**Prueba técnica Data Engineer**

**Mercado Libre**

**Nombre**: Juan David Espitia Aguillón

**Objetivo:**

Esta prueba consta de tres puntos de desarrollo, cada uno con fines y objetivos diferentes

con los cuales buscamos poner a prueba sus habilidades técnicas en diseño y

programación.

**Stack Tecnológico empleado para la prueba:**

1. Lenguaje de programación: Python 3.9
2. Control de Versiones: Git + GitHub + GitFlow
3. Framework: PyCharm
4. Diseños: Draw.io
5. Nube: Google Cloud Platform

**Dirección del repositorio de código:**

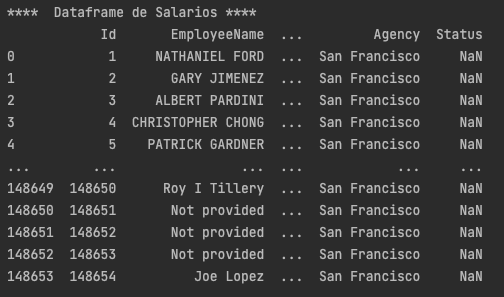
<https://github.com/JuanDavidEspitia/meli-tech-test-data-engineer>

**Punto 1:**

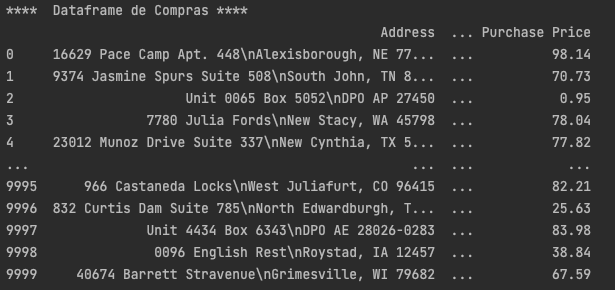
1. De acuerdo al set de datos SALARIOS ¿Cuantos cargos estaban ocupados solamente por una persona en 2011?

Inicialmente cargamos los set de datos a dataframes con la librería de Pandas

Dataframe de Salarios



Dataframe de Compras

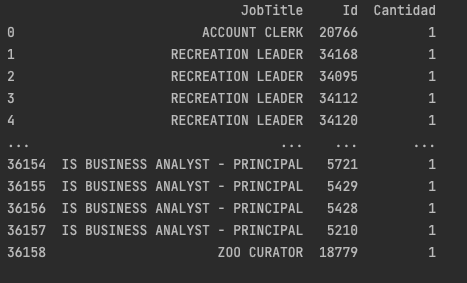


Imprimimos en pantalla la cantidad de registros por cada uno



Filtramos el set de datos de salarios para que solo estén los datos pertenecientes al año 2011.

Luego agrupamos el set de datos con los campo JobTitle e ID y obtenemos la cantidad de cargos ocupados por una única persona y posteriormente nos quedamos con aquellos cuyo conteo haya sido de uno solo.

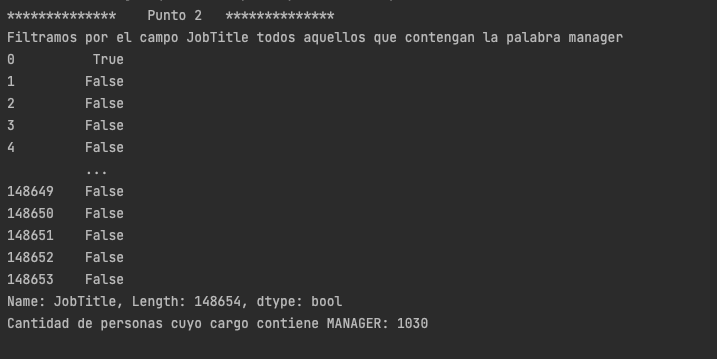


La cantidad final es de:



1. De acuerdo al set de datos SALARIOS ¿Cuánta gente tiene la palabra 'MANAGER' en su cargo?

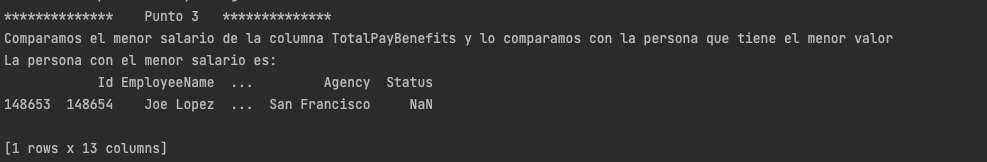
Inicialmente filtramos por el campo JobTitle todos los que contengan la palabra manager y luego los sumamos para obtener la cantidad.



La cantidad de personas con la palabra MANAGER son: 1030

1. De acuerdo al set de datos SALARIOS ¿Cual es el nombre de la persona que menos gana (incluyendo beneficios - TotalPayBenefits)?

Se debe comparar el menor salario de la columna TotalPayBenefits con la persona que obtiene ese mismo valor

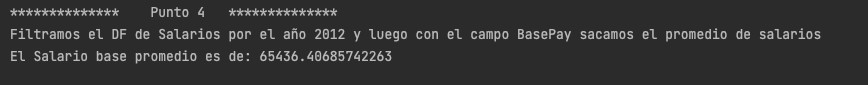


La persona con el menor salario es:

Joe Lopez ID: 148654

1. De acuerdo al set de datos SALARIOS ¿cuál es el salario base (BasePay) promedio de todos los empleados para el año(2012)?

