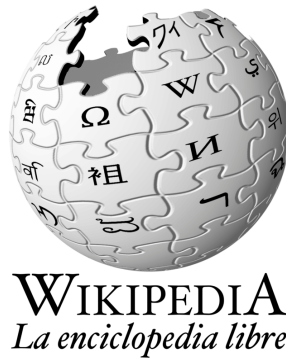


# TP1 - Pandas y Visualización de datos

## Primera parte - Pandas (15 pts)



Utilizamos el dump Wikipedia Español [al día 01/09](#) de 178gb, extrayendo [los siguientes CSVs](#):

### contents.csv

Tabla con datos de todos los contenidos de Wikipedia en su versión más reciente.



title	Título del contenido
id	Identificador único del contenido
namespace	👤
revision_id	Id de la última revisión realizada
parent_revision	Id de la revisión anterior a la actual
revision_timestamp	Timestamp de la última revisión
revisor_username	Username del autor de la última revisión
revisor_id	Id del revisor_username
revisor_ip	IP del revisor (en caso de que no tuviera usuario)
revisor_comment	Comentario de la revisión

## contents\_text\_sample.csv

Tiene una muestra aleatoria del 5% de los contenidos de texto de wikipedia

id	Id del contenido
title	Título del contenido
text	Texto

## geo\_tags.csv

gt_id	Id del geo tag
gt_page_id	Id del contenido al que corresponde
gt_globe	En qué globo se encuentra
gt_primary	
gt_lat	Latitud
gt_lon	Longitud
gt_dim	
gt_type	Tipo de locación
gt_name	Nombre
gt_country	País
gt_region	Región

## logs.csv

Todo el log de acciones realizadas.

item_id	ID del ítem afectado
timestamp	Timestamp del log
contributor_username	Username que realizó la acción
contributor_id	ID del user que realizó la acción
contributor_ip	IP (en caso de que no tuviera usuario)
comment	Comentario
logtype	Tipo de log
action	Acción realizada
title	Título del log




## languages.csv

Contiene información sobre qué idiomas habla cada usuario

babel_user	User id
babel_lang	Código de idioma (ISO 639-2)
babel_level	<a href="#">Nivel</a> en el lenguaje

## redirect\_list.csv

Algunos de los contenidos de Wikipedia son redirecciones a otros contenidos, esta tabla contiene esa información.

rd_from	ID del contenido que redirige
rd_namespace	
rd_title	Título del contenido al que redirige
rd_interwiki	
rd_fragment	

## categorylinks.csv

cl_from	ID del contenido
cl_to	Categoría a la que pertenece el contenido
cl_sortkey	👤
cl_timestamp	Timestamp de la asociación de la categoría
cl_sortkey_prefix	👤
cl_collation	👤
cl_type	El tipo de contenido que se asignó a esa categoría

## pagelinks\_sample.csv

Tabla con links que van de una página interna a otra. Es una muestra de dos tercios.

pl_from	ID del contenido donde está el link
pl_namespace	👤
pl_title	Título del contenido al cual va el link
pl_from_namespace	👤

