

Sprint #2

Data Engineering



Hitos

- **ETL:**
 - Criterio.
 - ETL completo.
 - Pipeline automatizado alimentando WareHouse.
- **Pipeline:**
 - Extracción: API Yelp.
 - Transformaciones: Aplicación de criterio, normalización, optimización.
 - Carga: Conexión con RDS automatizada.
- **Documentación:**
 - Stack tecnológico.
 - Modelo ER detallado.
 - Diccionario de datos.
 - Pipeline detallado.
- **Proof of Concept de Power BI**

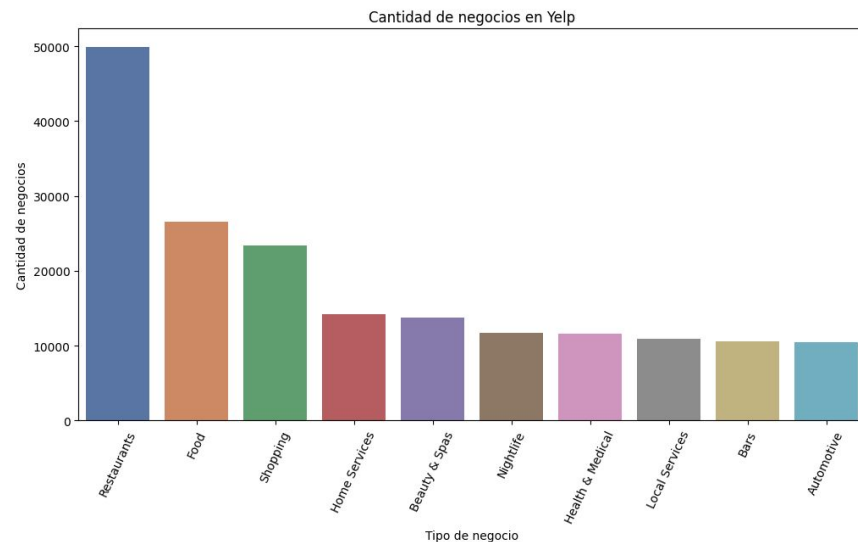
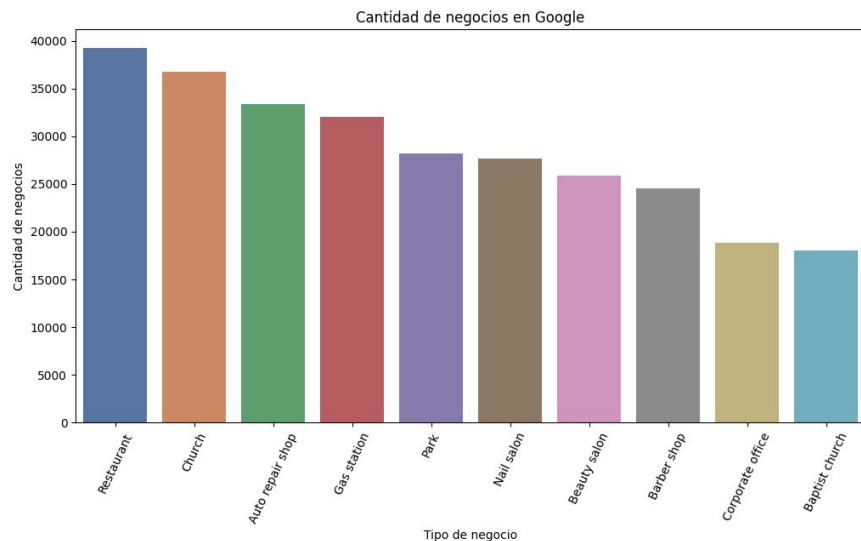
Criterio:

Según disponibilidad y alcance del dataset existente, nos adaptamos al contexto.

