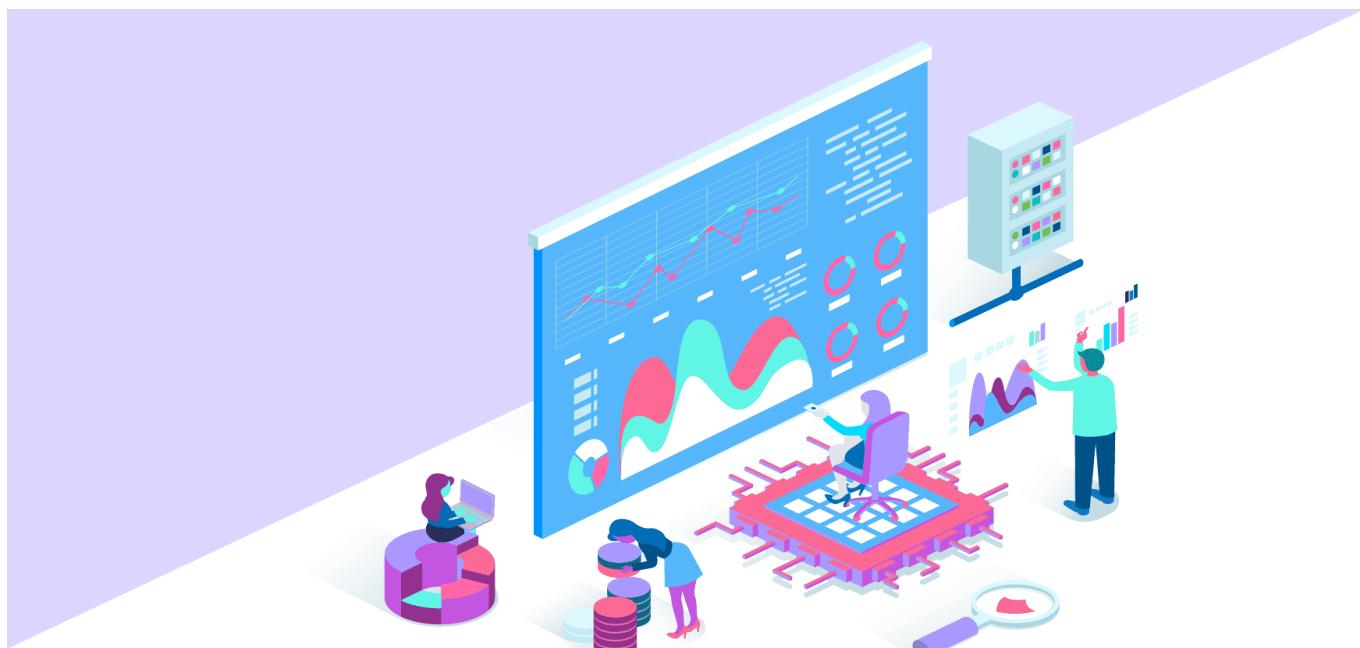


# ANÁLISIS DE DATOS

## Desafío Final



Alejandro Hernández Artiles

---

15 de diciembre de 2021

---

# Índice

<b>Introducción</b>	<b>4</b>
Estructura del proyecto	4
<b>Herramientas utilizadas</b>	<b>5</b>
Python	5
Pandas	5
Kepler	6
Pandas Profiling	6
Google Drive	6
Google Colaboratory	6
<b>Gestión del Proyecto</b>	<b>7</b>
Metodología	7
Conocer el negocio	8
Definición de objetivos	9
Equipos	9
Identificación de los orígenes de los datos	10
Estructura de almacenamiento	10
Adquirir y comprender los datos	11
Modelar	12
Diseño de características	12
Entrenamiento de los modelos	12
<b>Adquisición y análisis de datos</b>	<b>13</b>
Obtención	13
Filtrado	13
Segmentación	15
Agregación de atributos relevantes	16
<b>EDA. Análisis Visual de Datos</b>	<b>17</b>
Análisis Visual Inicial	17
Mapa: Id's	17
Mapa: Rating	19
Mapa: Negocios Abiertos	19
Mapa: Número de Reviews	20
Mapa: Contraste	21

---

Mapa: Zipcodes	22
Mapas Filtrados	23
Mapa: Restaurantes	23
Mapa: Rating	24
Mapa: Número de Reviews	26
Análisis Visual después del Filtrado	27
<b>Aprendizaje Automático</b>	<b>37</b>
Preparación de los datos	37
Elección de algoritmos	41
Generación de modelos	42
Resultados	43
3 clases - Bajo, Medio y Alto	43
2 clases - Bajo y Alto	45
<b>Conclusiones y dificultades encontradas</b>	<b>47</b>
<b>Apéndice</b>	<b>49</b>
Estructura de Notebooks	49
Filtrado-AD	49
Segmentación-AD	49
MapasSinFiltrar-AD	50
MapasFiltrados-AD	50
Auxiliar-AD	50
AprendizajeAutomatico-AD	51
AnalisisVisual-AD	51
<b>Referencias</b>	<b>51</b>

---

# 1. Introducción

El **objetivo principal** de este proyecto es la aplicación de técnicas de análisis de datos a unos datos proporcionados de negocios, usuarios y opiniones. Con la aplicación de estas técnicas se tratará de obtener algún valor añadido de los datos.

## Estructura del proyecto

La estructura elegida para el proyecto es la siguiente:

- **Herramientas utilizadas.** En esta sección se presentan las herramientas utilizadas para el desarrollo del proyecto, tanto herramientas de implementación de código como herramientas de control de versiones.
- **Gestión del proyecto.** En esta sección se expone la organización del proyecto, el rol de cada integrante y los equipos creados para el desarrollo del mismo.
- **Adquisición y análisis de datos.** En esta sección se describen las técnicas de filtrado de datos y segmentación con la finalidad de establecer un proceso de limpieza para acto seguido realizar un análisis visual exhaustivo.
- **EDA. Análisis Visual de Datos.** En esta sección se incluyen mapas, gráficos, diagramas, matrices, etc. para dar una visión general de los datos.
- **Aprendizaje Automático.** En esta sección se describe el proceso de predicción elegido, así como las técnicas y algoritmos seleccionados.
- **Conclusiones y dificultades encontradas.** En esta última sección se exponen las conclusiones extraídas de los resultados obtenidos durante todo el proceso de análisis de datos.

---

## 2. Herramientas utilizadas

### Python

Python [1] es uno de los lenguajes de programación más utilizados y demandados a nivel mundial. Se trata de un lenguaje de programación con muchas ventajas y cualidades, entre las que se destacan:

- Es de código abierto
- Es multiparadigma
- Es orientado a objetos
- Es un lenguaje de alto nivel

En el ámbito de la ciencia de datos, Python es un lenguaje muy utilizado porque permite realizar complejas operaciones sin tener grandes conocimientos de análisis de datos. Por otro lado, proporciona una gran cantidad de librerías que apoyan al proceso de ciencia de datos, tales como:

- Numpy
- Pandas
- Scikit-learn
- Matplotlib
- TensorFlow
- ...

Por estas razones se ha decidido realizar este proyecto con Python. Las librerías utilizadas se describen a continuación.

### Pandas

---

Pandas [2] es una librería que proporciona una serie de estructuras muy flexibles que facilitan enormemente el tratamiento de los datos. En concreto, Pandas permite cargar datos, modelar, analizar, manipular y prepararlos.

## **Kepler**

Kepler [3] es un software desarrollado por Uber para la visualización de datos espaciales. Permite importar datos en múltiples formatos y usar sus funciones integradas de visualización de datos espaciales en la ventana interactiva embebida en el cuaderno.

## **Pandas Profiling**

Pandas Profiling [4] es una librería perteneciente a Pandas que permite realizar análisis exploratorios automáticos en formato HTML. Es especialmente útil cuando se tienen datos de gran dimensionalidad.

## **Google Drive**

Para el control de versiones se ha decidido utilizar Google Drive [5] tanto por su facilidad de uso como por su gran compatibilidad con Google Colaboratory.

Google Drive es un servicio de almacenamiento en la nube que permite crear y compartir todo tipo de archivos y documentos.

## **Google Colaboratory**

Google Colaboratory o comúnmente conocido como Google Colab [6] es un entorno de programación que permite codificar y ejecutar código en Python fácilmente. Además, es muy útil cuando se trabaja con Google Drive, puesto que permite montar todo el directorio personal dentro del entorno de trabajo.

---

### 3. Gestión del Proyecto

En el presente proyecto se utilizará un **Proceso de Ciencia de Datos en Equipo (TDSP)** [7], un ciclo de vida que se puede usar para estructurar proyectos de análisis de datos.

#### Metodología

El proceso consta de 5 fases:

- Conocimiento del negocio
- Adquisición y comprensión de los datos
- Modelado
- Implementación
- Aceptación del cliente

A continuación se muestra una representación gráfica del ciclo de vida TDSP.

## Data Science Lifecycle

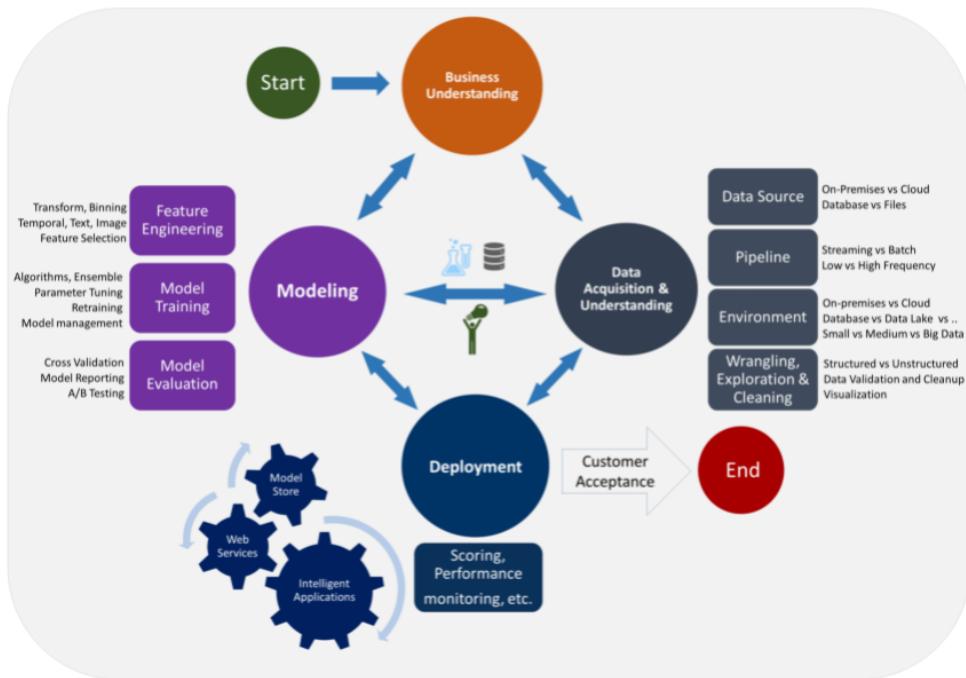


Ilustración 1. Ciclo de vida del proyecto

**TDSP** se modela como una serie de pasos iterativos que los científicos de datos utilizan para definir las tareas necesarias para crear modelos predictivos. De esta manera, las tareas se reparten entre los distintos equipos con la finalidad de alcanzar un punto final de interacción claro. Por otra parte, la comunicación de ideas es esencial en TDSP puesto que de esa manera se evitan la mayor parte de los malentendidos.

Sin embargo, en este proyecto tan solo se alcanzará la tercera fase, es decir, no se llegará a implementar un dispositivo o aplicación final que permita la interacción con los usuarios, sino que se limitará a lograr un buen modelo de predicción. Por tanto, las tareas a realizar son las siguientes.

## Conocer el negocio

---

El objetivo de esta fase es especificar las variables fundamentales que servirán como objetivos para los futuros modelos e identificar los orígenes de los datos proporcionados.

## Definición de objetivos

En esta subsección se exponen los objetivos clave elegidos para el proyecto. En concreto, se pretende crear un modelo que permita **predecir el rating según los atributos del restaurante dado y la zona en la que se le quiere colocar**.

Para ello, se llevará a cabo un proceso de **clasificación** en cada una de las concentraciones de negocios detectadas (ver [EDA: Análisis Visual de Datos](#)). Por tanto, el objetivo principal es crear un modelo para cada una de estas zonas.

## Equipos

Con el fin de organizar y comunicar tareas eficientemente, se han creado 3 equipos de trabajo. En esta subsección se exponen los miembros de los equipos, sus responsabilidades y roles.

- **Equipo de limpieza, filtrado y segmentación de datos.** Este equipo se compone de 3 integrantes, puesto que el preprocesado de los datos es probablemente la fase más importante del proceso de ciencia de datos. Los miembros del equipo son:
  - Franco Exequiel Schuler Allub
  - Aarón Úbeda-Portugués Cano
  - Enrique Ángel Arrabal
- **Equipo de análisis descriptivo y visual de datos.** La representación gráfica de distribuciones y datos de interés es crucial para comprender, analizar y poner en contexto los datos de los que disponemos. Los miembros del equipo son:
  - Roberto Cano García
  - Alejandro Hernández Artiles
- **Equipo de entrenamiento y validación de modelos.** En esta última fase, se generarán los entrenamientos de los modelos necesarios para realizar las

---

predicciones que satisfacerán los objetivos definidos previamente. Por ello, este equipo estará formado por todo el equipo del proyecto:

- Franco Exequiel Schuler Allub
- Aarón Úbeda-Portugués Cano
- Enrique Ángel Arrabal
- Roberto Cano García
- Alejandro Hernández Artiles

## **Identificación de los orígenes de los datos**

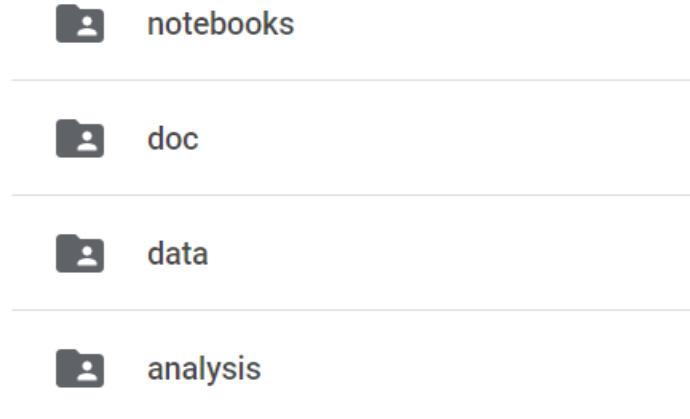
Los datos proporcionados se corresponden con datos reales acerca de una gran cantidad de negocios incluyendo usuarios, opiniones de clientes, etc. En concreto, contiene:

- Casi 7 millones de opiniones de usuarios
- Casi 200 mil negocios
- Más de 1 millón de usuarios

Puesto que el propósito de este proyecto consiste en tratar de predecir las mejores localizaciones para abrir restaurantes, parece que estos datos resultan útiles para el proceso de ciencia de datos.

