# Toma de decisiones:
# ¿Podemos mejorar las valoraciones de nuestros productos?

Aitziber Atutxa

Marzo 2023

## 1. Contexto de evaluación

A continuación se presenta el enunciado de un ejercicio grupal, cuyo entregable permitirá conseguir puntuaciones que van desde el 1,5 al 4 puntos dependiendo de los hitos que se alcancen. La entrega se realizará antes del día 5 de Mayo. Se podrá encontrar una descripción precisa de los hitos y su puntuación asociada al final de este documento. Las tareas se realizarán en grupos de 4 o 5 personas (dependiendo de cuanta gente siga en la evaluación continua) pero para poder conseguir esos puntos cada alumno deberá someterse a una pequeña prueba individual que se hará la última semana en las horas de clase (y posiblemente alguna hora adicional más) y que permitirá asegurar de que todos los participantes del grupo han contribuido al trabajo conjunto. Si es el caso, y las contribuciones son equitativas y todos los participantes saben desarrollar todas las tareas necesarias para alcanzar los hitos no debería de haber problema para conseguir los puntos. Solo en los casos en los que las contribuciones hayan sido muy desiguales, las personas que no hayan trabajado lo suficiente y/o hayan delegado demasiado en sus compañeros podrían no conseguir ningún punto.

## 2. Preguntas mínimas a las que queremos responder

1. ¿Existe alguna fecha/hora en la que las valoraciones hayan sido peores?
2. ¿Podemos identificar patrones relacionados con las críticas negativas y las positivas? Identificar estos aspectos pueden permitir a la empresa reducir las valoraciones negativas, relacionar los aspectos negativos como el servicio a clientes
3. ¿Qué aspectos de tu negocio/producto generan menciones positivas?
4. ¿Están estos relacionados con elementos mejorables en mi compañía o están relacionados con aspectos que no dependen de la empresa?.
5. ¿Cuál es el comportamiento de mi competidor en lo que respecta a la mismas preguntas?

Tareas que nos pueden ayudar a responder a las preguntas:

- **Análisis de Sentimientos:** ¿Por qué el análisis de sentimientos? Las reseñas de Amazon o Yelp están clasificados, pero Twitter u otras redes sociales no, los comentarios que pertenecen a el conjunto de datos objeto de este proyecto han sido claficados por humanos posteriormente, aprender un analizador de sentimientos puede ayudarnos a reconocer las valoraciones positivas o negativas de fuentes como Twitter puede ser muy interesante para la empresa.
- **Topic-modeling:** ¿Por qué topic-modeling? Una tarea que se llama topic-modeling y que no es más que un clustering no supervisado puede ofrecernos la clave. En estos laboratorios emplearéis LDA y NMF para llevar acabo el topic modeling. Se probará con distintos números de clusters, números de palabras clave etc hasta obtener algún dato que permita responder a las preguntas formuladas inicialmente de forma que ayuden a tomar decisiones identificando los aspectos a mejorar y aquellos que son bien valorados y podemos quizás potenciar.

## 3. Datos para el estudio

Se emplearán los datos de Twitter disponibles, con la información sobre las valoraciones de los comentarios que los usuarios han hecho en Twitter. Los datos se pueden descargar desde nuestro servidor:
`http://lsi.bp.ehu.es/asignaturas/SAD-2022-2023/datos`

## 4. Entregables

Los entregables asociados a esta tarea consistirán en:

1. Un informe descriptivo de los datos: Los tipos de cada descriptor, un parrafo explicativo de la información que almacenan. Una explicación de cúal ha sido el empleo que se le ha dado a cada uno (si ha sido empleado).
2. Un clasificador de Tweets (opinión positiva, negativa o neutra)
   - `https://www.overleaf.com/read/tfgmhrfmhhxw`

     El emtregable y el código se deberán subir para el 21/03/2023
3. Un Dashboard mostrando breve análisis exploratorio a través de visualizaciones de los más relevantes de los datos ([1]).
4. Una historia contada con datos (Data Story Telling) que permita responder a las preguntas iniciales.

## 5. Evaluación

Todo grupo que entregue la documentación requerida obtendrá:

**1 punto** si cumple los siguientes mínimos:

1. Un informe describiendo los datos. Los tipos de datos, si son cuantitativos o cualitativos, si son discretos o continuos, si son ordinales etc.
2. Un clasificador de opiniones de Twitter, empleando el sistema que le corresponda a tu grupo.
3. (0.5 adicional) Un clasificador de Twitters que supere el FScore obtenido por dataiku aplicando KNN.
4. Aplicación del clasificador obtenido a los Twitters de la empresa que ha contratado a tu grupo.

