Assessment Data Scientist

Daniel Alejandro Gonzalez Salamanca

Enero 29 de 2024

Contenido

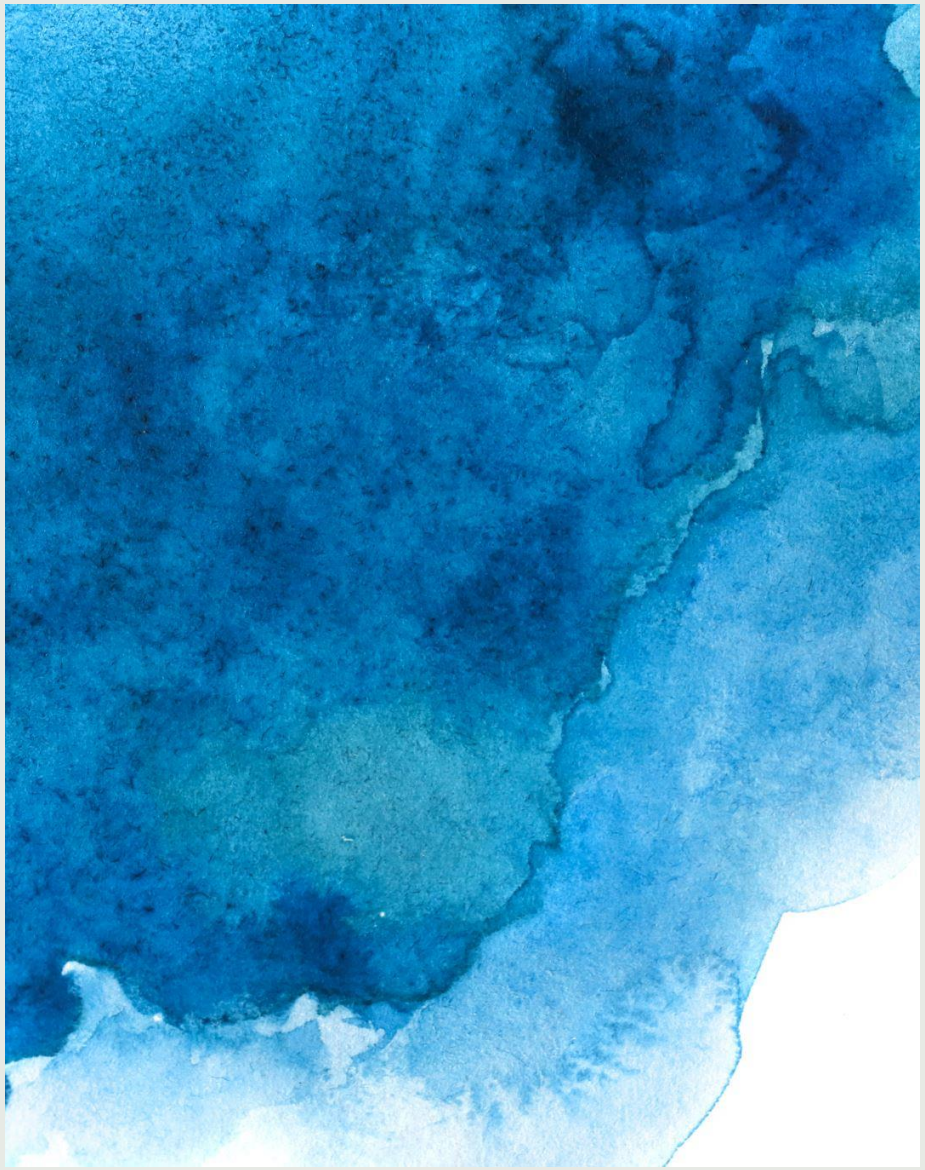
Introducción

A continuación, se presentarán los desarrollos realizados para los puntos propuestos en el assessment para el puesto de Data Scientist.

Todo lo que se encuentra en la presentación fue desarrollado en código Python, si se desea ver los notebooks de cada uno de estos, pueden ser consultados en el siguiente link

https://github.com/daagonzalezsa/Media.Monks_Assesment

Tasa de Desempleo EEUU



Contexto:

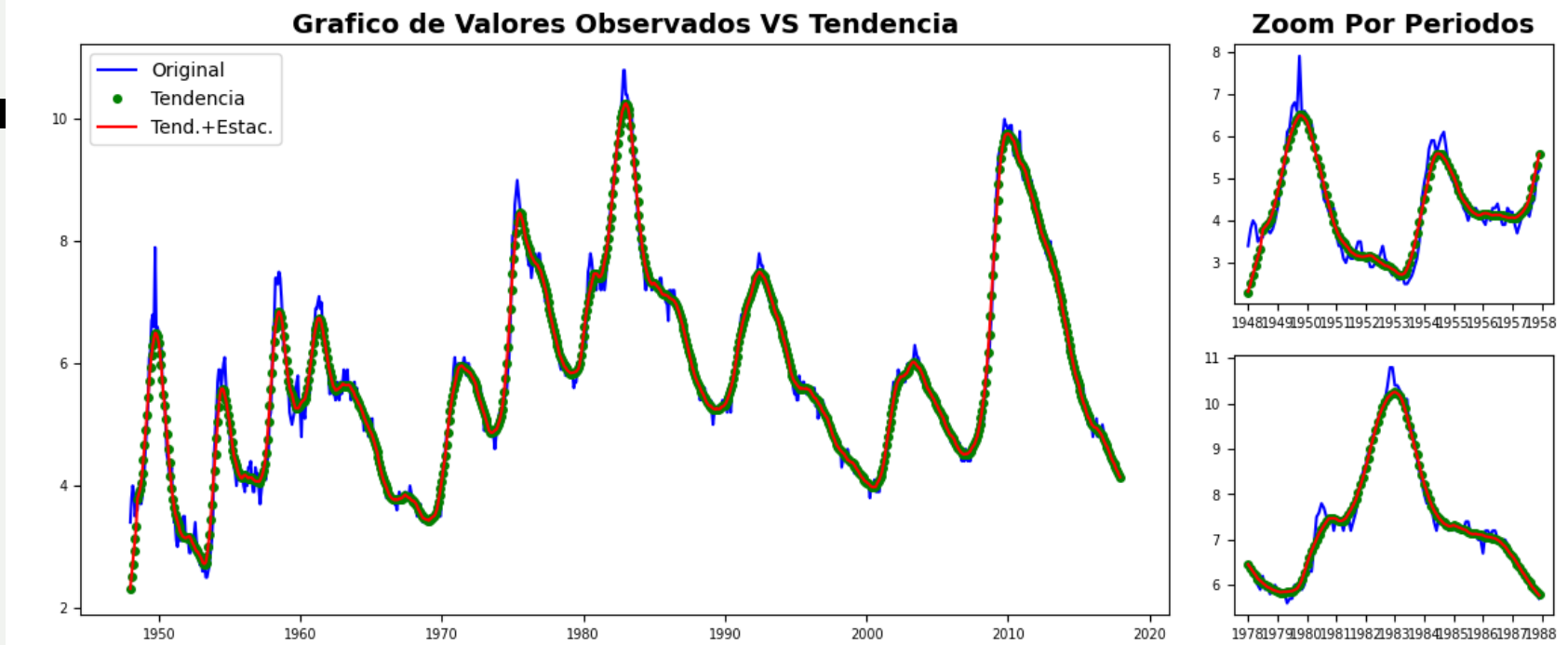
Se cuenta con una base de datos del evolutivo mensual de la Tasa de Desempleo en Estados Unidos, teniendo en cuenta desde 1948 hasta 2017. Ya que esta es una serie de tiempo, se buscaba que esta se descompusiera en 2 partes, y explicar cada uno de sus componentes.

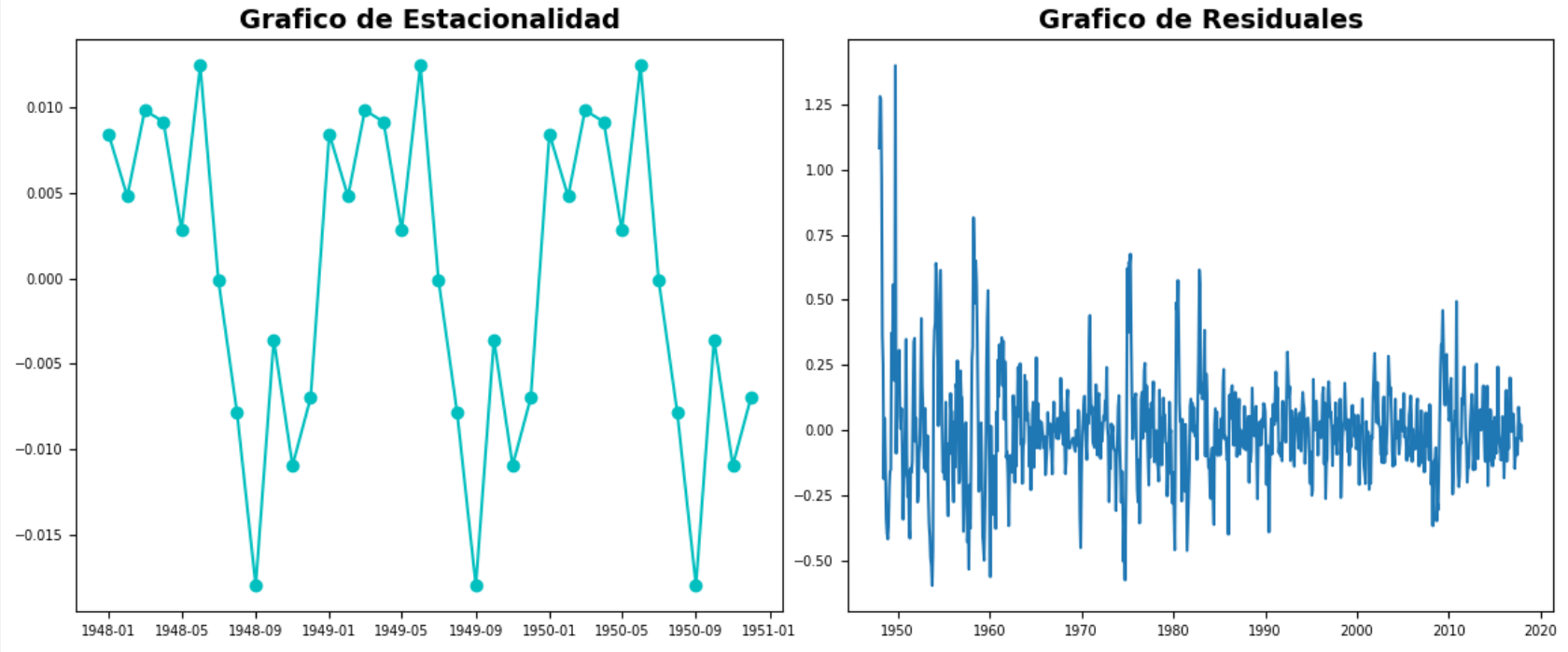
Desarrollo:

Se implementó un modelo de suavizamiento exponencial para explicar el comportamiento de la serie de tiempo con una periodicidad de 12 meses y de 10 años.

