TP01

* Identificación de tipos de Variables (numéricas, atributos, binarios)
* Medias, Medianas, Moda:
  + Datos Usuarios: Followers\_count, Friend\_count, Statuses\_count, Favorites\_counts
  + Datos Tweets: Favorite\_count, Retweet\_count
* Gráfico de Torta. Ej: Cantidad de Tweets agrupadas por Location
* Gráfico de Barra.
  + Protected (cuenta protegida o no)
  + Verified
  + Location: Se hace una primera visualización de los datos crudos, tabla de contingencias de por medio, sin poder obtener mucha información al respecto. Luego se hace una transformación de los datos a escala logarítmica, viendo que aun asi los datos están visualmente sesgados. Esta visualización se complementa con un Boxplot (también en escala logarítmica). Al visualizar la tabla de frecuencias de los datos ordenadas, se decide trabajar con la locación por Pais, aumentando la jerarquía por agrupamiento (Han). Se realiza un preprocesamiento y limpieza de los datos (descripto abajo en Preprocesamiento).

Una vez que podemos generar la variable “país”, hacemos un Histograma (en escala logarítmica?)

* + Hashtags (group por día?). Se hace para aquellos tweets que si tienen hashtags (datos crudos)
  + Hashtags (group por dia?) luego de hacer el preprocesamiento y obtención de la variable nueva de Hashtags.
  + Tweets que son originales y que son retweet
  + Source de tweet
* Histograma (Necesaria la Transformacion de datos para evitar Sesgo).
  + Followers Count, Friend Count (seguimiento mutuo)
  + Cantidad de Tweets por dia (por franja horaria?)
* Boxplot – Primera noción de Outliers.
  + Location (datos crudos, luego boxplot de país)
  + Followers: se hace una primera visualización global, luego se hace un plot de followers, pintando los puntos por cuentas verificadas y no verificadas y se identifican 2 clusters de datos. Luego se decide realizar los boxplots nuevamente, pero agrupando los datos por Verificados y no Verificados.
* Scatterplot/ Plot 2D:
  + Cantidad de Seguidores vs Retweets
  + Cantidad de seguidores vs likes
  + Cantidad de seguidores vs likes vs Antigüedad
* Scatter 3D
* Matriz de Correlación de variables:
  + Cantidad de Tweets vs Antigüedad de la cuenta
  + Cantidad de Likes vs Cantidad de Seguidores
  + Cantidad de seguidores vs likes vs Antiguedad
* Visualización agrupando por Atributo:
  + Cantidad de Tweets vs Verified T o F. Twittean mas las cuentas verificadas? En la description podemos encontrar si son periodistas, medios, diarios, etc?
  + Cantidad de retweets/likes si Media o URL asociado
* Normalización de datos para hacerlos comparables (para la matriz de correlación)

Integración de datos:

* A través de la variable “user\_id” . Y tomando XXX variables de cada dato para evitar redundancia. (followers count vs Friends count con Corr). Aca decidimos trabajar directamente con la table de “tweets” que ya incluye los datos de la tabla “users”. Se hace una verificación cruzada de estos datos?
* Verificar diferencias de granularidad (Ej: Formatos de fechas y Location. En el caso de Location quedarnos con el Pais)
* Estudio de Redundancia:
  + Prueba de Chi2 (Atributos Nominales, a través de Tabla de Contingencias). Hashtags con Location
  + Matriz de Correlación (atributos Numéricos)

Preprocesamiento:

* Detección de Outliers: Univariados (Boxplot, IQR max, Zscore, Zscore Mod) y Multivariados (Mahalanobitz y LOF)
* Datos Faltantes:
  + Contabilizar los datos faltantes: que no supere el 10-15%
  + Podemos elegir trabajar con los registros completos
  + Reemplazos: Sustitución por la media, por regresión, HotDeck o MICE
    - Habra algún usuario que tenga datos faltantes para Followers\_count, statuses\_count o favorites\_count? Podriamos inferirlo a partir de un MICE?

Decisiones de Diseño:

* Luego de consultar el significado y contenido de cada una de las variables (Fuente), decidieron eliminarse algunas de ellas, siendo que no eran relevantes para el estudio a desarrollar:
* Status\_id
* Display\_text\_width
* Symbols
* Urls\_url
* Urls\_t\_co
* Media\_url
* Media\_t\_co
* Ext\_media\_url
* Ext\_media\_t\_co
* Geo\_coords
* Coords\_coords
* Bbox\_coords
* Listed\_count
* Profile\_banner\_url
* Profile\_background\_url
* Profile\_image\_url
* retweet\_status\_id

Algunas variables parecían ser de interés en un comienzo, pero detectamos que tenían valor 0 para todos los registros:

* Quote\_count
* Reply\_count

Preprocesamiento de variables/ Tratamiento:

* Location: Para unificar los países, se realiza el siguiente procedimiento:
  + - Con los datos crudos, se genera una tabla de frecuencias.
    - Se decide tomar aquellos valores cuya frecuencia sea mayor o igual a XXX
    - Se identifican los países de estos valores (a mano?) y se arma una tabla auxiliar con los países de interés
    - Se buscan los strings correspondientes a los países en el resto de los datos de menor frecuencia, teniendo en cuenta las diferencias en mayúsculas, tildes, etc (tiene un nombre esto?)
* Hashtags:
  + Se dividen los datos en aquellos que tienen hashtags y aquellos que no tienen hashtags.
  + Se buscan los hashtags que hay en cada tweet y se genera una tabla con los valorse y sus frecuencias
  + Aquellos de mayor frecuencia (criterio, mayora a XX) se toman como relevantes para el estudio
  + Para aquellos tweets sin hashtags, se procede a buscar los hashtags generados en el paso anterior dentro del string del texto del tweet 🡪 Con esto se genera una NUEVA variable de “hashtags\_limpios”.
* Fecha Creacion de Tweet (created\_at):
  + Para las mismas fechas los rangos de hora son chicos, por lo cual se decide trabajar con la fecha de creación en formato día-mes-año. Se genera una variable nueva de “Fecha” con el dato en este formato (o se limpia la hora de la original?)

Variables auxiliares/Creadas:

* Proporción retweets: Tweets que son retweet / cantidad total de tweets.   
  Esta variable se visualiza vs la cantidad de seguidores de la cuenta. Hay alguna correalcion?
* Actividad en Twitter: Una combinación (lineal?) entre “statuses\_count” y “favorites\_count”. Que correlación tenemos de esta variable de “actividad en la cuenta” vs cantidad de followers/ antigüedad de la cuenta.

Informacion acerca de las variables/ otros comentarios:

* Is\_quote: si el tweet está citando a otro tweet
* Alguna variable que la descartamos por la cantidad excesiva de NA
* Geo\_coords, coords coords descartamos por muchos NA
* Statuses\_count: cantidad de cosas que twitteo o retwettio el usuario en toda su vida
* Favourites\_count: Cantidad de likes que dio en toda su historia
* Quote\_count / Replay\_count esta todo en 0