**Analítica de Grandes Datos**

**Departamento de Ciencias de la Computación y la Decisión**

**Facultad de Minas**

**Universidad Nacional de Colombia**

Trabajo Nociones de Arquitectura de la Información

Versión: 2021.05.19 23:27

*Observación: Cada vez que agregue nuevos elementos al documento, o que modifique algún componente del informe, revise la coherencia y consistencia con los otros elementos que hacen parte del mismo.*

**Responsables**

|  |
| --- |
| Nombre Completo – Documento de Identificación |
| 1. Juan Guillermo Ardila Giraldo - 1152203627 |
| 2. Diego Eusse Naranjo - 1037655233 |
| 3. Camilo Andres Laiton Bonadiez - 1016093055 |
| 4. Andrés Orrego Pérez - 1216725315 |
| 5. Julián David Pérez Navarro - 1037660483 |
| **REPO EN GITHUB: https://github.com/Anorpe/agd-grupal** |

*Realiza este trabajo considerando los datos que generan los sistemas transaccionales e información no estructurada de tu dominio (si trabajas por ejemplo para TCC tu dominio es la mensajería; también puedes explorar en la página* [*https://www.kaggle.com/datasets*](https://www.kaggle.com/datasets) *o* [*https://arxiv.org/*](https://arxiv.org/)*). Considera tener acceso a esta información, de al menos 10 MB (puede ser uno o varios archivos de texto), y* ***tener al menos cuatro clases conceptuales. Este documento también debe almacenarse en el REPO. Plazo Máximo de Entrega 23 de Mayo, NO SE recibirá por correo electrónico, envío por*** [***https://forms.gle/h7ty3yZykaUq5m7y6***](https://forms.gle/h7ty3yZykaUq5m7y6)

1. **Comprensión del negocio**
   1. **Descripción del contexto del negocio.**

Describa en máximo 250 palabras el contexto en el cual se generan los datos y cuál es el proceso que los genera.

Olist es una Startup brasileña dedicada al sector de comercio electrónico, cuya misión es apoyar a los pequeños y medianos empresarios a llegar a los mejores y más grandes marketplaces nacionales e internacionales. Estos comerciantes pueden vender sus productos a través de Olist Store y enviarlos directamente a los clientes mediante los socios logísticos de Olist. Después de que un cliente compra el producto en Olist Store, un vendedor recibe una notificación para cumplir con ese pedido. Una vez que el cliente recibe el producto, o vence la fecha estimada de entrega, el cliente recibe una encuesta de satisfacción por correo electrónico donde puede dar una calificación por la experiencia de compra y anotar algunos comentarios.

El flujo de información se va capturando en diferentes etapas. Inicialmente se captura la información demográfica de los vendedores y compradores vinculados. También, se registra la información completa de los productos que ofrecen dichos comerciantes. Posteriormente, al existir un comprador en línea, se registra la información de la solicitud incluyendo datos personales, de envío y formas de pago. Luego, se tiene toda la información referente al transporte y estatus del servicio. Es decir, se registra la evolución de la solicitud hasta su etapa final. Por último, se captura la información de experiencia del usuario la cual consta de la retroalimentación del servicio y la puntuación de este.

* 1. **Identificación del problema:**

Delimite en máximo 150 palabras la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios de la solución de este.

Olist, una de las Startup más grandes de Brasil, requiere recopilar una serie de indicadores que le permitan tener una estimación y control de la calidad de su servicio. Para ello, cuenta inicialmente con variables como: la retroalimentación de los clientes al finalizar la experiencia de compra, tiempos estimados contrastados con tiempos finales de entrega, clasificación de los productos según su categoría, entre otros.

Inicialmente se cuenta con una muestra de pedidos que cumplen con la condición de haber recibido una retroalimentación del servicio por parte de los clientes. Se asume que esta muestra es estadísticamente significativa para abordar el estudio de calidad deseado. Esto también puede considerarse como una restricción para el estudio, ya que es probable que algunos pedidos que no tengan retroalimentación correspondan a una buena experiencia de usuario.

Con esto, la compañía busca mejorar el servicio al cliente y la cantidad de ventas.

* 1. **Determinación de objetivos:**

Describa en máximo 150 palabras las metas a lograr al proponer una solución basada en un modelo de datos o de analítica (cómo y qué tipo de ventaja competitiva se ganará).

* Cuantificar en qué medida se están cumpliendo los plazos de entrega de los pedidos, con el fin de que la empresa tenga una medida sobre la eficiencia del área logística.



Con la anterior consulta SQL se puede obtener el tiempo promedio en días que demora un pedido en ser entregado después de ser enviado al socio logístico.

Según los datos obtenidos, los socios logísticos demoran casi 9 días en entregar los productos cuando estos presentan retrasos para lo cual la empresa deberá verificar cuales son estos pedidos y el socio logístico para realizar análisis más profundos sobre que podría estar sucediendo.

En los datos obtenidos, no tenemos acceso a los socios logísticos ni tampoco a la información personal de los clientes que podrían dar indicios de porqué existen los retrasos, por lo cual es difícil obtener resultados más precisos aparte del expuesto anteriormente.

**CONSULTA SQL:**

SELECT AVG(DAYSDIFF)

FROM (

SELECT

order\_id, order\_delivered\_carrier\_date, order\_delivered\_customer\_date,

CAST( (julianday(order\_delivered\_customer\_date) - julianday(order\_delivered\_carrier\_date)) AS INTEGER) AS DAYSDIFF

FROM olist\_orders

WHERE

order\_delivered\_carrier\_date NOT NULL AND

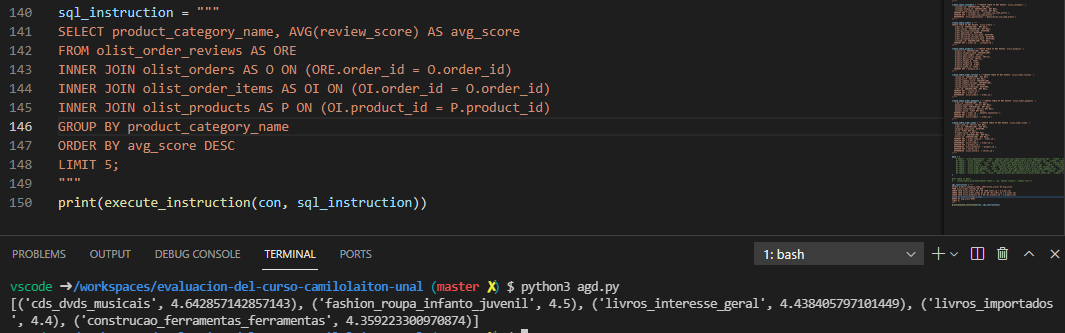
order\_delivered\_customer\_date NOT NULL AND