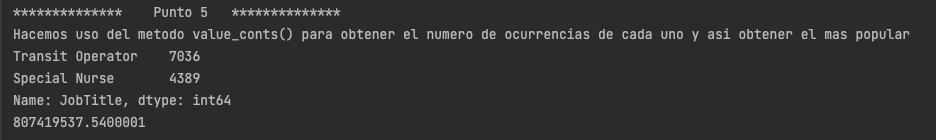
Filtramos el DF de salarios por el año 2012 y luego con el campo de BasePay (Salario Base) obtenemos el promedio.



El Promedio de salario es: 65436,407

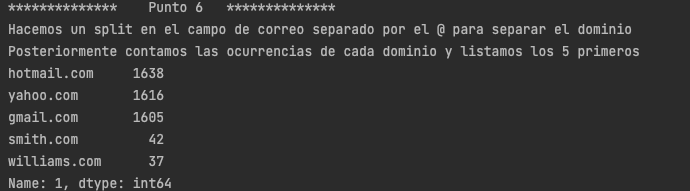
1. De acuerdo al set de datos SALARIOS ¿cuál fue la suma total pagada con beneficios por los dos trabajos más populares?.

Hacemos uso del método value\_conts() para obtener el número de ocurrencias de cada uno y así obtener el más populares y Posteriormente usamos la sentencia para sumar el total de beneficios pagados de ambos trabajos.



El total pagado por los 2 trabajos más populares es: 807419537.540001

1. De acuerdo al set de datos COMPRAS ¿Cuales son los 5 proveedores de correo electrónico más comunes, con cuantos usuarios está asociado cada uno? (hotmail.com,gmai.com, etc)

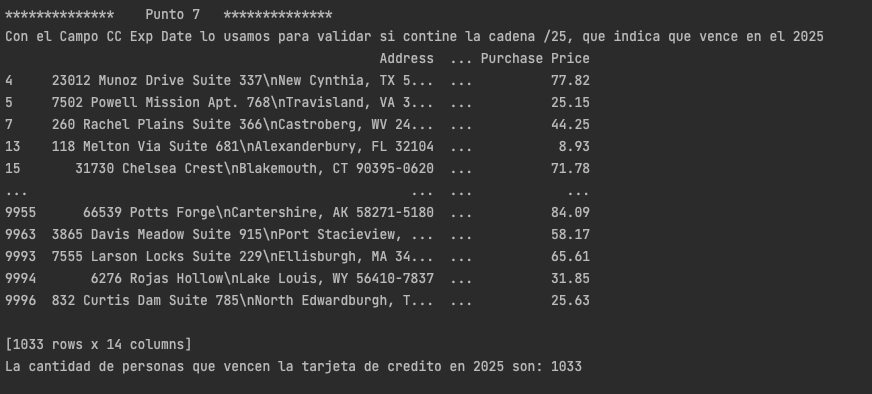


Los 5 correos más comunes son:

|  |  |
| --- | --- |
| **Proveedor** | **Ocurrencias** |
| Hotmail.com | 1638 |
| Yahoo.com | 1616 |
| Gmail.com | 1605 |
| Smith.com | 42 |
| Williams.com | 37 |

1. De acuerdo al set de datos COMPRAS ¿Cuantas personas tienen una tarjeta de crédito que expira en 2025?

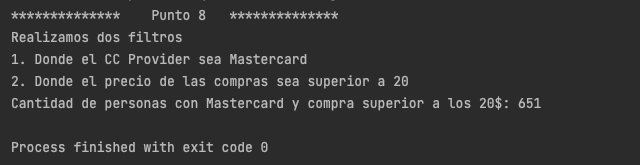
Con el Campo CC Exp Date lo usamos para validar si contine la cadena /25, que indica que vence en el 2025 y luego usamos la función len para validar la cantidad de registros reportados



La cantidad de personas que vence su tarjeta de crédito en 2025 son: 1033

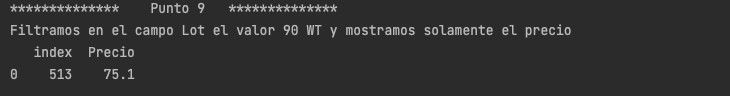
1. De acuerdo al set de datos COMPRAS ¿Cuántas personas tienen tarjetas Mastercard e hicieron una compra por más de $20?

Debemos realizar dos filtros, el primero sobre el campo CC Provider cuyo valor sea igual a Mastercard y el segundo sobre el campo Purchase Price cuyo valor sea superior a los 20$



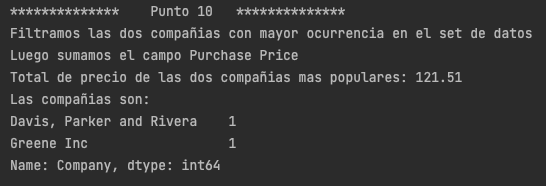
La cantidad de personas con Mastercard y compra superior a los 20$ son: 651

1. De acuerdo al set de datos COMPRAS ¿Alguien hizo una compra desde Lot: "90 WT", ¿cual fue el precio de compra de esta transacción?



El precio final es de 75.1

1. De acuerdo al set de datos COMPRAS ¿ Cuánto suma el total de precio de compras para las dos compañías menos populares?, ¿Cuáles son esas dos compañías?



