**Manipulación y Análisis de datos con Pandas y Python**

**Introducción al documento**

El contenido de este documento son **apuntes teoricos** del [Curso de Manipulación y Análisis de Datos con Pandas y Python](https://platzi.com/cursos/pandas/) y busca ser una guía para futuros trabajos personales. El mismo está dictado por [David Torres](https://twitter.com/davinci137), Data Scientist en [Platzi](https://platzi.com/).

Pandas es la librería de software libre para manipulación de datos con Python más usada en Data Science. Manipula grandes sets de datos numericos, tablas y series de tiempo. Trabaja con múltiples formatos de archivos de datos como csv o xls.

**Objetivos del documento**

* Ejecutar operaciones básicas de Pandas.
* Comprender el significado y uso de los DataFrames en Ciencia de Datos.
* Usar Computational Tools de Pandas para realizar cálculos básicos.
* Trabajar con Statisticals tools.
* Hacer preprocesamiento de datos para crear modelos.
* Extraer información de Data Sets usando Time Series y Date Functionality.

**Tabla de contenido**

* [Comenzando con pandas](https://github.com/francomanca93/analisis-de-datos/tree/master#Comenzando-con-pandas)
  + [¿Qué es pandas?](https://github.com/francomanca93/analisis-de-datos/tree/master#%C2%BFQu%C3%A9-es-pandas?)
  + [Configurando Google Colab](https://github.com/francomanca93/analisis-de-datos/tree/master#Configurando-Google-Colab)
  + [Series e Indexación y selección de datos](https://github.com/francomanca93/analisis-de-datos/tree/master#Series-e-Indexaci%C3%B3n-y-selecci%C3%B3n-de-datos)
  + [DataFrames](https://github.com/francomanca93/analisis-de-datos/tree/master#DataFrames)
  + [Indexado y manejo de archivos CSV](https://github.com/francomanca93/analisis-de-datos/tree/master#Indexado-y-manejo-de-archivos-CSV)
  + [Conexión con bases de datos tipo SQL](https://github.com/francomanca93/analisis-de-datos/tree/master#Conexi%C3%B3n-con-bases-de-datos-tipo-SQL)
  + [Diferentes formatos para datasets](https://github.com/francomanca93/analisis-de-datos/tree/master#Diferentes-formatos-para-datasets)
* [Funcionalidades básicas y esenciales de pandas](https://github.com/francomanca93/analisis-de-datos/tree/master#Funcionalidades-b%C3%A1sicas-y-esenciales-de-pandas)
  + [Formatos de lectura para cargar y guardar DataFrames](https://github.com/francomanca93/analisis-de-datos/tree/master#Formatos-de-lectura-para-cargar-y-guardar-DataFrames)
  + [Meteorite Landings Dataset](https://github.com/francomanca93/analisis-de-datos/tree/master#Meteorite-Landings-Dataset)
    - [Tipos de Variables que componen un data frame](https://github.com/francomanca93/analisis-de-datos/tree/master#Tipos-de-variables-que-componen-un-data-frame)
    - [Estructuras de dataframes en detalle](https://github.com/francomanca93/analisis-de-datos/tree/master#Estructuras-de-dataframes-en-detalle)
    - [Borrar filas, columnas y copiar información](https://github.com/francomanca93/analisis-de-datos/tree/master#Borrar-filas-columnas-y-copiar-informaci%C3%B3n)
* [Aplicando pandas](https://github.com/francomanca93/analisis-de-datos/tree/master#Aplicando-pandas)
  + [London bike sharing dataset](https://github.com/francomanca93/analisis-de-datos/tree/master#London-bike-sharing-dataset)
    - [Funciones matemáticas](https://github.com/francomanca93/analisis-de-datos/tree/master#Funciones-matem%C3%A1ticas)
    - [Funciones más complejas y lambdas](https://github.com/francomanca93/analisis-de-datos/tree/master#Funciones-m%C3%A1s-complejas-y-lambdas)
  + [Population dataset](https://github.com/francomanca93/analisis-de-datos/tree/master#Population-dataset)
