad de Consumo Potencial" (DCP).
3. Luego de obtener las métricas generales procedemos a calcular las métricas internas, las cuales dependen de la "segmentación" del mercado de cada estado. Esta es la actividad de mayor impacto en el análisis general. Procedemos a exponer entonces la secuencia de acciones o actividades y características del proceso.

Proceso de segmentación del mercado por cada estado.

Antes de exponer las distribuciones de frecuencia obtenidas en el proceso destacamos las principales observaciones:

1. Todas las distribuciones fueron sesgadas a la izquierda, es decir, el consumo tiende a agruparse hacia los comercios de menor operación. Lo cual resulta bastante lógico y confirma la tesis de inferencia con la cual asumimos el comportamiento del consumo.
2. También como es de esperarse, las distribuciones muestran valores extremos hacia el lado desconcentrado. Esto explica el consumo de alto nivel en este renglón comercial.
3. Los estadísticos habituales se ven relevantemente afectados por estos comportamientos extraordinarios, razón por la cual se procede a una estratificación que consiste en lo siguiente:
    a. Reducir la proporción de la población hasta que no se identifiquen comportamientos extraordinarios (outliers).
    b. Segmentar la distribución resultante de dicha reducción y evaluar estadísticos. En este sentido también se dieron particularidades generales:
        i. Todas las distribuciones resultantes de la reducción también fueron sesgadas a la izquierda.
        ii. Por esta razón anterior, se decidió utilizar como medida de tendencia central la "mediana".
        iii. Consecuentemente, la medida de dispersión seleccionada fue el "rango intercuartil".
        iv. Y la separación de la distribución (en dos segmentos) se hizo como resulta ser lo adecuado, mediante el uso de la mediana.
        v. Finalmente, todas las distribuciones de consumo por cada estado se dividieron en tres segmentos:
            1. Segmento 1: Desde el consumo cero (0) hasta el consumo indicado por la medina.

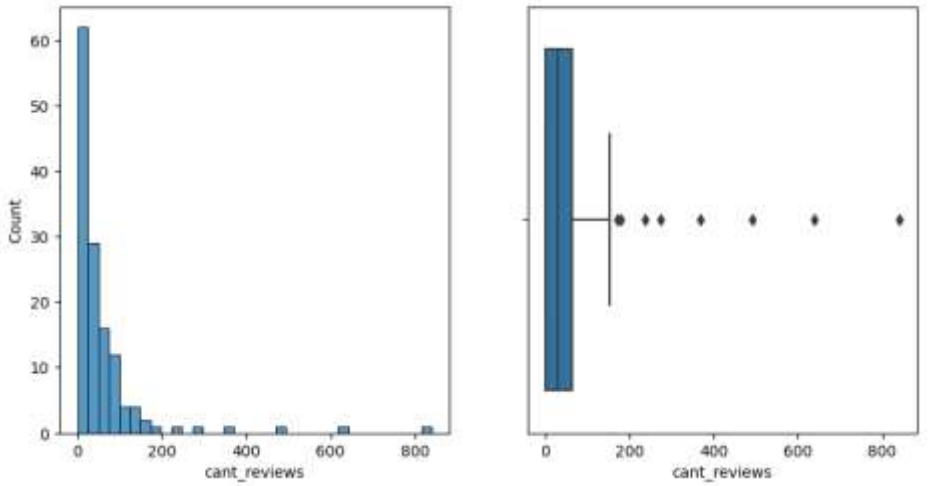
2. Segmento 2: Desde el valor de consumo indicado por la mediana hasta el cierre del cuartil 4 de la población reducida.
3. Segmento 3: Finalmente este segmento lo conforma el conjunto de restaurantes que fue excluido de la población con la cual se calcularon los estadísticos que permitieron establecer las dos primeras segmentaciones. Es decir, el segmento 3 lo conforman los "outliers" de cada estado.

A continuación, se presentan por estados lo siguiente:

1. Distribución de frecuencia y diagrama de caja con el 100% de la población.
2. Estadísticos de dicho 100% de la población.
3. Distribución de frecuencia y diagrama de caja con éln porcentaje reducido de la población.
4. Estadísticos inherentes a la población reducida.