El total de precio de las dos compañías más populares es de: 121.51 y las dos compañías son:

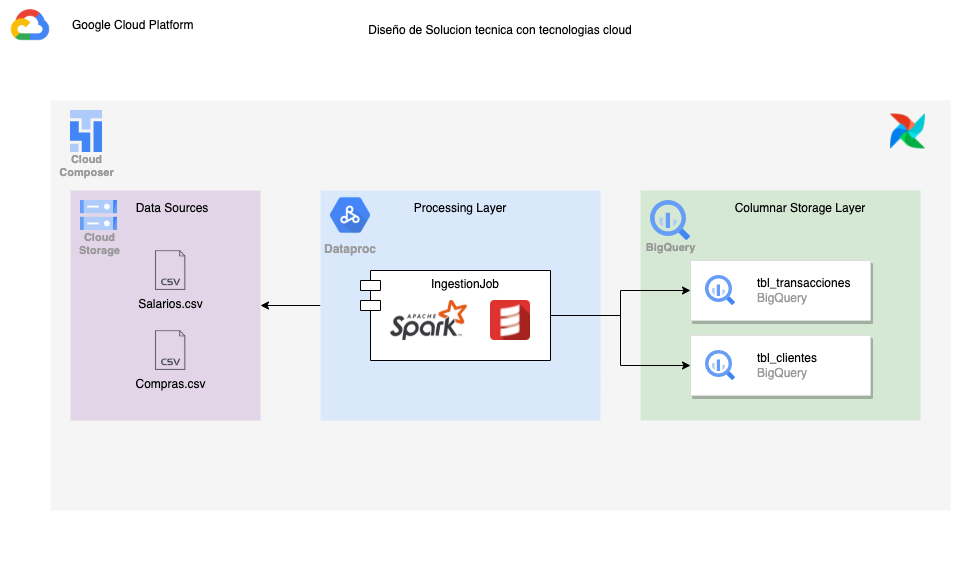
* Davis, Parker and Rivera
* Greene Inc

**Punto 2:**

**Diseño Cloud – GCP**

Diseño de la solución con tecnologías Cloud. Para este ejercicio se uso la nube de Google con sus servicios para procesar datos.

La arquitectura general de solución es la siguiente:



**Nota:**

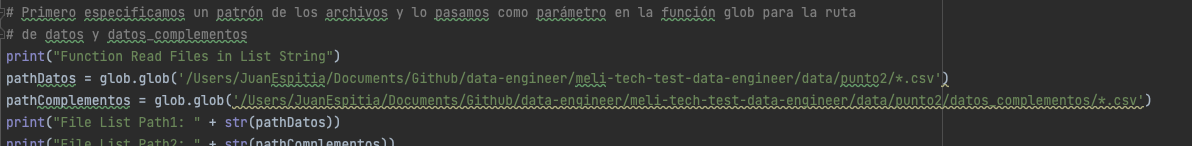
El diseño está pensado en un pipeline de datos con un alto volumen, por tal razón se necesita de servicios idóneos para la transformación de los datos.

El Componente de IngestionJob, sería un desarrollo construido en el framework de procesamiento de datos Apache Spark, bajo el lenguaje de programación Scala, donde se implementaría la lógica de lectura de los archivos, las transformación de concatenación, hash, homologación y persistencia de los datasets finales en BQ  
  
Si se piensa un diseño con set de datos de menor volumen, podemos reemplazar en la capa de procesamiento por el artefacto de Scala por una Cloud Function, desarrollada con Python y podemos utilizar el código desarrollado para la prueba en el punto 2.

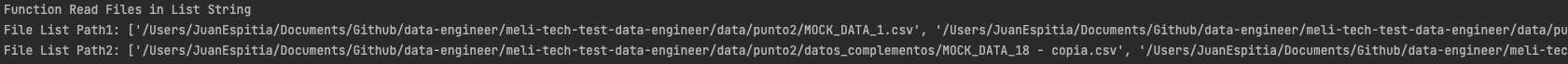
El desarrollo se realizó de la necesidad se implementó en lenguaje Python por efectos de agilidad con la entrega de la prueba técnica

**Implementación**

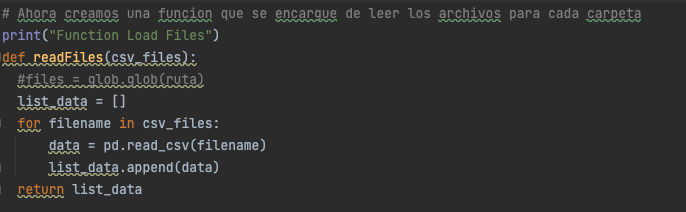
El primer paso es establecer las rutas donde se encuentran los datos a ser cargados.

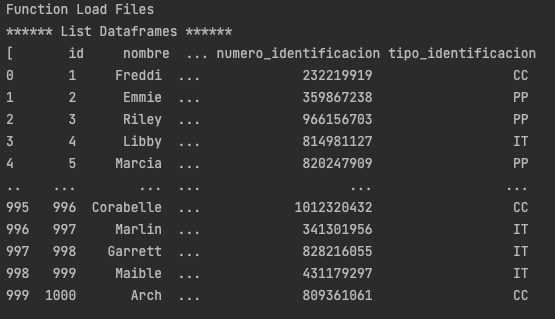


Resultado:

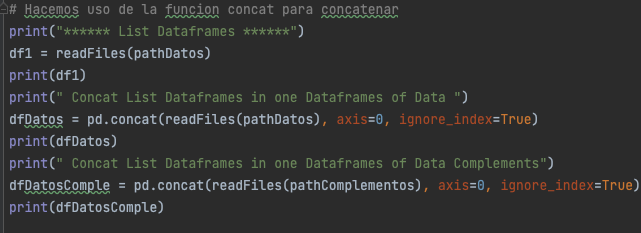


Seguido, creamos una función que itere cada una de las rutas con sus respectivos archivos .csv y vamos cargando uno a uno de los archivos .csv en una lista de dataframes y retornamos la lista de lista de dataframes.