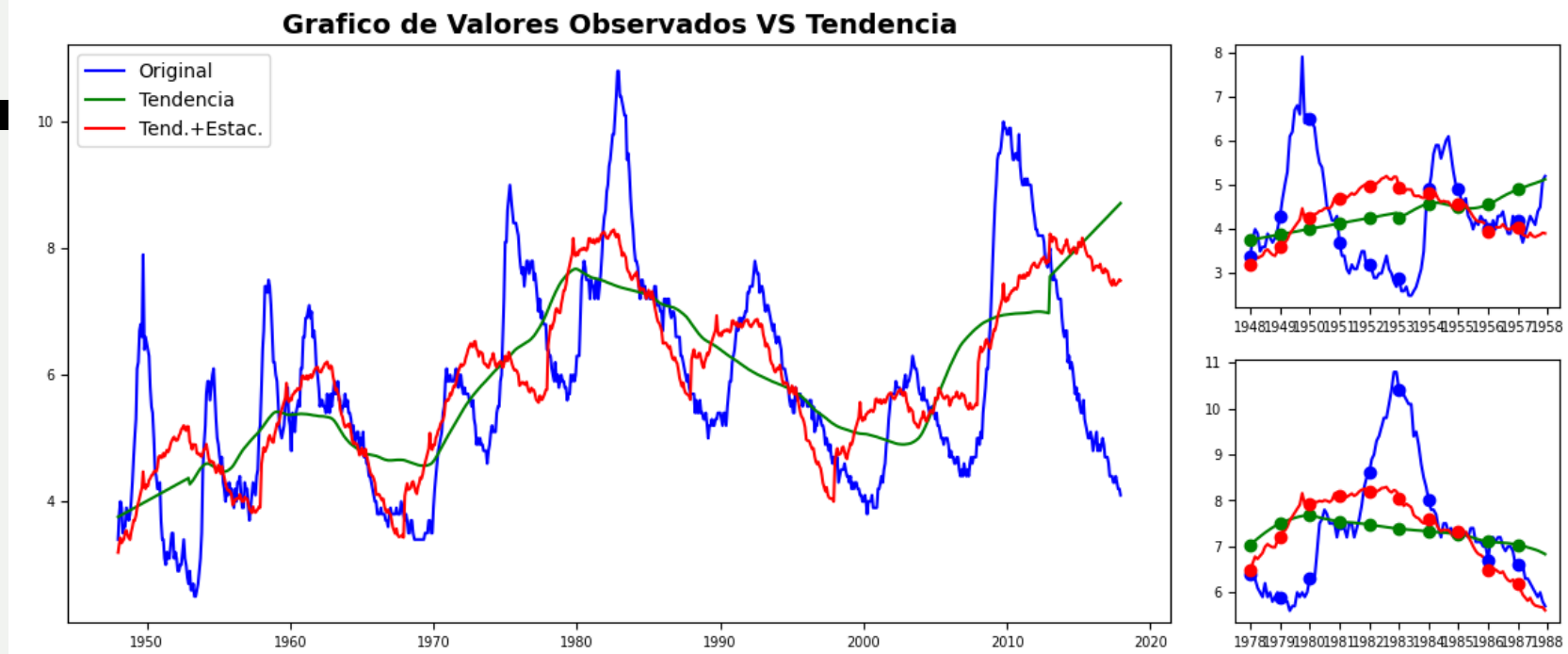
Los resultados obtenidos en cada una de las periodicidades se presentan a continuación:

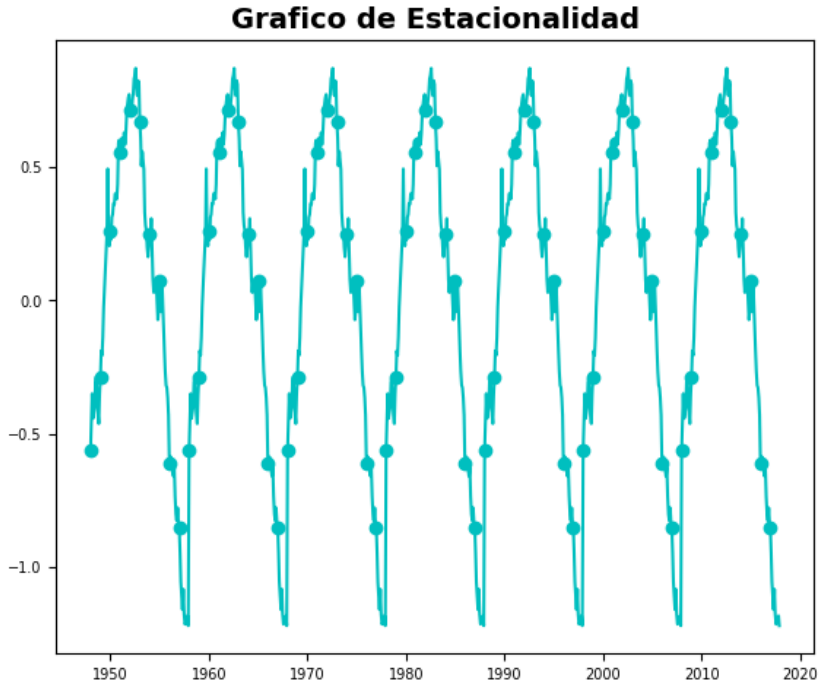
La gráfica presenta la serie de tiempo original en la parte superior, en la parte inferior izquierda la estacionalidad que le estamos atribuyendo y la de la parte derecha los residuales del modelo. Para una periodicidad de 12 meses, la gran mayoría del comportamiento se le atribuye a la tendencia, debido a que los valores de la estacionalidad son muy bajos.

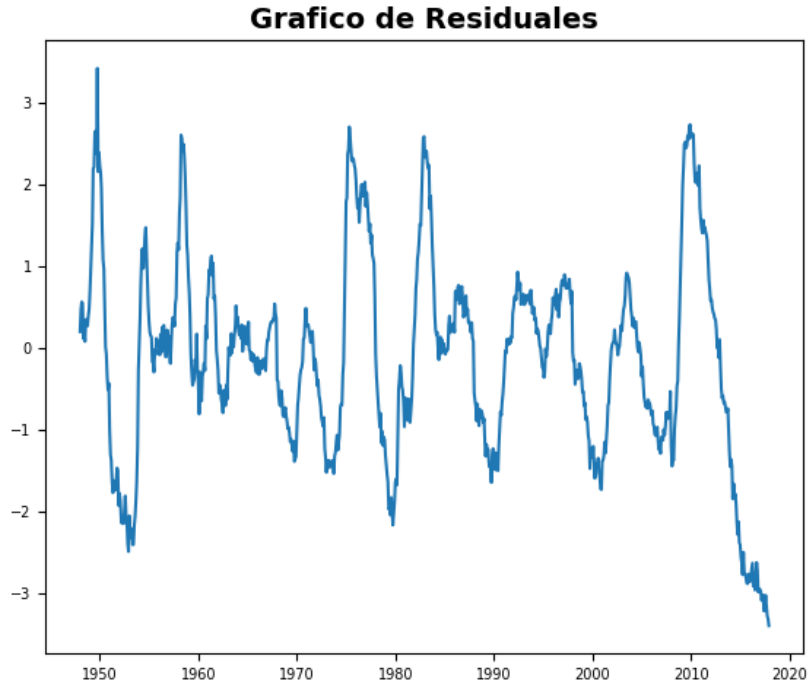




En este caso se presenta una periodicidad de 10 años. Se evidencia que para este caso la estacionalidad está presentando comportamientos cíclicos, donde los años donde finaliza y comienzan las décadas son donde se suelen presentar los picos en la Tasa de Desempleo, mientras que a mediados de las décadas tiende a decaer.