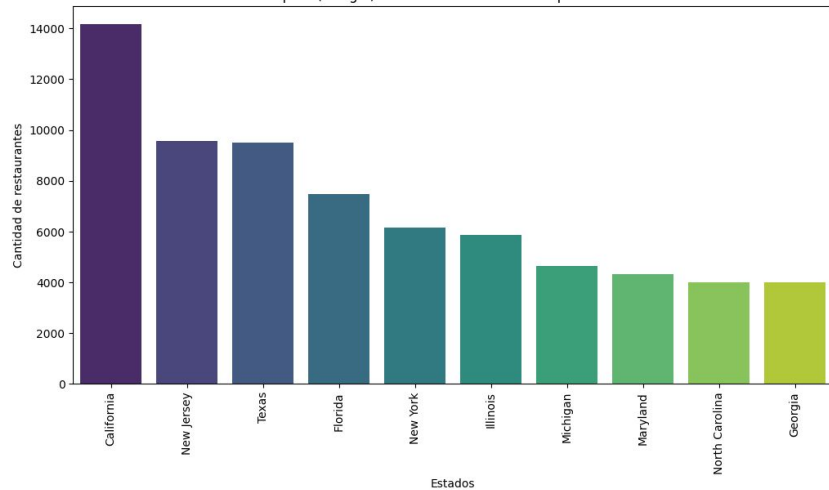
Rubro: Restaurantes.

Lugar: California, Illianois, New Jersey y Florida.

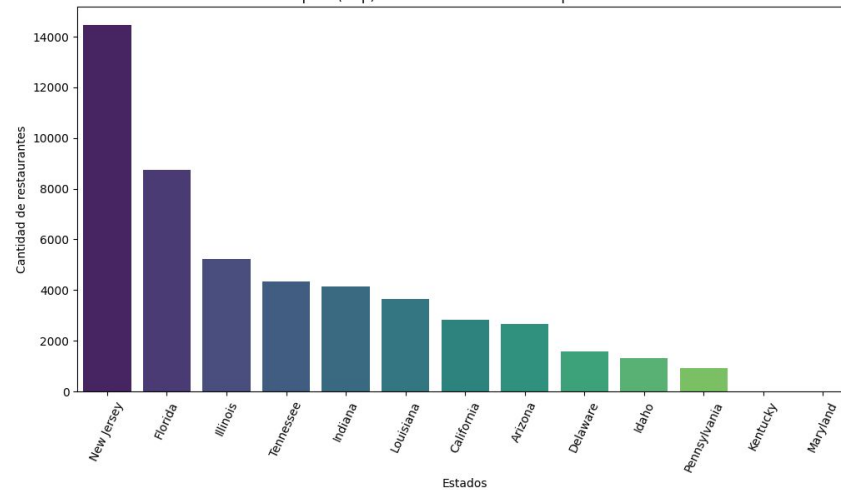
Tiempo; 2015 en adelante.



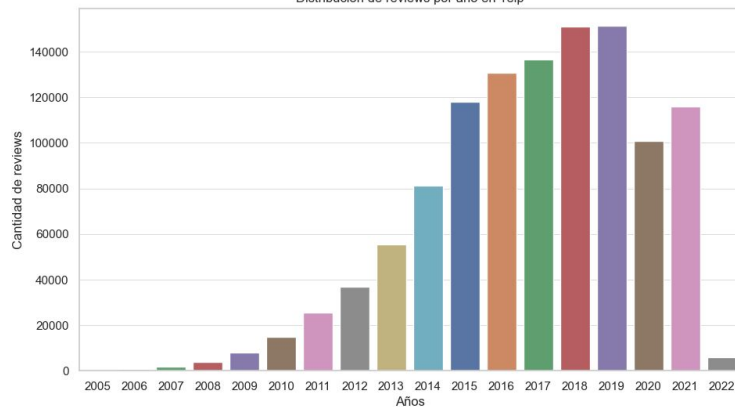
Top 10 (Google) - Cantidad de restaurantes por estados



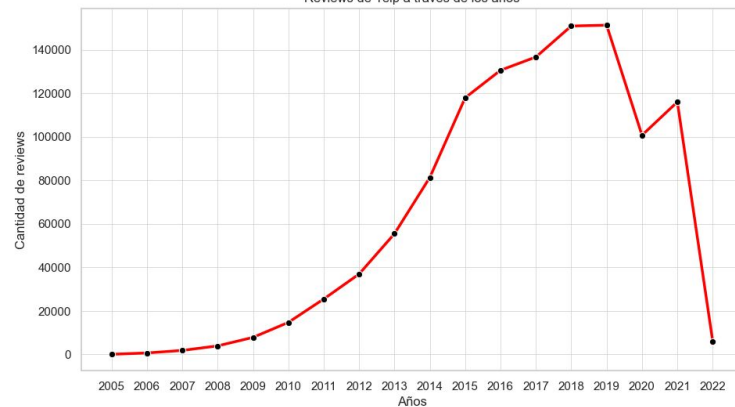
Top 10 (Yelp) - Cantidad de restaurantes por estados



Distribucion de reviews por año en Yelp



Reviews de Yelp a través de los años



ETL estático, Google

Aplicamos el criterio en torno a:

Rubro; carga incremental, filtrado y normalización.

