Guía 1: Pandas

1- Dado el registro de notas de los alumnos de la forma (padrón, materia, nota, fecha). Se pide resolver utilizando Pandas:

A. Cuál es el promedio general de notas.
B. Cuál es la nota más alta y la nota más baja registrada durante el año 2019.
C. Cuál es el padrón con mayor cantidad de materias aprobadas durante el último cuatrimestre.
D. Cuál es la nota promedio por materia.
E. Cuál es la nota promedio por padrón.

2- Se tiene un registro de transacciones bancarias, de la forma (nro de transacción, tipo de transacción, cuenta origen, cuenta destino, fecha, hora, monto). Se pide resolver en Pandas:

A. Validar que todas las transacciones cuenten con un tipo de transacción.
B. Validar que para las transacciones del tipo transferencia, exista siempre tanto cuenta origen como cuenta destino.
C. Verificar que todas las transacciones del tipo transferencia, depósito y extracción cuenten con montos distintos de cero.
D. Indicar cuáles fueron las 10 transacciones de mayor monto.
E. Indicar cuál es el tipo de transacción que registra mayor monto promedio.
F. Indicar cuáles son las 5 cuentas con mayor cantidad de transacciones.
G. Indicar cuáles son las 5 cuentas con mayor monto involucrado.
H. Para el tipo de transacción con mayor cantidad de monto promedio, indicar cuales son las 5 cuentas con más transacciones.

3- Dado un dataframe de la forma:

Código producto	Código Fabricante	Mes	Ventas Mensuales

Generar los siguientes dataframes conteniendo:

A. Promedio por producto con el siguiente formato:

Código producto	Promedio de ventas

B. Mínimo, Máximo y Promedio de ventas por producto con el formato:

Código producto	Minimo ventas mensual	
	Maximo ventas mensual	
	Promedio ventas mensual	

C. Mínimo, Máximo y Promedio de ventas por producto con el formato:

Código producto	Mínimo ventas mensual	Máximo ventas mensual	Promedio ventas mensual

D. Mínimo, Máximo y Promedio de ventas por producto con el formato:

	Codigo Producto 1	Codigo Producto 2	Código Producto 3	. . .
Mínimo ventas				
Máximo ventas				
Promedio ventas				

E. Qué productos son provistos por los distintos fabricantes, indicando True/False con el siguiente formato:

	Código Fabricante 1	Código Fabricante 2	Código Fabricante 3	...
Código Producto 1				
Código Producto 2				
Código Producto 3				
...				

4- Un sitio de Ebooks tiene información sobre los reviews que los usuarios hacen de sus libros en un DataFrame con formato (user_id, book_id, rating, timestamp). Por otro lado tenemos información en otro DataFrame que bajamos de GoodReads: (book_id, book_name, avg_rating). Podemos suponer que los Ids de los libros son compatibles. Se pide usar Python Pandas para:

A. Obtener un DataFrame que indique el TOP5 de Ebooks en el sitio de Ebooks. (Para este punto se puede ignorar el segundo DataFrame).
B. Obtener un DataFrame que indique qué libros tienen una diferencia de rating promedio mayor al 20% entre el sitio de Ebooks y GoodReads.

5- La Agencia Nacional de Estadísticas de Buenos Aires recolecta información de nacimientos cuando los padres registran a sus hijos en el registro civil a partir de una encuesta. Esa información se encuentra disponible para su análisis en un csv con el siguiente formato (dia_nacimiento, mes_nacimiento, anio_nacimiento, peso_al_nacer, longitud_al_nacer, id_hospital, tipo_parto), donde el tipo de parto 1 es natural y 2 es cesárea.
Por otro lado la agencia cuenta con información histórica de los hospitales en otro csv con siguiente formato (id_hospital, dirección, promedio_nacimientos_mensual). Se pide usar Pandas para:

A. Calcular la cantidad de nacimientos para cada uno de los hospitales para el mes de Octubre de 2017 e indicar aquellos hospitales que superan el promedio de nacimientos mensuales.
B. Comparando el mes de Octubre de 2017 indicar programáticamente si se incremento el % de cesáreas con respecto a ese mes del año 2016.

6- El GCPD (Gotham City Police Dept) recolecta la información de casos policiales que acontecen en Ciudad Gótica. Esta información se encuentra guardada en un dataframe con el siguiente formato: (fecha, id_caso, descripcion, estado_caso, categoria, latitud, longitud). Los posibles estados que puede tener un caso son 1: caso abierto, 2: caso resuelto, 3: cerrado sin resolución. Las fechas se encuentran en el formato YYYY-MM-DD. Por otro lado el comisionado Gordon guarda un registro detallado sobre en cuáles casos fue activada la batiseñal para pedir ayuda del vigilante, Batman. Esta información se encuentra en un Dataframe con el siguiente formato (id_caso, respuesta), siendo campo respuesta si la señal tuvo una respuesta positiva (1) o negativa (0) de parte de él El sector encargado de las estadísticas oficiales del GCPD quiere con esta información analizar las siguientes situaciones:

A. Tasa de resolución de casos de la fuerza policial por categoría de caso (considerando aquellos casos en los que no participó Batman).
B. Tasa de resolución de casos con la ayuda de Batman (considerando que aquellos casos en los que fue llamado con la batiseñal, participó en la resolución).
C. Indicar el mes del año pasado en el que Batman tuvo mayor participación en la investigación de casos.