**0.75 punto** si cumple los siguientes mínimos:

1. Un algoritmo de clustering que permita encontrar las razones tras las opcniones negativas.
2. (0.75 adicional) Si vuestro algoritmo encuentra alguna de las razones subyacentes que están etiquetadas en el corpus y lo justificáis apropiadamente.

**0.75 punto** si cumple los siguientes mínimos:

1. Un Dashboard mostrando breve análisis exploratorio a través de visualizaciones de los aspectos más relevantes de los datos (por ejemplo las valoraciones, representar la media y la desviación standar, comparativa con nuestros competidores.
2. Una historia contada con datos (Data Story Telling) que permita responder a las preguntas iniciales y que muestre los tópicos que el modelado de tópicos implementado haya obtenido y/o los del Gold Standard

# 6. Recursos útiles

## 6.1. Anaconda Navigator

Anaconda-navigator es una interfaz gráfica de usuario (GUI) de escritorio incluida en la distribución de Anaconda® que le permite iniciar aplicaciones y administrar fácilmente paquetes, entornos y paquetes de conda sin usar comandos de línea de comandos. Navigator puede buscar paquetes en Anaconda.org o en un repositorio de Anaconda Local.

conda está en la base de anaconda-navigator, y conda es un sistema de gestión de paquetes de código abierto y un sistema de gestión del entorno que se ejecuta en Windows, macOS y Linux. Conda instala, ejecuta y actualiza rápidamente paquetes y sus dependencias. Conda crea, guarda, carga y cambia fácilmente entre entornos en tu máquina local. Fue creado para programas Python, pero puede empaquetar y distribuir software para cualquier lenguaje.

Instalar en anaconda-navigator.

`https://www.anaconda.com/products/individual#Downloads`

Para redimensionar Anaconda-navigator 1. Anaconda-navigator File → Preferences

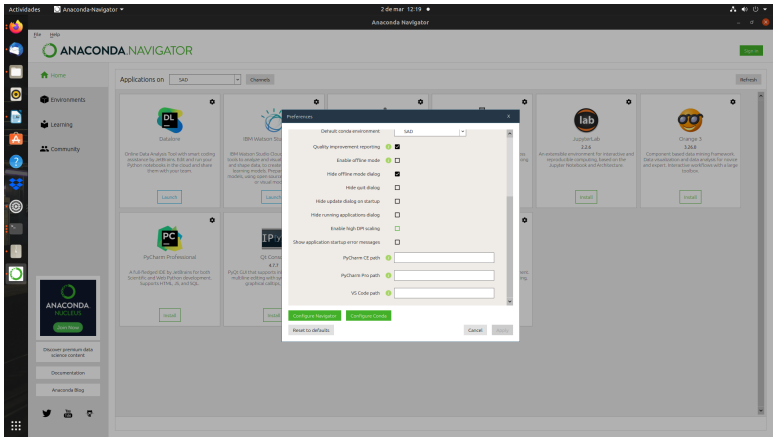


Figura 1: Adecuando anaconda a la pantalla actual.

Deshabilitar la opción Enable high DPI scaling (cuando lo deshabilitamos el boton de Apply se activa así que con darle al enter es suficiente mirar figura 4).

Una vez dentro podemos generar entornos. Vamos a hacer la prueba y vais a generar uno para SAD. Hay que fijarse en la versión de python que va a tomar como referencia, y si no es la que nos interesa modificarla.

En Enviroments a través del + creamos un nuevo enviroment, en este caso SAD (Figura 4).

Una vez generado, volvemos al Home y seleccionamos dicho entorno.

E instalamos desde el propio **anaconda-navigator** el jupyter notebook si no lo tuviesemos instalado.

Una vez instalado lanzamos el **jupyter notebook** que se abrirá en nuestro navegador de manera local. El **jupyter-lab** es la versión on-line de esta herramienta y es similar al colab (google). El problema de estas herramientas "lab" es que necesariamente necesitas una buena conexión de red para poder ejecutarlas. La ventaja es que si disponemos de una máquina poco potente la versión lab nos puede sacar del apuro.

En **jupyter-notebook** cargaréis los ficheros .nbpy especialmente preparados para los laboratorios de **Sentiment Analysis** y de **Topic modeling**.