Estado " Carolina del Norte"

Cantidad de Restaurantes: 136



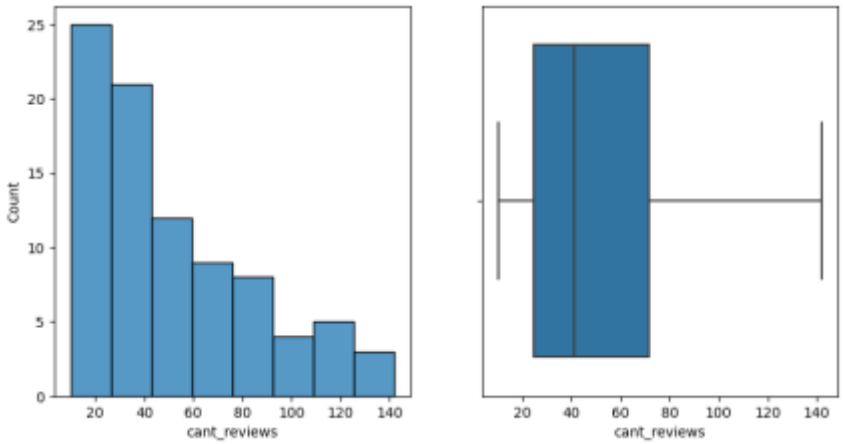
Estadísticos de NC SIN REDUCCION

el 25% de count_reviews está por debajo de: 0.0
el 50% de count_reviews está por debajo de: 29.0
el 75% de count_reviews está por debajo de: 64.5
El rango intercuartil (dispersión) es: 64.5

el promedio de Cantidad de revisiones está en: 57.42
Y la desviación estándar en: 107.89

Maximo: 840, y minima: 0

Gráficos de NC CON REDUCCION al 94%



Estadísticos de NC con una reducción al 94%

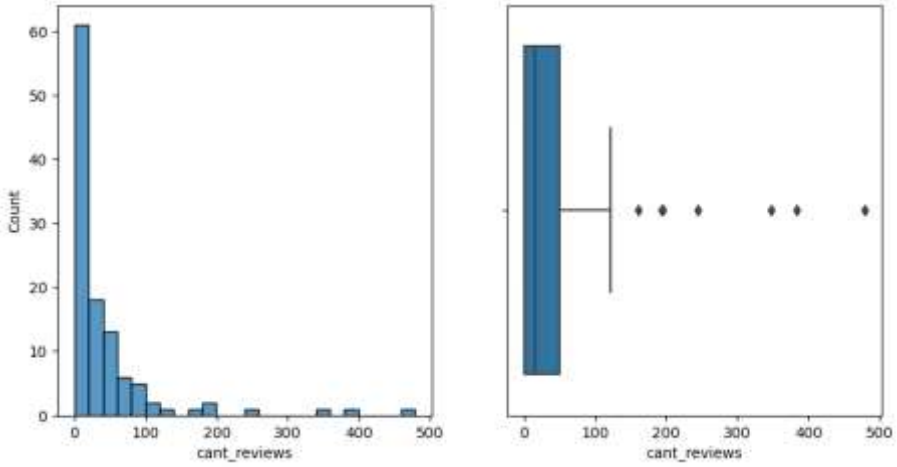
el 25% de count_reviews está por debajo de: 24.5
el 50% de count_reviews está por debajo de: 41.0
el 75% de count_reviews está por debajo de: 71.5
El rango intercuartil (dispersión) es: 47.0

el promedio de Cantidad de revisiones está en: 51.2
Y la desviación estándar en: 33.08

Máximo: 142, y mínima: 10

Estado ” Carolina del Sur”