    - [Múltiples índices](https://github.com/francomanca93/analisis-de-datos/tree/master#M%C3%BAltiples-%C3%ADndices)
  + [Datasets propios](https://github.com/francomanca93/analisis-de-datos/tree/master#Datasets-propios)
    - [Como trabajar con variables de tipo texto en Pandas](https://github.com/francomanca93/analisis-de-datos/tree/master#Como-trabajar-con-variables-de-tipo-texto-en-Pandas)
    - [Concatenación de DataFrames, concat y append](https://github.com/francomanca93/analisis-de-datos/tree/master#Concatenaci%C3%B3n-de-DataFrames,-concat-y-append)
    - [Merge de DataFrames](https://github.com/francomanca93/analisis-de-datos/tree/master#Merge-de-DataFrames)
    - [Lidiar con datos faltantes en DataFrames](https://github.com/francomanca93/analisis-de-datos/tree/master#Lidiar-con-datos-faltantes-en-DataFrames)
    - [Cómo lidiar con datos duplicados en Pandas](https://github.com/francomanca93/analisis-de-datos/tree/master#C%C3%B3mo-lidiar-con-datos-duplicados-en-Pandas)
  + [Diamond dataset](https://github.com/francomanca93/analisis-de-datos/tree/master#Diamond-dataset)
    - [Group by](https://github.com/francomanca93/analisis-de-datos/tree/master#Group-by)
  + [Tips dataset](https://github.com/francomanca93/analisis-de-datos/tree/master#Tips-dataset)
    - [Group by - Refozando su uso](https://github.com/francomanca93/analisis-de-datos/tree/master#Group-by---Refozando-su-uso)
    - [Group by - extraer valor con variables categóricas](https://github.com/francomanca93/analisis-de-datos/tree/master#Group-by---extraer-valor-con-variables-categ%C3%B3ricas)
    - [Tablas dinámicas con Pivot Table](https://github.com/francomanca93/analisis-de-datos/tree/master#Tablas-din%C3%A1micas-con-Pivot-Table)
  + [COVID-19 dataset](https://github.com/francomanca93/analisis-de-datos/tree/master#COVID-19-dataset)
    - [Series de Tiempo](https://github.com/francomanca93/analisis-de-datos/tree/master#Series-de-Tiempo)
    - [Series de Tiempo - variables nulas](https://github.com/francomanca93/analisis-de-datos/tree/master#Series-de-Tiempo---variables-nulas)
    - [Visualización y graficación de datos](https://github.com/francomanca93/analisis-de-datos/tree/master#Visualizaci%C3%B3n-y-graficaci%C3%B3n-de-datos)
* [Proyecto - Cambio climático](https://github.com/francomanca93/analisis-de-datos/tree/master#Proyecto---Cambio-clim%C3%A1tico)
  + [Iniciando una rutina típica de manejo de datos](https://github.com/francomanca93/analisis-de-datos/tree/master#Iniciando-una-rutina-t%C3%ADpica-de-manejo-de-datos)
  + [Preprocesamiento de datos - terminando de preparar y limpiar los datasets](https://github.com/francomanca93/analisis-de-datos/tree/master#Preprocesamiento-de-datos---terminando-de-preparar-y-limpiar-los-datasets)
  + [Análisis de datos](https://github.com/francomanca93/analisis-de-datos/tree/master#An%C3%A1lisis-de-datos)

**Comenzando con pandas**

**¿Qué es pandas?**

[Pandas](https://pandas.pydata.org/) es una herramienta de análisis y manipulación de datos de código abierto rápida, potente, flexible y fácil de usar, construida sobre Python.

[Dibujo con letras blancas

Descripción generada automáticamente con confianza media](https://github.com/francomanca93/analisis-de-datos/blob/master/readme_img/pandas.png)

Pandas significa **Pan**el **Da**ta. Los paneles de datos son estructuras simples donde se puede organizar por categorias los datos en donde podemos tener variables tipo texto, numericos o booleanos. Cuenta con un eje de filas donde los datos pueden estar organizados temporalmente.