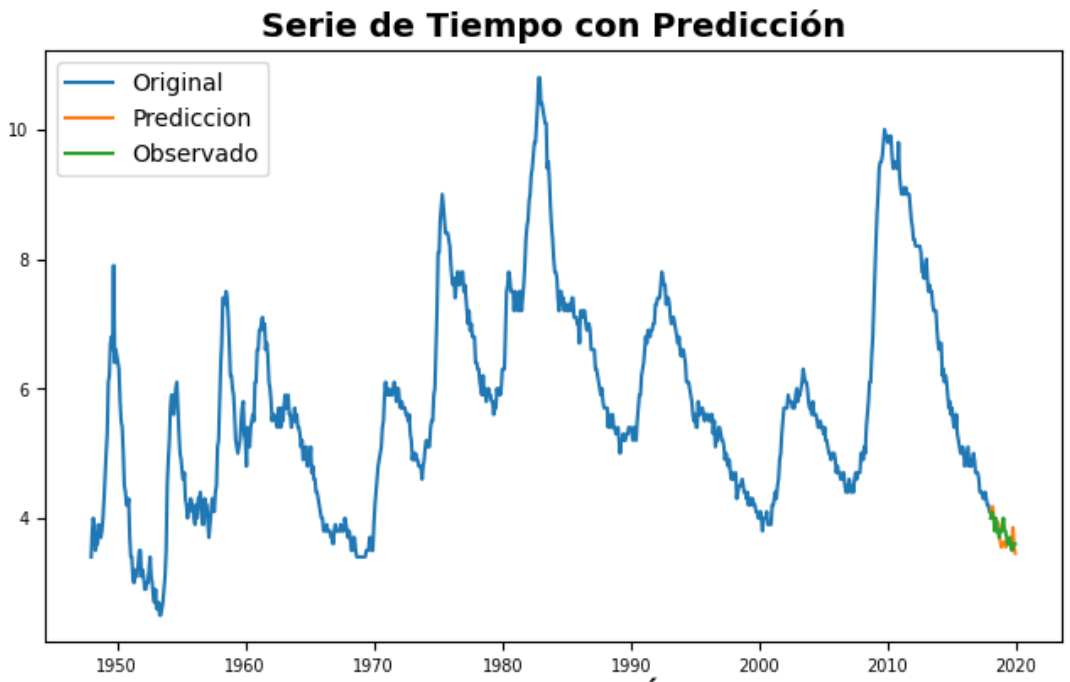


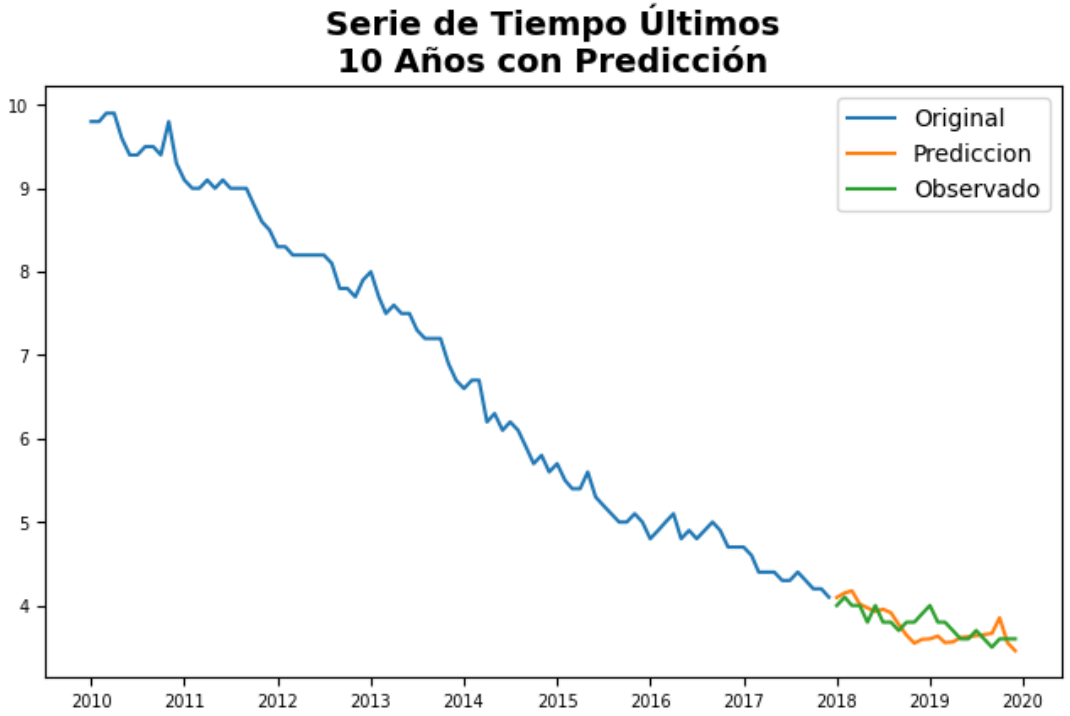


Finalmente, se emplea un modelo de suavizamiento exponencial con un periodo de 10 años para poder predecir los siguientes 2 años de la Tasa de Desempleo.

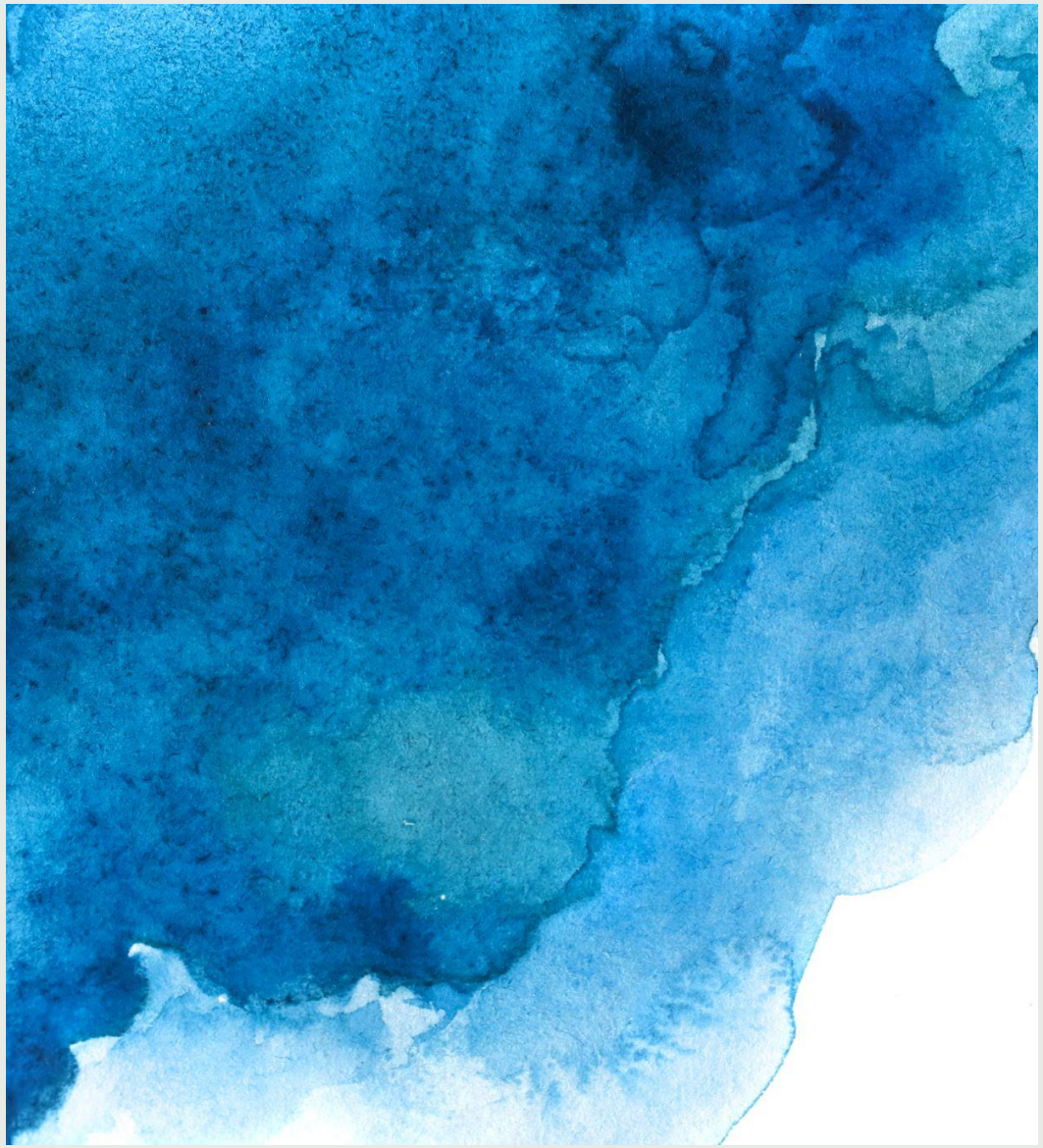
Los resultados obtenidos se presentan en el siguiente gráfico.

Para este modelo, el resultado del Error Cuadrático medio fue de 0,0286.





Mes	Prediccion
2018 - 01	4.1
2018 - 02	4.15
2018 - 03	4.18
2018 - 04	4.02
2018 - 05	3.97
2018 - 06	3.93
2018 - 07	3.96
2018 - 08	3.91
2018 - 09	3.77
2018 - 10	3.65
2018 - 11	3.55
2018 - 12	3.59
2019 - 01	3.6
2019 - 02	3.63
2019 - 03	3.56
2019 - 04	3.56
2019 - 05	3.62
2019 - 06	3.63
2019 - 07	3.64
2019 - 08	3.65
2019 - 09	3.67
2019 - 10	3.85
2019 - 11	3.55
2019 - 12	3.46



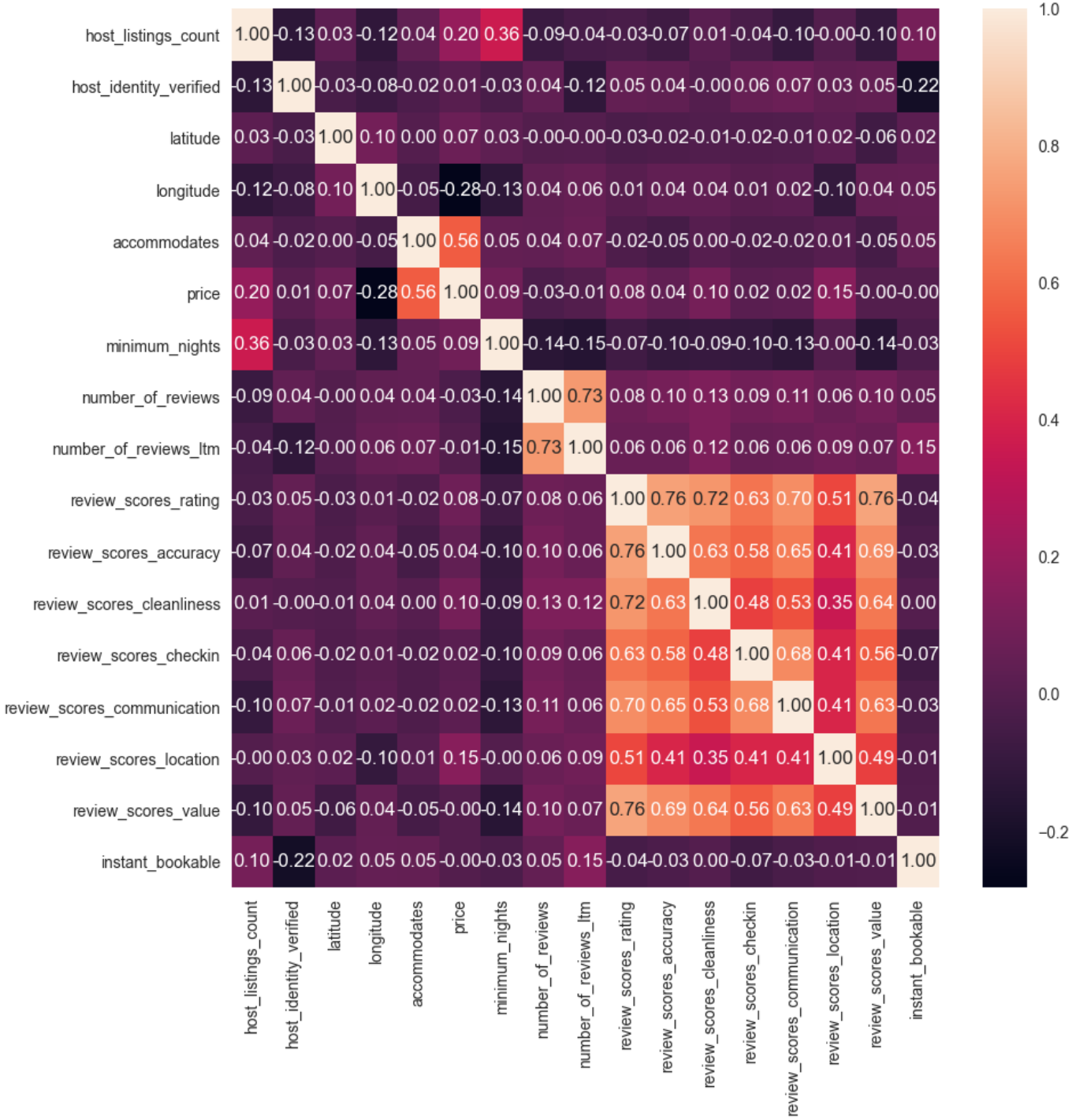
AIRBNB

Contexto:

Se tiene una base de datos de algunos AIRBNB registrados en la ciudad de Nueva York. Se nos solicita revisar si podemos ayudar a generar mayores ingresos para AIRBNB, evaluar los Scores de Review y generar agrupaciones que permitan la mejora para esta empresa.