Cantidad de Restaurantes: 113



Estadísticos de SC SIN REDUCCION

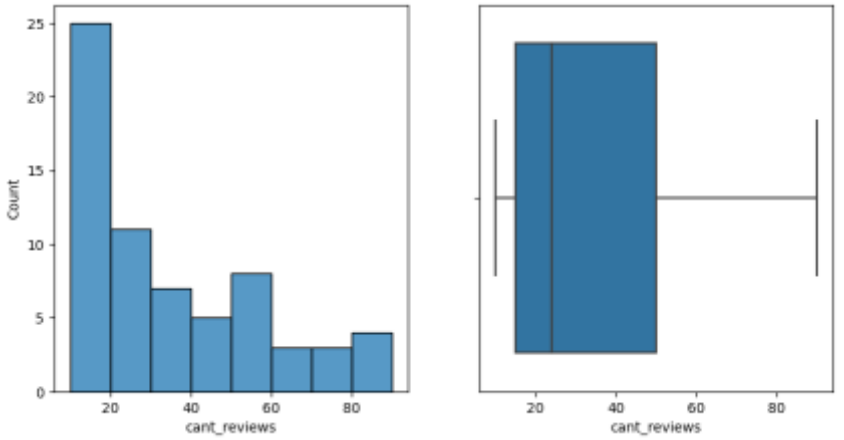
el 25% de count_reviews está por debajo de: 0.0
el 50% de count_reviews está por debajo de: 16.0
el 75% de count_reviews está por debajo de: 50.0

El rango intercuartil (dispersión) es: 50.0

el promedio de Cantidad de revisiones está en: 41.45
Y la desviación estándar en: 74.22

Maximo: 479, y minima: 0

Gráficos de SC CON REDUCCION al 91%



Estadísticos de SC con una reducción al 91%

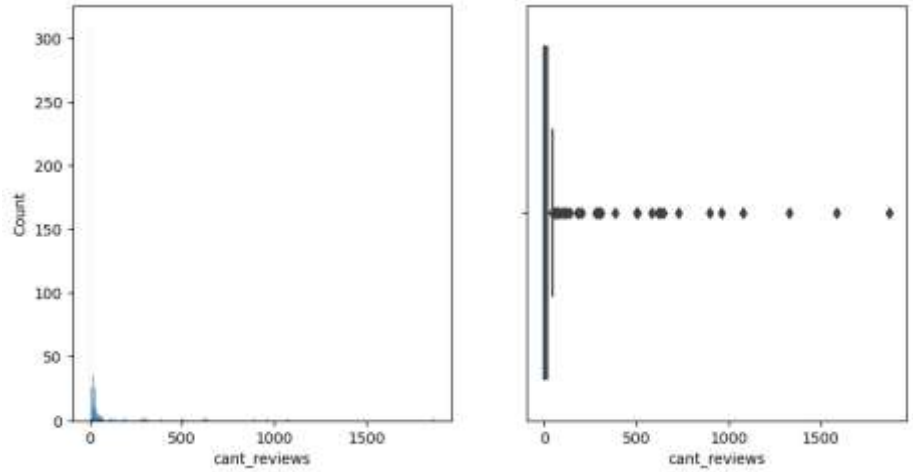
el 25% de count_reviews está por debajo de: 15.0
el 50% de count_reviews está por debajo de: 24.0
el 75% de count_reviews está por debajo de: 50.0
El rango intercuartil (dispersión) es: 35.0

el promedio de Cantidad de revisiones esta en: 33.94
Y la desviación estándar en: 22.72

Maximo: 90, y minima: 10

Estado ” La Florida”