No se asume que se vayan a emplear Modelos del Lenguaje dado que no se han explicado en clase, pero si alguien quisiera probar seguidamente la presentan referencias para comenzar. Si se fuesen a emplear (los Language Models) conviene:

- Abrir una cuenta en Colab para poder hacer uno de GPUs (https://colab.research.google.com/)
- Consultar estos tutoriales de hugginface:
  - https://www.youtube.com/watch?v=-QH8fRhqFHM
  - https://huggingface.co/course/chapter7/3?fw=tf
  - https://huggingface.co/blog/sentiment-analysis-python

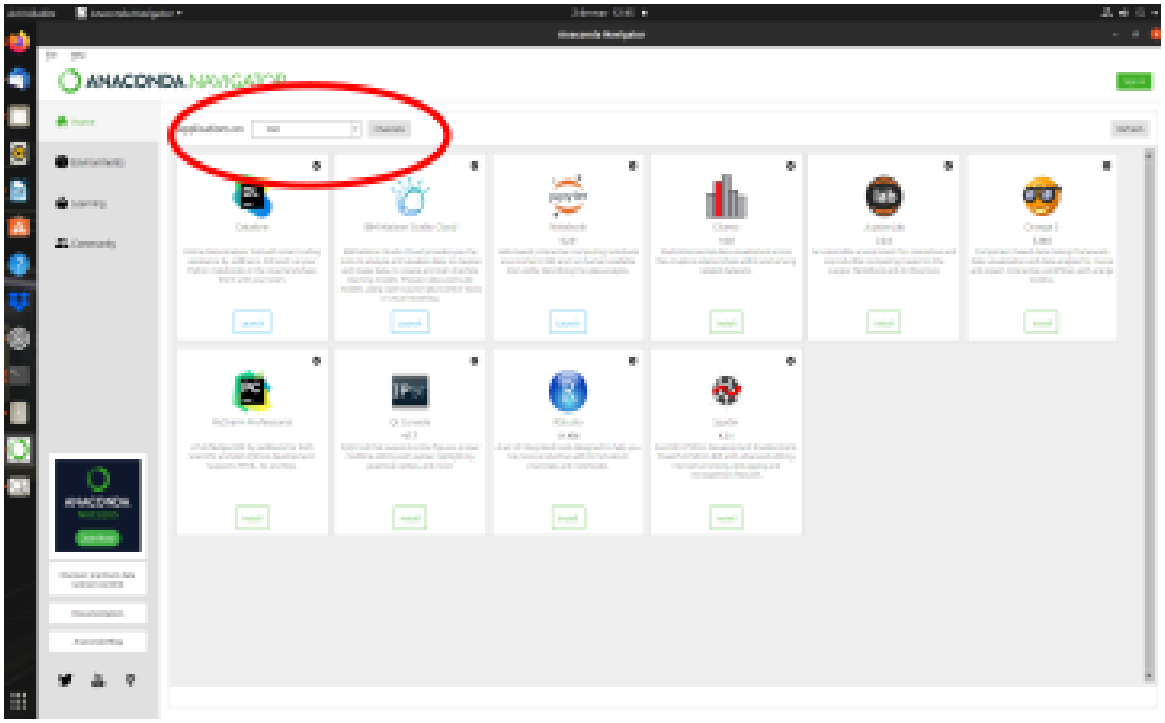

Figura 2: Entornos en Anaconda.

## 6.2. Tableau

En lo referente a las visualizaciones emplearéis Tableau (referencias: Datanyze,googleTrends)
Configuracion regional: Muy importante por lo que considere como float.
Para emplearlo online y poder compartir con vuestros compañeros el trabajo:
`https://sso.online.tableau.com/` Vuestro usuario será vuestro email y las pass inicial será sad1

Para instalarlo (solo sbre windows)
`https://www.tableau.com/products/desktop/download?signin=academic`
La clave de activación es: TCAR-F789-6360-9F68-0C17