Desarrollo:

Se implementó un modelo de Aprendizaje No Supervisado para buscar posibles agrupaciones de viviendas y parones dentro de los datos. También se generó un análisis descriptivo sobre estos grupos y finalmente se desarrolló un modelo de Aprendizaje Supervisado para la predicción del precio de un AIRBNB acorde a sus características.

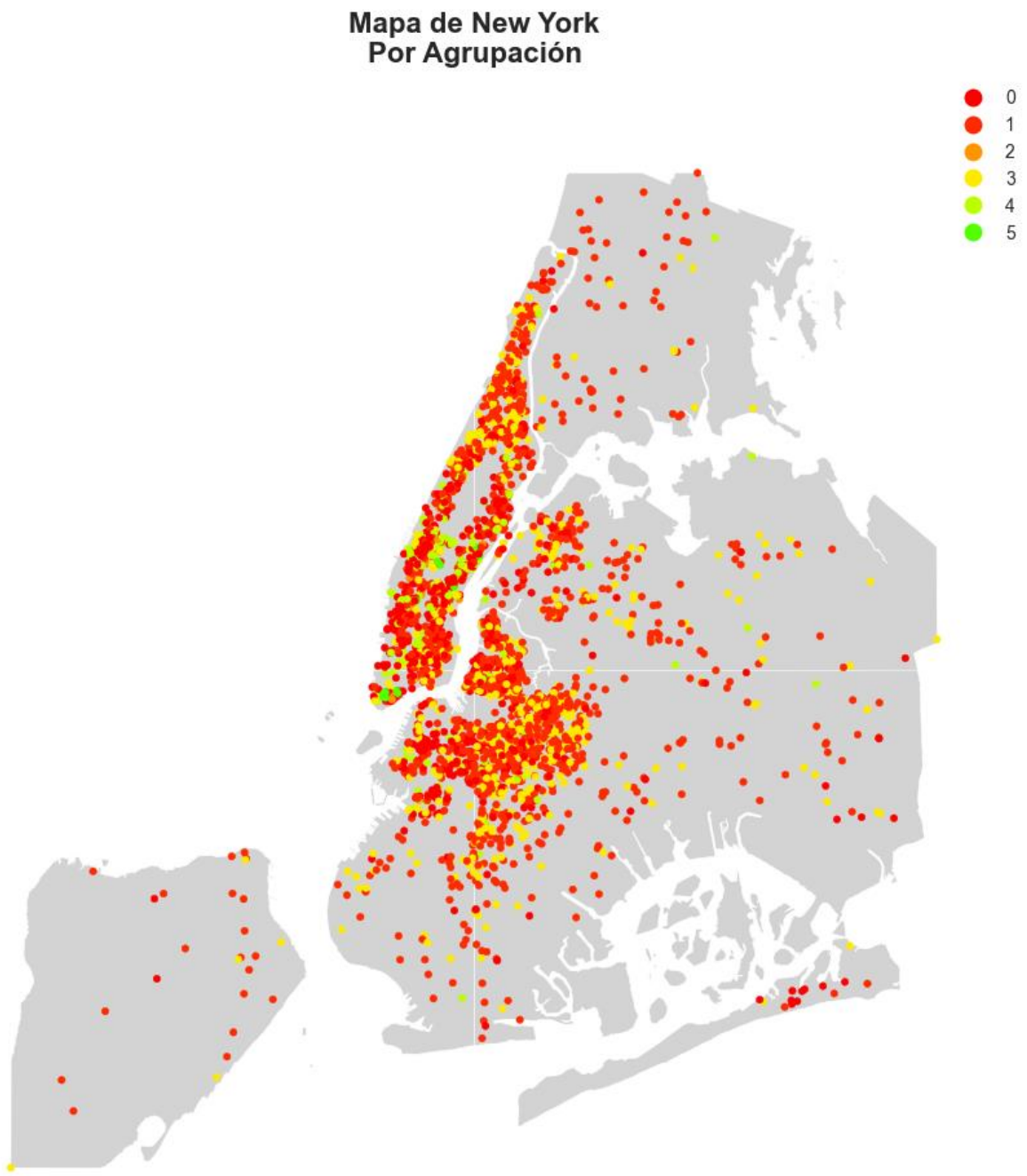


Correlación

Para las variables numéricas que encontramos dentro de la base de datos, realizamos un análisis de correlación para ver si se encuentran relaciones significantes.

En cuanto al precio, observamos que tiene una alta correlación con la cantidad de personas que pueden hospedarse (accommodates).

También encontramos una relación negativa entre la longitud y el precio, es decir, que a medida que nos dirigimos hacia el Oeste, tiende a aumentar el precio.



Kmeans

El algoritmo implementado para la segmentación de los AIRBNB fue Kmeans. Se utilizaron un total de 6 grupos teniendo en cuenta la mayoría de las variables compartidas.

El gráfico de la izquierda presenta la acomodación en la ciudad de Nueva York con respecto a cada una de las coordenadas de los AIRBNB.

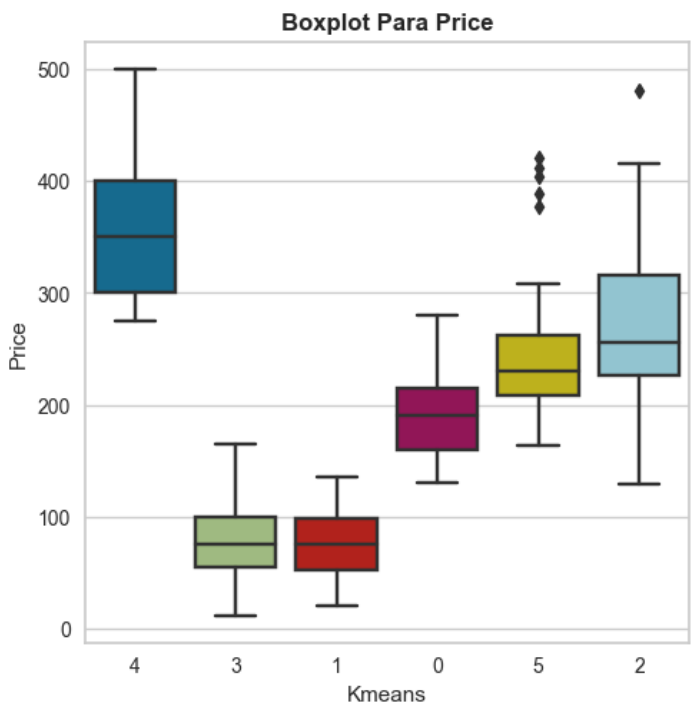
En este caso no se observa patrones claros de manera geográfica.

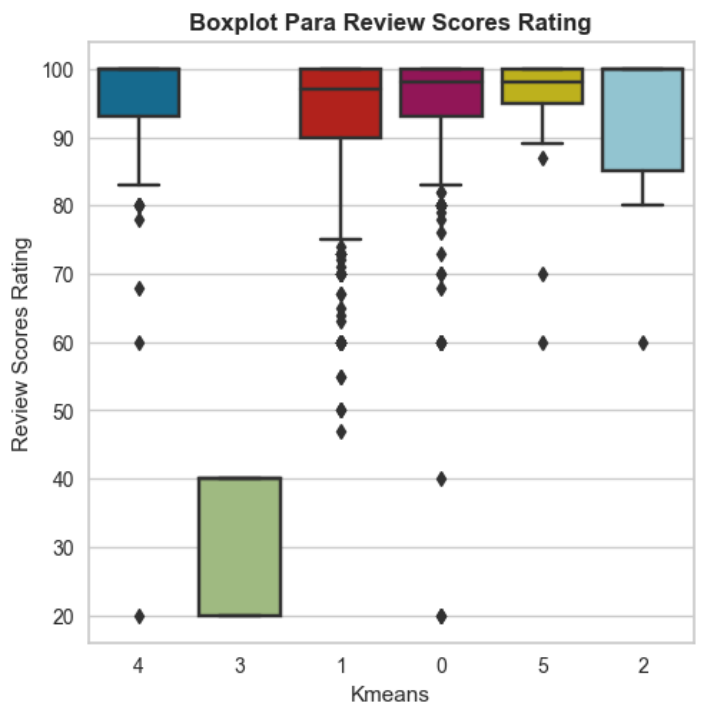
Kmeans

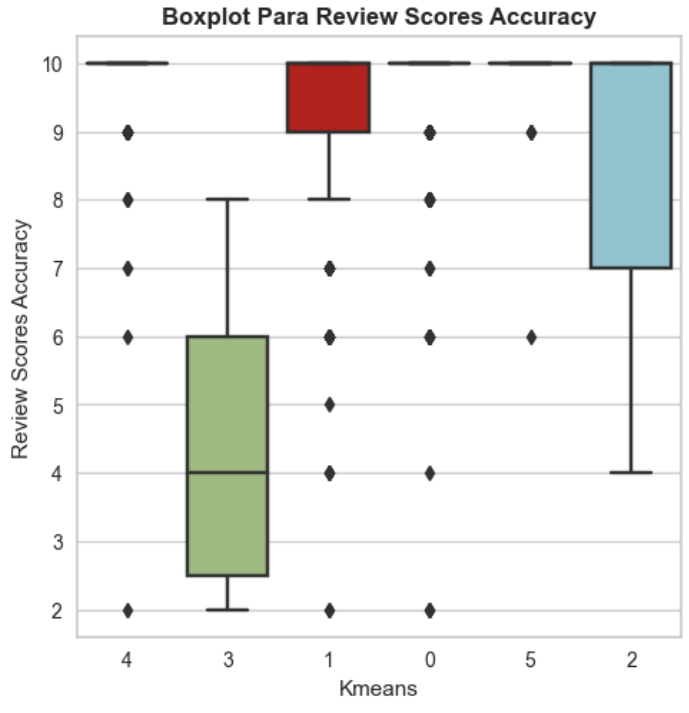
Encontramos que los grupos 3 y 1 son los que tienen menor precio, y los del grupo 4 son los que mayor precio presentan.