Lugar; normalización y rectificación por sesgo.

Tiempo; normalización de formato y tipo de dato.

Lenguaje natural; optimización, análisis de sentimiento.

Rectificación y especificidad; calidad de datos, calcular columnas para valor agregado para usuarios y negocios.

ETL estático, Yelp



Fueron brindados 5 datasets sobre Yelp: **business**, **review**, **user**, **tips** y **checkin**.

No utilizados:

Tips

- Da información sobre pequeñas reseñas de algunos negocios, que por la brevedad de las mismas decidimos no utilizarlo.

Checkin

- Da datos sobre los negocios y las fechas donde hubo alguna interacción con ellos, esto no daba ninguna información.

Business

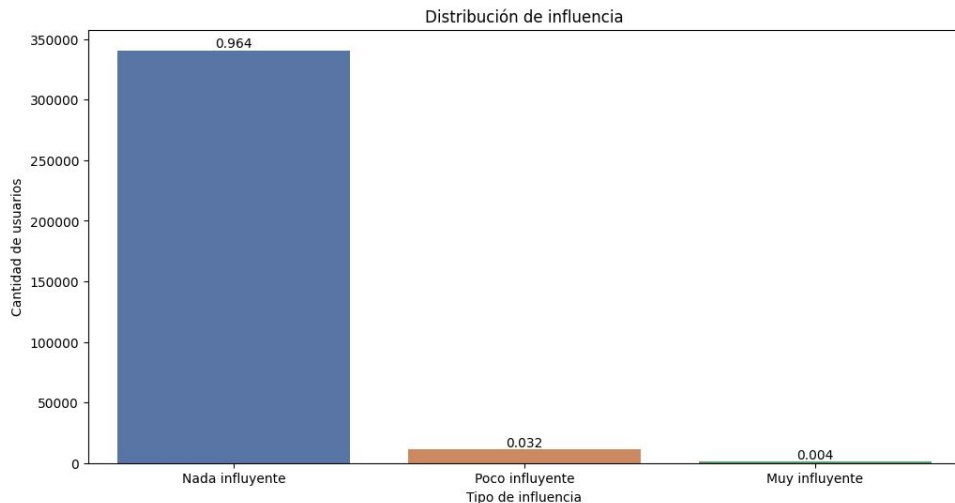
- Información sobre los negocios.
- El nombre del negocio, su latitud y longitud, promedio de estrellas de las reviews, y categorías.

Reviews

- Información sobre las reviews realizadas por los usuarios sobre el negocio.
- Calificación en estrellas, fecha en el cual se realizó la reseña y el texto de la review en sí.
 - Con el texto se realizó un análisis de sentimiento con la librería de python NLTK.

Users

- Información sobre los usuarios.
- Su nombre, cantidad de reviews que realizo, cuando se creó la cuenta, los fans que determinan tan influyente es, y el promedio de estrellas de las reviews del usuario.
 - Con los fans realizamos un criterio en el cual vemos que tan influyente es, esto es según la cantidad de fans que posee. Podemos ver la distribución de estos usuarios:



Problemática coordenadas

Identificamos discrepancias en la asignación de estados en la columna "estado" de nuestro conjunto de datos de empresas de Google Maps. La información de la columna "estado" no coincidía con la calculada a partir de las coordenadas geográficas de las empresas. Abordamos el problema incorporando un conjunto de datos geojson de los Estados Unidos, centrado en los Estados criterio: ("New Jersey", "California", "Florida", "Illinois").