Fue inventada en 2008 por [Wes McKinney](https://en.wikipedia.org/wiki/Wes_McKinney) como una necesidad para analizar grandes volumenes de datos en los mercados financieros.

**Ventajas**

* Reduce lineas de código
* Diseñada especialmente para análisis
* API fácil y concisa
* Multiples funciones

**Desventajas**

* Compatibilidad con matrices 3 (numpy)
* Curva de aprendizaje lenta

**Configurando Google Colab**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/comienzo/1-Comenzando-con-pandas/1_Configurando_Google.ipynb)

En esta sección se ve como configurar **Google Colaboratoty**, tambien se trabaja con operaciones básicas, variables, listas y numpy para experimentar con Colab.

**Colab** el cuál se basa en los cuadernos de Jupyter Notebook, por lo tanto la forma de trabajar es muy similar.

**Google Colab** te brinda una maquina virtual con 100 GB y 12 GB de RAM, estas especificaciones te permiten ejecutar codigos mas complejos en la nube. Ademas colab viene con la mayoria de librerias para el uso en ciencia de datos y machine learning.

**Series e Indexación y selección de datos**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/comienzo/1-Comenzando-con-pandas/2_Series.ipynb)

En esta sección vemos las ventajas que tiene Pandas con series. Vemos como crear series simples y definir ciertas caracteristicas como labels, etc.

Tambien se estudia como obtener los valores de esta, como crear series con diccionarios de Python.

Se finaliza la sección como trabajar con valores nulos, filtrar datos, y crear nuevas series con estos filtros.

**DataFrames**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/comienzo/1-Comenzando-con-pandas/3_Dataframes.ipynb)

En esta sección se estudian como trabajar con Dataframes en Pandas. Se crean dataframs a través de diccionarios de Python. Luego se estudian **métodos** que nos permite **analizar el set de datos inicialmente**. Esto nos permite conocer a nuestro DataFrame.

Seguido a lo anterior se aprende como **extraer datos** del dataframe una vez lo conocemos.

Se finaliza conociendo como **filtrar nuestros datos** gracias a metodos de Pandas y operadores relacionales. Con esto podemos crear dataframe personalizados especificando que es lo que queremos del inicial.

**Indexado y manejo de archivos CSV**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/comienzo/1-Comenzando-con-pandas/4_Guardar_y_cargar.ipynb)

Los que aprenderemos en esta notebook será **como conectar** nuestro ambiente de trabajo **con nuestro nube personal** de google drive, **generar** nuestro primeros **archivos en csv** y aprender a **leerlos**.

Nota: **csv** es el formato universal mas simple que se usa para compartir informacion.

**Conexión con bases de datos tipo SQL**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/comienzo/1-Comenzando-con-pandas/5_Conexi%C3%B3n_con_bases_de_datos_tipo_SQL.ipynb)

En esta sección se estudia cómo usar Pandas y Python para conectarte con tu base de datos SQL. Pandas cuenta con una funcionalidad que facilita el acceso a tus bases de datos tipo SQL.

Tenemos diferentes formas para conectarnos tanto a **PostgreSQL**, **SQL Server**, **MySQL**, **Oracle**, otras.

**Diferentes formatos para datasets**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/comienzo/1-Comenzando-con-pandas/6_Diferentes_formatos_para_datasets.ipynb)

En este notebook se estudia las ventajas y desventajas de trabajar con otros formatos para datasets. Como guardar nuestro dataframe utilizando otros tipos de formatos.

Existen diferentes formatos, entre estos estan los siguientes:

* **CSV** - Es muy versatil ya que solo tiene comas y saltos de linea.
* **JSON** - Tiene un formato muy similar al de un diccionario de Python.
* **Excel** - Permite guardar el archivo en formato .xls para trabajar con el en Excel o Spreadsheets.
* **Pickle** - Permite comprimir la información, es util cuando se tienen tablas grandes.
* **Parquet** - Permite darle un formato que puede usarse en ambientes de Big Data como Hadoop.