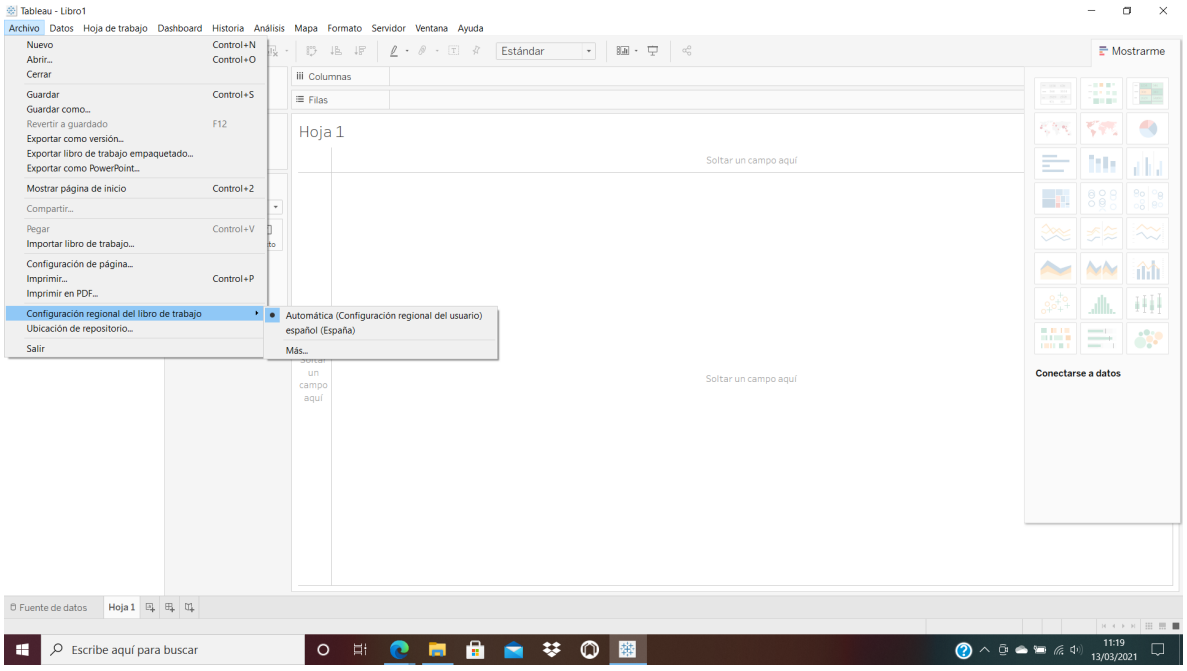


Figura 3: Tableau: Configuración Regional.

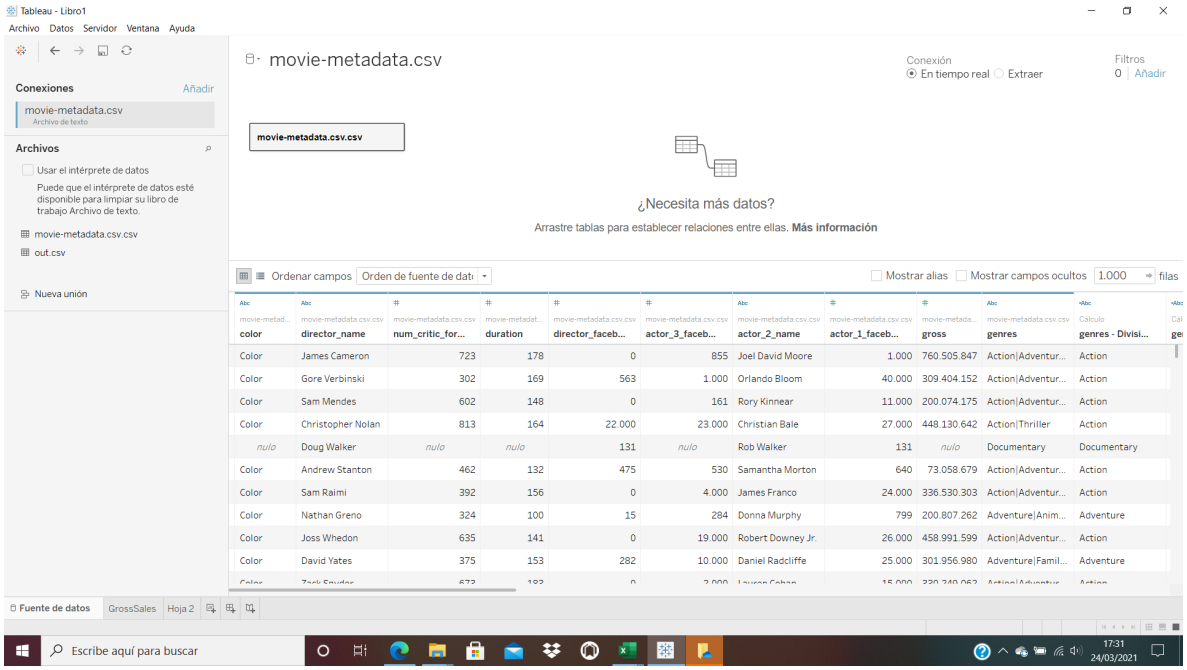


Figura 4: Fuente de Datos Tableau.