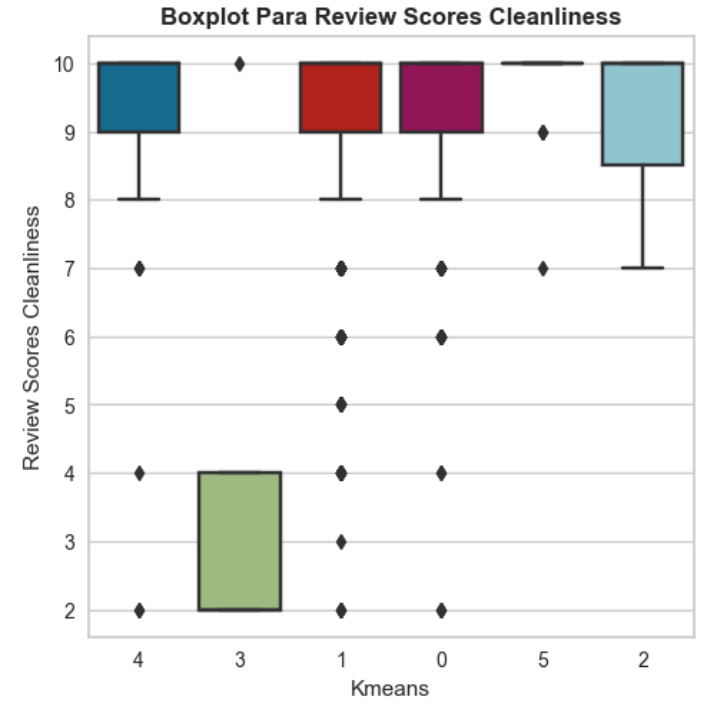
Adicionalmente, sobre todo para el caso de los AIRBNB del grupo 3, se encontró que la gran mayoría de Review Scores son mucho más bajas en comparación a los demás grupos.

Algo que vemos es que la limpieza de estos sitios tiende a ser muy baja debido a las Review Obtenidas.





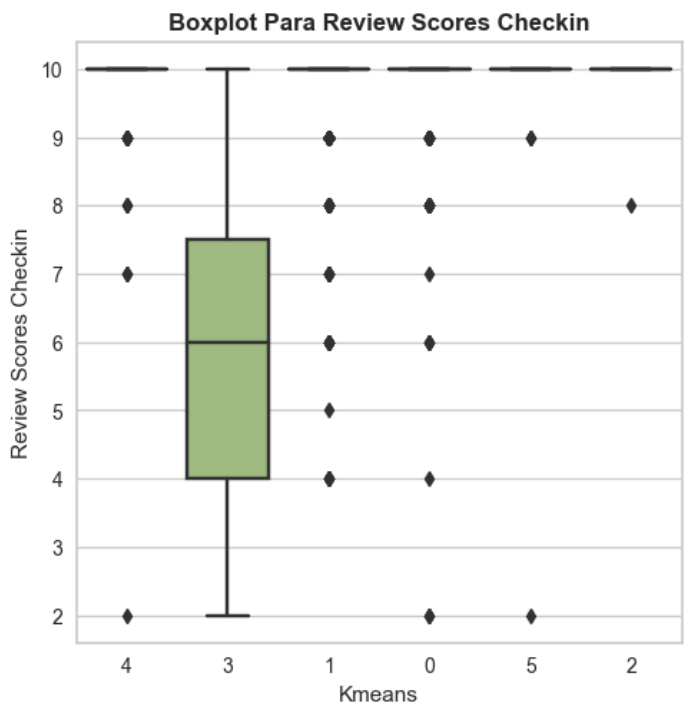


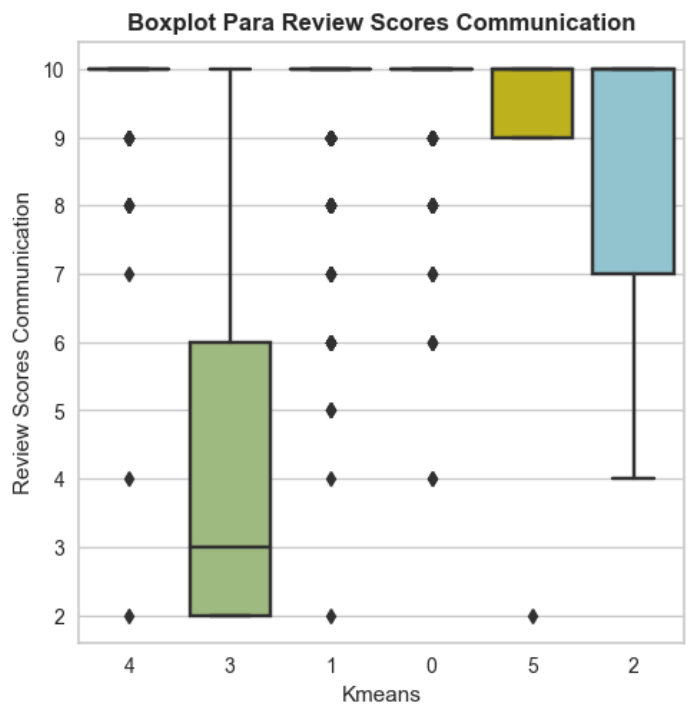


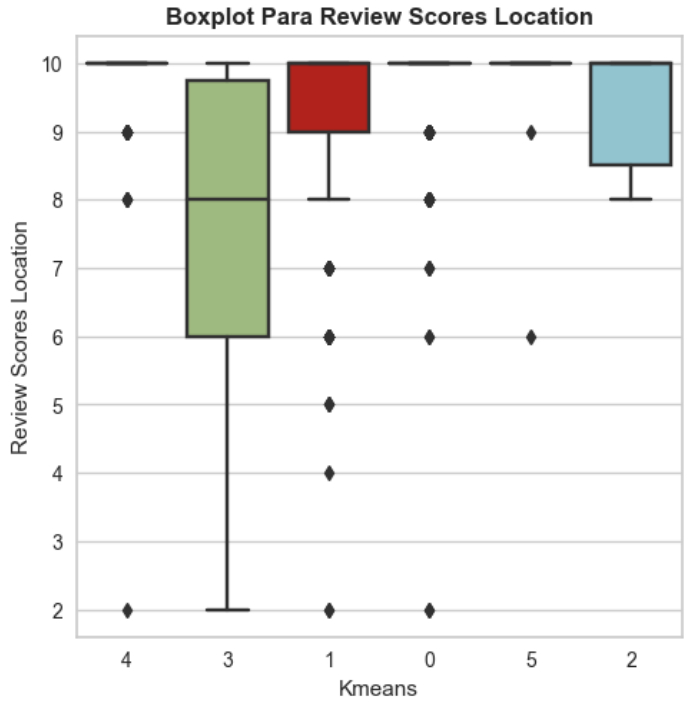
Kmeans

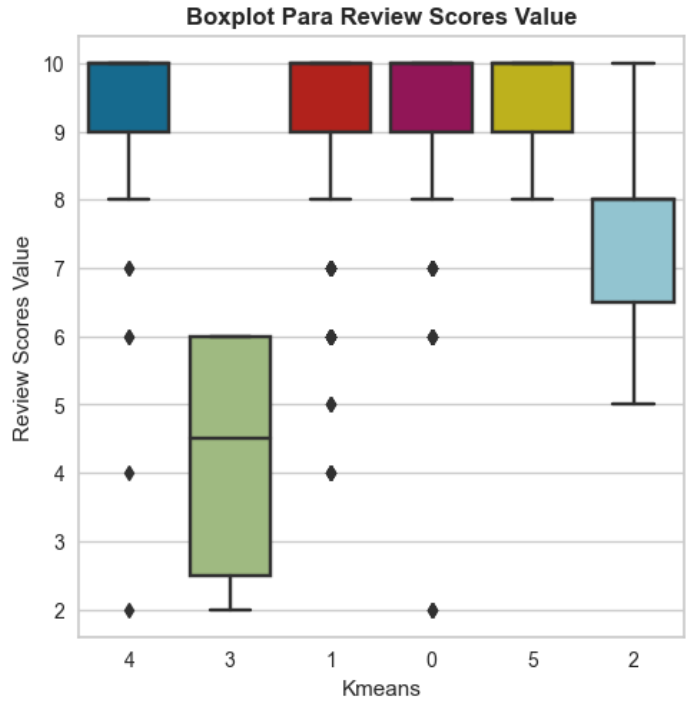
También se halló que para este grupo 3, el Score del Check In son muchos más bajos que el resto de los grupos; además, sucede algo muy similar con los demás Scores.