order\_delivered\_carrier\_date < order\_delivered\_customer\_date

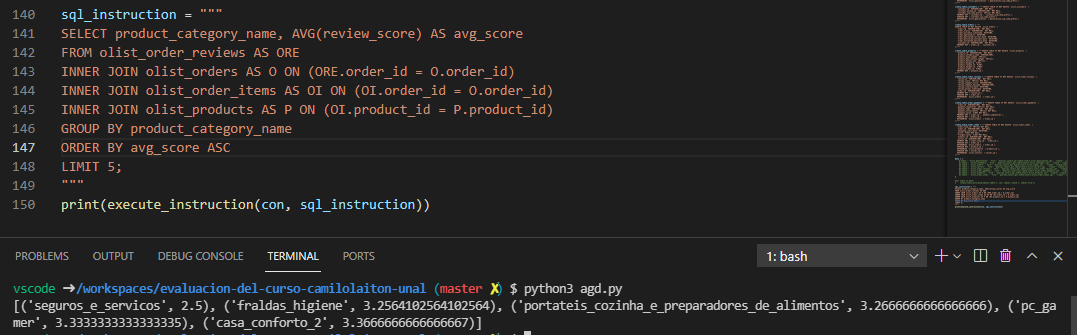
);

* Determinar el top 5 de las categorías de productos mejores y peores calificadas, con el fin de que la empresa promocione dichos productos o los intervenga, respectivamente.

El top 5 de las categorías de productos mejor calificadas permite conocer a la empresa las categorías en las cuales debería invertir en promocionar con el objetivo de obtener mejores resultados económicos. Estás categorías son:



Las 5 categorías en las que la empresa debería mejorar el marketing y sus productos son:



**CONSULTA 1 SQL:**

SELECT product\_category\_name, AVG(review\_score) AS avg\_score

FROM olist\_order\_reviews AS ORE

INNER JOIN olist\_orders AS O ON (ORE.order\_id = O.order\_id)

INNER JOIN olist\_order\_items AS OI ON (OI.order\_id = O.order\_id)

INNER JOIN olist\_products AS P ON (OI.product\_id = P.product\_id)

GROUP BY product\_category\_name

ORDER BY avg\_score DESC

LIMIT 5;

**CONSULTA 2 SQL:**

SELECT product\_category\_name, AVG(review\_score) AS avg\_score

FROM olist\_order\_reviews AS ORE

INNER JOIN olist\_orders AS O ON (ORE.order\_id = O.order\_id)

INNER JOIN olist\_order\_items AS OI ON (OI.order\_id = O.order\_id)

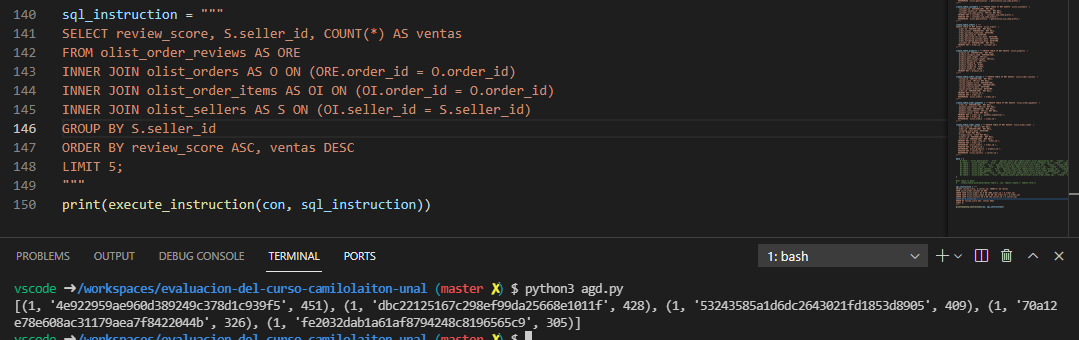
INNER JOIN olist\_products AS P ON (OI.product\_id = P.product\_id)

GROUP BY product\_category\_name

ORDER BY avg\_score ASC

LIMIT 5;

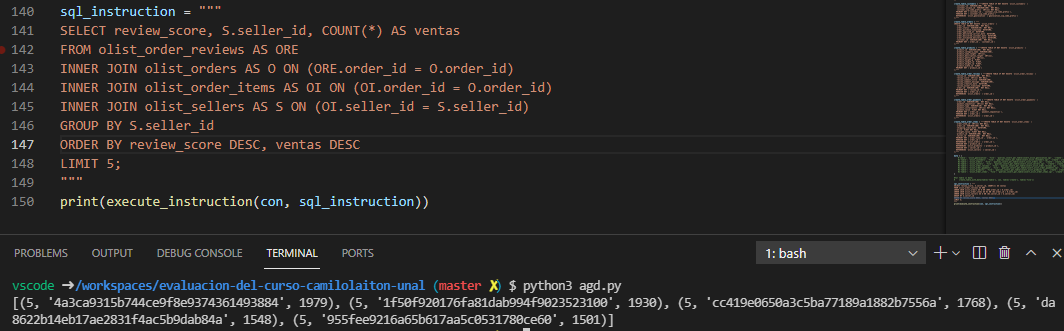
* Generar una clasificación de los vendedores basada en las calificaciones obtenidas y la cantidad de ventas, con el fin de que la empresa implemente estrategias que incentiven a sus vendedores.



En la consulta anterior se tienen los peores vendedores con la mayor cantidad de ventas, lo cual indica que estos vendedores deben mejorar la forma en como tratan a los clientes al momento de realizar la transacción.

También es posible encontrar los peores vendedores con la menor cantidad de ventas simplemente cambiando el ordenamiento de ventas a ascendente. Sin embargo, se debe también tener en cuenta el tiempo que lleva asociado el vendedor con la empresa porque puede que apenas haya comenzado a trabajar. Con esto la empresa también tendría información para tomar medidas correctivas (despido ó planes de mejoramiento) con los vendedores.