**Funcionalidades básicas y esenciales de pandas**

**Formatos de lectura para cargar y guardar DataFrames**

En este notebook estudiamos sobre diversos formatos de almacenamiento de datos y cómo elegir el formato eficiente que se adapte a nuestras necesidades.

En la clase se preparo un [script para generar un DataFrame](https://github.com/francomanca93/analisis-de-datos/blob/funcionalidades-basicas/2-Funcionalidades-basicas-de-pandas/3_1_Save_Load_formats.ipynb) de mas de **100 mil registros** y **30 columnas** (las primeras 15 con formato numérico,las restantes tipo texto).

Conclusiones:

**CSV y formatos String** : Son simples, requieren alto costo computacional y algo lentos.

**HDF** : Gran soporte, adecuado para grandes cantidades de datos, rápido a costo de alto costo computacional.

**Parquet** : Puede igualar a hdf e inclusive trabajar por chunks y en paralelo.

**Pickle** : Es práctico pero lento con grandes cantidades de datos.

**Meteorite Landings Dataset**

Utilizaremos un datasets de meteoritos de la NASA llamado [Meteorite Landings](https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh) para estudiar **tipos de variables**, **estructuras de dataframes**, **borrar y copiar información de dataframes**.

**Tipos de Variables que componen un data frame**

[Notebook del contenido](https://github.com/francomanca93/analisis-de-datos/blob/funcionalidades-basicas/2-Funcionalidades-basicas-de-pandas/7_Tipos_de_datos.ipynb)

En esta sección se estudia cual es la **rutina de preprocesamiento de datos** con la que debemos comenzar a analizar un dataframe. Vamos a estudiar la composicion de las diferentes formatos de variables que podemos encontrar en un dataframe.

Vamos a utiilzar [Google DataSearch](https://datasetsearch.research.google.com/) para buscar datasets. Utilizaremos un dataset de meteoritos de la NASA llamado [Meteorite Landings](https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh).

**Estructuras de dataframes en detalle**

[Notebook del contenido](https://github.com/francomanca93/analisis-de-datos/blob/funcionalidades-basicas/2-Funcionalidades-basicas-de-pandas/7_Tipos_de_datos.ipynb)

En esta sección estudiamos en detalle las diferentes variables que podemos encontrar en un mismo dataframe. Estas pueden ser categoricas, de tiempo, tipo texto, numericos (float e int).

**Una ventaja de trabajar con variables categoricas es que reducimos el tamaño de uso de la memoria RAM y el tamaño del archivo en si.**

**Borrar filas, columnas y copiar información**

[Notebook del contenido](https://github.com/francomanca93/analisis-de-datos/blob/funcionalidades-basicas/2-Funcionalidades-basicas-de-pandas/7_Tipos_de_datos.ipynb)

En esta sección estudiamos como borrar los registros de un dataframe usando la función drop.

Tambien aprenderemos como copiar correctamete un dataframe para trabajar con él. En la rutina de preprocesamiento de datos es muy importante que se mantenga la fuente de datos original.

**Aplicando pandas**

**London bike sharing dataset**

Utilizaremos un dataframe descargado de [kaggle](https://www.kaggle.com/), el mismo se llama [London bike sharing dataset](https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset) para estudiar funciones matemáticas, funciones complejas y lambdas aplicados a datasets.

**Funciones matemáticas**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/8_Funciones_matem%C3%A1ticas.ipynb)

En esta sección se estudia como hacer **operaciones matematicas** en un dataframe, tambien a realizar **operaciones entre diferentes columnas** de un mismo dataframe y utilizar funciones matematicas de otra libreria como **numpy**.

Utilizamos un dataframe descargado de [kaggle](https://www.kaggle.com/), el mismo se llama [London bike sharing dataset](https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset).

**Funciones más complejas y lambdas**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/8_Funciones_matem%C3%A1ticas.ipynb)

En esat sección el objetivo es aprender a usar **funciones** mucho mas **avanzadas** (**creandolas**) en el dataframe utilizando **.apply()** y **lambda**.

En esta sección seguimos utilizando el dataset [London bike sharing dataset](https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset).