## **Estructura de almacenamiento**

Como ya se ha comentado, se utilizará Google Drive para almacenar los set de datos y para llevar un control de versiones de todos y cada uno de los ficheros que se vayan generando. En concreto, la estructura de carpetas que se va a seguir se muestra en la ilustración 2:



*Ilustración 2. Estructura del proyecto en Drive*

Donde:

- **notebook** almacena todos los cuadernos de Colab utilizados para el proceso de análisis de datos
- **doc** contiene toda la documentación realizada, así como apuntes y aspectos a tener en cuenta para el proyecto
- **data** contiene los datos proporcionados, así como un sistema de subcarpetas correspondiente al filtrado, segmentación, limpieza, etc.
- **analysis** contiene los análisis tanto descriptivos como visuales de los datos.

## Adquirir y comprender los datos

Esta tarea tiene varios objetivos principales. En primer lugar, generar un conjunto de datos limpio y de alta calidad, de manera que tenga una gran relación con las variables objetivo del proyecto. En segundo lugar, se debe desarrollar una serie de componentes que permitan mantener actualizados los datos.

Para lograr esto, se deben seguir una serie de pasos:

- 
- **Extracción de los datos** e integración en nuestro entorno de trabajo.
  - **Exploración de los datos** para determinar si son de una calidad suficiente como para continuar con el proceso.
  - **Desarrollar los componentes** necesarios para mantener actualizados los datos.

## Modelar

Con la finalidad de crear modelos precisos y robustos, es necesario seguir una serie de pasos muy bien definidos.

### Diseño de características

El **diseño de características** consiste en combinar y transformar variables sin procesar para crear las características finales que se utilizarán para el entrenamiento. Resulta crucial entender cómo se relacionan las variables entre sí y cómo deberán utilizarlas los algoritmos de Machine Learning.

Este proceso requiere una mezcla de creatividad, experiencia del dominio y de los conocimientos obtenidos durante la fase de exploración de datos. Por tanto, el trabajo consistirá en conseguir variables informativas que mejoren los resultados y evitar las variables que introducen ruido en el entrenamiento.

### Entrenamiento de los modelos

El proceso de entrenamiento de los modelos incluye los siguientes pasos:

- **Separar los datos** en un conjunto de entrenamiento que se encargará de entrenar a los modelos y un conjunto de test que se encargará de evaluarlos.
- **Entrenar** los modelos obteniendo los modelos necesarios.
- **Evaluar** los modelos con el conjunto de test.
- **Determinar la mejor solución** posible para dar solución a los objetivos planteados comparando las métricas elegidas para cada algoritmo.

---

## 4. Adquisición y análisis de datos

### Obtención

Los datos con los que se han trabajado han sido proporcionados por los profesores de la asignatura, todos ellos almacenados en distintos ficheros con formato csv. Debemos añadir que no se ha trabajado con todos los ficheros proporcionados, ya que algunos de ellos no han resultado útiles para alcanzar el objetivo propuesto. Como se ha mencionado previamente, la tarea es predecir el rating de un restaurante en base a sus atributos y a su localización. Sabiendo esto, se ha utilizado, tan solo, el siguiente fichero: *business\_data.csv*. El resto de ficheros han sido excluidos porque no contienen datos útiles para cumplir el propósito.

### Filtrado

Los ficheros de datos proporcionados para realizar el análisis contenían una cantidad muy extensa de datos y además algunos de ellos no resultaban interesantes para materializar los objetivos. Por ello, y con el fin de tener una visión general más clara de los datos (tipos de datos, organización de estos...), ha sido necesario realizar un proceso de filtrado de datos, el cual ha sido realizado en el [notebook de filtrado de datos](#).

Para desempeñar esta tarea se ha utilizado la librería *Pandas Profiling* [6], que permite obtener un análisis completo de los sets de datos, indicando entre otros aspectos variables conflictivas en cuanto a cardinalidad, uniformidad, distribución, etc. Una vez realizado este filtrado inicial, se volvió a examinar los datos con el fin de determinar qué variables no eran interesantes o cuáles podrían resultar conflictivas a la hora de hacer el análisis.

Se ha dividido el filtrado de los datos en distintas etapas dependiendo del tipo de información que se estuviera tratando en cada momento. Han aparecido situaciones en las que primero se ha realizado el filtrado y después se han examinado los datos mediante el

profiling ya que la librería Pandas Profiling consume mucha memoria RAM y se demora una gran cantidad de tiempo cuando se tiene un conjunto de datos grande. Las etapas en las que se ha dividido el proceso de filtrado, han sido las siguientes:

## Filtrado de Negocios

En este apartado se ha utilizado el fichero *business\_data.csv* para proceder al filtrado de datos. Al realizar el profiling se puede observar que muchas de las variables tienen una alta cardinalidad y tres de ellas tienen celdas en las que faltan datos.

<code>business_id</code> has a high cardinality: 192609 distinct values	High cardinality
<code>name</code> has a high cardinality: 145046 distinct values	High cardinality
<code>address</code> has a high cardinality: 151976 distinct values	High cardinality
<code>city</code> has a high cardinality: 1203 distinct values	High cardinality
<code>zipcode</code> has a high cardinality: 17540 distinct values	High cardinality
<code>attributes</code> has a high cardinality: 93976 distinct values	High cardinality
<code>categories</code> has a high cardinality: 93385 distinct values	High cardinality
<code>hours</code> has a high cardinality: 51566 distinct values	High cardinality
<code>address</code> has 7682 (4.0%) missing values	Missing
<code>attributes</code> has 28836 (15.0%) missing values	Missing
<code>hours</code> has 44830 (23.3%) missing values	Missing
<code>business_id</code> is uniformly distributed	Uniform
<code>business_id</code> has unique values	Unique

Ilustración 3: Profiling de *business\_data.csv*

También se puede observar en la Ilustración 3 que la variable *business\_id* está uniformemente distribuida y además sus valores son únicos (es lógico que esto ocurra ya que, habitualmente, en los conjuntos de datos se suele utilizar una variable id para identificar únicamente cada registro). Se ha eliminado la variable *name* ya que se dispone el atributo *id* que es identificativo de cada negocio y que permitirá hacer un *join* con el resto de tablas. Por otra parte, se puede tener una idea de dónde se localiza un negocio a través de la variable *zipcode*, por lo que se han eliminado las columnas de la dirección y el estado del negocio. La variable *hours* también se ha eliminado del dataset ya que no resultaba interesante para lograr el objetivo que se tenía en mente. También se han eliminado las coordenadas en las que se localizan los negocios, ya que para esta sección no serán útiles.

---

No obstante, estas dos variables se recuperarán para realizar la representación en mapas de los negocios.

Si se presta atención a la imagen mostrada previamente, que contiene todas las características obtenidas mediante el profiling, se observa que algunas de las variables tienen valores incompletos. Estos valores se traducen en el dataset como valores vacíos o como valores NaN. Con el fin de almacenar tan solo la información que interesa, se procedió a eliminar todas esas filas que contienen valores vacíos o NaN.

## Segmentación

En este apartado se va a realizar una segmentación de los distintos negocios en función de varios parámetros. Para ello, se va a utilizar el fichero *business\_data.csv* que contiene toda la información correspondiente a los distintos negocios. Además, toda la segmentación realizada se llevará a cabo en el [cuaderno de segmentación de datos](#).

Este fichero fue almacenado en un dataframe y contenía los negocios correspondientes a los países de Canadá y Estados Unidos. Se debe destacar que el atributo *zipcode* de los distintos negocios era de vital importancia para realizar el predictor y Canadá contenía códigos postales alfanuméricos. Por todas estas razones, se han almacenado los negocios de cada país en dataframes. Es decir, los negocios de Canadá en un dataframe y los de Estados Unidos en otro. Se trabajará con los últimos mencionados, puesto que como se ha explicado, aparte de que existen muchos más negocios, contienen códigos postales únicamente numéricos.

Una vez se tienen los códigos de Estados Unidos, visualizando los distintos negocios se pueden observar 7 grandes concentraciones de puntos que concuerdan exactamente con 7 ciudades en 7 estados diferentes. Por ello, se ha decidido volver a segmentar los negocios en distintos data frames, cada uno correspondiente con la ciudad del estado que contiene la concentración de puntos. Esto se ha realizado investigando acerca de las distribuciones de zipcodes que se utilizan en Estados Unidos. La información proporcionada por [12] ha facilitado mucho esta tarea.

---

## Agregación de atributos relevantes

Como último paso antes del análisis visual, se llegó a la conclusión de que tan solo se tenía un atributo que podía servir como medida de calidad o popularidad de los restaurantes: el rating. Por ello, se procedió a añadir otros atributos que caracterizaran a dichos restaurantes y que aportaran más información en lo que a calidad se refiere.

Para ello, en el [cuaderno auxiliar](#) se añadió una columna procedente del set de datos original: los atributos. Estos consistían en diccionarios con varias claves que representaban los atributos de cada restaurantes y los valores asociados a cada atributo. Para poder utilizar estos atributos tanto en el análisis visual como en los procesos de aprendizaje automático, debimos seguir una serie de pasos bien definidos:

1. Realizar un 'merge' con el dataset original para añadir la columna de atributos en forma de diccionarios.
2. Eliminar los NaN sustituyéndolos por diccionarios vacíos
3. Aplicar una evaluación de literales utilizando la librería AST de Python [9]. Esta librería permite realizar una evaluación de objetos con la finalidad de convertirlos a texto
4. Utilizar la librería Json\_normalize de Python [10] para convertir dichos diccionarios a columnas, de tal manera que cada columna se corresponda con un atributo distinto.
5. Filtrar dichos atributos, obteniendo sólo aquellos que son relevantes para nuestro estudio.

Tras estos pasos, el resultado es el siguiente:

	business_id	city	num_reviews	open	rating	zipcode	GoodForKids	NoiseLevel	RestaurantsDelivery	Caters	WiFi	RestaurantsGoodForGroups
0	gnKjwL_1w79qoiV3IC_xQQ	Charlotte	170.0	1.0	4.0	28210.0	True	average		False	False	no
1	1Dfx3zM-rW4n-31KeC8sJg	Phoenix	18.0	1.0	3.0	85016.0	True	None		False	None	no

OutdoorSeating	HasTV	RestaurantsReservations	RestaurantsPriceRange2
False	True	True	2
False	False	False	1

*Ilustración 4. Conjunto de atributos de negocios*

## 5. EDA. Análisis Visual de Datos

En esta sección se describirá el análisis visual realizado a los datos, incluyendo gráficas, mapas, diagramas, etc. Todo este análisis se ha realizado en los cuadernos que se especifica en cada subsección.

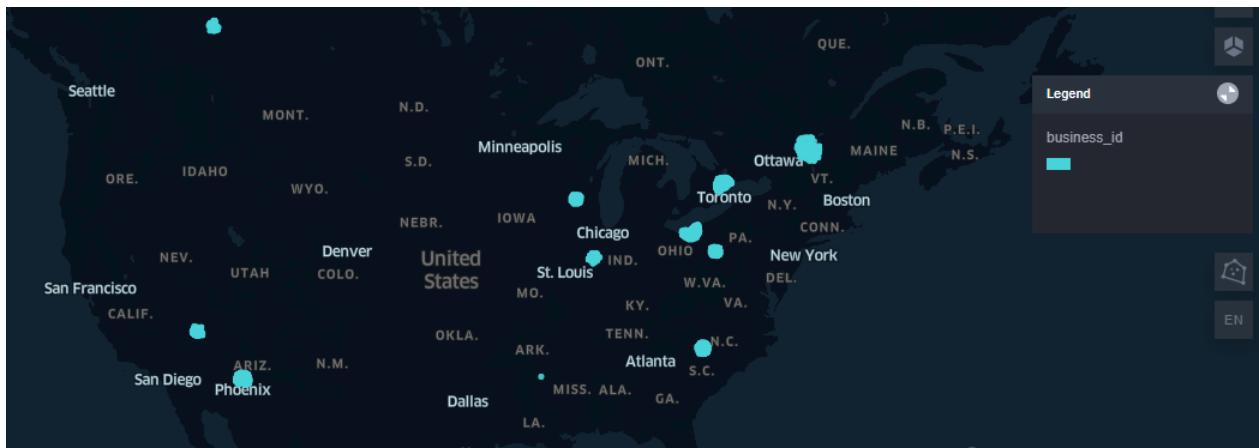
### Análisis Visual Inicial

Como análisis visual inicial, es decir, con los datos sin filtrar, se muestran los datos a través de mapas mediante la librería Kepler [3]. Dado que el objetivo del estudio está relacionado con la localización de los diferentes negocios y su relación con la puntuación de los usuarios, un primer acercamiento a los datos mediante mapas es adecuado. Este trabajo se ha llevado a cabo en el [cuaderno de mapas sin filtrar](#).