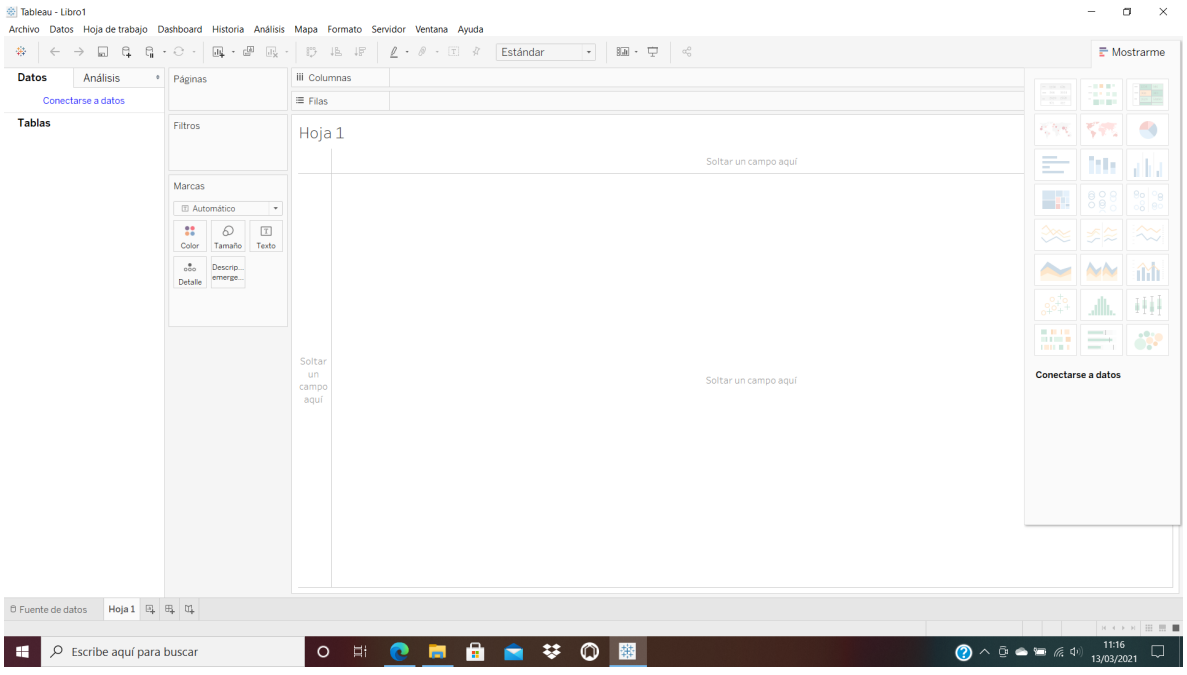


Figura 5: Hoja Tableau.