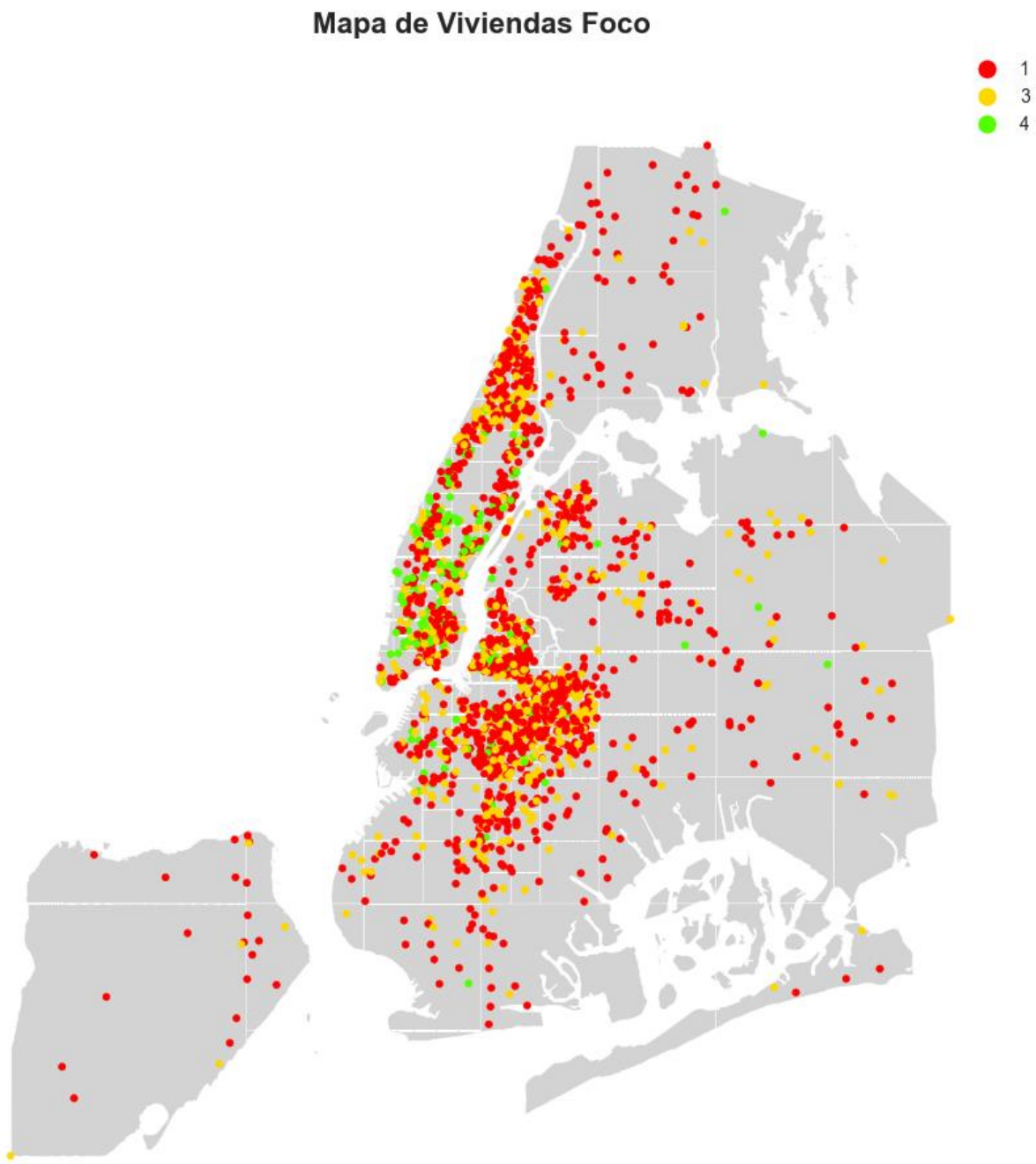
En cuanto a comunicación y locación también el Score tiende a ser más bajo. Lo que es extraño es que este grupo, a pesar de que presenta los precios más bajos, en el Review Score Value también presento los valores más bajos.







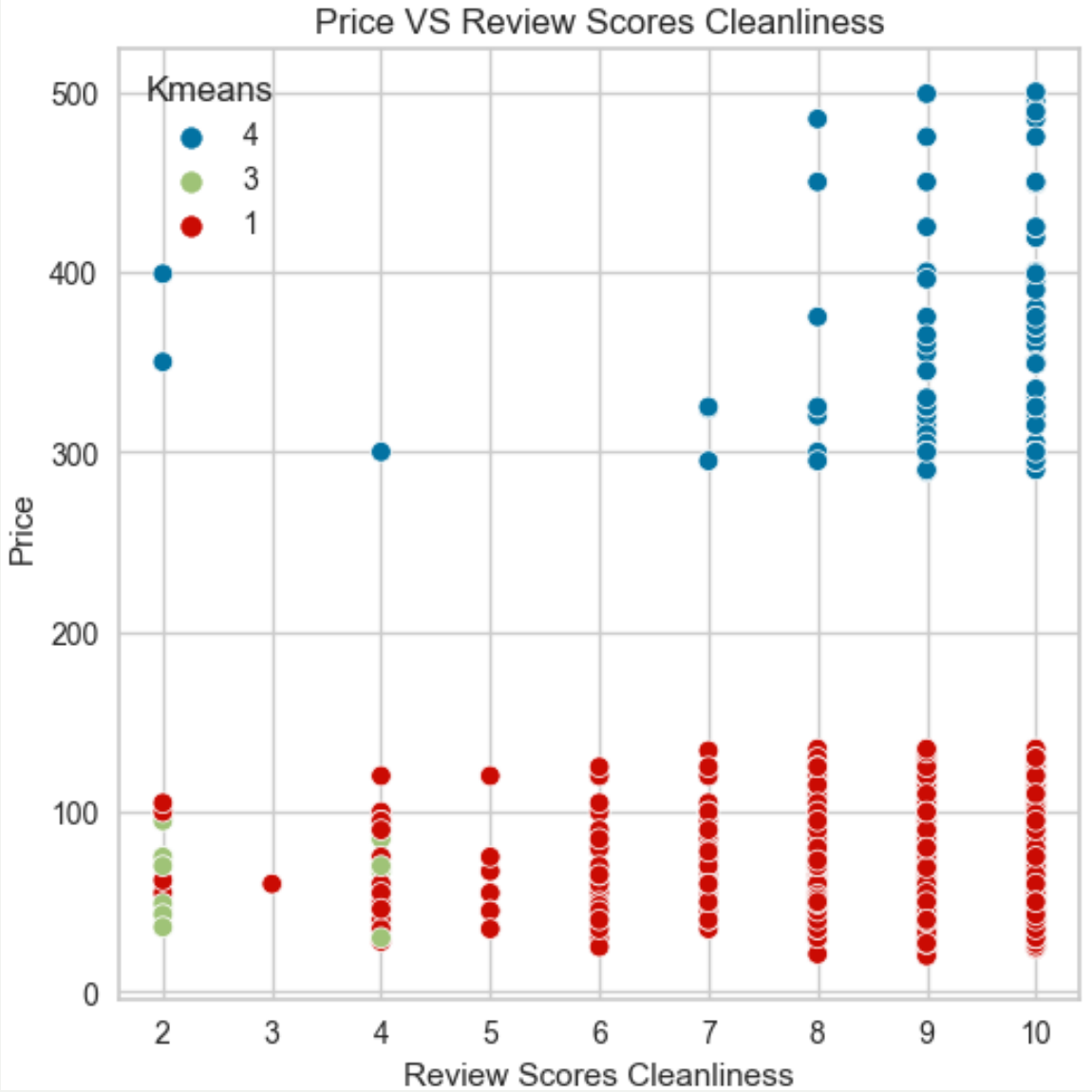




Grupos Foco

Se tomaron los grupos 3, 4 y 1 como grupos objetivo, ya que el 3 y el 1 tienden a tener precios bajos, mientras que el grupo 4 presenta precios muy altos.

Para el grupo 4 estamos encontrando que la gran mayoría de estos se encuentran hacia el centro de New York, mientras que el 3 y el 1 tienden a estar dispersos a lo largo de la ciudad.



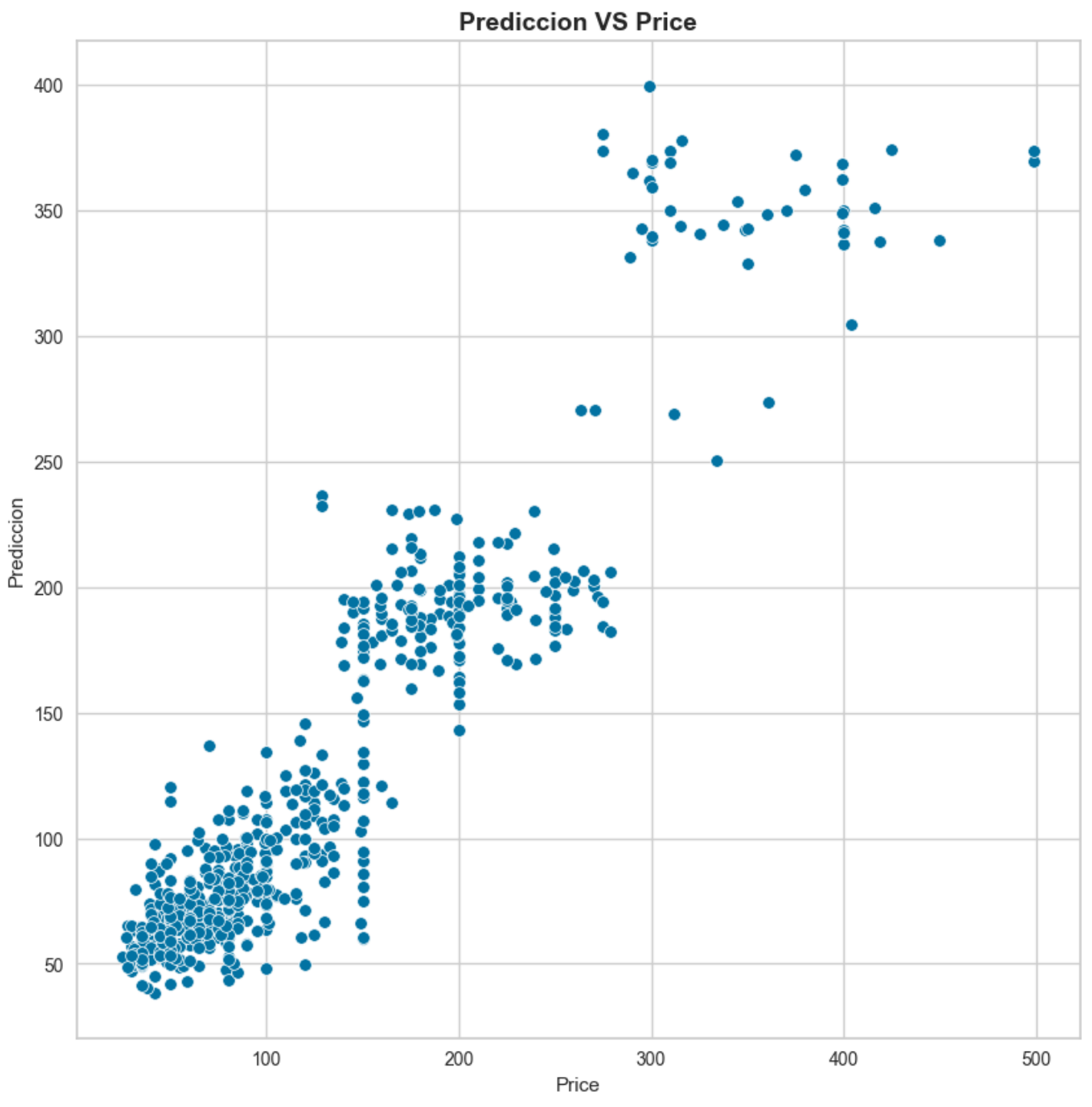
Grupos Foco

Se observa que el Score de limpieza es uno de los que más afecta a los grupos, en el caso de los que tienen un costo alto, presenta un Score alto de limpieza, mientras que el grupo 1 y 3 son los que presentan un valor mucho más bajo, sobre todo para el caso de este último.