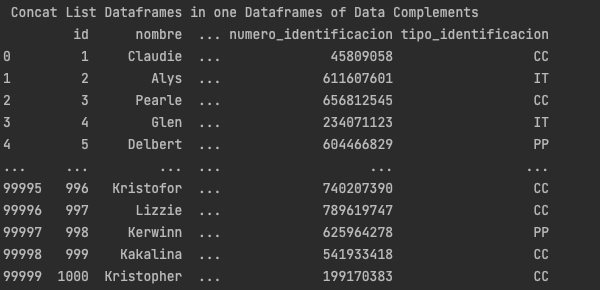




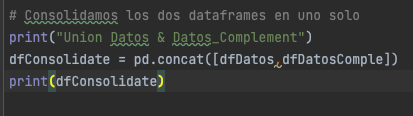
Luego hacemos uso de la función de Python concat() de dataframes para concatenar los elementos que contiene la lista que nos retorna la función anterior y lo convertimos en un dataframe para cada ruta de datos.



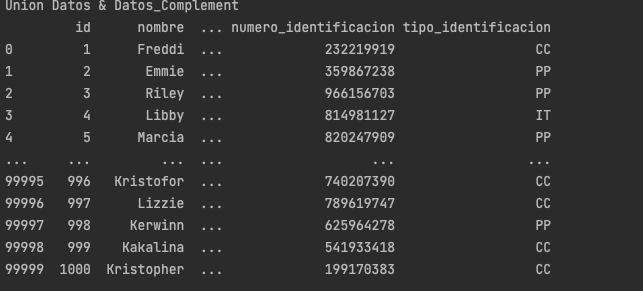
Output



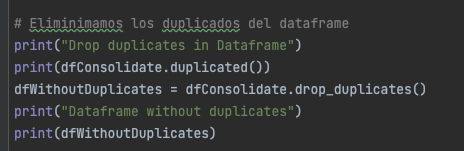
Luego consolidamos los dataframes de datos y datos\_complementarios en un solo dataframe, nuevamente con la función concat()



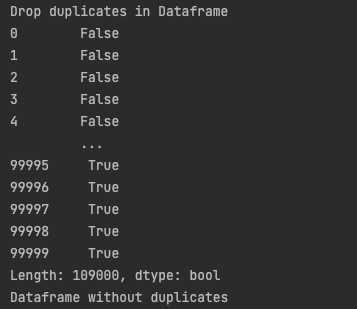
Output



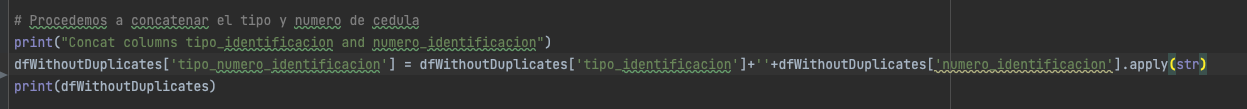
Posteriormente procedemos a eliminar los registros duplicados del dataframe consolidado mediante la función de drop\_duplicates()



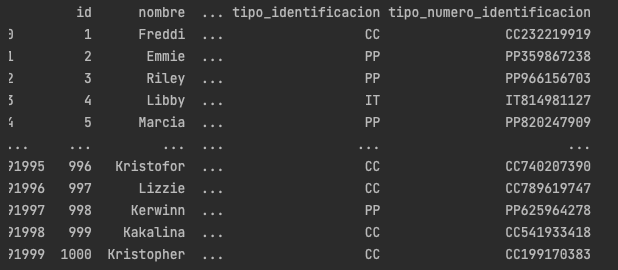
Output



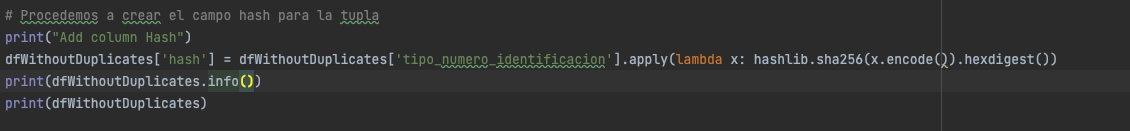
Luego procedemos a concatenar los campos tipo y numero de documento mediante las siguientes sentencias de código



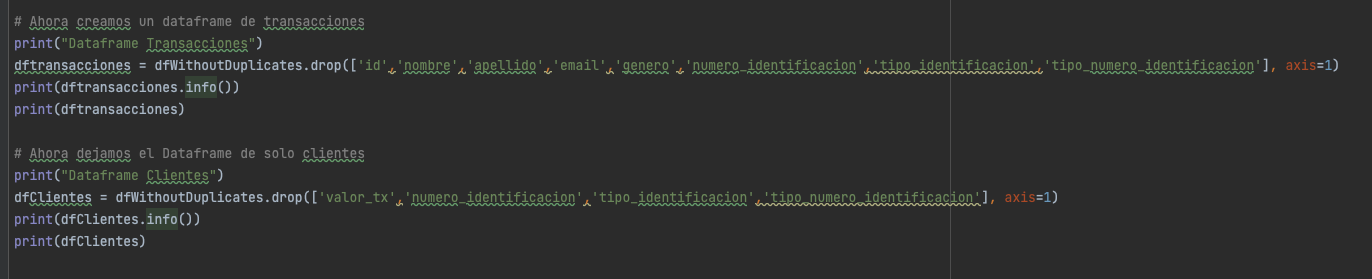
Y obtenemos la siguiente salida



Posterior a la concatenación de columnas realizamos el proceso de encriptar la columna generada de la concatenación de tipo y numero de documento, esto mediante las siguientes sentencias.