**Population dataset**

Utilizamos un dataset descargado de [world bank](https://data.worldbank.org/), buscamos [population](https://data.worldbank.org/indicator/SP.POP.TOTL), con el estudiaremos **múltiples índices**.

**Múltiples índices**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/9_Multiples_indices.ipynb)

En esta seccion aprendemos a como trabajar un dataframe con **múltiples índices**.

Un **índice** es un artificio que nos permite en pandas encontrar la ubicación de un archivo o dato. Son las coordenadas en una dataframe.

La ventajas de utilizar multiples índices es que podemos aplicar **funciones matemáticas en diversos niveles**.

Utilizamos un dataframe descargado de [world bank](https://data.worldbank.org/), buscamos [population](https://data.worldbank.org/indicator/SP.POP.TOTL).

**Datasets propios**

En las siguientes secciones se crearán datasets para el fin con el cual se estudien. Veremos **como trabajar con variables de tipo texto**, **concatenación**, **merge**, **datos faltantes**.

**Como trabajar con variables de tipo texto en Pandas**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/10_Variables_tipo_texto.ipynb)

Pandas cuenta con una gran funcionalidad a la hora de interactuar con texto, es super versatil si uno esta interesado en crear modelos de análisis de lenguaje natural.

Cuando queremos manejar texto utilizamos la función **.str** y con esta podemos concatenar otros métodos para **hacer mayúsculas o minúsculas todo el texto**, y otras funcionalidades. Podemos **contar los caracteres** del string o **separar por palabras**. Tambien podemos **buscar coincidencias** y **crear filtros**. Inclusive podemos utilizar **expresiones regulares** para realizar busquedas mas complejas y exhaustivas.

**Concatenación de DataFrames, concat y append**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/11_Concatenacion_de_DataFrames.ipynb)

En este notebook se estudia como unir dos dataframes en uno solo. Esto es muy usual cuando se hace análisis de datos y a estos solo los podemos obtener en pequeños paquetes. Utilizando la función **for y append** podemos unirlos en unos solo rapidamente.

Aprenderemos a como **concatenar vectores y matrices con numpy**, **concatenar filas y columnas (series) con pandas**, y como **concatenar dataframe con pandas**, esta última con **.concat()** y **.append()**.

**Merge de DataFrames**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/12_Merge_de_DataFrames.ipynb)

Aprenderemos a como unificar dos dataframe utilizamos parámetros en común utilizando la función **merge**. En merge veremos atributos como **inner**, **left**, **right**, **outer**.

En general cuando un **cientifico de datos** analiza muchas fuentes de información, las cuales son ideales unificarlas a través de los parámetros en común. Los conceptos relacionados con merge que se verán en **pandas** funciona exactamente igual en **SQL**

En una base de datos, las tablas están o deben estar correlacionadas con llaves primarias y secundarias.

**Lidiar con datos faltantes en DataFrames**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/13_Datos_faltantes_DataFrames.ipynb)

Es muy común que nuestros DataFrames presenten datos faltantes, por eso en esta sección veremos como procesar nuestros **datos faltantes** en DataFrames y en qué consisten los **objetos NaN** (Not a Number).

**Cómo lidiar con datos duplicados en Pandas**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/15_Datos_duplicados_DataFrames.ipynb)

Es muy usual que los registros de una base de datos aparezcan más de una vez, así que en esta sección veremos cómo pandas puede ayudarnos a lidiar con estos casos. Utilizaremos funciones como **duplicated()** y **diferentes parámetros** que esta puede tomar para trabajar con datos duplicados.

**Diamond dataset**

Vamos a utilizar una base de datos que se encuentra en una importante libreria de visualizacion de datos, **seaborn**. La misma se llama **diamonds dataset** para estudiar **groupby**.

**Group by**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/14_Group_by.ipynb)

En esta sección aprenderemos a como agrupar nuestro dataframe **bajo una categoria** y sobre esta aplicar algunas **funciones estadísticas**. Tambien aprenderemos a **crear nuestra propia función y agregarla al estudio**.