Implementamos la creación de un GeoDataFrame determinando el área de cada estado mediante sus coordenadas de latitud y longitud máximas. Además, empleamos Geopy como biblioteca para comparar las coordenadas de las empresas y llevamos a cabo una unión espacial con el conjunto de datos de los estados. Esta estrategia soluciona la asignación incorrecta de estados, lo que mejoró significativamente la precisión de la información geográfica. Esta corrección asegura una representación más precisa de la ubicación de las empresas en nuestro análisis.



Streaming ETL

1. Extracción de datos mediante la API de Yelp:

- Extracción de restaurantes por estado.
- Extracción de reseñas por restaurantes.
- Máximo volumen de datos de 1400 .

2. Transformación de los datos ingresados:

- Transformación de los negocios.
Normalización de estados.
Normalización de categorías.
Selección de features.
- Transformación de reseñas.
Análisis de sentimiento.
Cálculo de características de usuarios.
Selección de features.

3. Carga en la base de datos:

- Conexión a la base de datos RDS.
- Validación de datos a ingresar.
- Carga y actualización de los datos.

4. Orquestación de flujos:

- EC2.
- Cron.

Estructura de datos implementada

- **Escalabilidad y Administración Simplificada:** RDS gestiona tareas operativas complejas, como el aprovisionamiento de hardware, el escalado automático, las copias de seguridad y las actualizaciones de software. Esto libera al usuario de tareas de mantenimiento y permite escalar fácilmente la capacidad de almacenamiento y la potencia de procesamiento según las necesidades del proyecto.
- **Rendimiento y Fiabilidad:** RDS está diseñado para ofrecer un rendimiento confiable y consistente. Utiliza infraestructura altamente disponible y redundante, lo que reduce la posibilidad de tiempos de inactividad y garantiza una mayor confiabilidad de los datos.
- **Compatibilidad con múltiples motores de bases de datos:** RDS admite varios motores de bases de datos, como MySQL, PostgreSQL, SQL Server, etc. Esto permite seleccionar el motor que mejor se adapte a las necesidades específicas del proyecto.
- **Seguridad:** RDS proporciona características de seguridad avanzadas, como la encriptación de datos en reposo y en tránsito, gestión de accesos y la posibilidad de implementar redes virtuales privadas (VPC) para controlar el acceso a la base de datos.

Servicio base de datos relacional

Despliegue más rápido
Administración eficiente
de los datos
Durabilidad y disponibilidad
Seguridad



Workflow detallando tecnologías

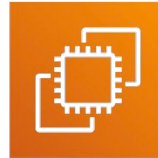
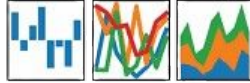