Cantidad de Restaurantes: 496



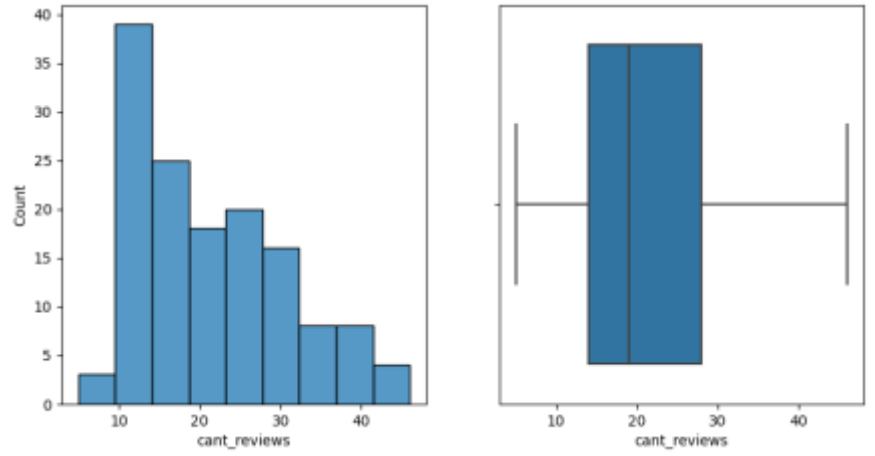
Estadísticos de FL SIN REDUCCION

el 25% de count_reviews está por debajo de: 0.0
el 50% de count_reviews está por debajo de: 0.0
el 75% de count_reviews está por debajo de: 17.0
El rango intercuartil (dispersión) es: 17.0

el promedio de Cantidad de revisiones está en: 38.33
Y la desviación estándar en: 164.27

Maximo: 1869, y minima: 0

Gráficos de FL CON REDUCCION al 91%



Estadísticos de FL con una reducción al 91%

el 25% de count_reviews está por debajo de: 14.0
el 50% de count_reviews está por debajo de: 19.0
el 75% de count_reviews está por debajo de: 28.0

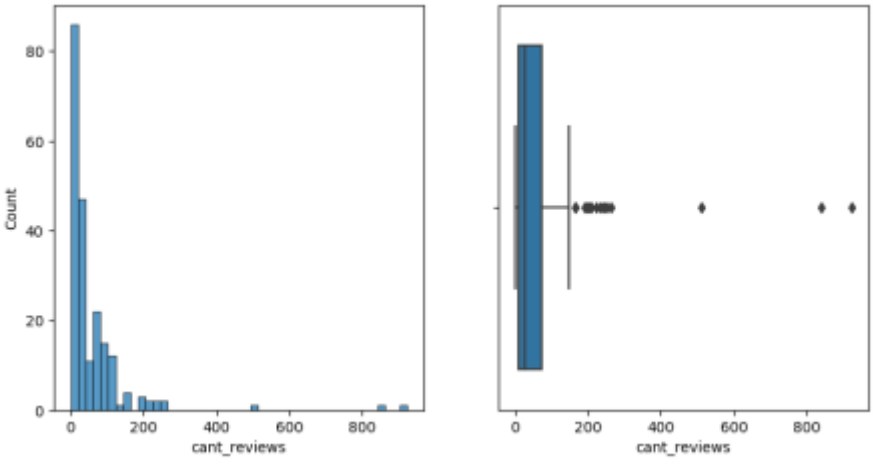
El rango intercuartil (dispersión) es: 14.0

el promedio de Cantidad de revisiones esta en: 21.23
Y la desviación estándar en: 9.23

Maximo: 46, y minima: 5

Estado ” Giorgia”