## Realizar [sus correspondientes consultas](#) en Pandas

1. Para el usuario que más versiones actuales de contenido de wikipedia editó, calcule la fecha promedio, mínima y máxima en que lo hizo (★)
2. Qué porcentaje de las versiones actuales son páginas que se editaron una sola vez (★)
3. Cual es el porcentaje de títulos de contenidos de wikipedia cuya longitud es menor a 20 (★)
4. La probabilidad de que un usuario edite sin dejar comentario y la misma probabilidad para un editor que no está logueado para los artículos que están visibles(★)
5. La palabra más común entre los títulos que no sea una stopword del inglés ni español (★★)
6. El porcentaje de contenidos que están publicados cuya última edición no tiene comentario para los usuarios que realizaron 1, >10 y >100 de las últimas ediciones (★★)
7. La antigüedad promedio de la última edición de los artículos cuyo título contenga tu apellido (si no hay, tu nombre y si tampoco hay usa Cafferata) (★★)
8. La mediana de la antigüedad para las últimas ediciones vigentes agrupado por el primer carácter del título (★★)
9. Cuales son los contenidos de wikipedia cuyo título empieza o termina con un emoji (★★)
10. Para los contenidos visibles en wikipedia, cuales son los artículos que tienen la máxima y mínima distancia entre ids de su revisión actual y la anterior (★★)
11. Para todos los comentarios de revisión de contenido que tengan más de 20 ocurrencias realice una matriz cuyas columnas sean esos comentarios y de índice los usuarios/ips con valores: True si ese usuario realizó ese comentario, sino False (★★)
12. Cuantos comentarios de revisión de artículos usan la palabra "mejor" (sin incluir sus variaciones) (★★)
13. Realice una consulta en los contenidos actuales que le permita identificar algún artículo que este vandalizado utilizando los datos de la revisión (★★)
14. Qué porcentaje de contenido geolocalizado de wikipedia NO está en la tierra (★)
15. Obtenga la matriz de distancias euclídeas para todos los contenidos que están en Marte. ¿Cuáles son los dos contenidos que están a menor distancia? (★★)
16. Calcule la probabilidad de las palabras para los textos, luego encuentre el documento que más se desvie de esas probabilidades utilizando la divergencia de Kullback-Leibler (★★)
17. Utilice los textos del contenido para realizar consultas por texto utilizando las técnicas vistas en la clase de NLP (BOW o TF-IDF) de modo que la query "retablo iglesia" devuelva alguna página acerca del retablo de alguna iglesia (★★)
18. Divida la tierra en bloques de latitud y longitud de 5x5, ¿Cuál es el bloque con menos (o ninguna) referencias? (★★)
19. Calcule la latitud y longitud promedio de los contenidos con referencias en la tierra y diga dónde está eso (★)
20. ¿Cuál es el segundo contenido con más referencias geográficas asignadas? (★★)
21. ¿Dónde está la referencia geográfica más repetida en la tierra de toda la Wikipedia Español? (★)
22. Elija su lugar favorito en el mundo y tome su latitud y longitud, ¿cuál es el título de la

- página de wikipedia más cercana? (★★)
23. ¿Qué porcentaje de los contenidos contienen a su mismo título en el texto? (★★)
  24. Calcule el porcentaje de nulos para todas las columnas de geo\_tags.csv (★)
  25. ¿Quién es el usuario que más ha bloqueado a otros? (★)
  26. ¿Cuál es el usuario o IP más bloqueado? (★)
  27. ¿Cuál es el mínimo que ha durado desde su registro un usuario bloqueado en la plataforma? (★★)
  28. ¿Cuál es la antigüedad promedio para cada usuario según su última actividad? (★★)
  29. Utilice los logs para crear una matriz cuyas columnas sean los logtypes, los índices los actions y las celdas la cantidad de la intersección de ambas (★)
  30. La 3-upla de palabras más común en los comentarios de los logs (★★)
  31. El día con más y menos actividad que tuvo el sitio (★)
  32. El usuario que más agradece y el que más agradecimientos tiene (★)
  33. La primera discusión creada (★)
  34. ¿Cuántos usuarios son nativos en un idioma que no sea español? (★)
  35. Para los usuarios nativos (o superior) en español obtenga una serie cuyo índice sea el idioma y valor sea el nivel promedio (★★)
  36. Quien es el usuario que más idiomas domina con un nivel de 2 o superior (★)
  37. Obtenga un dataframe que tenga como índice al user\_id, como columnas a los idiomas y el nivel de cada usuario para cada idioma como valor con -1 en caso de no tenerlo cargado. (★★)
  38. Obtenga la matriz de correlación para saber idiomas distintos considerando que un usuario sabe un idioma si indicó un nivel de 1 o superior (★★)
  39. ¿Cuál es el contenido al que más se hacen redirecciones? (★)
  40. Para los contenidos geolocalizados: ¿Cuál es el contenido más cercano del que fue editado más recientemente? ¿Y la diferencia entre sus tiempos de edición? (★★★★)
  41. Para los contenidos geolocalizados, según la última versión de cada contenido: ¿Cuál es la latitud y longitud promedio del contenido editado según qué idioma sabe el editor? (★★★★)
  42. Si la experiencia de un usuario es la cantidad de logs en los que participó, ¿cuál es la tasa de contenidos cuya última revisión no tiene comentario en función de la experiencia de su revisor? (★★★★)
  43. ¿Cuántos usuarios o ips han sido bloqueados al menos una vez y la vez son los revisores de una última versión de un contenido? Calcule la diferencia entre la primera fecha de bloqueo y el promedio de las fechas de revisión correspondientes para cada usuario. (★★★★)
  44. Si decimos que la ubicación de un usuario es el promedio de la latitud y longitud de los contenidos geolocalizados para los cuales editó la última versión (ignorar usuarios que no editaron contenido geolocalizado). ¿Cuáles son los dos usuarios más cercanos? (★★★★)
  45. ¿A qué contenido se asignó por primera vez una categoría? (★)
  46. Si decimos que la ubicación de una categoría es el promedio de la latitud y longitud de sus contenidos geolocalizados que son miembros de ella (si es que tiene): ¿Cuales son las dos categorías más cercanas? (★★★★)
  47. La mediana de cantidad de links internos que tienen todos los contenidos que existen. (★★)