Google Maps



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



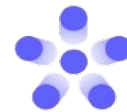
Amazon
EC2



Power BI



Streamlit



monday work
management

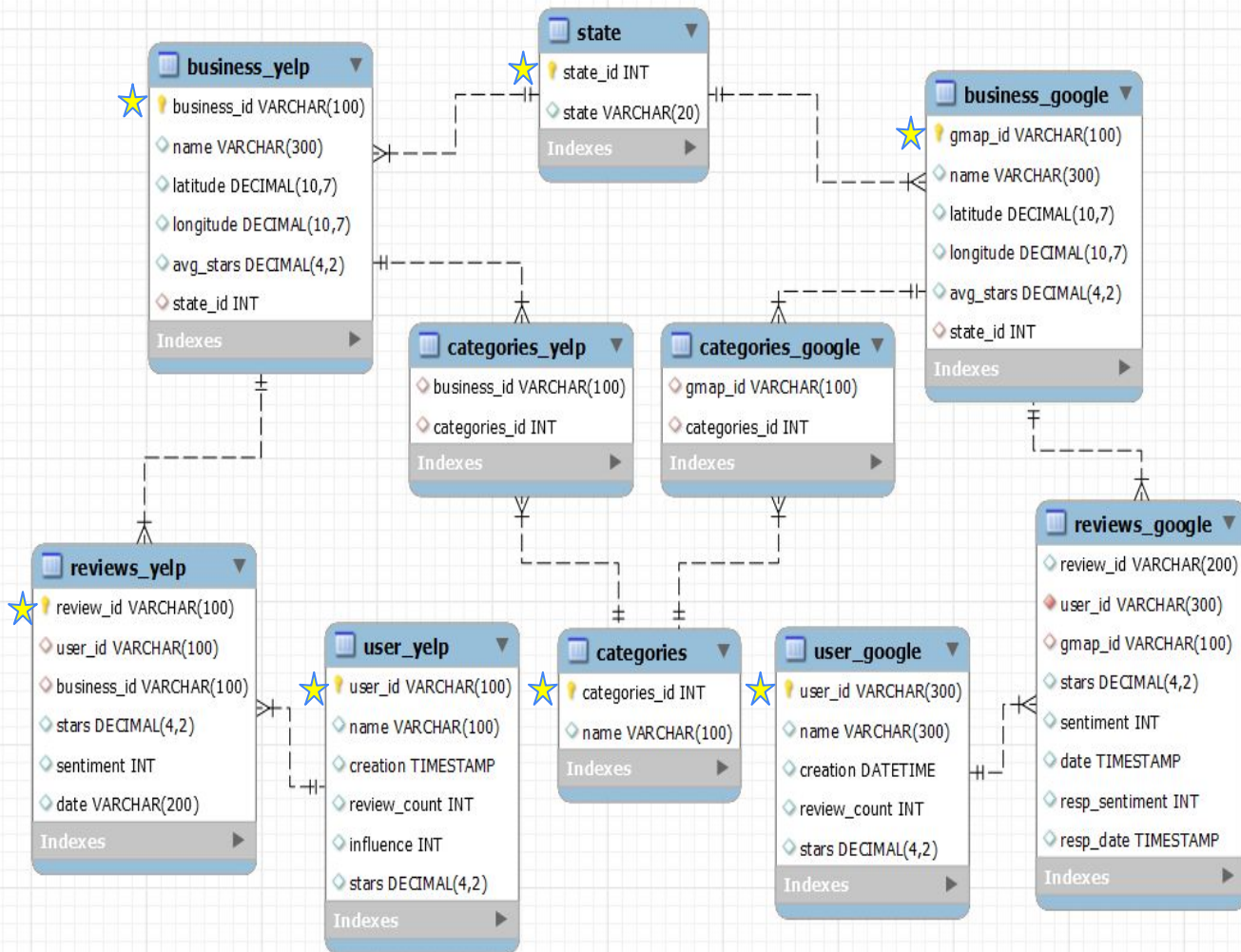
Diseño del Modelo ER (Entidad - Relación)

Representa las entidades en nuestro sistema y cómo se relacionan entre ellas, permitiendo la visualización de las tablas de datos y sus relaciones.

facilitando la comprensión de cómo se almacenan, su manipulación y extracción, su relación e identificación de elementos clave.



“ El ID en un modelo ER es crucial para la gestión eficiente de datos, ya que proporciona una identificación única para cada entidad, facilitando la organización, búsqueda y manipulación de los datos.”



Resumen del diccionario de datos:

reviews_yelp
review_id VARCHAR(100)
user_id VARCHAR(100)
business_id VARCHAR(100)
stars DECIMAL(4,2)
sentiment INT
date VARCHAR(200)
Indexes

business_google
gmap_id VARCHAR(100)
name VARCHAR(300)
latitude DECIMAL(10,7)
longitude DECIMAL(10,7)
avg_stars DECIMAL(4,2)
state_id INT
Indexes

business_yelp
business_id VARCHAR(100)
name VARCHAR(300)
latitude DECIMAL(10,7)
longitude DECIMAL(10,7)
avg_stars DECIMAL(4,2)
state_id INT
Indexes

categories
categories_id INT
name VARCHAR(100)
Indexes

categories_google
gmap_id VARCHAR(100)
categories_id INT
Indexes

categories_yelp
business_id VARCHAR(100)
categories_id INT
Indexes

reviews_google
review_id VARCHAR(200)
user_id VARCHAR(300)
gmap_id VARCHAR(100)
stars DECIMAL(4,2)
sentiment INT
date TIMESTAMP
resp_sentiment INT
resp_date TIMESTAMP
Indexes

state
state_id INT
state VARCHAR(20)
Indexes

user_google
user_id VARCHAR(300)
name VARCHAR(300)
creation DATETIME
review_count INT
stars DECIMAL(4,2)
Indexes

user_yelp
user_id VARCHAR(100)
name VARCHAR(100)
creation TIMESTAMP
review_count INT
influence INT
stars DECIMAL(4,2)
Indexes

Machine Learning

Modelos de recomendación :

- Filtro colaborativo : usuarios, restaurantes, sentiment.