Se muestran 6 diferentes mapas:

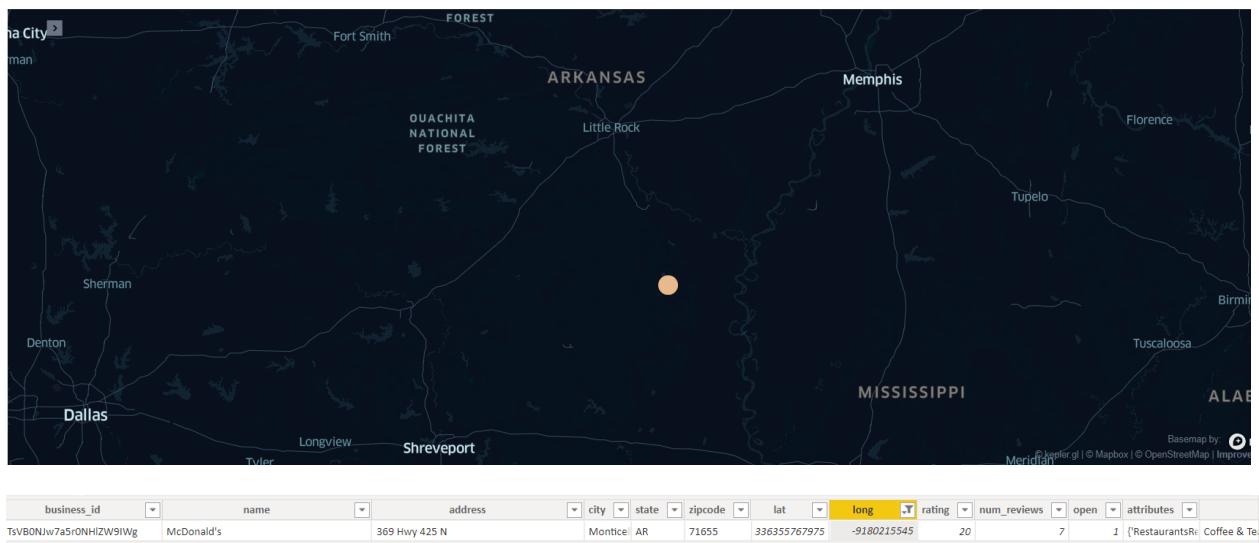
#### Mapa: Id's

El primero muestra todos los negocios con su id. Esto sirve para ver la distribución de todos los negocios.



*Ilustración 5. Mapa de negocios mostrados por business\_id*

Se aprecia en la Ilustración 5 cómo los negocios se concentran en 10 puntos del mapa, los cuales suelen ser ciudades como por ejemplo Las Vegas o Phoenix. Aparte, si se sube el radio de los puntos que representan un negocio, se puede encontrar un **outlier posicional** que puede apreciarse en la Ilustración 6, es decir, un valor atípico en cuanto a localización se refiere. Este negocio se encuentra en el estado de Arkansas, y filtrando usando su longitud, la cual es única en el dataset, se encuentra que es un McDonald's.



*Ilustración 6. Outlier posicional en Arkansas*

Mediante este mapa también se realizó la comprobación de los fallos al atribuir un estado a la dupla latitud y longitud como se comenta en el filtrado.

En estos puntos de agrupaciones de negocios serán en los que posteriormente se crearán los modelos, ya que para poder precisar, se considera más apropiado crear un modelo individual para cada zona.

## Mapa: Rating

En este mapa se muestran los negocios con una escala de color que depende de su rating, siendo el verde más claro la puntuación más baja, 1, y el verde más oscuro la más alta, 5.

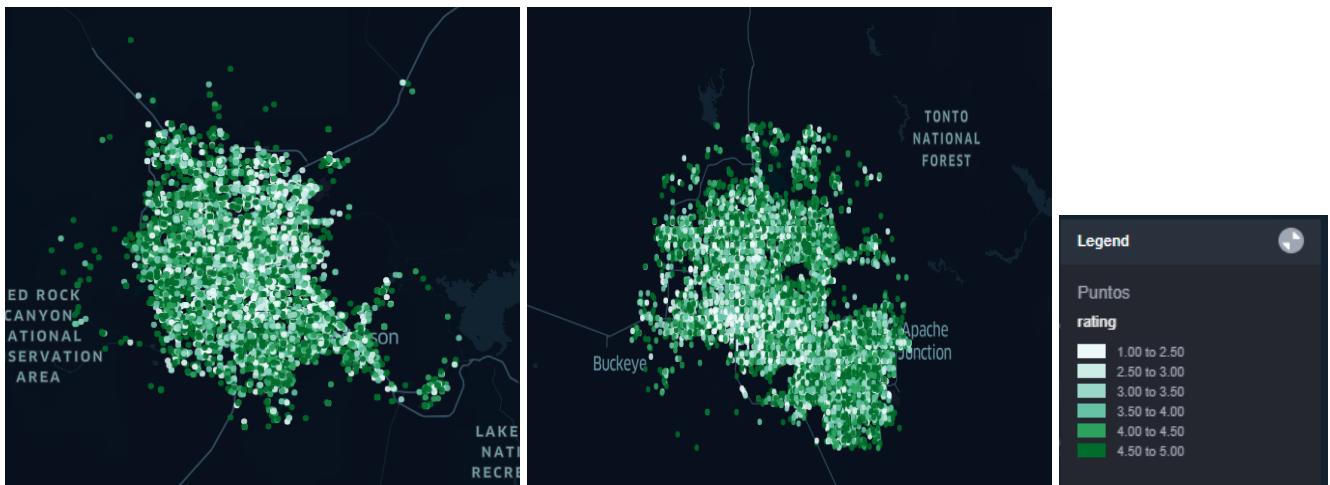


Ilustración 7. Mapa de rating en dos zonas distintas

Por ejemplo, se pueden observar en la Ilustración 7 la zona de Las Vegas a la derecha y la zona de Phoenix a la izquierda. Mediante el análisis de este mapa se formuló la pregunta de si los negocios en zonas más apartadas del centro o menos pobladas eran valorados mejor que los que estaban en zonas más concurridas, ya que se aprecia mayor acumulación de puntos claros en el centro de las zonas, y más oscuros en los bordes.

## Mapa: Negocios Abiertos

En este mapa simplemente se muestra los colores de los negocios en función de si están abiertos o cerrados, indicando que está abierto el color azul y rojo que está cerrado.



*Ilustración 8. Mapa de los negocios cerrados y abiertos. (En azul cerrados y en rojo abiertos)*

Como ejemplo, se puede observar en la Ilustración 8 la zona de Phoenix en el mapa. Se aprecia que la mayoría de los negocios están abiertos, y que los negocios cerrados no siguen ningún patrón posicional perceptible.

### **Mapa: Número de Reviews**

Mediante un mapa de calor se van a representar las zonas donde se han realizado más y menos reviews. En el mapa las zonas de color amarillo son dónde más abundan las reseñas y las más rojas dónde menos.

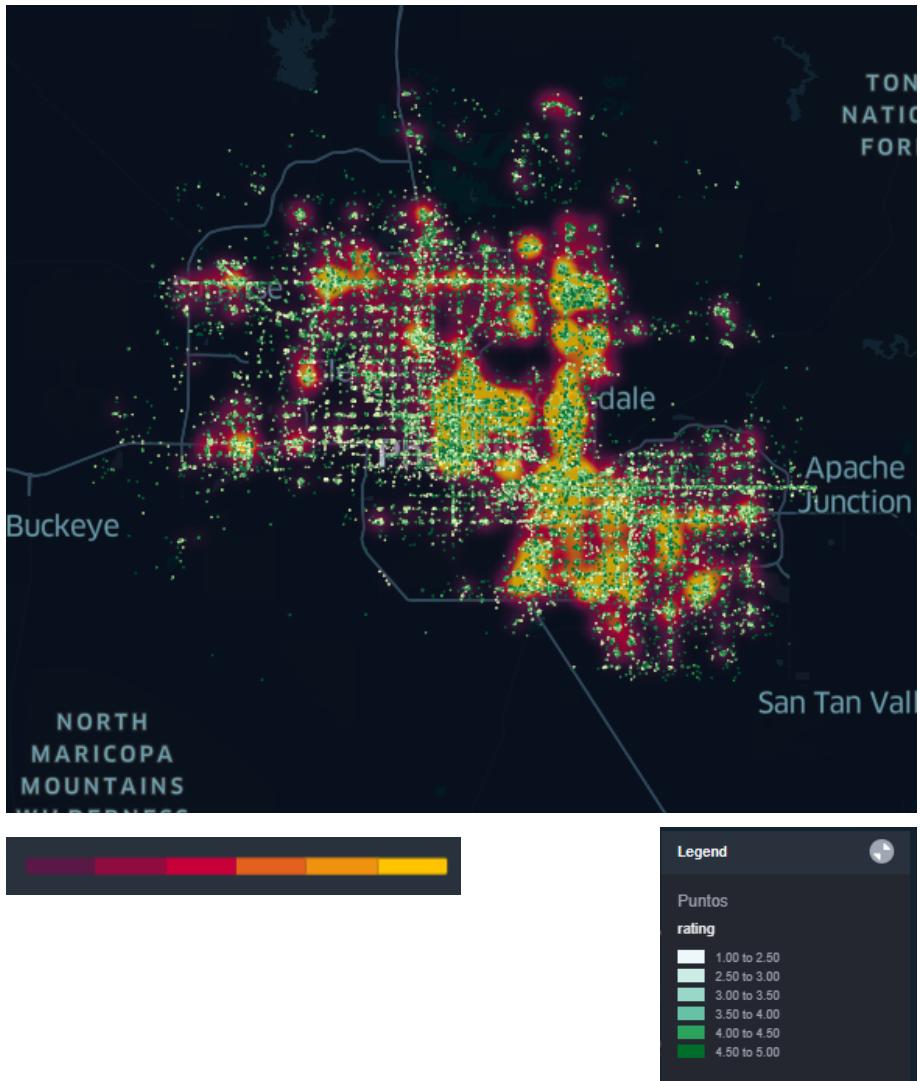


*Ilustración 9. Mapa de calor según las reseñas en dos zonas distintas. La leyenda representa la intensidad del número de reviews (a más brillo, más reviews).*

De nuevo, se puede ver en la Ilustración 9 a la izquierda el territorio de Las Vegas y a la derecha Phoenix, donde se aprecia claramente gracias al mapa, las zonas con negocios con más reseñas, y por tanto, se podría inferir que más visitados.

### Mapa: Contraste

Este mapa es una mezcla del mapa de rating y el de número de reviews. En él se ven superpuestos el mapa de color y el de puntos para poder apreciar como se distribuye el rating en relación con el número de reseñas.



*Ilustración 10. Representación de número de reseñas y ratings de los negocios. La intensidad del calor representa el número de rating (a más brillo, más rating)*

En la Ilustración 10 se puede ver el territorio de Phoenix. En este caso se aprecia una concentración sin distribución aparente en el centro. Sin embargo, en una zona caliente en la esquina superior derecha de la zona se ve una concentración de buenos ratings.

## Mapa: Zipcodes

Este mapa sirve básicamente para observar las zonas asociadas a un código postal dentro de un territorio, ya que como se indicó anteriormente, se usarán para agrupar los datos de

los negocios. En la Ilustración 11 se puede observar de nuevo Phoenix dividido por colores por el zipcode.

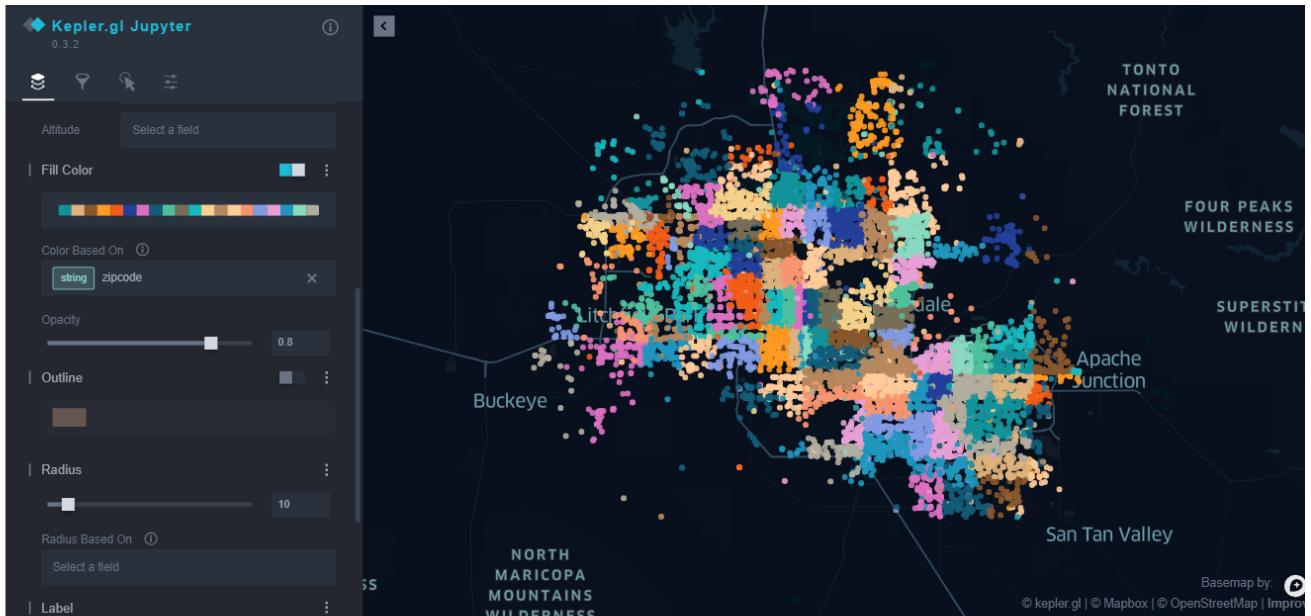
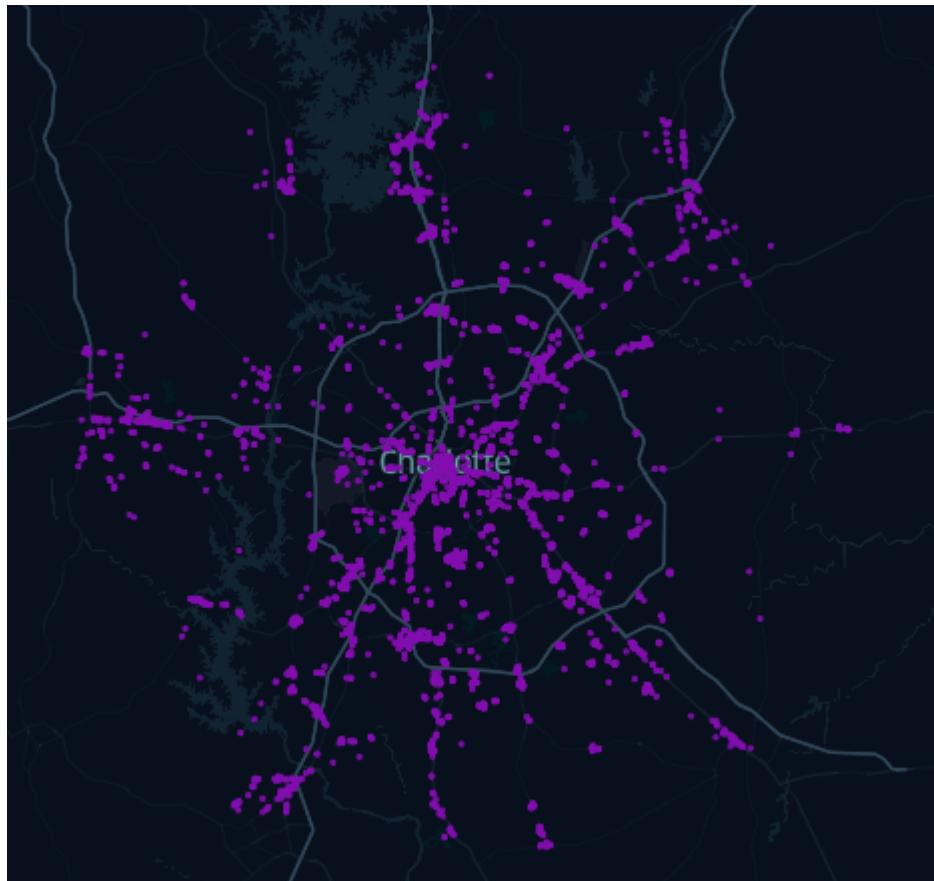