48. Si decimos que la ubicación de una página linkeada por otra es el promedio de la latitud y longitud de los contenidos geolocalizados que la referencian: ¿Cuales son las dos páginas que están más lejos? (★★★★)
49. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para aquellos que editaron una de las versiones actuales del contenido, ¿Cuál es la tasa de revisiones sin comentario que realizan en función de los idiomas que saben? (★★★★)
50. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1 consiga un dataframe cuyas columnas son tipos de logs, el índice es la cantidad de idiomas que sabe un usuario y las celdas la probabilidad de que esos usuarios generen ese tipo de log. (★★★★)
51. Si la experiencia de un usuario es la cantidad de logs en los que participó, queremos saber que tanto nos sirve para predecir el futuro vandalismo: ¿Cuál es la probabilidad de que un usuario sea bloqueado según experiencias: <10, 10-40, 40-100, >100? Tener en cuenta que esta experiencia debe ser PREVIA al bloqueo del usuario. (★★★★)
52. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para cada grupo de usuarios que sabe una determinada cantidad de idiomas, ¿Cuántos de esos usuarios fueron bloqueados al menos una vez? (★★★★)
53. Si para un usuario tenemos la cantidad de acciones que realizó para cada tipo de log y la cantidad de veces que fue bloqueado: ¿Cuál es la acción que más y menos correlaciona con ser bloqueado? ¿Qué acción correlaciona más con saber algo (babel>=0) de inglés? (★★★★)
54. ¿Cuál es la acción más realizada por usuarios que no están registrados? (★★)
55. La cantidad promedio de modificaciones históricas que tuvieron los ítems cuya última versión fue editada por un usuario registrado o no registrado. (★★★★)
56. Calcule la cantidad de acciones realizadas por usuarios según día de la semana (★)
57. Calcule la probabilidad de que una acción en general se realice según día de la semana. Calcule también para los días de la semana la probabilidad de que la última edición de un contenido sea realizada ese día. Calcule la entropía de ambas y la divergencia de Kullback Leibler entre ellas. (★★★★)
58. Observe una muestra aleatoria de los comentarios de las acciones realizadas por usuarios o ips antes de ser bloqueados. Observe otra muestra de comentarios de acciones de todos. (★★)
59. ¿Cuál es el idioma para el cual sus usuarios realizan más agradecimientos en promedio? ¿Y el de menos agradecimientos? Calcule lo mismo para quienes reciben agradecimientos. (★★★★)
60. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para aquellos que editaron una de las versiones actuales del contenido, ¿Cuál es la cantidad de agradecimientos promedio que reciben en función de los idiomas que saben? (★★★★)

## Segunda parte - Visualización de datos (10 ptos)

1. (6 ptos) Elegir 3 de los siguientes datasets:

- [Proyectando el comportamiento de la soja](#)
- [¿Llevo paraguas? Pronosticando la lluvia](#)
- [Predicción de éxitos en oportunidades comerciales](#)
- [Clasificación de preguntas de clientes](#)
- [MELI Data Challenge 2021](#)
- [Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines](#)
- [DengAI: Predicting Disease Spread](#)

Realizar dos visualizaciones para cada uno que expliquen la variable a predecir conteniendo al menos cada uno de los siguientes tipos de plots:

- Bar plot
- Histograma
- Violin plot
- Box plot
- Heatmap




2. (4 ptos) Utilice alguna herramienta para realizar diagramas (por ejemplo Google Draw, draw.io, Google Slides, HTML, Illustrator, Photoshop, etc.) para crear una visualización **ORIGINAL** que no pueda realizarse de forma directa con las librerías más comunes de Python, puede utilizar las librerías de Python como paso intermedio. Puede realizar este punto sobre los datos de: cualquier dataset, estadística oficial, paper, estadística no oficial, encuesta, números sin ninguna fuente en un blog, etc. El objetivo es elegir un tema de su interés y comunicarlo de forma efectiva y agradable.



# Criterio de aprobación

El criterio general es que la totalidad del tp tiene que sumar 15 puntos de los 25, un 60%. Cada uno va a tener un ayudante asignado, pueden hacer consultas a por slack o piazza o a sus ayudantes.

## Primera parte - Pandas

- Todos los ejercicios valen lo mismo que las estrellitas que tienen asignadas, a cada uno le corresponde hacer según indiquemos cual les toca:
  - 1 ejercicio de 
  - 4 ejercicios de 
  - 2 ejercicio de 
- Cada ejercicio se considera 100% correcto si:
  - Resuelve lo pedido (¡cuidado con casos bordes! ¡revisen todo lo que pueda ser NULL!): Si el ejercicio no resuelve al 100% lo pedido, se considera que vale como máximo la mitad
  - Lo hace de la forma más eficiente posible: Si el ejercicio no está resuelto de la forma más óptima, se considera que vale la mitad
- La idea es que no lo hagan solos! Las consignas son complejas de entender en una sola lectura y necesitan pensarse lento, por esto es que es crucial consultar. Para esto hacemos lo siguiente según el tipo de duda:
  - Dudas de consigna:
    - Van a poder consultar en el canal de slack #consultas-tp1, es MUY importante que antes de consultar vean si su duda no fue resuelta.
    - En caso de no haber sido resuelta tienen que publicarla siguiendo el formato: “<NÚMERO DE CONSIGNA> - La pregunta...”. De esta forma todos podemos buscar fácil si ya se resolvió la duda o sumarnos a la discusión. **No** se debe incluir código de la resolución, ni en la pregunta ni interactuando con otros compañeros.
  - Dudas para saber si se puede usar alguna librería:
    - Se hacen en el mismo formato que las dudas de consigna.
  - Dudas de código y optimización:
    - Si son dudas generales de “cómo se hace algo en pandas” se puede consultar en las clases de consulta o en el canal #otras-consultas
    - El resto de las dudas se deben consultar con su ayudante asignado

## Segunda parte - Visualización de datos

1. Cada visualización vale un punto, y debe cumplir con las siguientes condiciones:
  - a. Debe explicarse por sí misma, sin necesidad de texto aclaratorio.
  - b. Debe tener rótulos en los ejes que corresponda y en el título.
  - c. Debe mostrar una relación con el target que sea clara.
  - d. El uso del color debe ser intencional, elegido por ustedes, no por la librería.
  - e. La visualización debe ser legible (Un bar chart de 40 barras por ejemplo es ilegible)
2. Debe cumplir el objetivo propuesto: Les recomendamos preguntar en clases de consultas o por slack, vamos a estar guiándolos en este punto. Dado que la elección de este dataset es personal, pueden ir compartiendo sus ideas/bocetos o consultando cosas en #consultas-tp1.

¡También valoramos que se ayuden entre ustedes, debatan y compartan ideas en el canal slack!