Por otra parte, es importante premiar a los vendedores que tengan un excelente desempeño laboral para así lograr motivar al resto de la empresa y esto se puede conseguir con la siguiente consulta SQL:



**CONSULTA 1 SQL:**

SELECT review\_score, S.seller\_id, COUNT(\*) AS ventas

FROM olist\_order\_reviews AS ORE

INNER JOIN olist\_orders AS O ON (ORE.order\_id = O.order\_id)

INNER JOIN olist\_order\_items AS OI ON (OI.order\_id = O.order\_id)

INNER JOIN olist\_sellers AS S ON (OI.seller\_id = S.seller\_id)

GROUP BY S.seller\_id

ORDER BY review\_score ASC, ventas DESC

LIMIT 5;

**CONSULTA 2 SQL:**

SELECT review\_score, S.seller\_id, COUNT(\*) AS ventas

FROM olist\_order\_reviews AS ORE

INNER JOIN olist\_orders AS O ON (ORE.order\_id = O.order\_id)

INNER JOIN olist\_order\_items AS OI ON (OI.order\_id = O.order\_id)

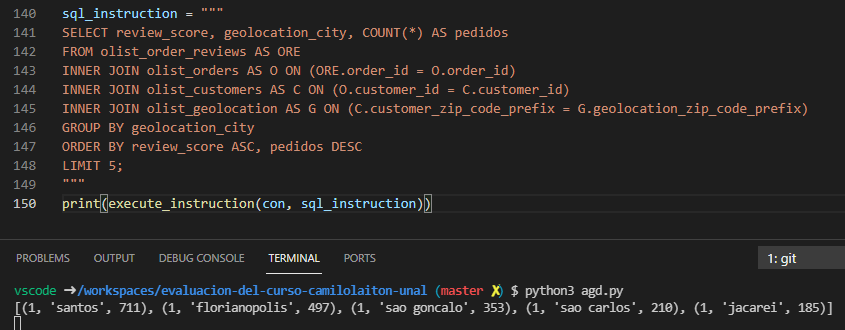
INNER JOIN olist\_sellers AS S ON (OI.seller\_id = S.seller\_id)

GROUP BY S.seller\_id

ORDER BY review\_score DESC, ventas DESC

LIMIT 5;

* Encontrar las ciudades con índices de calificación más bajos, con gran volumen de pedidos, con el fin de que la empresa implemente acciones para mejorar la calidad del servicio.



A partir de la respuesta obtenida a través de la consulta SQL, la empresa Olist debe tomar medidas para mejorar la forma en cómo atienden a sus clientes en estas ciudades dado que tienen los índices de calificación más bajos con una gran cantidad de pedidos.

Si la empresa Olist no toma medidas necesarias para mejorar la atención al cliente en estas ciudades es seguro que perderán sus clientes en estas locaciones lo que se traduce en menos dinero.

**Consulta SQL:**

SELECT review\_score, geolocation\_city, COUNT(\*) AS pedidos

FROM olist\_order\_reviews AS ORE

INNER JOIN olist\_orders AS O ON (ORE.order\_id = O.order\_id)

INNER JOIN olist\_customers AS C ON (O.customer\_id = C.customer\_id)

INNER JOIN olist\_geolocation AS G ON (C.customer\_zip\_code\_prefix = G.geolocation\_zip\_code\_prefix)

GROUP BY geolocation\_city

ORDER BY review\_score ASC, pedidos DESC

LIMIT 5;

* 1. **Evaluación de la situación actual:**

Describa en máximo 150 palabras el estado actual antes de implementar la solución de analítica, a fin de tener un punto de comparación que permita medir el grado de éxito de la solución.

Inicialmente se cuenta con una base de datos organizada por medio de 8 tablas con diferentes variables en cada una de ellas. Estas 8 tablas corresponden a distintos datos sobre el proceso de compra mediante la aplicación de Olist, tales como:

* Información del comprador (olist\_customers)
* Información del vendedor (olist\_sellers)
* Información del pedido realizado (olist\_orders)
* Información de los ítems incluídos en cada pedido (olist\_order\_items)
* Información de los productos en venta(olist\_products)
* Información de los pagos (olist\_order\_payments)
* Información de las calificaciones dadas por los compradores (olist\_order\_reviews)
* Información sobre la dirección o ubicación de compradores y vendedores (olist\_geolocation)

En algunas de estas tablas es posible identificar variables que se repiten y que dan indicio de generar una conexión entre ellas. Mediante este trabajo se busca organizar esta base de datos de la forma más adecuada posible para llevar a cabo consultas que permitan dar respuesta a los objetivos planteados en el numeral anterior.

1. **Comprensión de los datos**
   1. **recolección de datos**

Describa en máximo 150 palabras los datos a utilizar identificando las fuentes, las técnicas empleadas en su recolección, los problemas encontrados en su obtención y la forma como se resolvieron los mismos. Además, adjunte los datos (archivos de texto, etc.) agréguelos en el github **(REPO EN GITHUB)** en un solo archivo, por favor comprímalo(s). Llame el archivo **T1.2.1.Datos.zip**

Las fuentes de información disponibles para el estudio corresponden a las bases de datos que Olist genera desde su plataforma de servicio. En este caso, se cuenta con una muestra aleatoria anonimizada de ocho bases de datos. Estas bases comprenden datos de cerca de 100.000 pedidos realizados en múltiples mercados de Brasil entre 2016 y 2018 y que cumplen la condición de tener una calificación de servicio, brindada por el cliente. Esta información permite ver un pedido desde múltiples dimensiones: desde el estado del pedido, el precio, el pago y el costo de envío hasta la ubicación del cliente, los atributos del producto y finalmente las reseñas escritas por los clientes. Adicionalmente, se cuenta con un conjunto de datos de geolocalización que relaciona los códigos postales brasileños con las coordenadas latitud / longitud. El sistema de recolección o captura de información es automático y en tiempo real.

* 1. **Descripción de datos (diccionario):**

Diligencia la siguiente tabla, puede agregar otra columna si lo considera necesario.