Ilustración 11. Mapa de una zona separada por Zipcodes

## Mapas Filtrados

Para una mejor visualización de las distribuciones, se ha decidido mostrar nuevos mapas una vez realizado el filtrado de datos, para ver si se obtiene nueva información sobre los negocios de nuestro interés. Esta tarea se ha llevado a cabo en el [cuaderno de mapas filtrados](#).

### Mapa: Restaurantes



*Ilustración 12. Mapa de restaurantes en Charlotte*

Este mapa se hizo con la intención de ver si los restaurantes estaban distribuidos de alguna forma específica. En la Ilustración 12, se puede ver que siguen un mismo patrón (en todas las ciudades a parte del ejemplo). Suelen estar muy concentrados en el centro de la ciudad, seguidos por las calles principales que ramifican del centro.

### **Mapa: Rating**

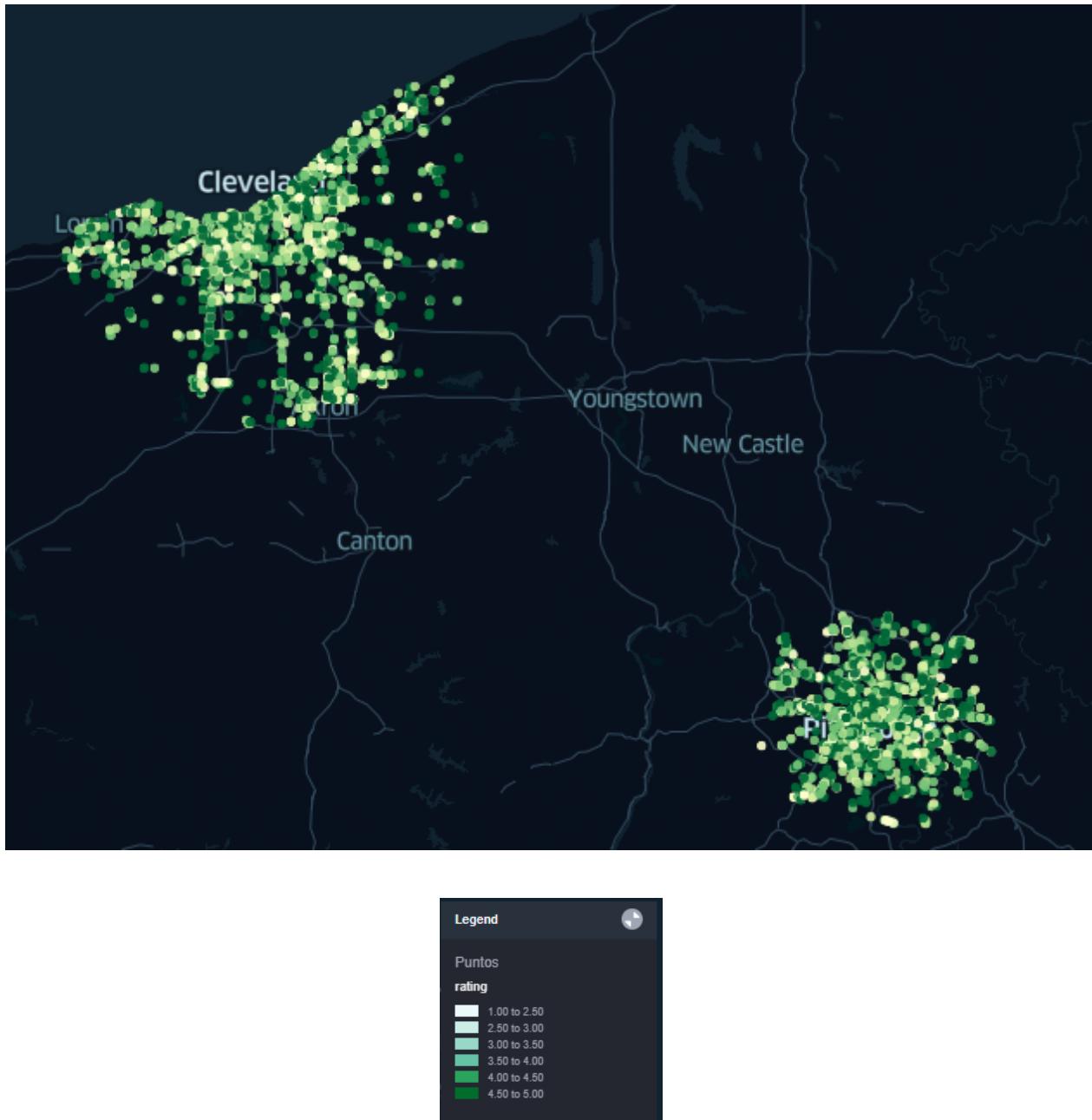


Ilustración 13. Mapa de rating en dos zonas distintas

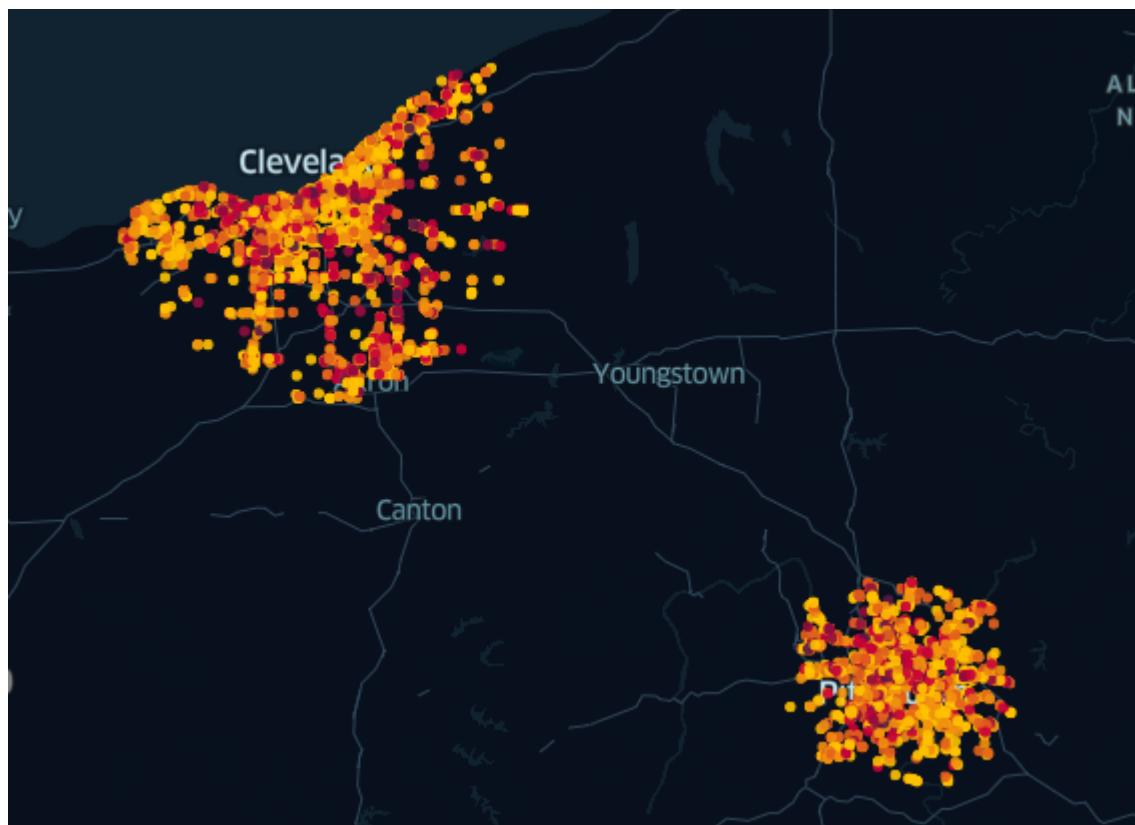
Este mapa se realizó con la intención de ver si las valoraciones de los restaurantes siguen alguna distribución específica. Para ello, se muestra en la Ilustración 13 cada punto con su

---

valoración en una escala de blanco a verde (menor valoración es más blanco, mientras que mayor sería más verde)

Como se puede apreciar en estas dos concentraciones de puntos de ejemplo, se pueden observar ligeras diferencias en los patrones de las valoraciones de los usuarios. Se pueden ver también distribuciones muy homogéneas. Por lo tanto, parece que la localización de un restaurante puede llegar a afectar directamente a su valoración.

### Mapa: Número de Reviews





*Ilustración 14. Número de revisiones en dos estados distintos*

En el mapa de la Ilustración 14, el objetivo era mostrar los restaurantes con su número de valoraciones, para poder ver si, de nuevo, siguen algún patrón.

En este caso, se puede ver que sí que parecen seguirlo, aunque es probable que sea independiente en cada ciudad. Por ejemplo, en el bulto de abajo, se pueden ver muchas menos valoraciones en la zona sur que en la norte. Sin embargo, en el bulto de la ciudad de Cleveland, se puede ver que en la zona izquierda y en ciertas partes del centro no hay muchas valoraciones.

Esto podría ser un indicativo de menos clientela en aquellos sitios con menor número de valoraciones, por lo tanto es algo que se considerará a estudiar a posteriori.

## Análisis Visual después del Filtrado

En esta sección se describe todo el análisis visual después de haber aplicado filtrado a los datos de los restaurantes. En concreto, se mostrarán análisis descriptivos y visuales que traten de esclarecer de qué datos se disponen y la utilidad de estos. Este trabajo se ha realizado en el [cuaderno de análisis visual de datos](#).

En primer lugar, se realizará un análisis descriptivo rápido de los datos de los restaurantes con la finalidad de detectar anomalías tempranas y de estudiar la dispersión de los datos.

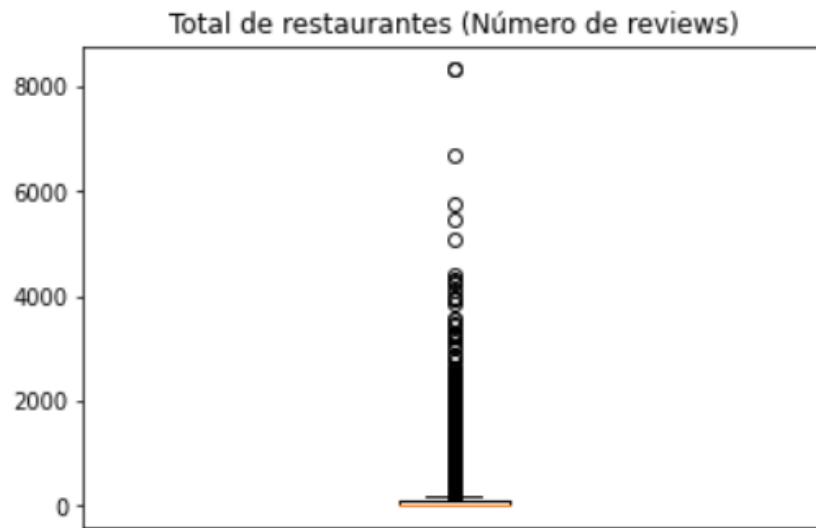
	<b>num_reviews</b>	<b>open</b>	<b>rating</b>	<b>zipcode</b>
<b>count</b>	35826.000000	35826.000000	35826.000000	35826.000000
<b>mean</b>	92.351002	0.697008	3.436568	62479.063418
<b>std</b>	218.222034	0.459558	0.817048	27891.888866
<b>min</b>	3.000000	0.000000	1.000000	5440.000000
<b>25%</b>	10.000000	0.000000	3.000000	44022.000000
<b>50%</b>	30.000000	1.000000	3.500000	85016.000000
<b>75%</b>	90.000000	1.000000	4.000000	85331.000000
<b>max</b>	8348.000000	1.000000	5.000000	93013.000000

*Ilustración 15. Análisis descriptivo de los restaurantes*

La Ilustración 15 muestra un análisis descriptivo simple de los negocios. De la información mostrada pueden sacarse varias conclusiones. En primer lugar, el número de reviews u opiniones (`num\_reviews`) muestra una desviación estándar de 218, lo cual es muy grande si se observa la media, la cual es de tan solo 92. Además, se puede observar que el número máximo de reviews que presenta el set de datos es de 8348, número considerablemente mayor que la media, por lo que se puede prever que habrá datos atípicos u outliers. En cuanto a las variables que aporta información sobre los negocios abiertos o cerrados (`open`), lo cierto es que no se logra detectar ninguna anomalía, aunque sí que será una variable útil para demostrar si los negocios abiertos son sinónimo de que tienen éxito y, por contrapartida, si los negocios cerrados implican que no han tenido éxito. En cuanto al rating, si se observa detenidamente la media y los percentiles, se puede ver que los valores están balanceados. Por último, el código postal o `zipcode` no aporta demasiada información, aunque sí que será útil para realizar ciertas segmentaciones.