Cantidad de Restaurantes: 210



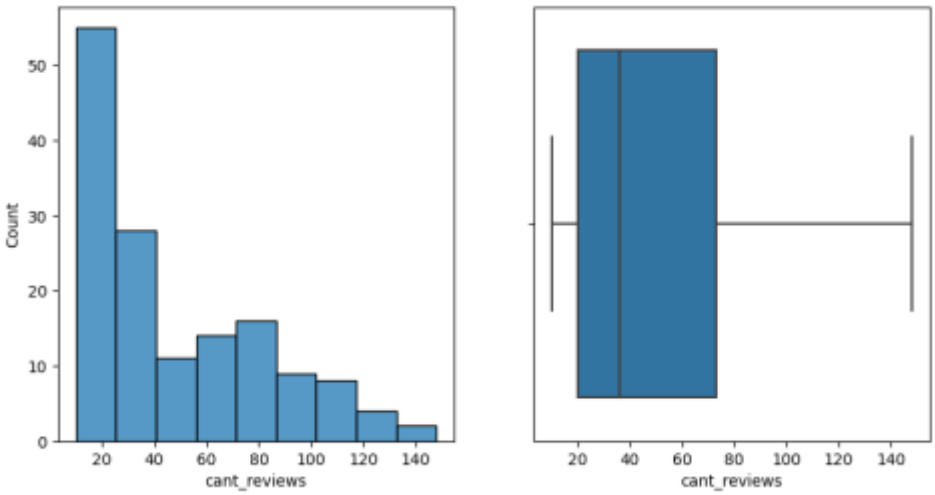
Estadísticos de GA SIN REDUCCION

el 25% de count_reviews está por debajo de: 11.0
el 50% de count_reviews está por debajo de: 26.0
el 75% de count_reviews está por debajo de: 72.0
El rango intercuartil (dispersión) es: 61.0

el promedio de Cantidad de revisiones esta en: 56.38
Y la desviación estándar en: 102.32

Maximo: 923, y minima: 0

Gráficos de GA CON REDUCCION al 93%



Estadísticos de GA con una reducción al 93%

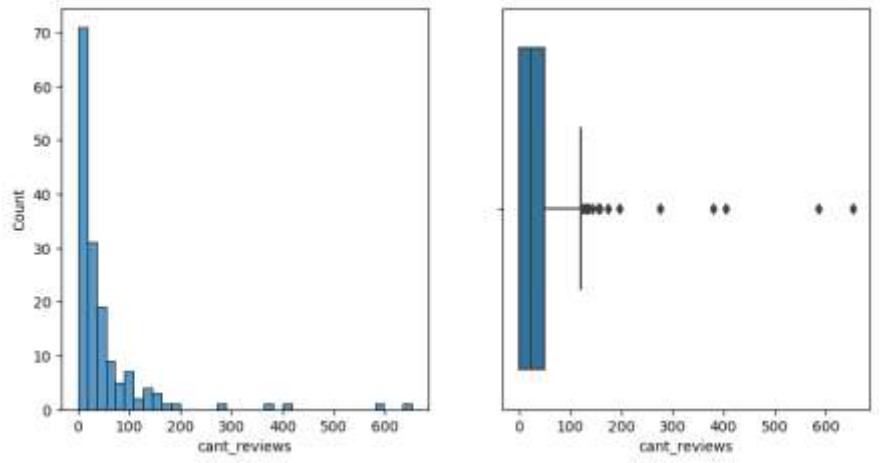
el 25% de count_reviews está por debajo de: 20.0
el 50% de count_reviews está por debajo de: 36.0
el 75% de count_reviews está por debajo de: 73.0
El rango intercuartil (dispersión) es: 53.0

el promedio de Cantidad de revisiones está en: 48.12
Y la desviación estándar en: 34.31

Maximo: 148, y minima: 10

Estado ” Maryland”

Cantidad de Restaurantes: 158



Estadísticos de MD SIN REDUCCION

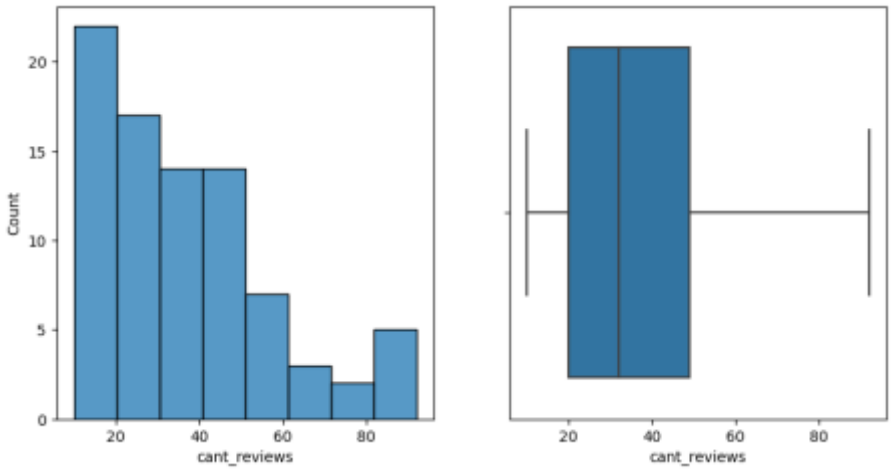
el 25% de count_reviews está por debajo de: 0.0
el 50% de count_reviews está por debajo de: 23.5

el 75% de count_reviews está por debajo de: 50.0
El rango intercuartil (dispersión) es: 50.0

el promedio de Cantidad de revisiones esta en: 47.87
Y la desviación estándar en: 88.76