Ahora separamos los dos set de datos requeridos, el de Clientes y el de Transacciones con las especificaciones de campos indicados en el documento de la prueba técnica.



Dataframe transacciones

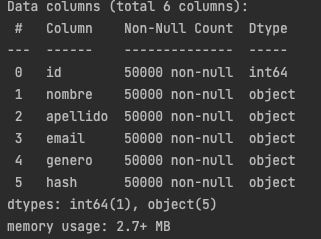


Dataframe de Clientes

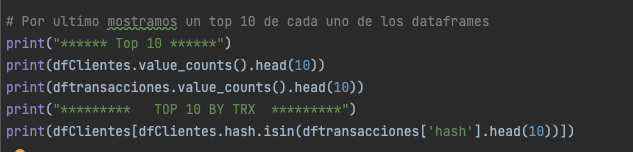
Nota: No se visualizan bien todas las columnas



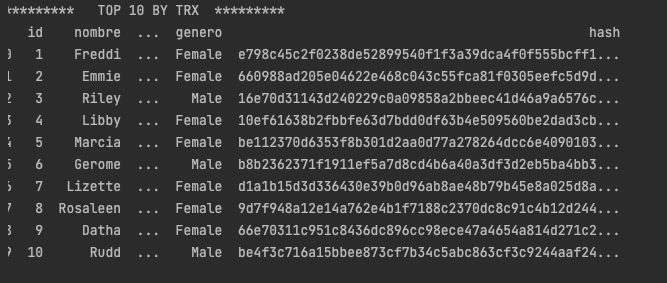
Complemento de la salida:



Por ultimo mostramos el Top 10 de los clientes con más transacciones realizadas

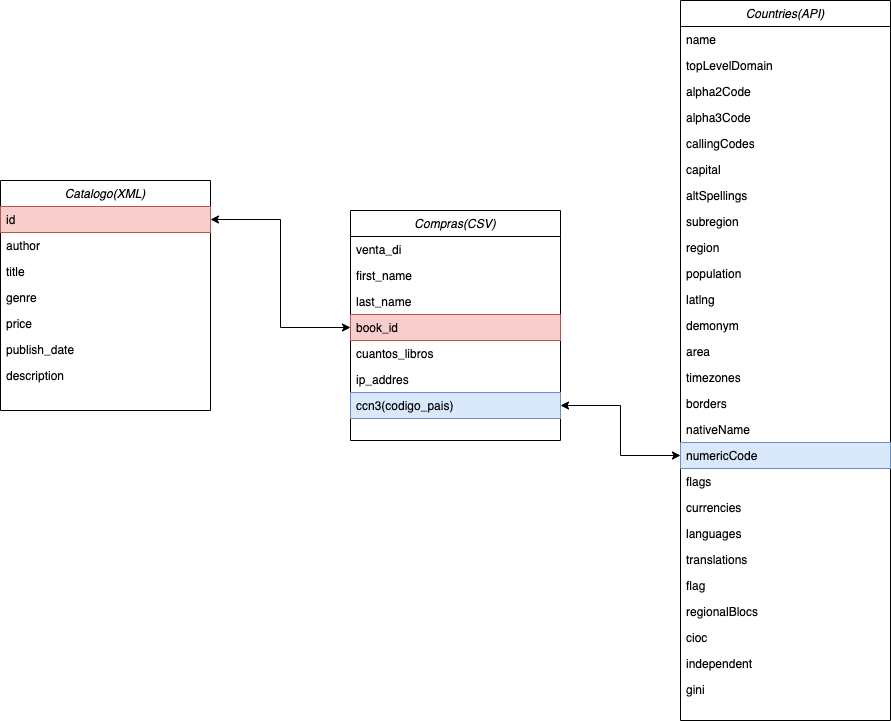


Salida por consola

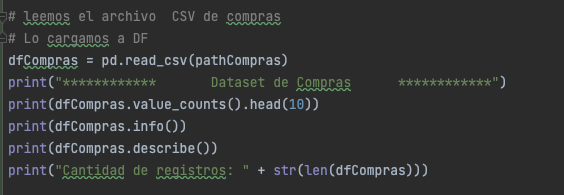


**Punto 3**

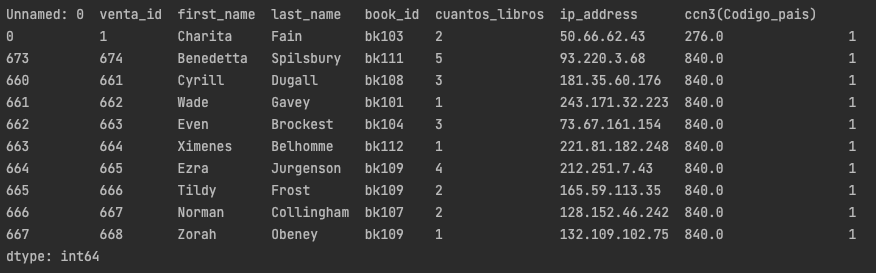
Lo primero que se debe realizar es un análisis de como están conformados los datos, comprender el dominio de cada uno de estos sets y como se pueden relacionar cada uno de ellos. Por tal razón la primera tarea que se realizo fue aterrizar los datos en un diagrama de relación de estos objetos e identificación de los campos llave para comprender como se pueden relacionar. La siguiente figura muestra la relación de cada uno de estos objetos.