Predicciones

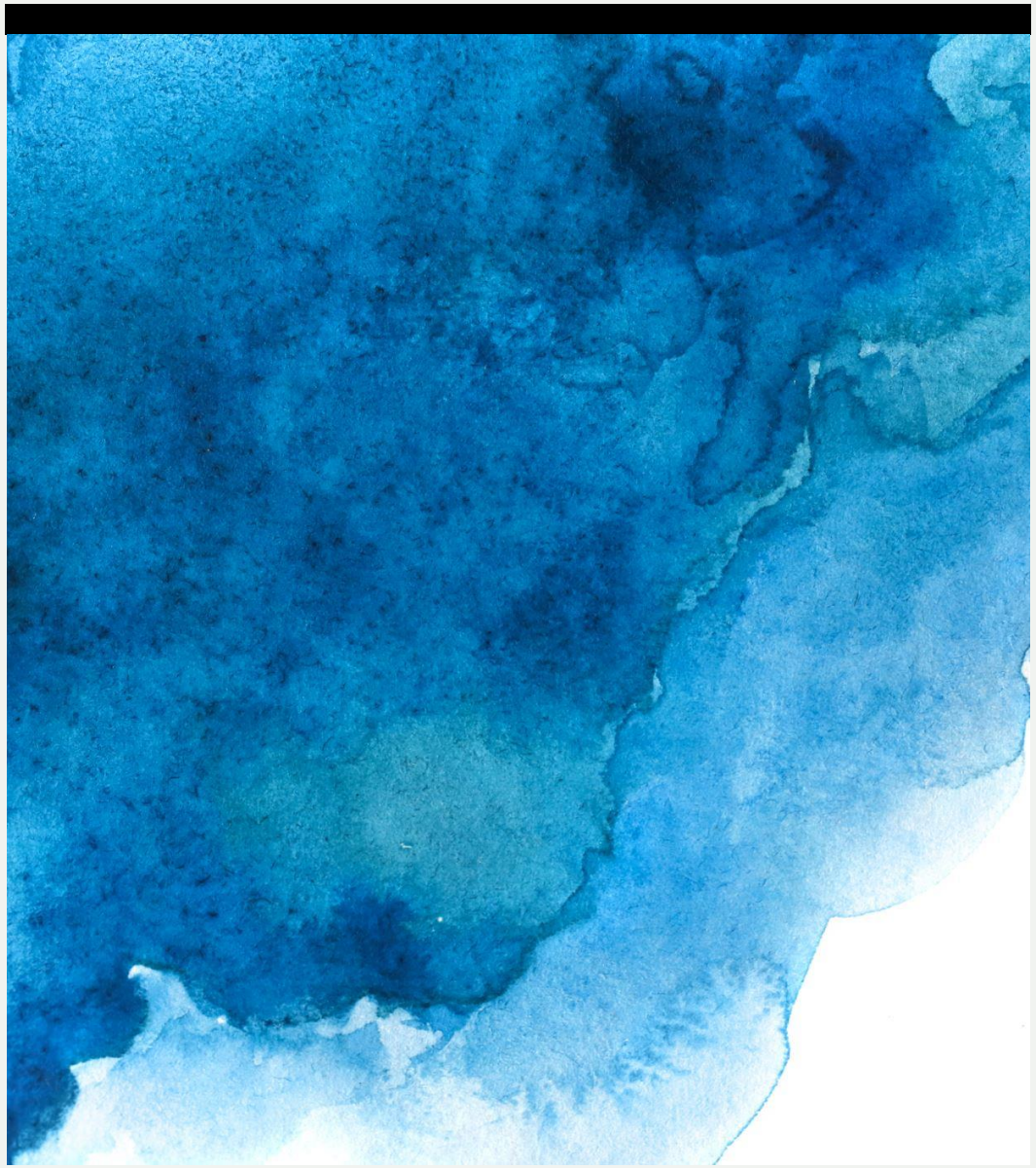
Se compararon diferentes modelos para predecir el precio que podría tener un AIRBNB acorde a las características que presentaba. El mejor modelo fue el Gradiente Boosting Regressor.

La siguiente imagen presenta los valores del precio original y predicho por el modelo para AIRBNB los cuales no fueron utilizados en el entrenamiento del modelo:



Hallazgos y Conclusiones:

- La gran mayoría de Review Scores presentan un comportamiento muy similar entre los AIRBNB, pero los que son relacionados al precio, limpieza y comunicación son los que más impactan y hacen segmentar las viviendas.
- El grupo 3 es objetivo para revisar qué está pasado con este, principalmente debido a que, aunque cuenta con valores bajos, tiene las calificaciones más bajas.
- Se puede evaluar el precio de las viviendas que están en el centro de la ciudad, ya que como se observó, el grupo 4 se encuentra cerca a este sector.
- El modelo de predicción puede ser implementado para tener una mayor acertividad acerca de los precios. Esto permitirá un valor más homogéneo y que no dependa de la subjetividad del dueño.



Food Brand

Context:

We have a data base related to social media. The aim is to give some recommendations according to the performance of the column Type.

Develop:

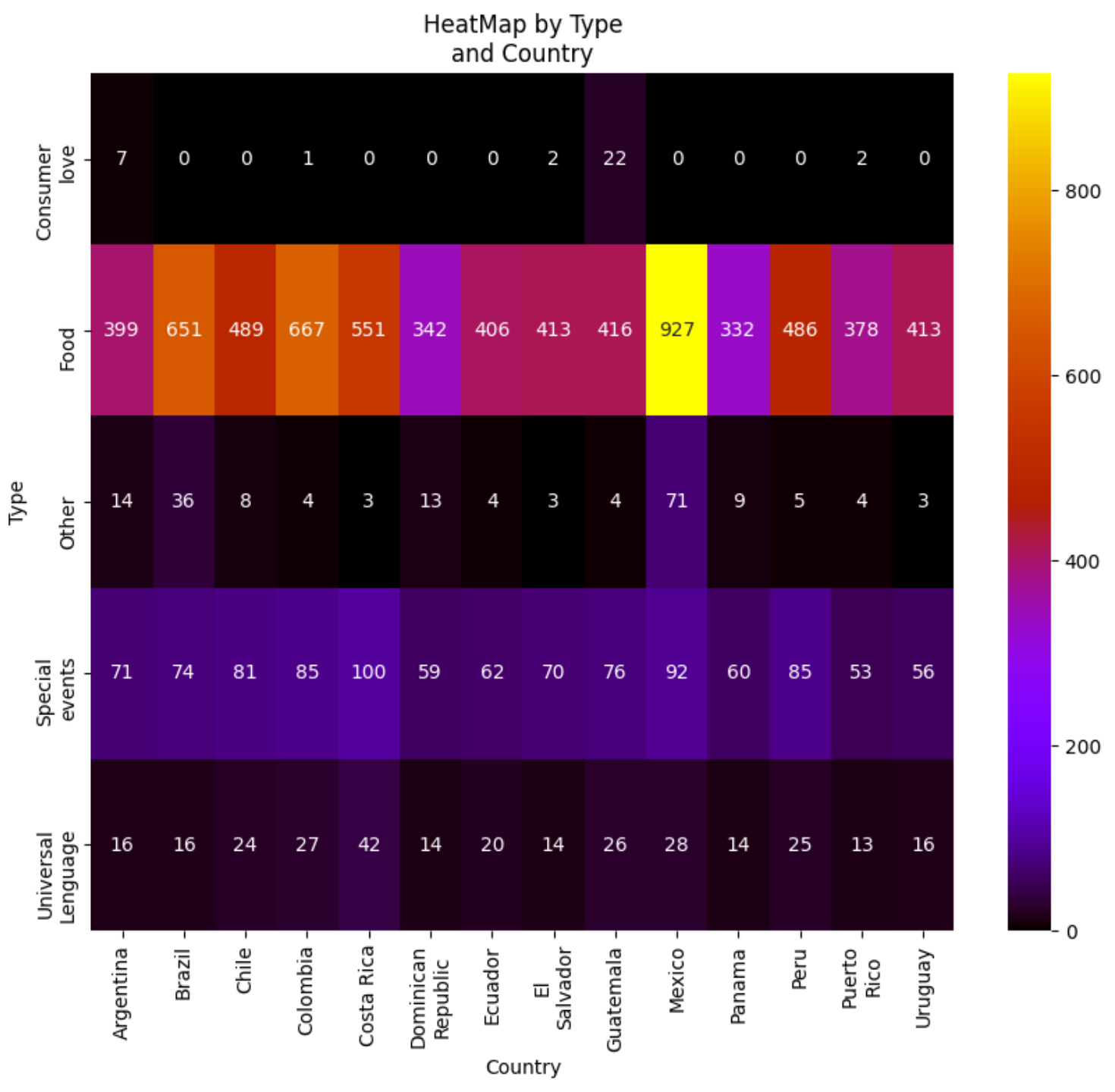
First I clean the data base, review the variables and change de types according to the metric. After this, do a descriptive analysis using the Type variable and the others, and finally, implement a WordCloud using the Brand post.