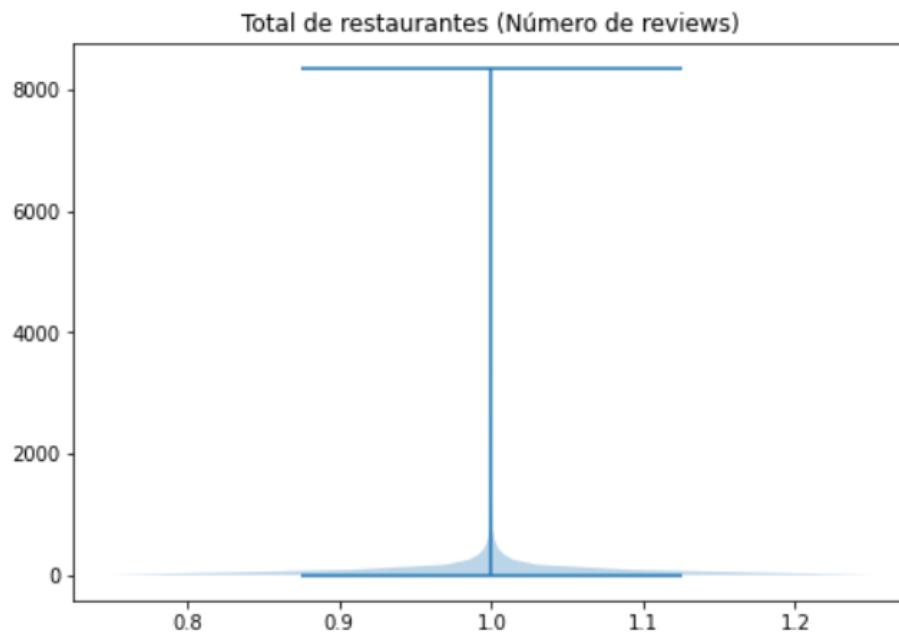
---

El número de reviews es una variable importante a la hora de estudiar el éxito de un restaurante o la puntuación de este. A priori es difícil saber si tendrá una gran correlación con el rating de los restaurantes; sin embargo, se hará un estudio individual.



*Ilustración 16. Boxplot generado del número de reviews de restaurantes*

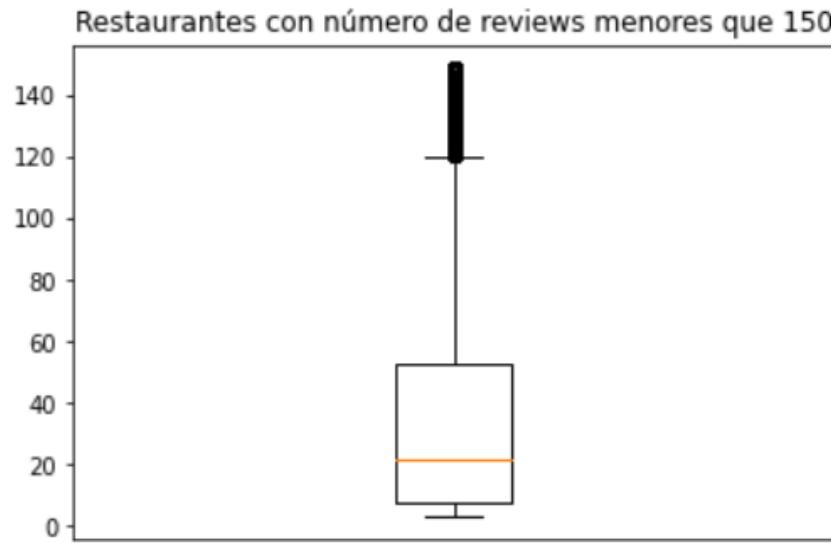
La Ilustración 16 muestra un diagrama de cajas relativo al número de reviews u opiniones de los restaurantes. Como puede verse, la forma del diagrama no aporta demasiada información. Los datos están distribuidos en rangos de número de reviews muy pequeños y, además, se tiene una gran cantidad de outliers o datos atípicos. Esto puede ser indicativo de que los datos presentan varias distribuciones o que simplemente hay varios outliers que deben ser estudiados. En cualquier caso, en los siguientes diagramas se analizará este caso.



*Ilustración 17. Diagrama de violín del número de reviews para el total de restaurantes*

El diagrama de violín mostrado en la Ilustración 17 no deja constancia de que existan otras distribuciones en los datos, por lo que se puede concluir que existe una gran cantidad de datos atípicos pero no siguen ninguna distribución.

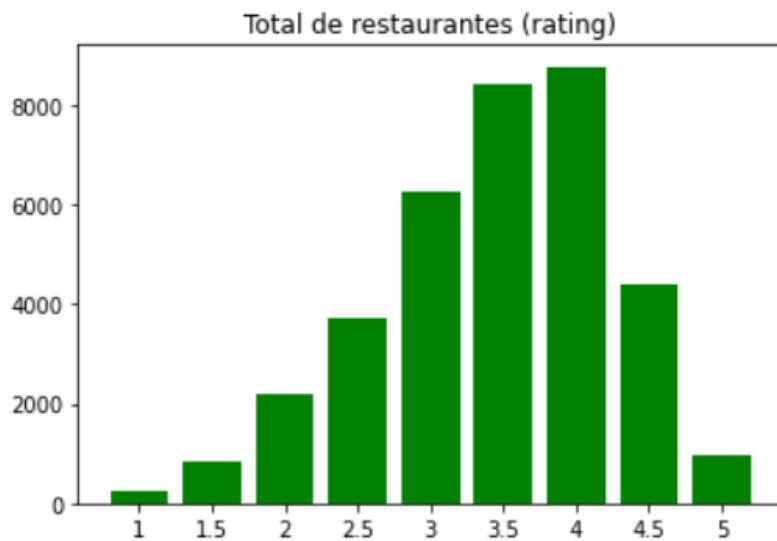
Por tanto, se debe proceder a realizar una pequeña segmentación de esta variable para estudiar el segmento de los datos donde se encuentran los números de reviews más comunes.



*Ilustración 18. Boxplot de restaurantes con número de reviews menores que 150*

A modo de experimento, se han extraído aquellos restaurantes en los que el número de reviews sea menor a 150. Como puede verse, ahora el diagrama es mucho más legible y puede extraerse la conclusión de que la mayor parte de los restaurantes de Estados Unidos tienen un número de reviews reducido.

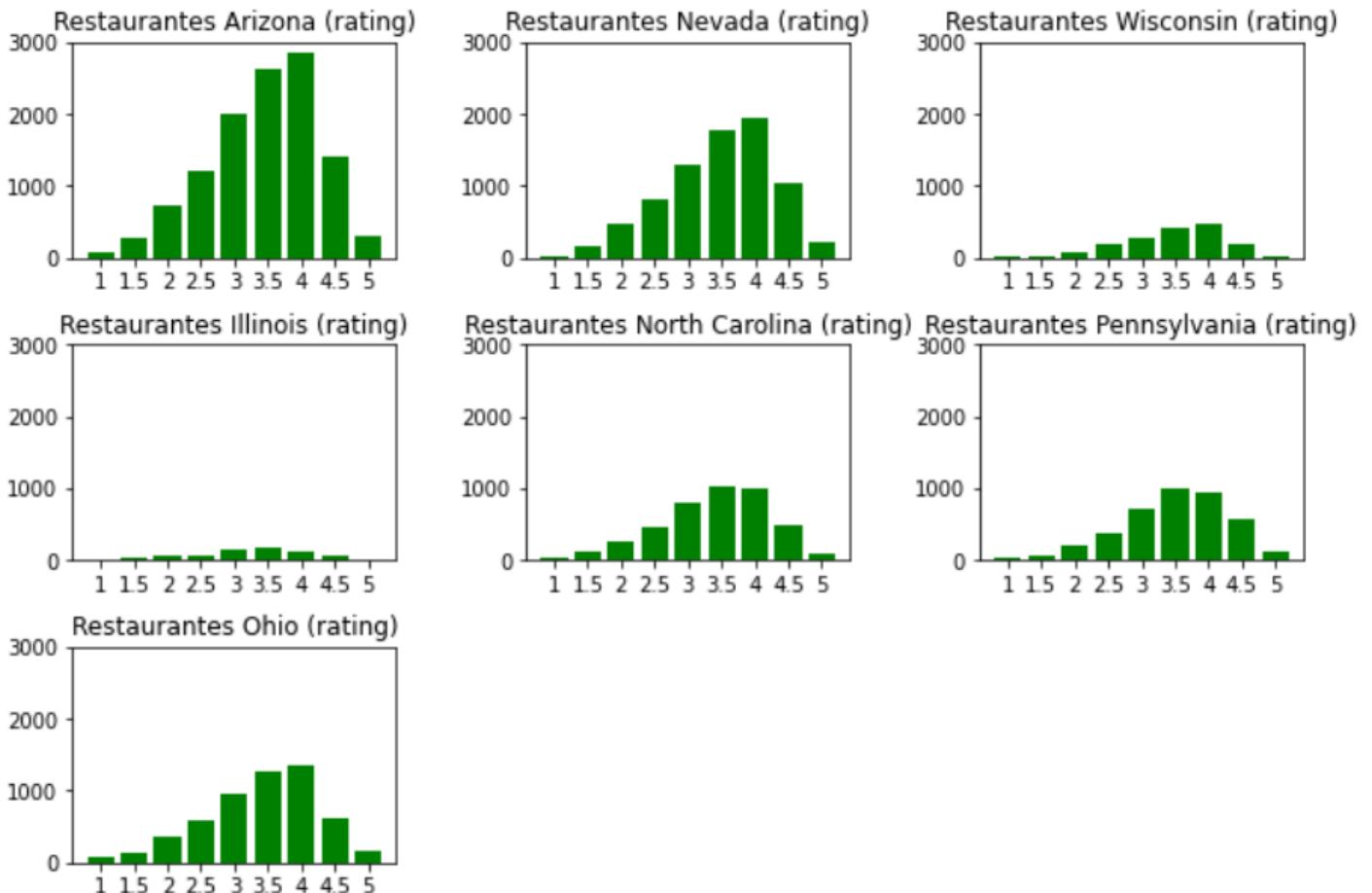
A continuación se estudiará el rating o puntuación de los restaurantes. Esta será la variable que se tratará de predecir en el futuro por lo que resulta crucial conocer su distribución.



*Ilustración 19. Diagrama de barras del rating para todos los restaurantes*

La Ilustración 19 muestra un diagrama de barras para el rating de todos los restaurantes extraídos para Estados Unidos. Como puede verse, los datos siguen una distribución normal ligeramente desplazada hacia la derecha, lo que parece indicar que hay cierta tendencia hacia reviews con puntuaciones altas.

No obstante, no tiene demasiado sentido analizar los ratings de esta forma, puesto que no se tienen datos que se distribuyen uniformemente a lo largo de todo Estados Unidos, aunque se ha realizado de todas formas para observar de manera general la distribución de los datos. Por tanto, se deben obtener concentraciones de datos en ciertas regiones, lo que ha podido verse en el [cuaderno de mapas previos al filtrado](#). Por ello, se compararán los ratings de cada una de estas zonas para ver si siguen alguna tendencia común o, por el contrario, deben ser analizadas de manera independiente.



*Ilustración 20. Histograma de los ratings de los estados estudiados.*

La Ilustración 20 muestra una comparación entre las distribuciones de ratings de las distintas zonas de Estados Unidos elegidas. Como puede observarse, las zonas muestran distribuciones similares, que pueden aproximarse a una distribución normal. Sin embargo, podemos ver que las zonas no contienen el mismo número de negocios. Por estas razones parece oportuno realizar las predicciones para cada una de estas zonas, dado el hecho de que poseen distribuciones parecidas.

Adicionalmente, se estudiará la relación entre la variable relativa a la puntuación de los negocios (rating) y al número de opiniones de los negocios. Para ello, se ha realizado el siguiente diagrama.