Maximo: 653, y minima: 0

Gráficos de MD CON REDUCCION al 86%



Estadísticos de MD con una reducción al 86%

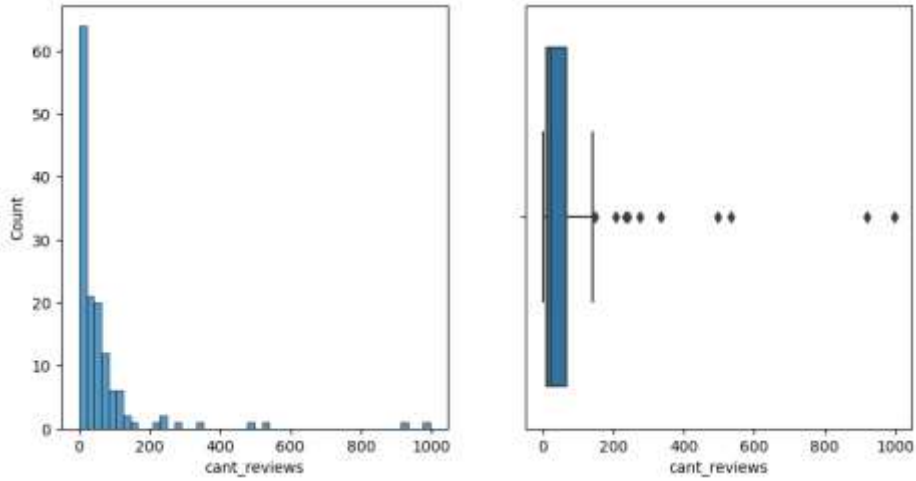
el 25% de count_reviews está por debajo de: 20.0
el 50% de count_reviews está por debajo de: 32.0
el 75% de count_reviews está por debajo de: 49.0
El rango intercuartil (dispersión) es: 29.0

el promedio de Cantidad de revisiones está en: 36.11
Y la desviación estándar en: 20.97

Maximo: 92, y minima: 10

Estado ” Virginia”

Cantidad de Restaurantes: 141



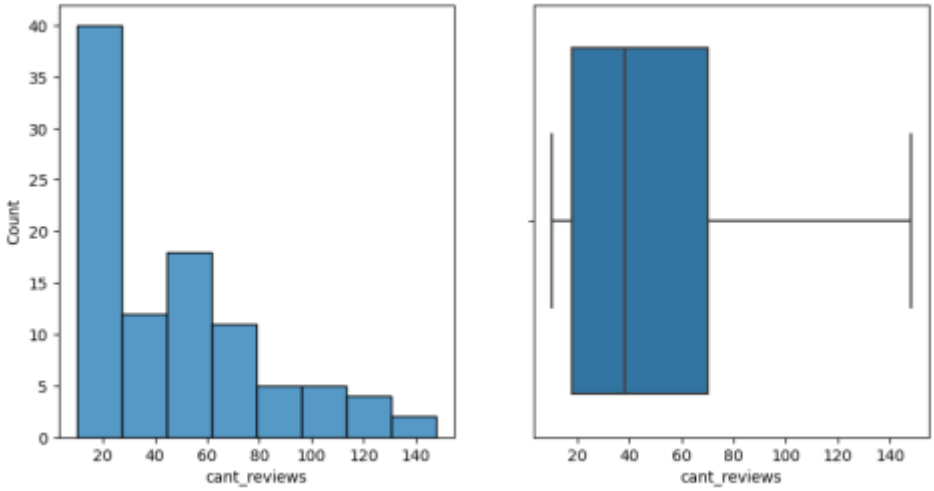
Estadísticos de VA SIN REDUCCION

el 25% de count_reviews está por debajo de: 10.0
el 50% de count_reviews está por debajo de: 24.0
el 75% de count_reviews está por debajo de: 65.0
El rango intercuartil (dispersión) es: 55.0

el promedio de Cantidad de revisiones está en: 62.43
Y la desviación estándar en: 132.87