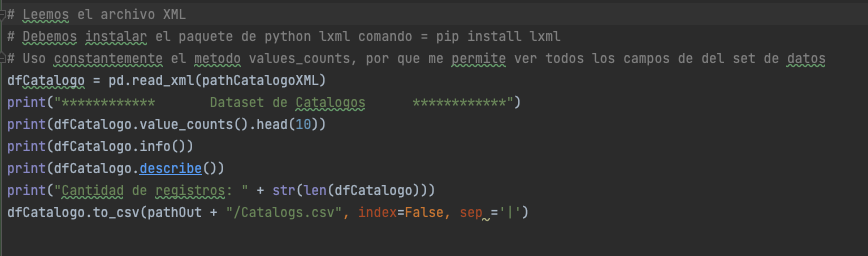
Posteriormente cargamos todos los sets de datos a sus respectivos dataframes, como lo evidencia las imágenes a continuación.



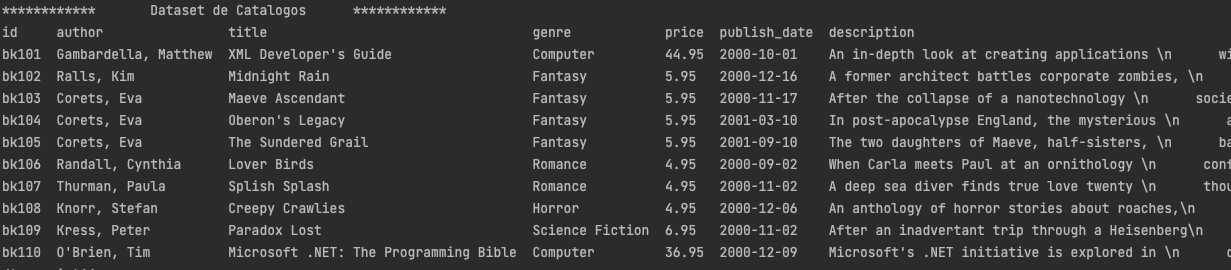
Salida:



Dataset Catálogos



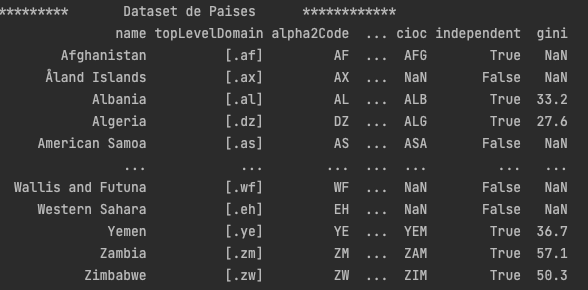
Salida:



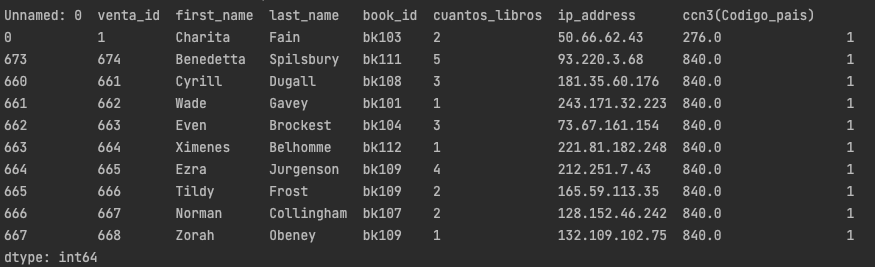
Dataset Countries (API Rest)



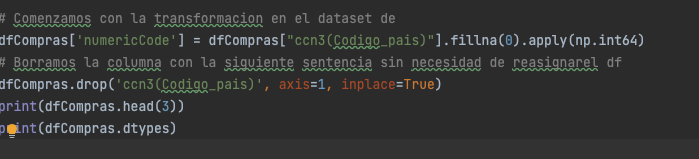
Salida:



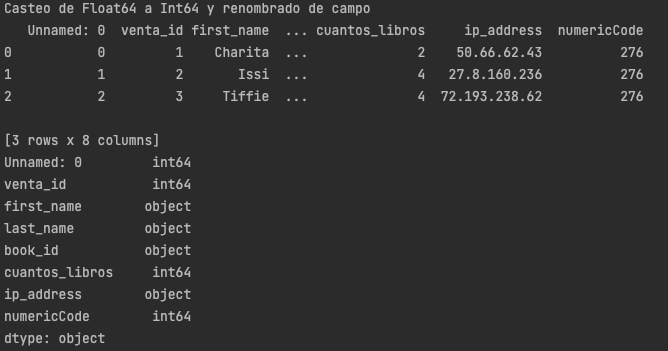
Luego entender cómo se relacionan cada uno de ellos, notamos que debemos hacer unas transformaciones de limpieza de datos en algunos campos, como es el caso del campo ccn3(Codigo del Pais) del Dataset de compras, como se puede ver en la siguiente imagen.



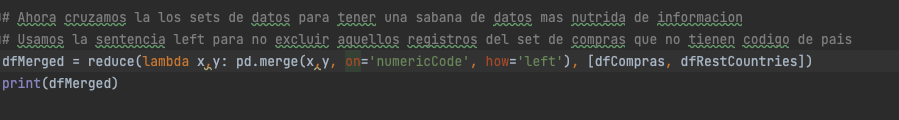
Ahora bien procedo a castear el tipo de dato de float64 a un tipo de dato Entero para realizar los cruces de tablas y en el mismo bloque de código renombro la variable para que tenga el mismo nombre de la llave del set de Countries. Esto mediante las siguientes sentencias de código.