Modelo ALS por MLlib.

- Filtro basado en contenido: Restaurantes, categorías.

Similitud coseno.

Vecinos más cercanos.

Redes Neuronales.

- Modelo Híbrido:

Ponderación o fusión de salidas.



Power BI

Promedio de analisis de sentimiento por Año

0,89

Reseñas en Google y su análisis de sentimiento

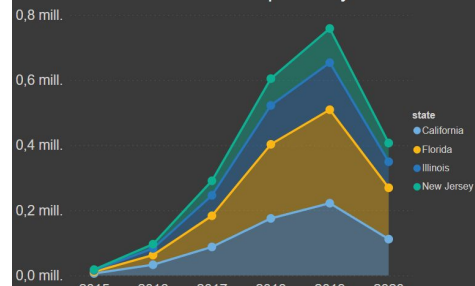
Promedio de estrellas por Año

4,30

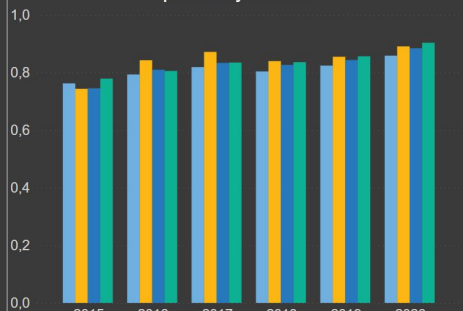
Seleccione año a comparar

2015 2016 2017 2018 2019 2020

Cantidad de reseñas por año y estado

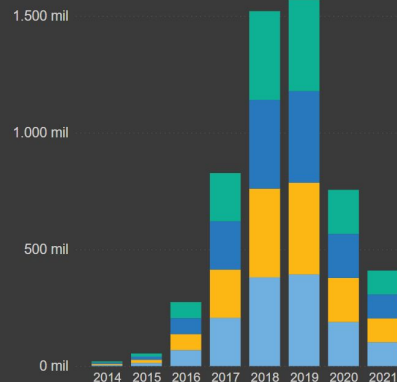


Promedio de analisis de sentimiento de reseñas por año y estado



Cantidad de usuarios nuevos por año

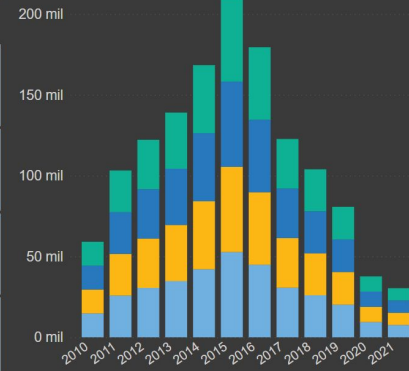
Google



Yelp

Seleccione Estados a comparar

California
Florida
Illinois
New Jersey



Promedio de analisis de
sentimento por Año

0,67

Reseñas en Yelp y su análisis de sentimiento

Promedio de estrellas
por Año

4,50

Seleccione año a comparar

2015

2016

2017

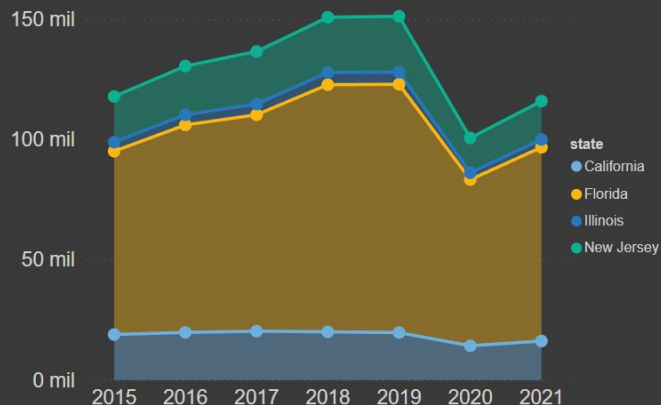
2018

2019

2020

2021

Cantidad de reseñas por año y estado



Promedio de analisis de sentimiento de reseñas
por año y estado