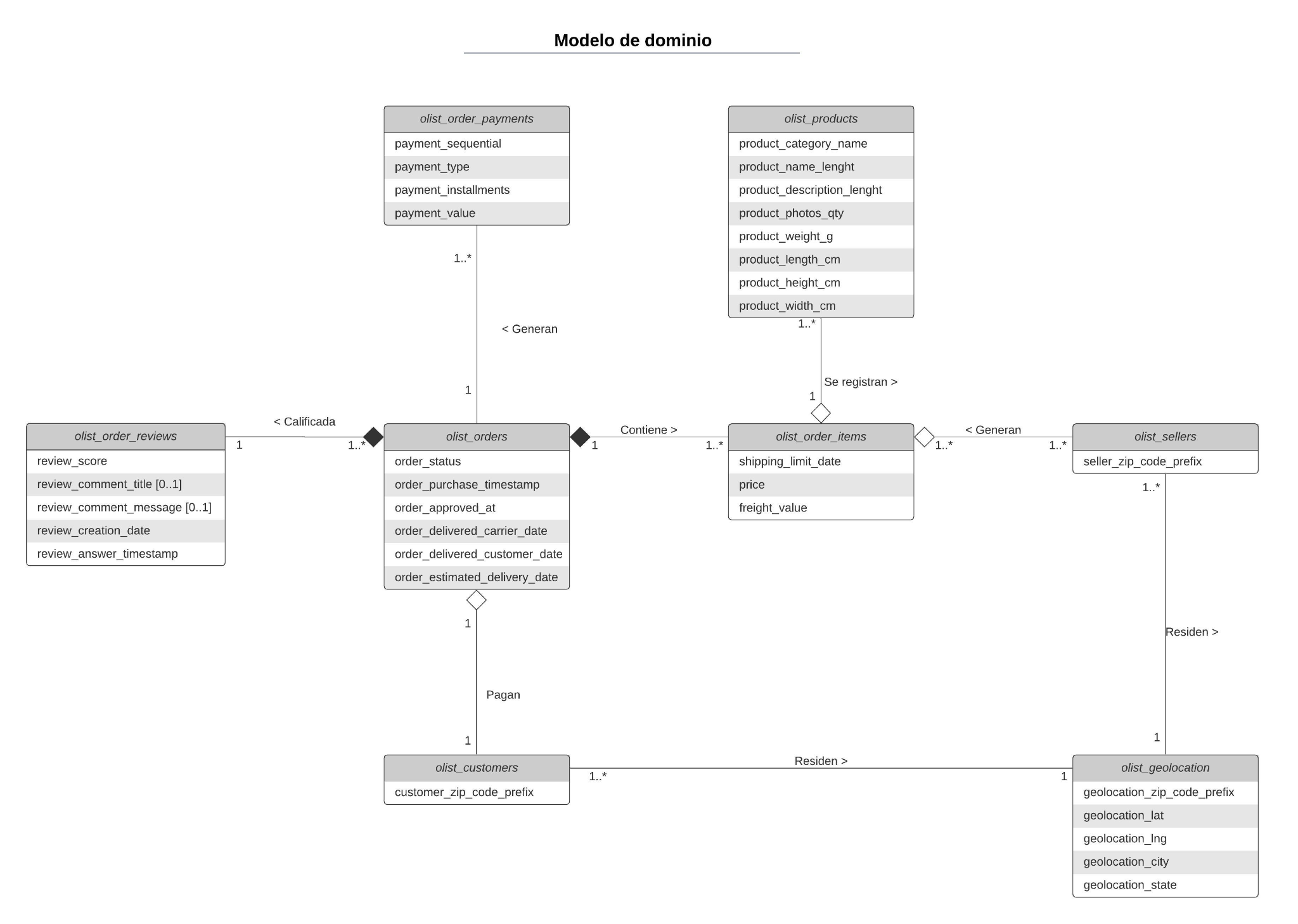
|  |  |  |  |
| --- | --- | --- | --- |
| **Entidad** | **Nombre del atributo / variable** | **Formato o Tipo de Dato** | **Descripción** |
| olist\_customers | customer\_id | NVarchar(100) | Clave principal para el conjunto ‘customer’ . Cada cliente tiene un único customer\_id. |
| customer\_unique\_id | NVarchar(100) | Identificador único de un cliente. |
| customer\_zip\_code\_prefix | int | Código postal del cliente. Clave foránea con la tabla olist\_geolocation. |
|  | | | |
| olist\_geolocation | geolocation\_zip\_code\_prefix | int | Primeros dígitos del código postal dada una localización específica. |
| geolocation\_lat | float | Latitud. |
| geolocatio\_lng | float | Longitud. |
| geolocation\_city | NVarchar(100) | Nombre de Ciudad. |
| geolocation\_state | CHAR(2) | Nombre del Estado. |
|  | | | |
| olist\_sellers | seller\_id | NVarchar(100) | Clave principal para el conjunto ‘sellers’. Cada vendedor tiene un único seller\_id. |
| seller\_zip\_code\_prefix | int | Primeros 5 dígitos del código postal del vendedor. |
|  | | | |
| olist\_orders\_reviews | review\_id | NVarchar(100) | Clave principal para el conjunto ‘review’ . Cada calificación tiene un único review\_id. |
| order\_id | NVarchar(100) | Identificación única de para el conjunto ‘orders’ . Clave foránea con la tabla olist\_orders. |
| review\_score | int | Calificación otorgada por el comprador para la orden de compra. |
| review\_comment\_title | NVarchar(100) | Título asociado a la calificación otorgada a la orden de compra. |
| review\_comment\_message | NVarchar(100) | Comentario asociado a la calificación otorgada a la orden de compra. |
| review\_creation\_date | DATETIME | Fecha y hora de creación de la calificación de la orden de compra. |
| review\_answer\_timestamp | DATETIME | Fecha y hora de respuesta a la calificación de la orden de compra. |
|  | | | |
| olist\_orders | order\_id | NVarchar(100) | Identificador único que representa a una orden. |
| customer\_id | NVarchar(100) | Identificador que se dirige a la tabla de órdenes en donde una orden tiene una clave única. |
| order\_status | NVarchar(50) | Cadena de texto que indica el estado de la orden. Ej: si este fue entregado, enviado, en camino, etc… |
| order\_purchase\_timestamp | DATETIME | Fecha que indica el momento de la compra. |
| order\_approved\_at | DATETIME | Fecha que indica el momento de aprobación del pago por parte del cliente. |
| order\_delivered\_carrier\_date | DATETIME | Fecha que indica el momento en que el producto fue entregado al socio logístico. |
| order\_delivered\_customer\_date | DATETIME | Fecha de entrega del pedido al cliente. |
| order\_estimated\_delivery\_date | DATETIME | Fecha estimada de entrega que se le comunicó al cliente al momento de la compra. |
|  | | | |
| olist\_order\_items | order\_id | NVarchar(100) | Identificador único que representa a una orden. |
| order\_item\_id | NVarchar(100) | Número secuencial que indica la cantidad de ítems incluidos en una misma orden. |
| product\_id | NVarchar(100) | Identificador único de un producto. |
| seller\_id | NVarchar(100) | Identificador único de un vendedor. |
| shipping\_limit\_date | DATETIME | Fecha límite del envío por parte del vendedor con el objetivo de que el socio logístico maneje la orden. |
| price | Float | Precio del ítem en moneda brasileña. |
| freight\_value | Float | Precio de transporte del producto en moneda brasileña. |
|  | | | |
| olist\_order\_payments | order\_id | NVarchar(100) | Identificador de la orden asociada al pago. |
| payment\_sequential | int | Valor entero que hace referencia a la secuencia de pagos que hace el cliente dado que un cliente puede pagar de diversas maneras (ej: Efectivo, tarjeta, etc). |
| payment\_type | NVarchar(30) | Método de pago escogido por el cliente. |
| payment\_installments | int | Número de cuotas elegido por el cliente para pagar la orden. |
| payment\_value | float | Valor de la transacción de la orden. |
|  | | | |
| olist\_products | product\_id | NVarchar(100) | Identificador único del producto. |
| product\_category\_name | NVarchar(100) | Nombre de la categoría a la cual pertenece el producto. Está se encuentra en portuges. |
| product\_name\_lenght | int | Número de caracteres que tiene el nombre del producto. |
| product\_description\_lenght | int | Número de caracteres que tiene la descripción del producto. |
| product\_photos\_qty | int | Número de fotos publicadas del producto. |
| product\_weight\_g | float | Peso del producto en gramos. |
| product\_length\_cm | float | Longitud del producto en centímetros. |
| product\_height\_cm | float | Alto del producto en centímetros. |
| product\_width\_cm | float | Ancho del producto en centímetros. |