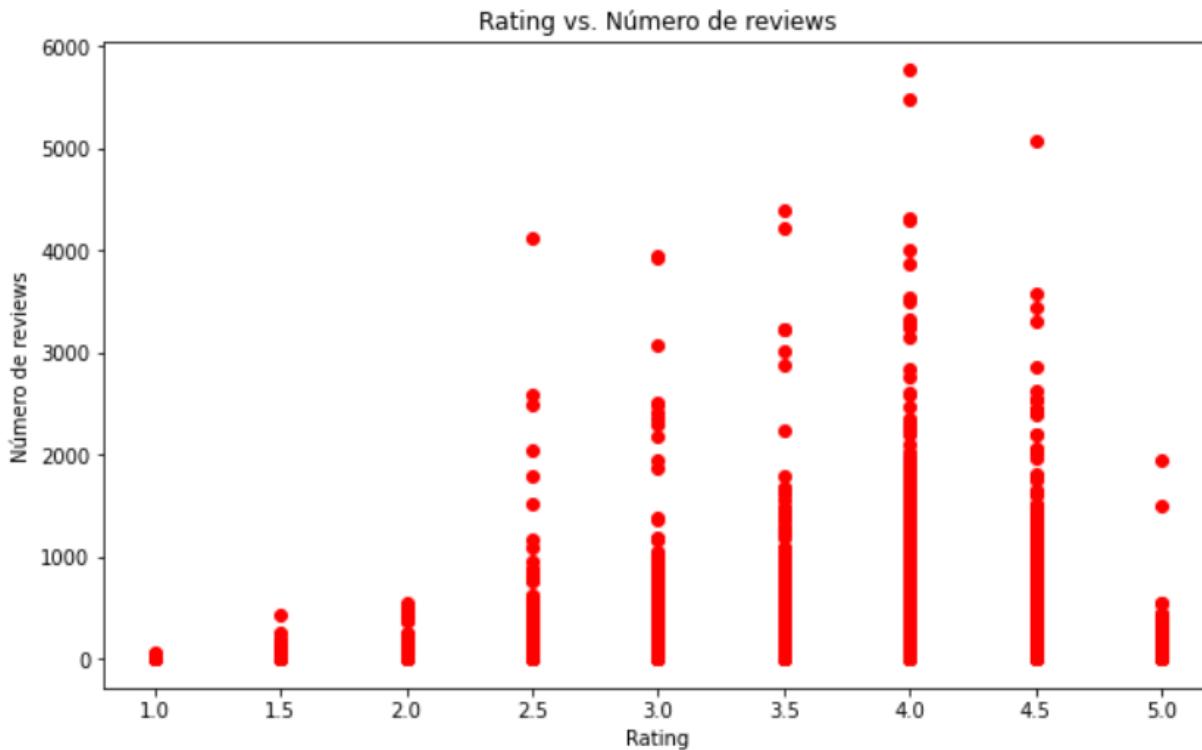
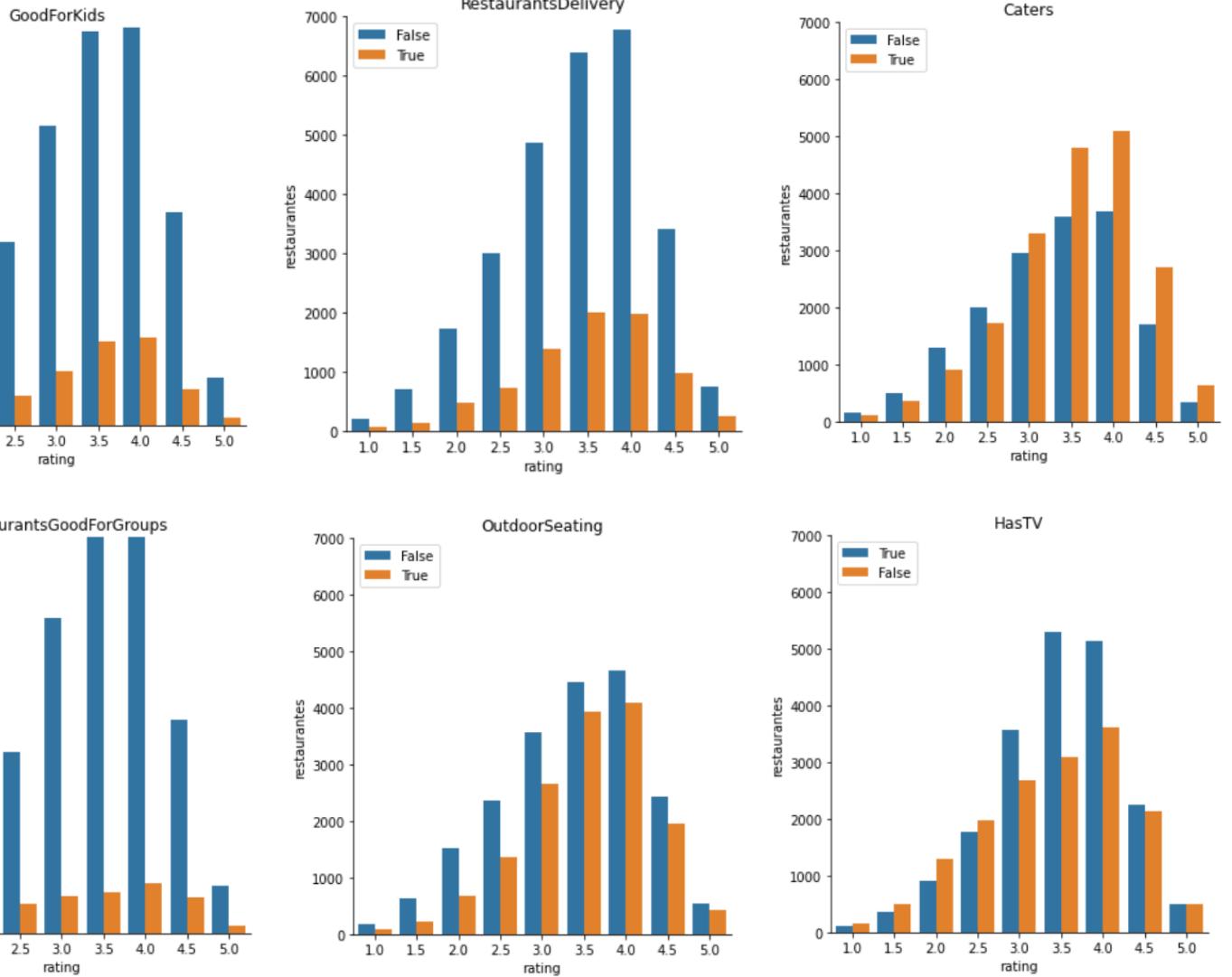
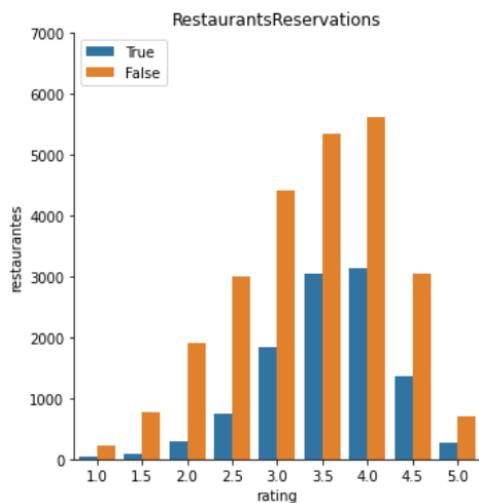


Ilustración 21. Diagrama de dispersión para rating y num reviews

La Ilustración 21 muestra un diagrama de dispersión que relaciona las variables `rating` y `num\_reviews`. El objetivo de esta evaluación es estudiar si ambas variables presentan alguna correlación, puesto que se suele pensar que cuantas más opiniones tiene un restaurante mayor será su puntuación. La teoría se confirma. Como se ve en el diagrama, el rating es mejor cuanto mayor es la concentración del número de opiniones de los restaurantes.

Los restaurantes proporcionan una gran cantidad de atributos interesantes que pueden llegar a servir como medida de calidad de los mismos. De entre todos, se han elegido 7 que se consideran los más significativos para las futuras tareas de predicción.





*Ilustración 22. Histogramas de los atributos de restaurantes*

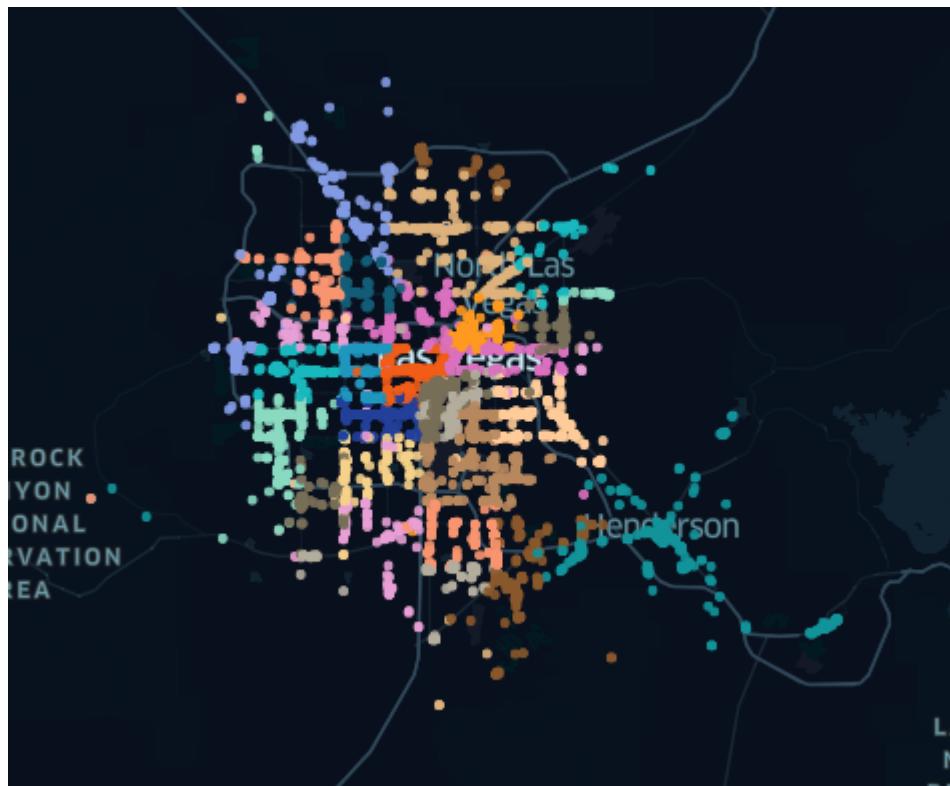
La Ilustración 22 muestra gráficos de barras para cada uno de los atributos, relacionando en función del número de restaurantes y el rating. Si se observa la variable `GoodForKids` se podrá ver la distribución de los restaurantes que son recomendados para niños y los que no lo son, en función del rating. Como puede observarse en este gráfico, la mayoría de restaurantes que son recomendados para niños se mueven en ratings bastante elevados. Es decir, la mayor concentración de restaurantes se alcanza en torno al 3.5 y 4 de rating, que se corresponden con puntuaciones que se consideran elevadas. Lo mismo ocurre con otros atributos como `RestaurantsDelivery`, `RestaurantsGoodForGroups` y `HasTV`; la razón que se ha deducido es que en general las personas prefieren restaurantes que sean recomendados tanto para niños (en el caso de parejas, por ejemplo) como para grupos (en el caso de grupos de amigos o cenas empresariales), y que además dispongan de televisores. Por otra parte, se puede ver que hay otras variables en las que ocurre lo contrario: es el caso de las reservas. Como se ve, la mayoría de los restaurantes que no necesitan reserva previa en general tienen mayor rating. Esto tiene sentido si se tiene en cuenta que a la mayoría de la gente le resulta mucho más cómodo ir a un restaurante que no necesita reserva previa. Por último, en variables concretas como `Caters` o `OutdoorSeating` no se ha logrado sacar conclusiones fructíferas, puesto que en ellas el contraste entre valores de variables no es tan notorio.

## 6. Aprendizaje Automático

En esta sección se describe todo el proceso de aprendizaje automático llevado a cabo para la generación de modelos de predicción. En concreto, este trabajo se ha realizado en el [cuaderno de aprendizaje automático](#).

## Preparación de los datos

Una vez se ha obtenido el dataframe deseado (solo con los restaurantes) se va proceder a hacer mapas por zonas. Al hacer ese mapa, se obtuvieron agrupaciones de restaurantes distribuidas por el mapa estadounidense en 7 grupos, cada uno en un estado distinto:



*Ilustración 23. Mapa de zipcodes para Las Vegas.*

---

Tras analizar e intentar hacer una división por zonas de cada una de las agrupaciones de restaurantes, se analizó que era imposible debido a que muchos de los zipcodes (o códigos postales) estaban incorrectos de acuerdo con la localización del restaurante como se puede ver en la Ilustración 23. Esto suponía un gran inconveniente debido a que la idea era dividir cada agrupación de puntos por zonas, dependiendo de su zipcode.

Por lo tanto, se decidió proceder a corregir estos zipcodes. Para ello, se hizo uso de la API Geoapify [8], la cual devolvía el zipcode de un punto dadas sus coordenadas en el mapa. Posteriormente, se corrigió toda la tabla de restaurantes.

Tras esto, ya se pudo proceder a separar las agrupaciones de puntos por zonas. Para ello, se hizo uso del previo análisis de mapas que se realizó al principio, para subjetivamente, decidir cuál de las siguientes divisiones era más adecuada para cada zona:

- Norte/Centro/Sur
- Oeste/Centro/Este

Para realizar esta división, se crearon varias listas para cada estado, donde se almacenaron los zipcodes correspondientes a cada zona. Todo esto se hizo manualmente, con ayuda de mapas de zipcodes de cada estado:

```
zip_north_NV = [89134,89128,89129,89124,89166,89143,89131,89130, 89108,89196,89032,89030,89031,89084,89085,89086,89087,89081,89115,89156,89149,89191,89136]  
zip_center_NV = [89138, 89144, 89145, 89107, 89106, 89101, 89118, 89135, 89117, 89147, 89146, 89103, 89102, 89169, 89121, 89104, 89142, 89122, 89158, 89154, 89152]  
zip_south_NV = [89161,89064,89124,89054,89148,89113,89118,89119,89178,89139,89141,89123,89183,89044,89052,89120,89011,89074,89012,89015,89002,89005,89124,89193,89014,89105,89016,89159,89111,89165,89199]
```

*Ilustración 24. Ejemplo de división de zipcodes para Nevada.*

Una vez conseguidas estas listas, simplemente se añadió a cada dataframe de estado una nueva columna 'zona', donde se asignó una zona a ese restaurante dependiendo de su Zipcode.