HeatMap

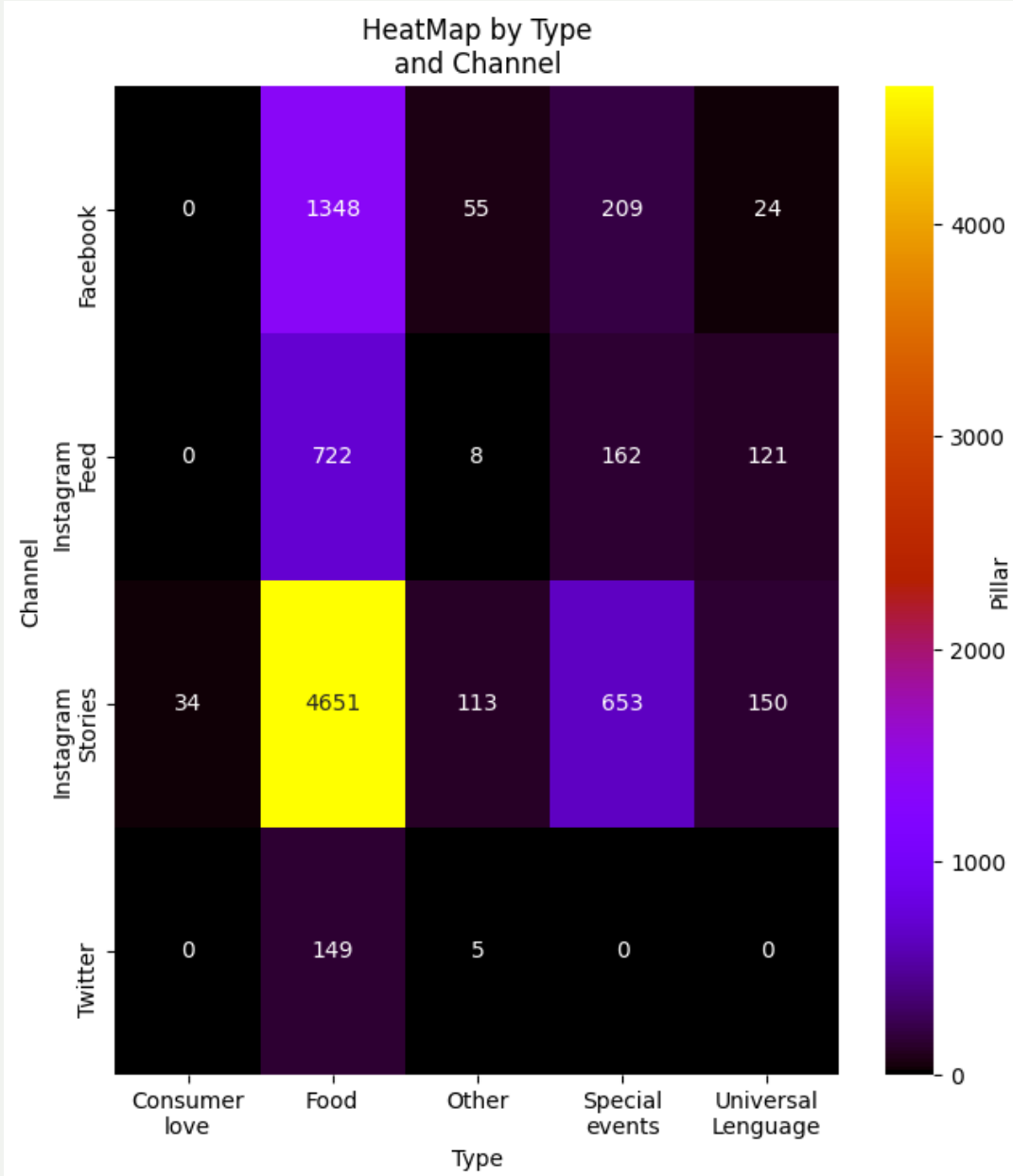
We found that Food is the most category present in the Type, after this, is Social Events.

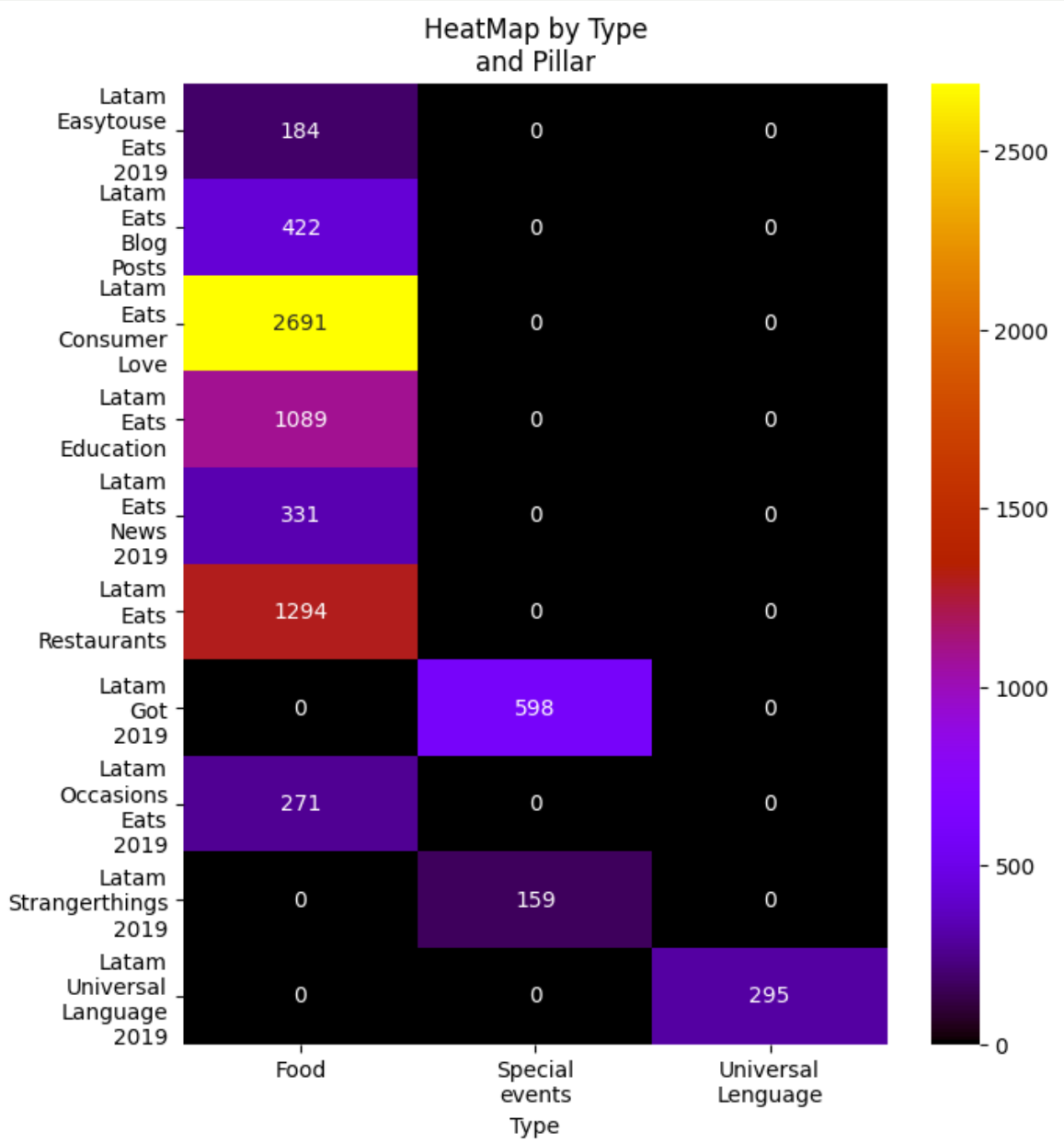
In the other hand, Mexico is the country with more participations, and Brazil is the second.

The Channel more implemented is Instagram and in Pillar, is Latam Eats Costumer Love that generates more interactions.



HeatMap


HeatMap by Type and Channel
Channel
Facebook
Instagram Feed
Instagram Stories
Twitter
0
1348
55
209
24
0
722
8
162
121
34
4651
113
653
150
0
149
5
0
0
Consumer love
Food
Other
Special events
Universal Lenguage
Type
Pillar
4000
3000
2000
1000
0


HeatMap by Type and Pillar
Latam Easytouse Eats 2019
Latam Eats Blog Posts
Latam Eats Consumer Love
Latam Eats Education
Latam Eats News 2019
Latam Eats Restaurants
Latam Got 2019
Latam Occasions Eats 2019
Latam Strangerthings 2019
Latam Universal Language 2019
184
0
0
422
0
0
2691
0
0
1089
0
0
331
0
0
1294
0
0
0
598
0
271
0
0
0
159
0
0
0
295
Food
Special events
Universal Lenguage
Type
2500
2000
1500
1000
500
0

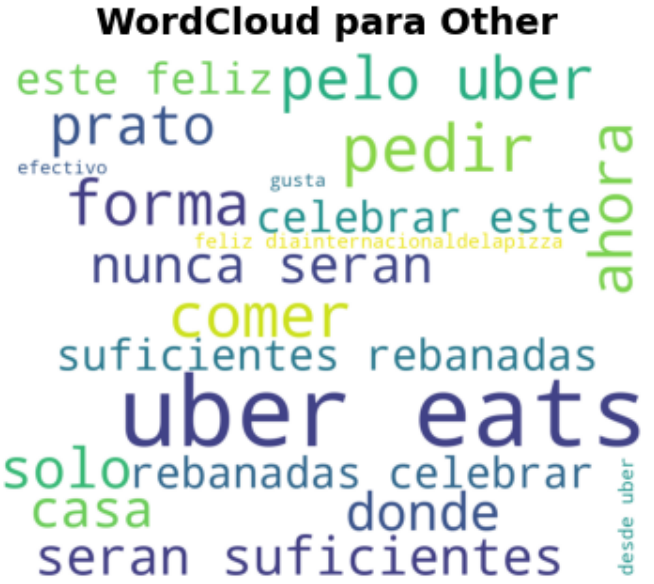
Total

This table present the total for each variable cross with type. We found that Costumer Love have not reactions and the Impressions and Cost are the lowest. For Food, this has the greatest values in general.

Variable	Consumer love	Food	Other	Special events	Universal Lenguage
Total Reach	13298.32	21094171.14	153045.31	7404006.2	0.0
Total Impressions CIRCUS	0.0	17520.0	69.0	352.0	0.0
Total Impressions UBER	14414.78	173162700.24	71382325.18	15628769.13	603359.0
Engagements	12816.0	7630046.0	462500.0	633758.0	176303.0
Likes	0.0	581059.0	54078.0	113581.0	132.0
Love	0.0	28141.0	4401.0	8286.0	17.0
Haha	0.0	15290.0	1689.0	7530.0	6.0
Wow	0.0	5949.0	619.0	1655.0	0.0
Angry	0.0	2508.0	217.0	442.0	33.0
Sad	0.0	742.0	157.0	193.0	1.0
Positive Reactions	0.0	614846.0	60787.0	119877.0	155.0
Negative Reactions	0.0	3250.0	374.0	635.0	34.0
Production Cost	666.4	234538.46	6263.88	37047.79	11675.72
Media Investment	0.0	182549.55	43266.34	34131.73	0.0
Total Cost	666.4	401867.49	48680.93	66235.28	11675.72
CPM	4070.99	287159.48	6801.8	26838.58	11404.1

WordCloud

In food, the words are related to hashtags to increase the reactions, and are similar to Special Events. In the Universal Lenguage, the words are related to greet.



Recommendations

- Evaluate if is necessary to work with Consumer love, because the metrics of this are the lowest in general.
- Food has the highest values of metrics, for this, the post in Instagram could be ingrease the reactions.
- Take in care the post for each country. If we can créate media that are related for each country, the impact could be highter.
- Also, the use of hashtags that are currently used in the social media can help us to impact in more people.
- Finally, in Special Events, we can use the festivities of each country to generate more reviews in this type.