* 1. **Modelo del dominio**

***Observación****: Incluya el gráfico del modelo del dominio que representa la estructura de datos de su problema.*

A continuación se adjunta la imagen del modelo de dominio realizado para la base de datos trabajada. Para una mejor visualización, se adjuntará este gráfico en formato PDF en el repositorio del proyecto. También es posible descargarlo mediante la siguiente URL:

https://lucid.app/lucidchart/invitations/accept/inv\_9457e01f-c66a-4013-96e3-bbe75cbf8778

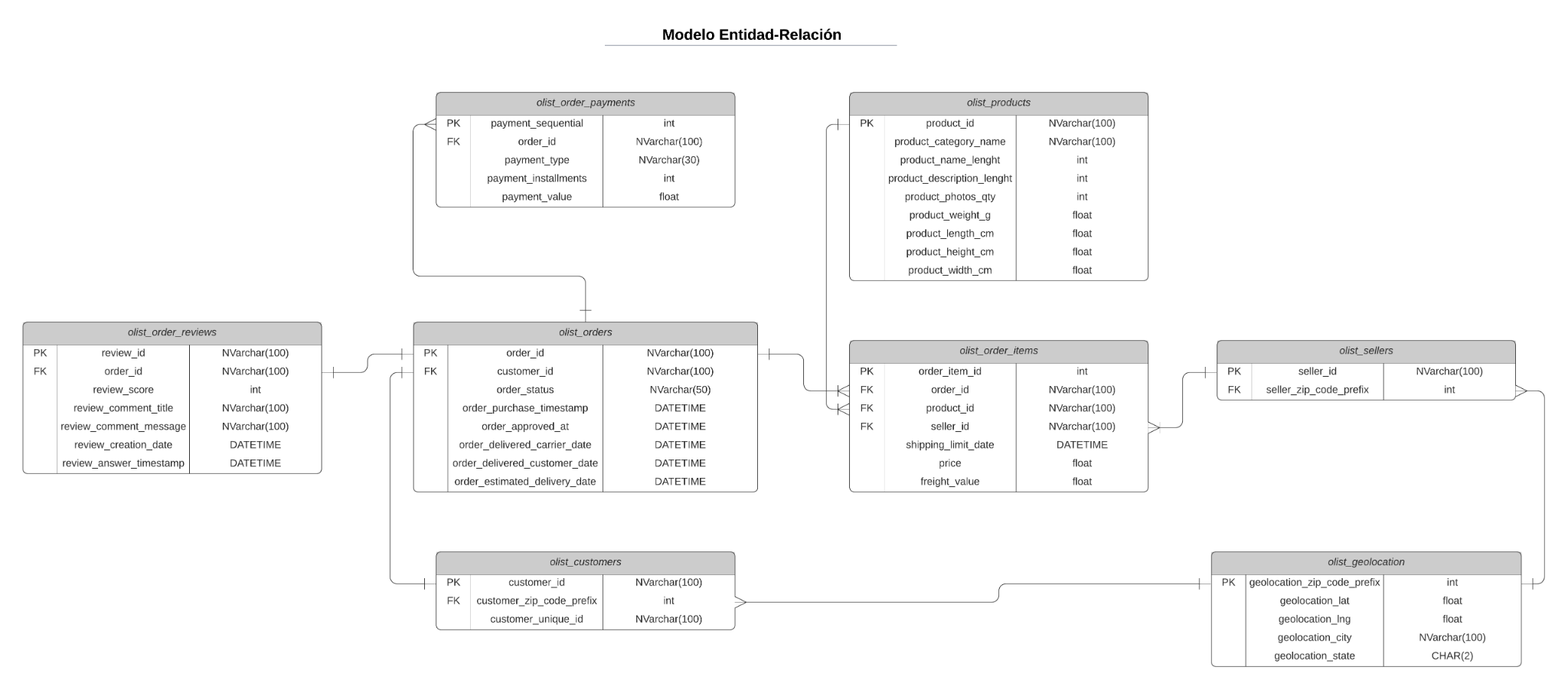
**

1. **Modelo Entidad-Relación**
   1. **Toma de pantalla del modelo E-R**

***Observación****: lo que se pide, puede usar https://draw.io o Microsoft Visio® y modele usando la notación de Barker.*

A continuación se adjunta la imagen del modelo Entidad Relación E-R realizado para la base de datos trabajada. Para una mejor visualización, se adjuntará este gráfico en formato PDF en el repositorio del proyecto. También es posible descargarlo mediante la siguiente URL:

https://lucid.app/lucidchart/invitations/accept/inv\_4605d247-dde9-442f-bd88-45fca6116efd

****

* 1. **Sentencia o consulta de creación del tabla(s)**

***Observación****: Escriba el código en el Sistema de Gestión de Bases de Datos Relacionales de su elección (se recomienda SQLite por simplicidad, mediante https://sqlitebrowser.org/) para crear las tablas que corresponda con su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.2.Creacion\_Tablas.sql*

***Se recomienda repasar SQL en https://www.w3schools.com/sql/default.asp***

* 1. **Sentencias para Insertar datos**

***Observación****: Escriba el código para insertar los datos en cada una de las tablas creadas. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.3.Insertar\_Datos.sql*

**NOTA:** El archivo T1.3.3.Insertar\_Datos.sql se encuentra en el archivo comprimido **“T1.3.3.Insertar\_Datos.rar”** dado que Github tiene problemas para almacenar archivos muy pesados.