Aquí un ejemplo de cómo quedó la división del estado de Nevada (NV):

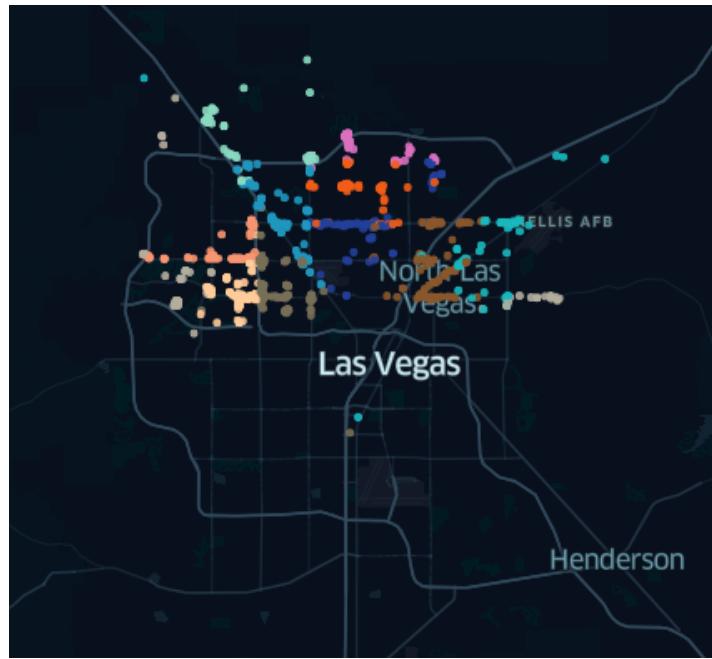


Ilustración 25. División por zipcodes Nevada (zona norte)

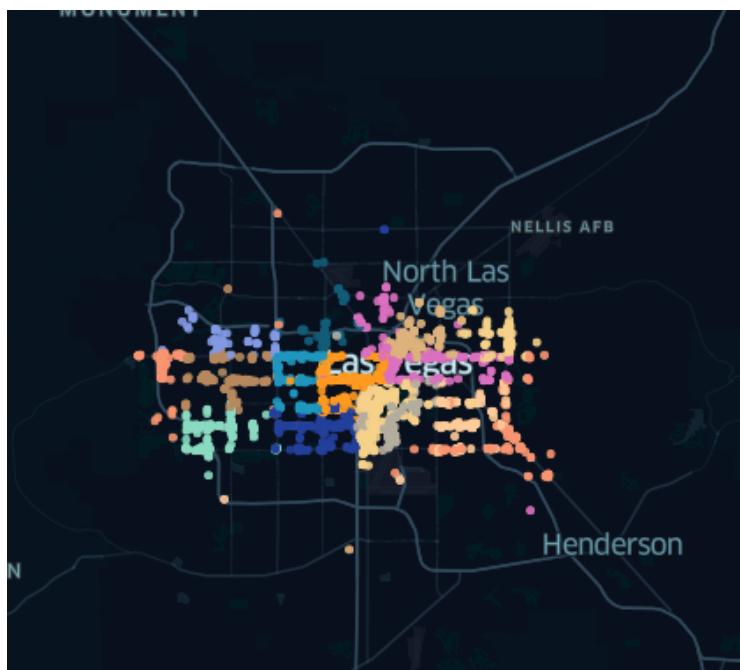


Ilustración 26. División por zipcodes Nevada (zona centro)

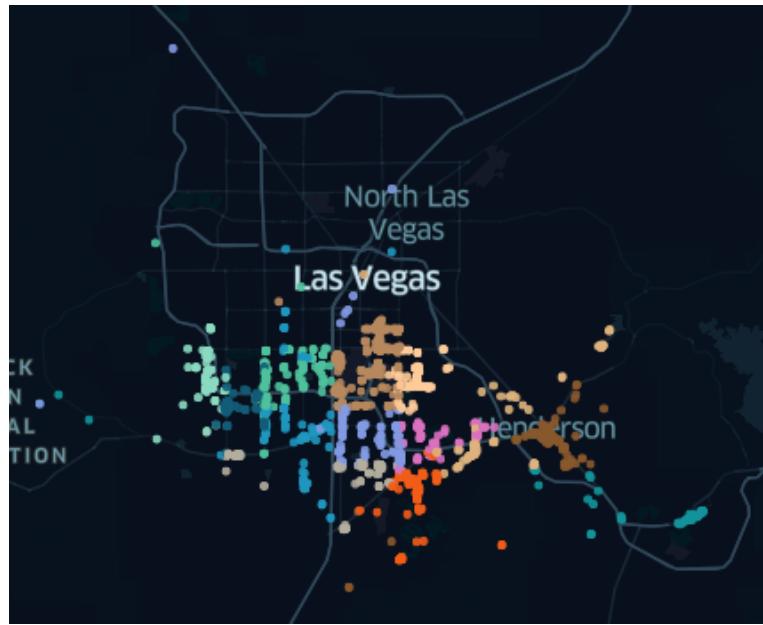


Ilustración 27. División por zipcodes Nevada (zona sur)

Como se puede observar, se aprecia claramente la división por zonas realizada en este estado. Esto mismo, se ha hecho con el resto de estados (Exceptuando IL, el cual no disponía de los suficientes datos para poder separarlo por zonas)

Una vez realizada la separación por zonas, se procedió a realizar la discretización de la variable ratings. En un principio se utilizaron tres clases para clasificar los ratings: "bajo" si el valor del rating es menor que 3, "medio" si el valor es mayor o igual que 3 y menor que 4, y "alto" si el rating es mayor que 4. Posteriormente, también se probó a hacer una discretización de esta misma variable, pero en esta ocasión, utilizando dos clases: "bajo" si el rating es menor o igual que 3,5 ó "alto" si el rating es mayor que 3,5.

Hay que destacar que también se discretizó la variable number of reviews ya que se presentaban una gran variedad en los valores con los que se trabajó. Se utilizaron cuatro clases: "baja" si el número de reseñas es menor o igual que 10, "medias" si el número de reseñas es mayor que 10 y menor o igual que 50, "altas" si es mayor que 50 o menor o igual que 200, y "muy altas" si el número de reseñas es mayor que 200.

---

## Elección de algoritmos

En este apartado se van a detallar los distintos algoritmos que se han seleccionado para generar los distintos modelos que realizarán la tarea de predicción.

En primer lugar se ha decidido utilizar un árbol de decisión con distintos parámetros. Se han probado los criterios de evaluación de caminos “gini” y “entropy” que proporcionarán resultados muy parecidos. Además, el criterio de expansión puede ser random o seleccionando aquel nodo que mejora los valores de los criterios de evaluación. Se han escogido este algoritmo por los siguientes motivos:

- **Fácil de entender e interpretar.**
- **Capaz de manejar tanto datos numéricos y categorizados.**
- **Robustez**
- **Funciona bien con grandes conjuntos de datos.**

También se ha utilizado Random Forest, que utiliza árboles de decisión pero de una manera mucho más compleja, incorporando varios árboles en lugar de un único árbol.

Este utiliza los parámetros de evaluación de un camino, que puede ser “entropy” y “gini”. Las ventajas que proporciona Random Forest son las siguientes:

- **Puede manejar una gran cantidad de variables de entrada e identificar las más significativas.**
- **Permite obtener la importancia de las variables**
- **Es posible usarlo como método no supervisado (clustering) y detección de outliers.**

En tercer lugar se decidió probar Boosting que permite combinar los resultados de clasificadores sencillos para obtener un clasificador más robusto. Para este algoritmo se ha utilizado el parámetro “estimators” o número de clasificadores que se refiere al número

---

modelos sencillos que se utilizan para construir el modelo final. Estos toman valores de 5, 30, 50 y 100. AdaBoost ha sido utilizado por las siguientes razones:

- **Tiene una precisión extremadamente alta**
- **La introducción de la aleatoriedad hace que el bosque aleatorio no sea fácil de sobreajustar, tiene una buena capacidad anti-ruido y no es sensible a valores atípicos de puntos anormales.**
- **Puede manejar tanto datos discretos como datos continuos**

También se ha usado knn, que utiliza los k vecinos más cercanos para seleccionar la clase mayoritaria en función de sus respectivas clasificaciones. En este caso, se utilizaron distintos números de vecinos, para ver si se obtienen resultados distintos en función de los ratings de los vecinos variando cuántos vecinos se utilizan. Este algoritmo se ha utilizado principalmente por su versatilidad y simplicidad.

Por último, se ha hecho uso del Perceptrón Multicapa combinando distintos hiper parámetros: se ha utilizado una tasa de aprendizaje adaptativa ("learning\_rate: adaptative) y se han utilizado distinto número de capas (se podrá ver con detalle en el apartado "Modelos Generados"). Los motivos por lo que se ha utilizado el Perceptrón Multicapa son los siguientes:

- **Tienen capacidad de aprender a realizar tareas basadas en un entrenamiento inicial**
- **Facilidad de inserción en la tecnología existente.**

## Generación de modelos

En este apartado se va a explicar brevemente el procedimiento seguido para generar los distintos modelos que realizan la tarea de predecir las valoraciones de los restaurantes.

En primer lugar, para generar los modelos se han definido los algoritmos mencionados anteriormente en el formato de un diccionario python. Para todos los restaurantes, se han

---

seleccionado aquellas columnas que se utilizarían para entrenar los modelos, que son: *num\_reviews*, *GoodForKids*, *NoiseLevel*, *RestaurantDelivery*, *WiFi*, *RestaurantsGoodForGroups*, *OutdoorSeating*, *RestaurantsReservations*, *RestaurantsPriceRange2* y *Zone*. Evidentemente, también se ha seleccionado el atributo que será la salida del modelo, *rating*. Posteriormente, se seleccionaron los valores de dichos atributos, se generaron las “folds” utilizadas para la validación. Para cada fold, se generaron los conjuntos de entrenamiento y validación, para luego ser balanceados mediante “undersampling”, es decir, las clases con más instancias se reducían hasta equiparar en tamaño a la clase con menos instancias. Luego se entrenaron distintos modelos en función de los clasificadores y sus hiper parámetros almacenados en el diccionario anteriormente. Esto quiere decir que para cada estado, se va a entrenar un modelo distinto, de forma que se tomarán todos los restaurantes pertenecientes a esa zona y se entrenarán tantos modelos como combinaciones de clasificadores con sus distintos hiper parámetros.

Por último, se han almacenado los resultados en un dataframe y se han mostrado para poder analizar los resultados obtenidos.

## Resultados

En esta sección se presentan los resultados finales obtenidos para los algoritmos utilizados. La técnica de evaluación utilizada será Cross Validation y, puesto que hemos generado modelos para cada una de las 7 zonas reconocidas durante el estudio, tan sólo se mostrarán los resultados relativos al estado de Arizona, a modo de ejemplo. Cabe destacar que se han obtenido dos tipos de modelos, los cuales se exponen a continuación.

### 3 clases - Bajo, Medio y Alto

En primer lugar, se ha realizado una discretización de la clase rating en 3 clases distintas:

- Bajo. Si el rating es menor que 3
- Medio. Si el rating es mayor o igual que 3 y menor que 4
- Alto. Si el rating es mayor o igual que 4

De esta manera, se intentará predecir el rating de los restaurantes de los restaurantes en función de sus características. Los resultados obtenidos para el caso de Arizona son los siguientes:

classifier	hyperparameters	cv_fold	accuracy	precision	recall	f1
AdaBoost	{'n_estimators': 100}	2	0.451406	0.451406	0.451406	0.451406
	{'n_estimators': 30}	2	0.452018	0.452018	0.452018	0.452018
	{'n_estimators': 50}	2	0.451144	0.451144	0.451144	0.451144
	{'n_estimators': 5}	2	0.423091	0.423091	0.423091	0.423091
Decision Tree	{'criterion': 'entropy', 'splitter': 'best'}	2	0.443803	0.443803	0.443803	0.443803
	{'criterion': 'entropy', 'splitter': 'random'}	2	0.441093	0.441093	0.441093	0.441093
	{'criterion': 'gini', 'splitter': 'best'}	2	0.441530	0.441530	0.441530	0.441530
	{'criterion': 'gini', 'splitter': 'random'}	2	0.444414	0.444414	0.444414	0.444414
K-Nearest Neighbors	{'n_neighbors': 10, 'p': 1}	2	0.418371	0.418371	0.418371	0.418371
	{'n_neighbors': 10, 'p': 2}	2	0.418633	0.418633	0.418633	0.418633
	{'n_neighbors': 15, 'p': 1}	2	0.428072	0.428072	0.428072	0.428072
	{'n_neighbors': 15, 'p': 2}	2	0.428072	0.428072	0.428072	0.428072
	{'n_neighbors': 5, 'p': 1}	2	0.413651	0.413651	0.413651	0.413651
	{'n_neighbors': 5, 'p': 2}	2	0.414088	0.414088	0.414088	0.414088
Multilayer-Perceptron	{'learning_rate': 'adaptive', 'hidden_layer_sizes': (5, 10)}	2	0.448347	0.448347	0.448347	0.448347
	{'learning_rate': 'adaptive', 'hidden_layer_sizes': 10}	2	0.453154	0.453154	0.453154	0.453154
	{'learning_rate': 'adaptive', 'hidden_layer_sizes': 5}	2	0.423703	0.423703	0.423703	0.423703
RandomForestClassifier	{'criterion': 'entropy'}	2	0.447299	0.447299	0.447299	0.447299
	{'criterion': 'gini'}	2	0.449483	0.449483	0.449483	0.449483

*Ilustración 28. Resultados de los modelos obtenidos para 3 clases*

Los resultados obtenidos se mueven entre el 40 y el 45 por ciento en todas las métricas utilizadas y para todos los algoritmos. A pesar de haber balanceado los datos, lo cierto es que los resultados son poco alentadores. Posiblemente esto se deba a que los atributos utilizados no sean lo suficientemente significativos para lograr predecir con exactitud el rating de los restaurantes. Por ello, a modo de experimento, se ha obtenido una gráfica que muestra las importancias de los distintos atributos aplicados al algoritmo de los árboles de decisión.



*Ilustración 29. Importancia de atributos para predicción de 3 clases*

Si se observa las escalas de importancia de la gráfica, se puede ver que los atributos son poco significativos en general, si bien es cierto que se puede detectar ciertos atributos que son más importantes que otros. Es el caso del número de reviews y zone, los más importantes, y luego el nivel de ruido, la existencia de WiFi en el restaurante, el precio o si tiene servicio a domicilio.