7- Dada la exitosa convocatoria de los Juegos Olímpicos de la Juventud por parte del público, sus organizadores realizan distintos análisis para planificar las jornadas finales del certamen. Es por ello que cuentan con información en los siguientes archivos csv: eventos.csv (id_evento, fecha, id_locacion, nombre_evento, id_categoria_deportiva, cantidad_espectadores) locacion.csv (id_locacion, nombre, capacidad, capacidad_extendida, sede, latitud, longitud) categorias_deportivas.csv (id_categoria_deportiva, nombre, año_de_adopcion)
El primer archivo cuenta con información de los eventos, indicando la fecha (en formato “YYYY-mm-dd hh:mm:ss”), el lugar donde ocurrió (id_locacion) y la cantidad de espectadores que asistieron. Además se aporta información sobre la categoría deportiva a la cual pertenece el evento.
Por otro lado se tiene información sobre las distintas locaciones en la sedes del certamen en las que ocurrieron los eventos. Contamos con información de su capacidad total de espectadores así como de su capacidad extendida (cuantos asientos extras se pueden brindar sobre la capacidad de la locación).
Se desea obtener:

A. Nombre de la sede que acumuló la mayor cantidad de espectadores en eventos durante el certamen del 14 al 15 de octubre inclusive. Esto es de vital importancia para distribuir el merchandising oficial del evento, para las fechas finales.

B. Nombre del evento y nombre de la categoría deportiva de aquellos eventos cuya cantidad de espectadores superó la capacidad de la locación, más allá de la capacidad extendida. Esto es de vital importancia para detectar problemas de seguridad o si es necesario realizar algún cambio de locación.

8- El dataframe (sales) lista las ventas de productos con los siguientes campos: Dia, Mes, Año, ProductID, Importe(USD). Para un mismo dia, mes y año puede venderse n veces el mismo producto. Por otro lado tenemos una descripción de los productos en el dataframe (products): ProductId, Title, Category, Description. Category puede ser "Men", "Women", "Kids"
Proponer un programa en Pandas que permita:

A. Indicar los títulos de los productos de la categoría "Men" para los cuales el Importe de venta supera el promedio mensual de ventas de los productos de la misma categoría. (por ejemplo si el promedio de Abril de "Men" es 120 dolares y un producto se vendió en Abril a 135 dolares lo tenemos que listar). Usar Transform.
B. Indicar el top-10 de productos que se vendieron mayor cantidad de días de forma consecutiva.

9- Un importante broker de compra y venta de vehiculos online se encuentra dando sus primeros pasos en la preparación de su algoritmo de pricing, es por eso que se encuentra generando algunos features iniciales para experimentar con distintos algoritmos de machine learning.
Para ello cuenta con un archivo con información de todas las transacciones que tuvo en su primer año de operación en el formato (transaction_id, timestamp, vehicle_model_id, price).
Por otro lado cuenta con información que fue extrayendo a partir de scrapping durante el último año en el formato (timestamp, source, vehicle_model_id, price). La información puede venir de múltiples fuentes (source), que pueden ser por ejemplo distintos sitios de marketplace o clasificados.
Luego de un intenso trabajo previo ha podido unificar los modelos de vehículos que utiliza para sus transacciones con la información que ha podido obtener de otros competidores mediante scrapping. Muchos de los modelos disponibles en la información de scrapping no han sido aún comercializados por la empresa, pero se sabe que se cuenta con precios scrapeados de todos los modelos que se vendieron.
Se pide generar utilizando Pandas un dataframe que tenga el siguiente formato (vehicle_model_id, ext_mean_price, ext_std_price, int_mean_price, int_std_price), siendo:
- mean_price: precio promedio para ese vehículo.
- std_price: desvío estándar del precio para ese vehículo.
y los prefijo ext_ y int_ indicando que deben ser calculado sobre respectivamente datos externos (los obtenido vía scraping) y datos internos (los de las transacciones).

10- Se tiene información diaria de la cotización de acciones en el NYSE en el archivo nyse_daily.csv en el siguiente formato (symbol, date, open, measure_midday, measure_afternoon, close, volume). Por cada acción cuyo nombre está indicado en el campo

symbol, tendremos una entrada por fecha con los valores open, measure_midday, measure_afternoon, y close indicando respectivamente a qué valor abrio la acción, cuál fue el valor que tuvo al mediodía, cual fue su valor que tuvo por la tarde y cual fue su valor al cierre del mercado. Asimismo en volume se indica el volumen operado ese día para esa acción.
Por otro lado se cuenta con el archivo s&p500.csv de formato (symbol, company_name) que indica aquellas acciones que deben ser consideradas para calcular el índice Standard & Poor's 500 (S&P 500).
Se pide calcular el valor diario del índice S&P 500, teniendo en cuenta que el mismo se calcula como el promedio del valor promedio de las mediciones que tuvo cada acción ese día (open, measure_midday, measure_afternoon, close), para las 500 acciones que se encuentran en el archivo s&p500.csv.
El resultado debe estar en un dataframe de la forma (date, index_name, value). Por ejemplo, una entrada del mismo sería ('2019-03-24', 'SP500', '35.46').
Nota: A los efectos prácticos del ejercicio consideraremos como estadísticamente significativo calcular el promedio con esas pocas mediciones.