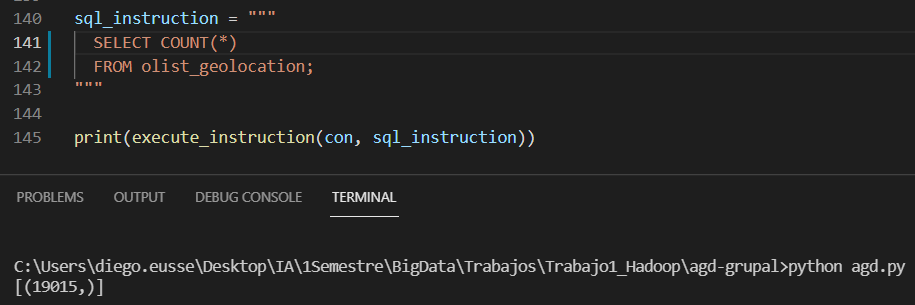
* 1. **Sentencia de consulta**

***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.4.Consultar\_Datos.sql*

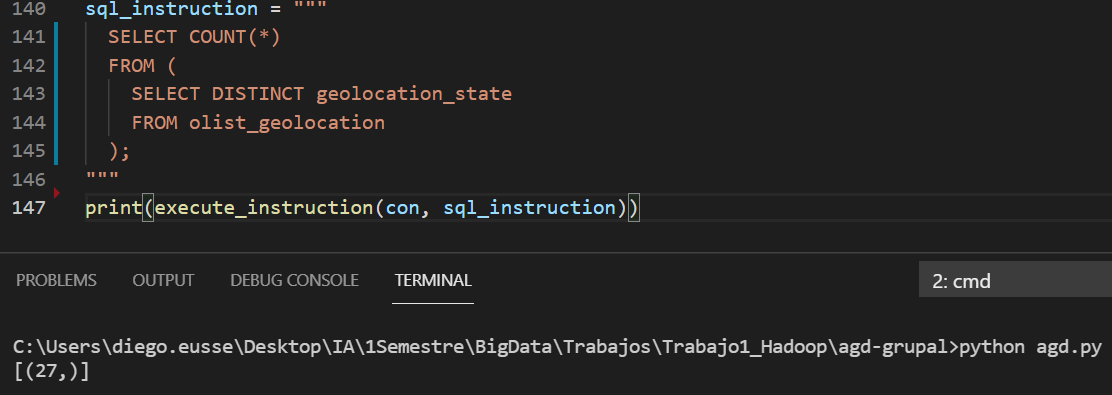
Para realizar las consultas se decide trabajar las tablas una por una:

**Tabla 1. olist\_geolocation**

* **Conteo cantidad total de registros**

****

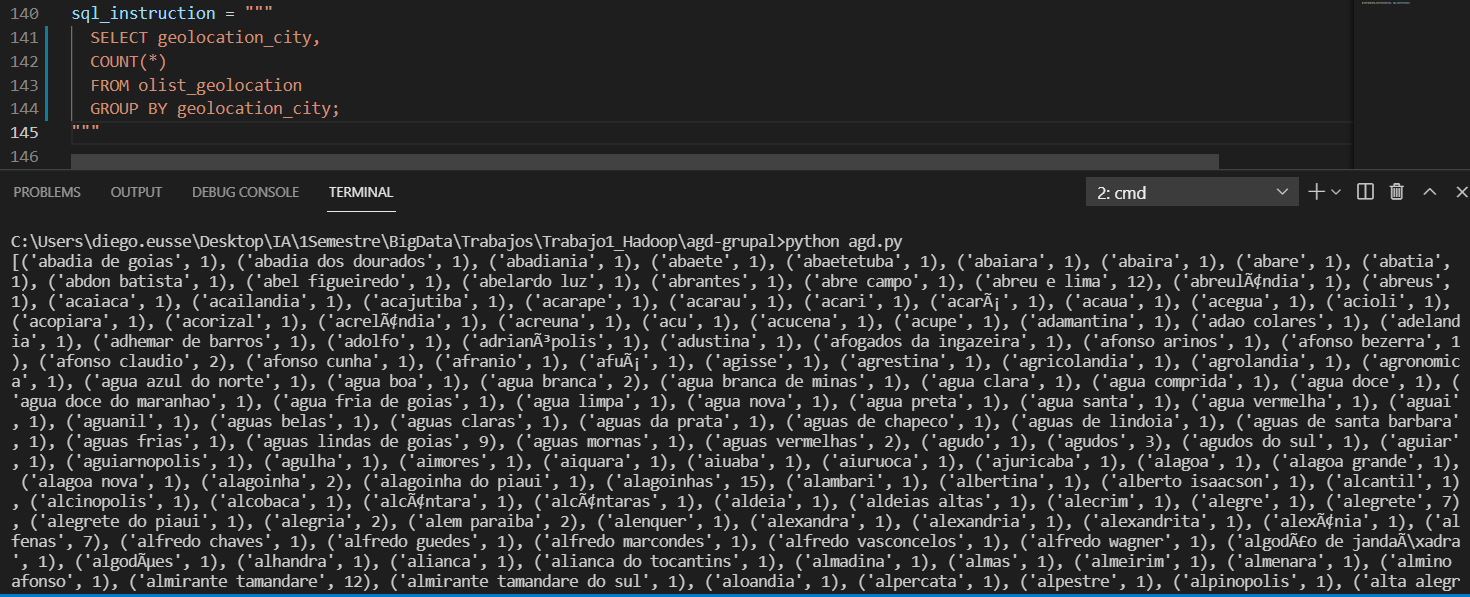
* **Conteo total de estados diferentes**

****

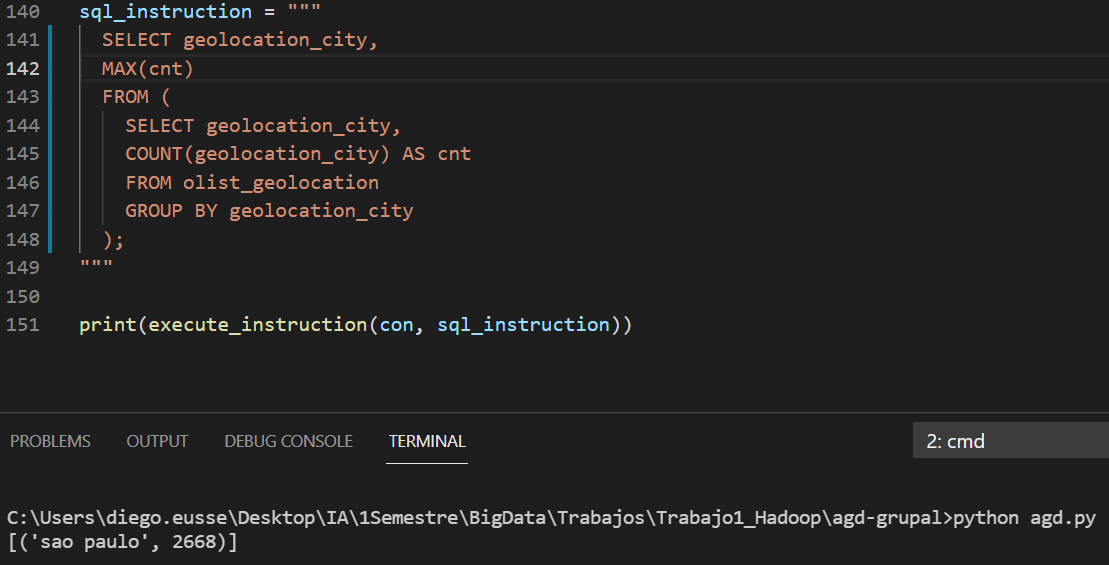
* **Conteo cantidad de registros por estado**