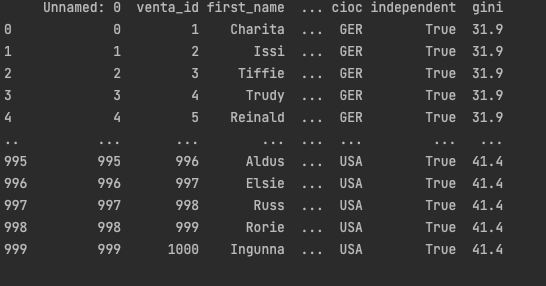
Salida:



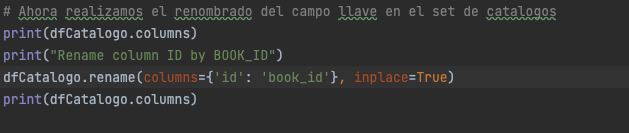
Ahora cruzamos dos sets de datos, el set de compras con el set de países para asi tener el detalle de la compra en que país se realizó. De esta forma el Set de datos nos queda más nutrido de información. Esto lo hacemos mediante la sentencia de reduce(), que es útil cuando necesita aplicar una función a un iterable y reducirlo a un solo valor acumulativo. Y la lambda con dos variables que corresponden al mismo nombre del campo llave para cruzar los dos dataframes.



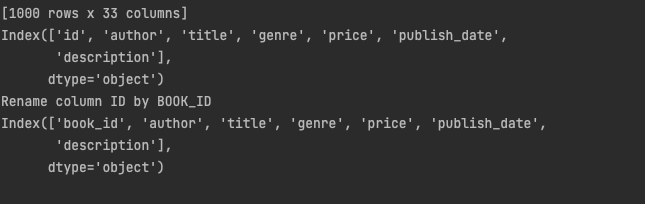
Salida:



Ahora hacemos el renombrado del campo llave del Set de Catalogos, para que sea fácilmente identificado por el mismo nombre del campo por el que aparece en el Set de Compras. Esto lo hacemos mediante la siguiente función.



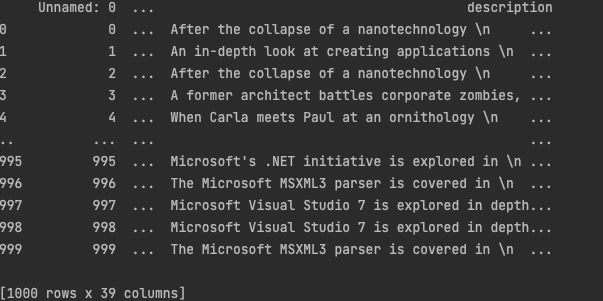
Salida:



Luego hacemos la unión del dataset resultante de merge anterior y lo relacionamos con el Dataframe de catalogos, mediante las líneas de codigo acontinuacion.



Salida



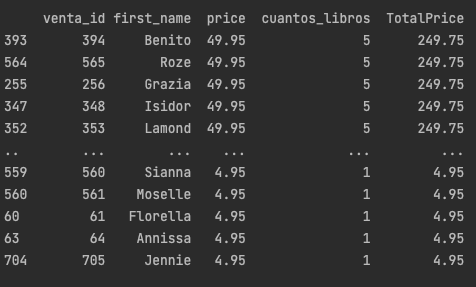
Finalmente teniendo el set de datos consolidado con las diferentes fuentes de información procedemos a establecer las preguntas de negocio o que posiblemente pueden surgir para tomar decisiones y que sirvan como insumos para las áreas analíticas.

**Consultas de negocio**

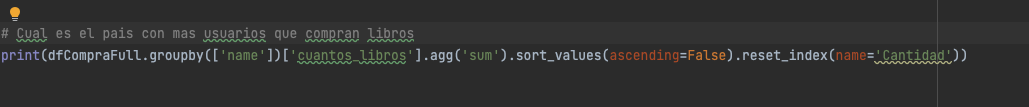
1. Cual es el nombre de los clientes con las compras más costosas y más economicas.



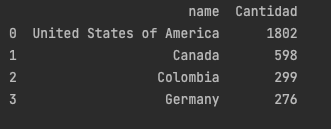
Salida:



1. Cual es el país que posee mas usuarios que compran libros?



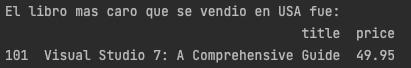
Salida:



1. El libro más costoso que se vendió en USA.

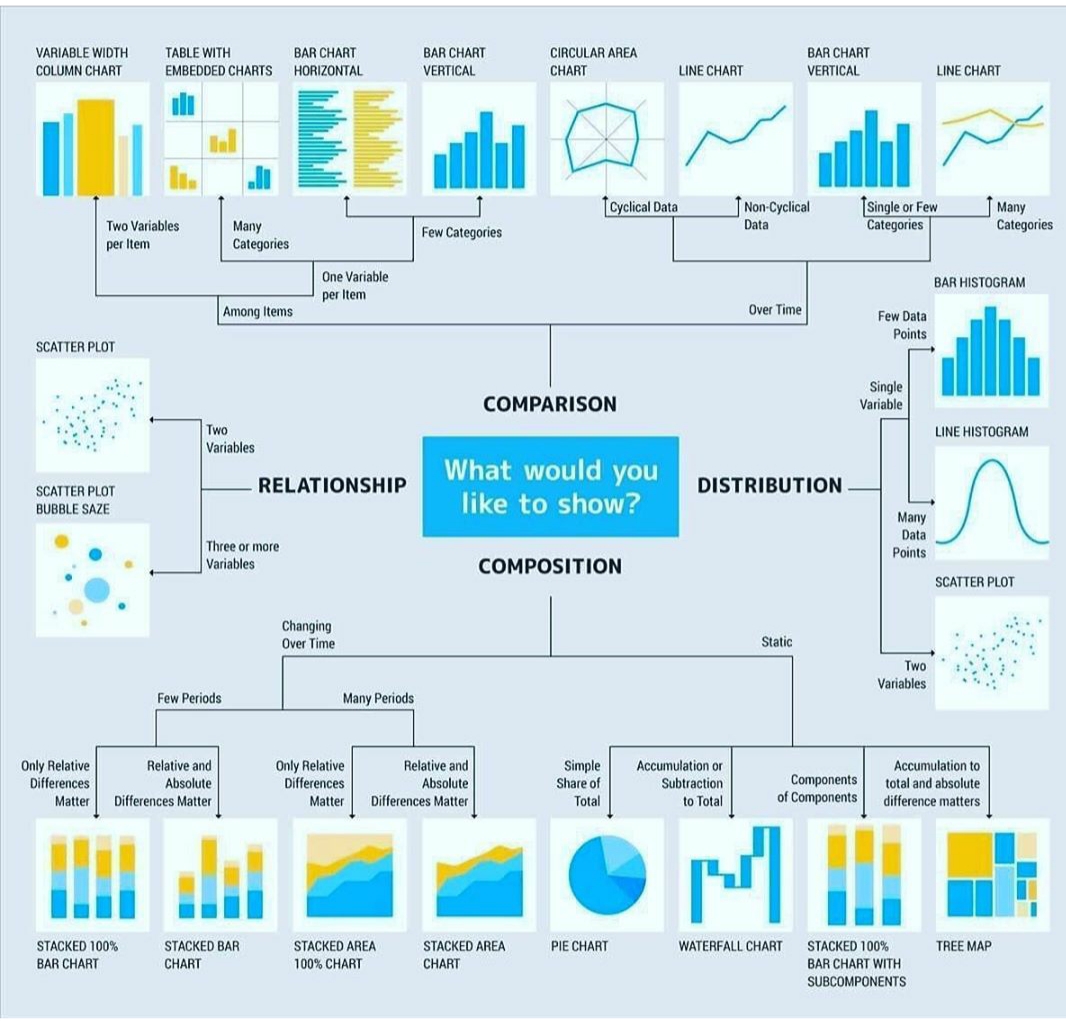


Salida:



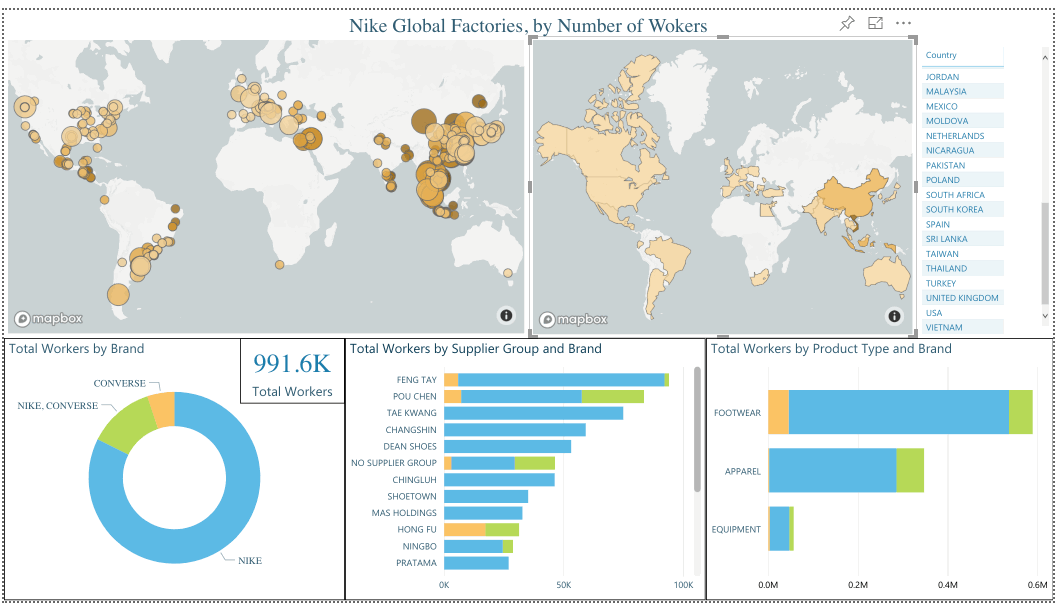
**Mock de Visualizaciones:**

Para los temas de dashboard y reportes que puedo inferir de los datos, principalmente me baso como referencia a esta imagen que me ayuda a identificar el tipo de grafico adecuado para mi necesidad de datos.



De acuerdo a este grafico determino el tipo de diagrama que mejor me ayuda visualizar mi información.

Como el set de datos hablamos de ventas en países, un gráfico bastante útil que me puede soportar la información que estoy brindando es el diagrama geográfico o un chart Maps, este me puede brindar información de cantidad de ventas de libros por países, como se puede ver en la siguiente imagen.



Otro grafico bastante interesante es un grafico de barras que me puede ayudar a informar cuales son los géneros de libros que mas les gusta a los lectores. La imagen acontinuacion me puede brindar un apoyo visual de como se puede ilustrar esta información. Imaginando en el eje x los géneros de libros y en el eje Y la cantidad de compras en ese genero.