Maximo: 997, y minima: 0

Gráficos de VA CON REDUCCION al 94%



Estadísticos de VA con una reducción al 94%

el 25% de count_reviews está por debajo de: 18.0
el 50% de count_reviews está por debajo de: 38.0

```
el 75% de count_reviews está por debajo de: 70.0
El rango intercuartil (dispersión) es: 52.0

el promedio de Cantidad de revisiones está en: 47.09
Y la desviación estándar en: 34.0

Maximo: 148, y minima: 10
```

Como se puede observar, Las reducciones de población a fin de lograr estadísticos sobre una "sub población" sin "outliers" varía según el estado. Así se implementó.

Estado	Reduccion de poblacion
Carolina del Norte	94%
Carolina del Sur	91%
La Florida	91%
Maryland	86%
Giorgia	93%
Virginia	94%

Teniendo de esta manera, los estadísticos con los cuales vamos a segmentar por estado, procedemos entonces a implementarlo.

Estado	Segmento	Cota minima (cant_reviews)	Cota maxima (cant_reviews)
Carolina del Norte	1	0	41
	2	41	142
	3	142	372.77
Carolina del Sur	1	0	24
	2	24	90
	3	90	222.18
La Florida	1	0	19
	2	19	46
	3	46	356
Maryland	1	0	32
	2	32	92
	3	32	205.66
Giorgia	1	0	36
	2	36	148
	3	148	317.66
Virginia	1	0	38
	2	38	148
	3	148	470.55

A partir de dicha segmentación se calcularon todos los siguientes índices y métricas sobre cada estado.

- Índices de densidad de consumo.

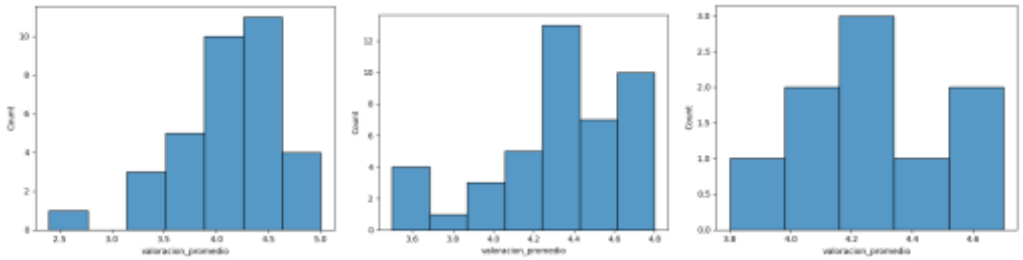
- Métricas de percepción.

Procederemos a obviar el cálculo de los respectivos índices y métricas de percepción dado que estos están plasmados en el archivo “EDA.google.ipynb”. Lo mas importante de este documento ha sido el criterio de segmentación, pues a partir de allí se generan los estadísticos y consecuentes valores de índices y métricas.

Procedamos a mostrar las distribuciones de frecuencia de las métricas de percepción por estado.

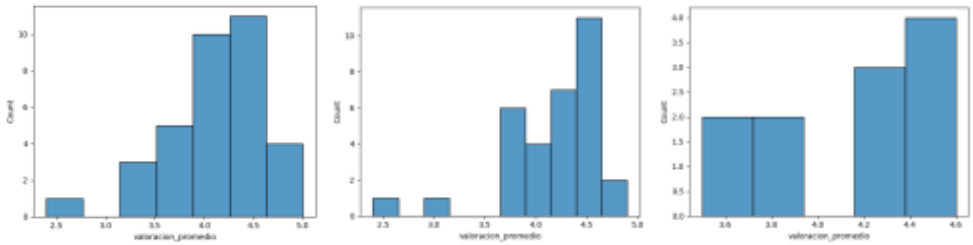
Estado “Carolina del Norte”.