Vamos a utilizar una base de datos que se encuentra en una importante libreria de visualizacion de datos, **seaborn**. La misma se llama **diamonds**, es una base de datos que muestra diferentes caracteristicas de diamantes como quilates, tipo de corte, color, claridad, precio y dimensiones.

**Tips dataset**

El dataset sobre el que trabajaremos es sobre **tips (propinas)**, de la librería **seaborn**. Con el estudiaremos **groupby**, **refozando su uso** y **extrayendo valor de variables categoricas**, y **tablas dinámicas**.

**Group by - Refozando su uso**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/16_Group_by_Extrayendo_valor_con_variables_categ%C3%B3ricas.ipynb)

En esta sección veremos nuevamente Group by y rezorzaremos con otro dataframe el uso de esta función. Veremos como agrupar nuetro dataframe bajo una categoria y sobre esta aplicar algunas funciones estadísticas.

Lo principal de esta sección es como **extraer valor** de las **variables numericas** con las funcion **groupby** y **aggregate**, usando categorias.

El dataset sobre el que trabajaremos es sobre **propinas**, de la librería **seaborn**. En él aparecen parámetros sobre la cuenta, la propina que dieron las personas, si fue un hombre o mujer, si era fumador, tambien si como era el día (soleado, nublado, etc), el horario de la misma (desayuno, almuerzo, cena, etc) y cuantas personas compartian la comida.

**Group by - extraer valor con variables categóricas**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/16_Group_by_Extrayendo_valor_con_variables_categ%C3%B3ricas.ipynb)

En la seccion anterior se aprendio como **extraer valor** de las **variables numericas** con las funcion groupby y aggregate, usando categorias.

En esta seccion se vera como **extraer valor** de los datos de las **variables categoricas**.

Haremos **conteo** y un **analisis** con las **variables categoricas**, aplicaremos funciones **lambdas**, **transformaremos** **variables numericas en categorica**, y **extraeremos valor** de estos datos nuevos creados.

**Tablas dinámicas con Pivot Table**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/16_Group_by_Extrayendo_valor_con_variables_categ%C3%B3ricas.ipynb)

En esta sección veremos como crear **tablas dinámicas**. Para eso pandas cuenta con una función llamada **pivot\_table**, esta nos permite extraer, cuando tenemos variables categoricas, gran informacion y valor de nuestro dataframe.

**COVID-19 dataset**

Se utilizará una base de dato que tiene información sobre el **coronavirus**. La misma la podemos encontrar en el siguiente enlace a la base de datos de [covid-19](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv). Con él estudiaremos **series de tiempo**, como manejar **variables nulas** es estas series y como **visualizar y graficar datos**.

**Series de Tiempo**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/17_Series_de_tiempo_y_visualizacioes.ipynb)

En las lecturas previas se aprendio a como trabajar con **variables nulas** y a como darle formato a las **variables de tipo tiempo**.

En esta sección se vera como utilizar las herramientas anteriores y como trabajar con **series de tiempo**. Daremos **formato de tiempo** a una columna y la tomaremos como **indice de nuestro dataframe**, agruparemos un nuevo dataframe del inicial usando **groupby**, **agregaremos datos faltantes** y finalmente extraeremos **valor estadístico** de nuestro dataframe.

Se utilizará una base de dato que tiene información sobre el **coronavirus**. La misma la podemos encontrar en el siguiente enlace a la base de datos de [covid-19](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv).

**Series de Tiempo - variables nulas**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/17_Series_de_tiempo_y_visualizacioes.ipynb)

Vamos a ver en mas profundidad las **variables nulas** y como trabajar con ellas.

Tambien trabajaremos con **graficos simples**. Utilizaremos herramientas de pandas para hacer analisis de series de tiempo, **como agrupar** cuando tenemos columnas de tipo tiempo, utilizando una función llamada **Grouper** y veremos la importancia de trabajar con **intervalos de tiempo** y calcular **suavizados de curvas**, usando **rolling()**.