****

* **Conteo cantidad de registros por ciudad**

****

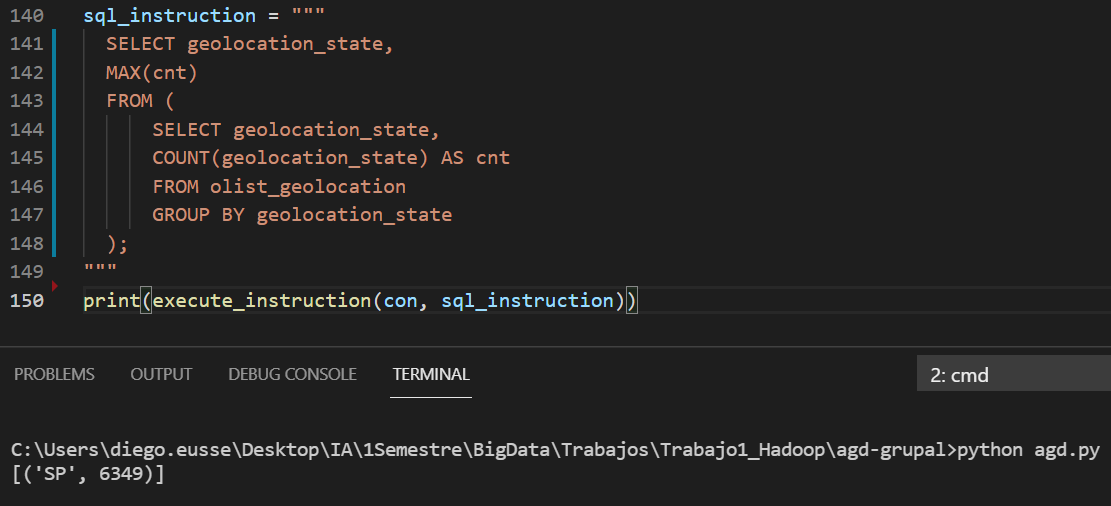
* **Cantidad máxima de registros en una ciudad**

****

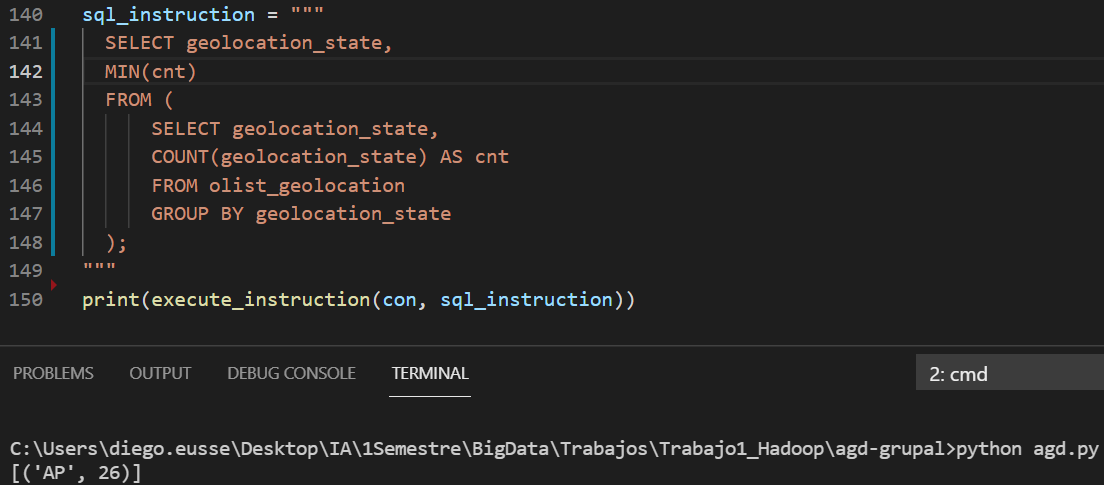
* **Cantidad mínima de registros en una ciudad**

****

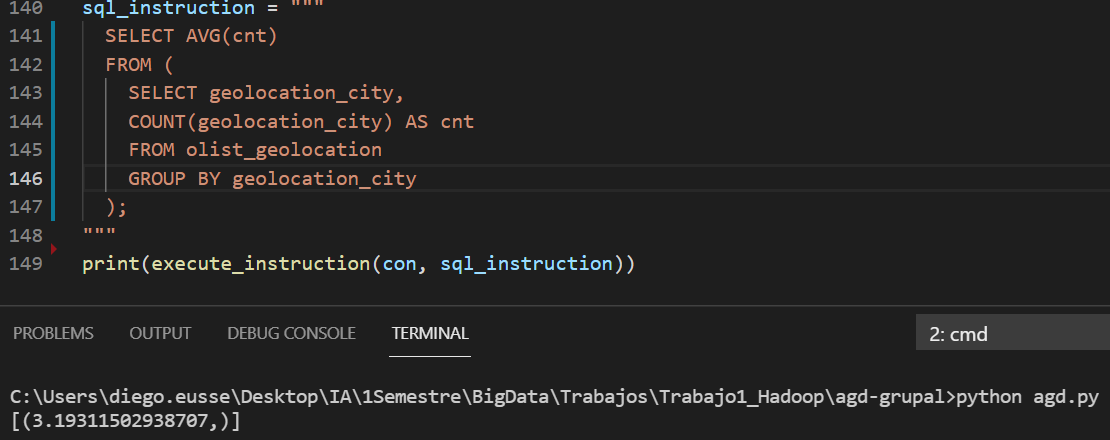
* **Cantidad máxima de registros en un estado**