Por Segmento (1, 2 y 3):



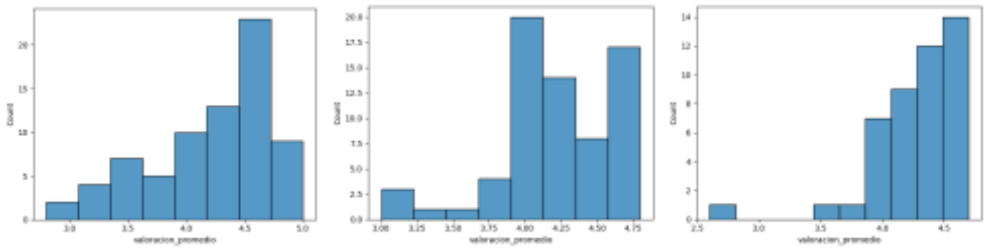
Estado “Carolina del Sur”.

Por Segmento (1, 2 y 3):



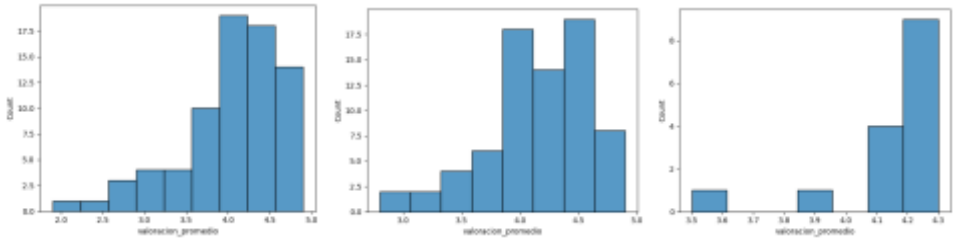
Estado “La Florida”.

Por Segmento (1, 2 y 3):



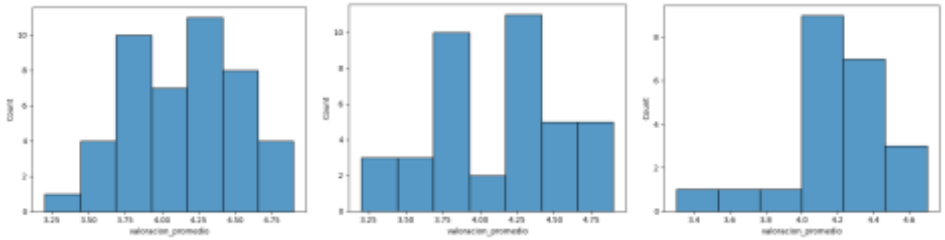
Estado “Giorgia”.

Por Segmento (1, 2 y 3):



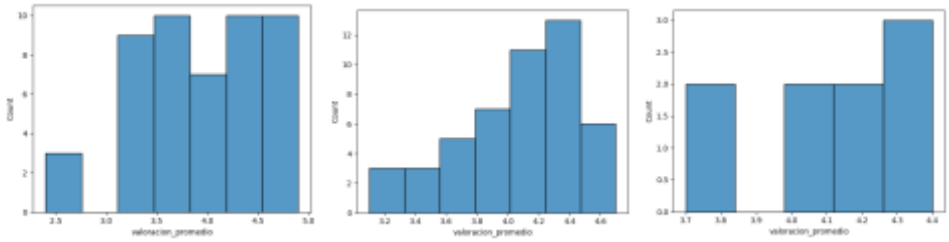
Estado “Maryland”.

Por Segmento (1, 2 y 3):



Estado “Virginia”.

Por Segmento (1, 2 y 3):



De esta manera y en forma honologa tanto para las metricas de percepcion, como los indices que permiten comparar el consumo entre los diversos segmentos de cada estado, se analizan. Todo en funcion de la segmentacion iniciar.