En general en la naturaleza haremos este tipo de analisis por ejemplo en:

* temperaturas a lo largo de un año.
* cuando realizamos proyecciones financieras.
* en nuestro caso, analisis de casos de COVID-19 a lo largo del tiempo desde que comenzo.

**Visualización y graficación de datos**

[Notebook de contenido](https://github.com/francomanca93/analisis-de-datos/blob/aplicaciones/3-Aplicando-pandas/17_Series_de_tiempo_y_visualizacioes.ipynb)

Tomaremos los datos de un **pais determinado** y **analizaremos sus casos** gracias a groupby. Luego **graficaremos** utilizando diferentes **[plot de pandas](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.html?highlight=plot)** y matplotlib.

**Proyecto - Cambio climático**

[Notebook del proyecto](https://github.com/francomanca93/analisis-de-datos/blob/proyecto/Proyecto_Cambio_clim%C3%A1tico.ipynb)

En este proyecto juntaremos diversas bases de datos para hacer un estudio del cambio climatico. Las bases de datos que utilizaremos son las siguientes:

* Base datos que habla sobre el cambio climático, [Global Land Temperature by Country](https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalLandTemperaturesByCountry.csv)
* Cantidad de emision de CO2 en kilotones, [CO2 (kt)](https://data.worldbank.org/indicator/EN.ATM.CO2E.KT).
* Cantidad de bosques deforestados en km2, [Forest (km2)](https://data.worldbank.org/indicator/AG.LND.FRST.K2).
* Cantidad de agricultura por pais en km2, [Agricultura (km2)](https://data.worldbank.org/indicator/AG.LND.AGRI.K2).
* La poblacion de cada uno de los paises, [Population (7mM)](https://data.worldbank.org/indicator/SP.POP.TOTL).
* Cantidad de energia electrica consumida en kWh per capita, [Electric power consumption (kWh per capita)](https://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC).

**Iniciando una rutina típica de manejo de datos**

En este proyecto se verá como analizar los datasets mediante una rutina practica en donde utilizaremos las funciones aprendidas.

En esta primer instancia utilizaremos el dataset [Global Land Temperature by Country](https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalLandTemperaturesByCountry.csv).

Una vez obtenidos los datos hay que hacer preprocesamiento y limpieza de datos, hay que consolidar los archivos en una base simple para analisis y extraer los insight. Para los insights utilizaremos siempre gráficas.

**Preprocesamiento de datos - terminando de preparar y limpiar los datasets**

En esta segunda instancia utiilzaremos datasets relacionados con el cambio climatico. Podemos encontrar todos los datasets en el siguiente enlace, [Topic](https://data.worldbank.org/topic/climate-change).

Los los que utilizaremos son: [CO2 (kt)](https://data.worldbank.org/indicator/EN.ATM.CO2E.KT), [forest (km2)](https://data.worldbank.org/indicator/AG.LND.FRST.K2), [agricultura (km2)](https://data.worldbank.org/indicator/AG.LND.AGRI.K2), [population (7mM)](https://data.worldbank.org/indicator/SP.POP.TOTL) y [electric power consumption (kWh per capita)](https://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC).

En esta sección se realizará el preprocesamiento de datos de los datasets mencionados, esto va desde **preparar los datasets**, **limpiarlos**, **transformarlos** y **unificarlos**. Este es el trabajo que nos lleva mas tiempo.

Normalmente este trabajo para un cientifico de datos lleva a consumir entre un **60% y 70% del tiempo de trabajo**. Luego de esto viene la parte mas interesante, el **análisis**.

**Análisis de datos**

Lo que se hizo fue preparar una serie de dataframes y condensarlos en uno solo, listo para ser analizado. En esta sección veremos diversas funciones de pandas que nos permiten **extraer insight** de una forma muy rápida.

Se estudia como **graficar diferentes variables en un unico gráfico**, como podemos ver la **correlacion entre variables** con un **heatmap** o mapa de colores, observaremos las **correlaciones con scatter plot** o graficos de puntos y finalmente es estudiarán **grafico de coordenadas paralelas**, para **comparar diferentes variables**.