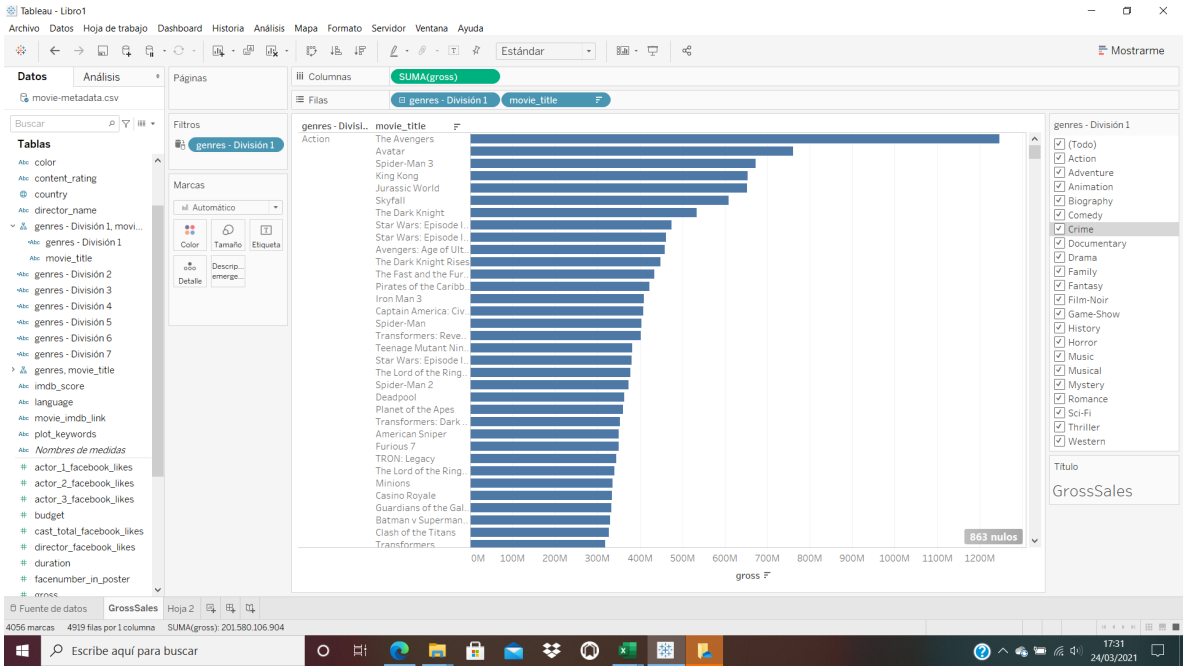


Figura 6: Ejemplo.

## 6.3. Librerías útiles

Obtener geolocalizaciones a partir de ciudades:

- Por ejemplo: https://amaral.northwestern.edu/blog/getting-long-lat-list-cities

## 6.4. Corpus de Twitter

Este corpus contiene información sobre comentarios de Twitter vertidos por usuarios en relación a líneas aereas americanas. Estos tweets estan etiquetados como positivos, negativos o neutros y por lo tanto permiten realizar experimentos de análisis de sentimientos y sobre los problemas de cada una de las principales aerolíneas estadounidenses. Los datos de Twitter se extrajeron de febrero de 2015 y se les pidió a los usuarios que primero clasificaran los tweets positivos, negativos y neutros, y luego clasificaran las razones negativas (como "vuelo tardío.º "servicio grosero"). Los datos precisarán de un pre-proceso porque puede que el .csv no esté perfecto y haya que limpiarlo y preprocesarlo (columnas que estén movidas, valores incorrectos, etc).

Este corpus dispone de información que te puede ayudar a responder las preguntas iniciales.

1. tweet_id (Integer): identificador del tweet
2. airline_sentiment (Text): sentimiento asociado al tweet, positive, negative o neutral.
3. airline_sentiment_confidence (Decimal): Un valor entre 0 y 1 que mide el nivel de confianza de la clasificación en positivo, negativo o neutro del humano que los ha anotado.
4. negativereason (Text): Razón asociada a la opinión y que en teoría tenéis que encontrar a través del clustering.
5. negativereason_confidence (Decimal): Un valor entre 0 y 1 que mide el nivel de confianza del humano que ha anotado esa razón.
6. airline (Text): Nombre de la línea aerea
7. retweet_count (Integer): Cuantas veces se ha retuiteado ese tweet.

8. text (Natural lang.): El texto del tweet
9. tweet_coord (Array): las coordenadas Latitud y longitud del GPS. Este atributo contiene muchos valores faltantes. Si falta el valor se empleará alguna librería que a partir de una ciudad nos devuelva unas coordenadas GPS.
10. tweet_created (Date): Fecha del tweet.
11. tweet_location (Text): ciudad en la que fue escrito es tweet. Este campo contiene mucho ruido y habrá que limpiarlo.
12. user_timezone (Natural lang.): Si no existe información de localización GPS, se empleará este atributo para recuperar el missing value del atributo tweet_coord. De forma que:
    - Eastern Time (US  Canada): será New York City, New York
    - Central Time (US  Canada): será Austin, Texas
    - Mountain Time (US  Canada): será Denver, Colorado
    - Pacific Time (US  Canada): será San Francisco, California
    - el resto podréis encontrarlo en: https://gisgeography.com/north-america-time-zone-map/
      En caso de no poder recuperarse el valor faltante a través del empleo del atributo user_timezone el valor faltante se recuperará a través del siguiente heurístico: Se considerarán los heasdquearters de la aerolínea como la ciudad para obtener su coordenada GPS.
      - United: Chicago, Illinois.
      - Southwest: Dallas, Texas.
      - Delta: Atlanta, Georgia.
      - US airways: Alexandria, Virginia.
      - Virgin America: Burlingame, California.

# Referencias

[1] The Big Book of Dashboards: Visualizing Your Data Using Real–World Business Scenarios, Jeffrey Shaffer, Steve Wexler, Andy Cotgreave, Wiley.

[2] Storytelling con datos. Visualización de datos para profesionales, Cole Nussbaumer Knaflic, Grupo Anaya.