## 2 clases - Bajo y Alto

En este caso, hemos decidido reducir el número de clases de la siguiente manera:

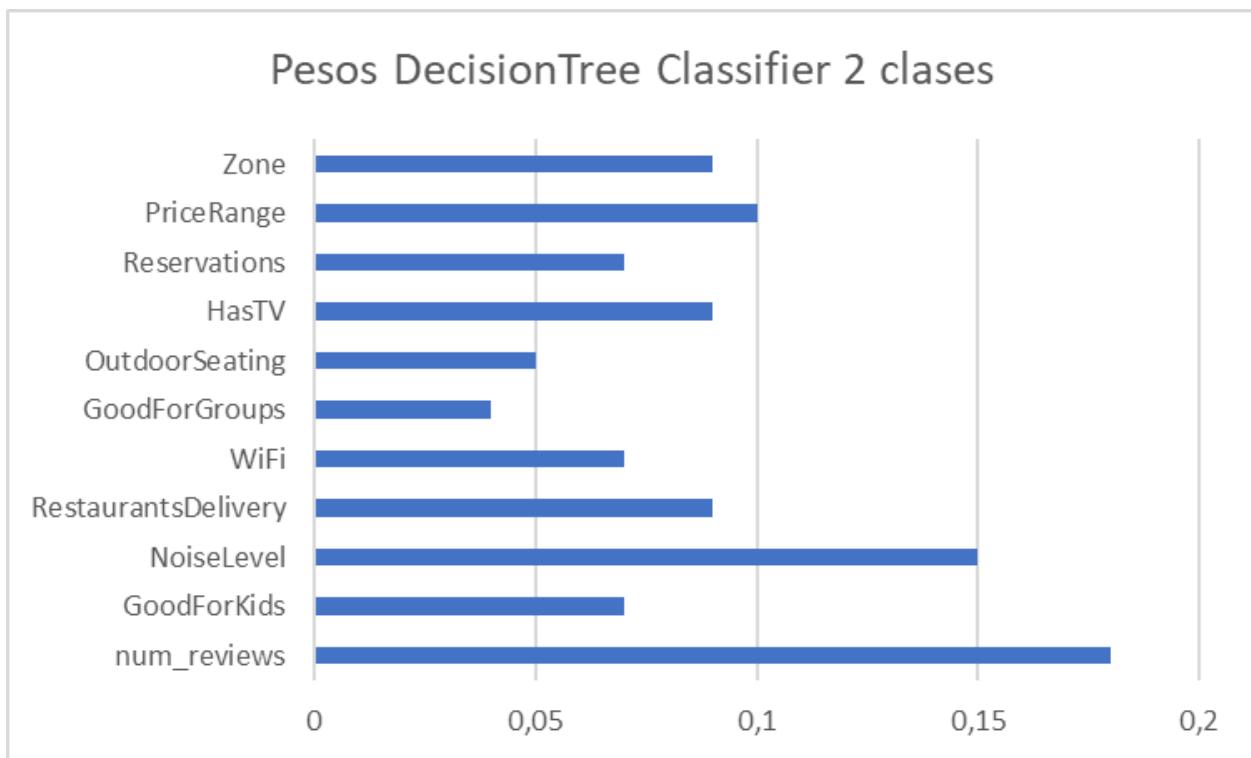
- Bajo. Si el rating es menor que 5.
- Alto. Si el rating es mayor o igual que 5

De nuevo, los resultados obtenidos para el caso de Arizona han sido los siguientes:

classifier	hyperparameters	cv_fold	accuracy	precision	recall	f1
AdaBoost	{'n_estimators': 100}	2	0.611432	0.611432	0.611432	0.611432
	{'n_estimators': 30}	2	0.611257	0.611257	0.611257	0.611257
	{'n_estimators': 50}	2	0.611345	0.611345	0.611345	0.611345
	{'n_estimators': 5}	2	0.606188	0.606188	0.606188	0.606188
Decision Tree	{'criterion': 'entropy', 'splitter': 'best'}	2	0.583989	0.583989	0.583989	0.583989
	{'criterion': 'entropy', 'splitter': 'random'}	2	0.582416	0.582416	0.582416	0.582416
	{'criterion': 'gini', 'splitter': 'best'}	2	0.584601	0.584601	0.584601	0.584601
	{'criterion': 'gini', 'splitter': 'random'}	2	0.583552	0.583552	0.583552	0.583552
K-Nearest Neighbors	{'n_neighbors': 10, 'p': 1}	2	0.459185	0.459185	0.459185	0.459185
	{'n_neighbors': 10, 'p': 2}	2	0.459534	0.459534	0.459534	0.459534
	{'n_neighbors': 15, 'p': 1}	2	0.482170	0.482170	0.482170	0.482170
	{'n_neighbors': 15, 'p': 2}	2	0.482782	0.482782	0.482782	0.482782
	{'n_neighbors': 5, 'p': 1}	2	0.460758	0.460758	0.460758	0.460758
	{'n_neighbors': 5, 'p': 2}	2	0.460845	0.460845	0.460845	0.460845
Multilayer-Perceptron	{'learning_rate': 'adaptive', 'hidden_layer_sizes': (5, 10)}	2	0.610387	0.610387	0.610387	0.610387
	{'learning_rate': 'adaptive', 'hidden_layer_sizes': 10}	2	0.607148	0.607148	0.607148	0.607148
	{'learning_rate': 'adaptive', 'hidden_layer_sizes': 5}	2	0.608635	0.608635	0.608635	0.608635
RandomForestClassifier	{'criterion': 'entropy'}	2	0.603829	0.603829	0.603829	0.603829
	{'criterion': 'gini'}	2	0.603742	0.603742	0.603742	0.603742

*Ilustración 30. Resultados de los modelos obtenidos para 2 clases*

En este caso los resultados son más prometedores. Adaboost obtiene los mejores resultados llegando a un 61% en todas las métricas. Después le sigue Random Forest obteniendo resultados algo peores. Sin embargo, para otros algoritmos como KNN o Árboles de decisión los resultados se han mantenido igual que en el caso de 3 clases.



*Ilustración 31. Coeficientes de importancia de las variables de entrada para el caso de 2 clases*

Con dos clases hay ciertos pesos que varían, pero la distribución se mantiene levemente. El atributo más importante sigue siendo el número de reviews, pero ahora la zona baja su relevancia, siendo el nivel de ruido, el precio y si el restaurante tiene televisión más relevantes.

## 7. Conclusiones y dificultades encontradas

En este apartado se van a explicar las conclusiones extraídas del proyecto, enfocadas en base a los resultados que se han obtenido por los modelos. Además se detallarán algunas posibles dificultades que se han encontrado a lo largo del desarrollo del proyecto.

En primer lugar, analizando en profundidad los resultados que se han obtenido, se puede observar que dichos resultados están muy lejos de los resultados que se esperaban para la tarea de clasificación. Esto quiere decir que se ha obtenido cerca de un 60% de accuracy en

---

los modelos cuando se esperaba una tasa de acierto mucho mayor, lo que hace dudar de la fiabilidad de este predictor de valoraciones.

Esto genera cierta inseguridad sobre los atributos utilizados y es que puede ser que dichos atributos no hayan sido lo suficientemente útiles como para predecir una valoración de un restaurante en una determinada zona.

Además, también se puede pensar que quizás el tratamiento filtrado o segmentación de los datos no ha sido el acertado, o al menos no ha sido del todo preciso cómo se quería. Esto puede ocurrir por ejemplo en el apartado de [Segmentación](#), donde la división por zonas ha sido subjetiva en base a las distintas agrupaciones de puntos y topología de cada una de las zonas.

Todo esto hace pensar en este proyecto como una guía de apoyo y opinión secundaria para el cliente para intentar predecir las valoraciones, pero está lejos de ser una guía precisa que se deba seguir al pie de la letra, pues proporciona unas tasas de acierto muy alejadas de lo esperado.

Por último se debe destacar que ha sido de gran utilidad encontrar aquellos atributos que influyen en mayor cantidad en la valoración de un restaurante. Estos han sido *num\_reviews*, *NoiseLevel*, *WiFi*, *RestaurantsPriceRange2* y *Zone*, lo cual tiene sentido, porque el número de valoraciones, el nivel de ruido, el wifi y los rangos de precio influyen de manera importante en la acogida social.

En cuanto a las dificultades encontradas, han sido varios obstáculos que se han encontrado durante la realización del proyecto.

En primer lugar se debe destacar la selección del objetivo del proyecto. No han sido pocas veces en las que el objetivo del proyecto no estaba claramente definido y han surgido dudas acerca de su viabilidad. Esto ha dado lugar a cambios en la línea del proyecto en cuanto a objetivos se refiere. Cabe destacar que el primer objetivo del proyecto era la predicción de la localización de un restaurante en base a sus características.

---

En segundo lugar, se han tenido ciertas dudas en la fase de segmentación, puesto que como se ha mencionado es un proceso subjetivo y existían varias soluciones para llevar a cabo este proceso.

En tercer lugar debemos mencionar las dificultades que se han encontrado en el tratamiento de los códigos postales o zipcodes de cada una de las zonas. Este atributo contenía valores incorrectamente clasificados, de manera que existían zonas con zipcodes que no apuntaban a la zona que debía. Esto principalmente ha sido difícil de ver, puesto que se confiaba en la veracidad y fiabilidad de los datos proporcionados. Sin embargo, tras pensarlo detenidamente se solucionó relativamente rápido gracias al uso de una API de google que permitía corregir esos zipcodes incorrectos.

En cuarto lugar, a la hora de predecir con los atributos de restaurante, como se comentó anteriormente, se obtuvieron malos resultados y no mejoraban de ninguna manera, por lo que se dedujo que los atributos no aportaban la suficiente información al modelo para poder predecir correctamente. Sería útil para mejoras futuras obtener atributos más informativos sobre este conjunto de datos.

## 8. Apéndice

### Estructura de Notebooks

En esta sección del apéndice se describe la estructura de los cuadernos de Google Colab utilizados así como un enlace a los mismos. Cabe destacar que estos enlaces han sido configurados para que sólo sean accesibles por el personal docente de la Universidad. En caso de que surja algún problema, el enlace al proyecto completo es el siguiente <https://drive.google.com/drive/folders/1sEBNDIfhNR-boz1x0jItaEekr9nsL6TT?usp=sharing>.

### Filtrado-AD

---

Este notebook se encarga de todo el filtrado de los datos iniciales, seleccionando atributos y realizando una limpieza general de dichos datos.

Enlace:

[https://colab.research.google.com/drive/1kBzzkzFZQhuVSxJMEnZt\\_4vRGu5uKKlo?usp=sharing](https://colab.research.google.com/drive/1kBzzkzFZQhuVSxJMEnZt_4vRGu5uKKlo?usp=sharing)

### **Segmentación-AD**

Este notebook se encargará de realizar la segmentación de los datos según las necesidades del estudio. Este proceso es fundamental para, posteriormente, lograr modelos de Machine Learning que sean significativos y sean capaces de presentar un tiempo de cómputo factible.

Enlace:

<https://colab.research.google.com/drive/1liKONfN xvZRLWlgOfTlvgGOKK7T9MlsQ?usp=sharing>

### **MapasSinFiltrar-AD**

Este notebook se va a encargar de imprimir los datos que se crean convenientes sin realizar todavía el filtrado, para poder sacar unas primeras conclusiones e hipótesis, sobre las cuales se podrán posteriormente basar nuestros análisis.

Enlace:

[https://colab.research.google.com/drive/1Hy7Lxqyk\\_M6z3F9kMARzFukOYzFID-wK?usp=sharing](https://colab.research.google.com/drive/1Hy7Lxqyk_M6z3F9kMARzFukOYzFID-wK?usp=sharing)

### **MapasFiltrados-AD**

Este notebook se va a encargar de imprimir todos los mapas correspondientes a los datos filtrados para poder tener una visión más clara y general sobre los mapas sobre los que se va a trabajar.

---

Enlace:

<https://colab.research.google.com/drive/11sedd8aPyPktU6JNoNEBAVaXxHBYTHdy?usp=sharing>

### **Auxiliar-AD**

Este notebook se encargará de realizar tareas de apoyo que fueron necesarias a medida que avanzaba el proyecto.

Enlace:

<https://colab.research.google.com/drive/1QqLOxhnWdt2LOrvH2HDkB2KmMNIGbNzM?usp=sharing>

### **AprendizajeAutomatico-AD**

Este notebook se encarga de generar las discretizaciones de ciertos atributos y de generar modelos de entrenamiento y de proporcionar una visualización de resultados.

Enlace:

<https://colab.research.google.com/drive/1BjPURXgCbOoF1cj7TXWNkwjpR4zxLtXO?usp=sharing>

### **AnalisisVisual-AD**

Este notebook se encarga de generar todo el análisis visual de los datos para posteriormente poder trabajar con ellos. Para ello, genera distintos profile reports, boxplots, histogramas muy útiles para el análisis de los datos.

Enlace:

<https://colab.research.google.com/drive/1LwyDkTbvWbuGf6lsyhqq6cCDbsS4Uo6?usp=sharing>

## **9. Referencias**

---

[1] Python. (2021) Acceso: 20 de noviembre de 2021. Disponible en:

<https://www.python.org/downloads/>

[2] Pandas. (2021) Acceso: 20 de noviembre de 2021. Disponible en:

<https://pandas.pydata.org/>

[3] Kepler. (2021) Acceso: 20 de noviembre de 2021. Disponible en:

<https://docs.kepler.gl/docs/keplergl-jupyter>

[4] Pandas Profiling. (2021) Acceso: 20 de noviembre de 2021. Disponible en:

<https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/>

[5] Google Drive. (2021) Acceso: 26 de noviembre de 2021. Disponible en:

[https://www.google.com/intl/es\\_es/drive/](https://www.google.com/intl/es_es/drive/)

[6] Google Cloab. (2021) Acceso: 26 de noviembre de 2021. Disponible en:

[https://colab.research.google.com/?utm\\_source=scs-index](https://colab.research.google.com/?utm_source=scs-index)

[7] TDSP. Proceso de Ciencia de Datos en Equipo . (2021) Acceso: 2 de diciembre de 2021.

Disponible en:

<https://docs.microsoft.com/es-es/azure/architecture/data-science-process/overview>

[8] Geoapify. (2021) Acceso: 9 de diciembre de 2021. Disponible en:

<https://www.geoapify.com/>

[9] AST - Abstract Syntax Tree (Python). (2021) Acceso: 1 de diciembre de 2021. Disponible

en: <https://docs.python.org/3/library/ast.html>

[10] json\_normalize (Pandas). (2021) Acceso: 1 de diciembre de 2021. Disponible en:

[https://pandas.pydata.org/docs/reference/api/pandas.json\\_normalize.html](https://pandas.pydata.org/docs/reference/api/pandas.json_normalize.html)

[11] How to balance a dataset in Python. (2021) Acceso: 14 de diciembre de 2021.

Disponible en:

<https://towardsdatascience.com/how-to-balance-a-dataset-in-python-36dff9d12704>

---

[12] Zipcodes by State. (2021) Acceso: 10 de diciembre de 2021. Disponible en:  
[http://www.structnet.com/instructions/zip\\_min\\_max\\_by\\_state.html](http://www.structnet.com/instructions/zip_min_max_by_state.html)