****

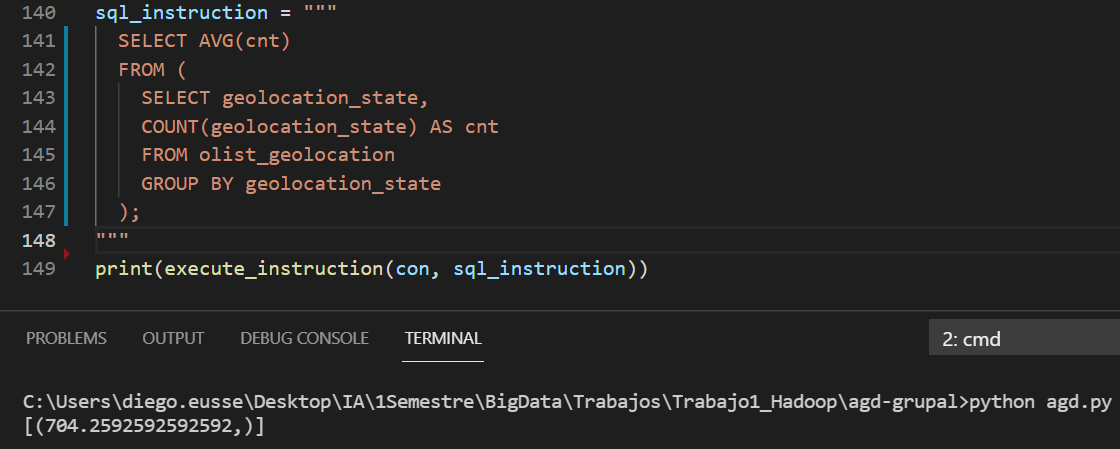
* **Cantidad mínima de registros en un estado**

****

* **Número promedio de registros por ciudad**

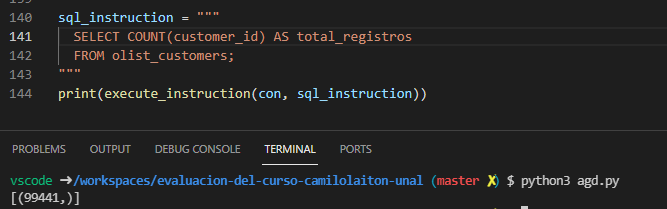
****

* **Número promedio de registros por estado**

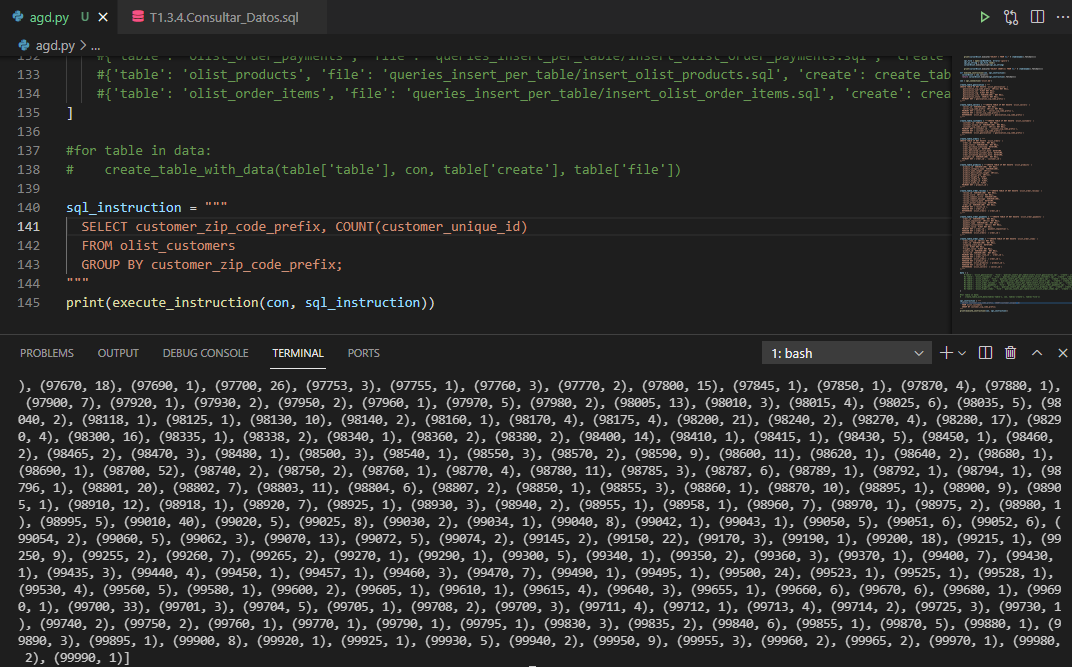
****

**Tabla 2: olist\_customers**

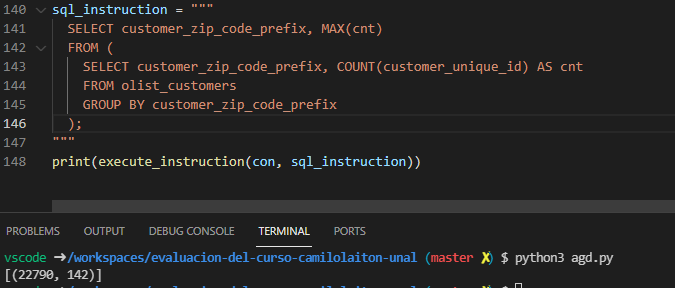
* **Cantidad total de registros**



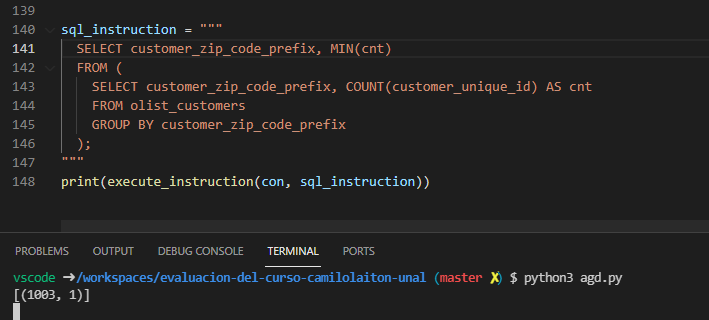
* **Cantidad de clientes registrados por código postal**



* **Código postal en donde hay la mayor cantidad de clientes registrados**

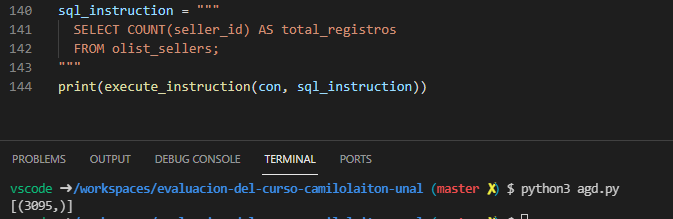


* **Código postal en donde hay la menor cantidad de clientes registrados**

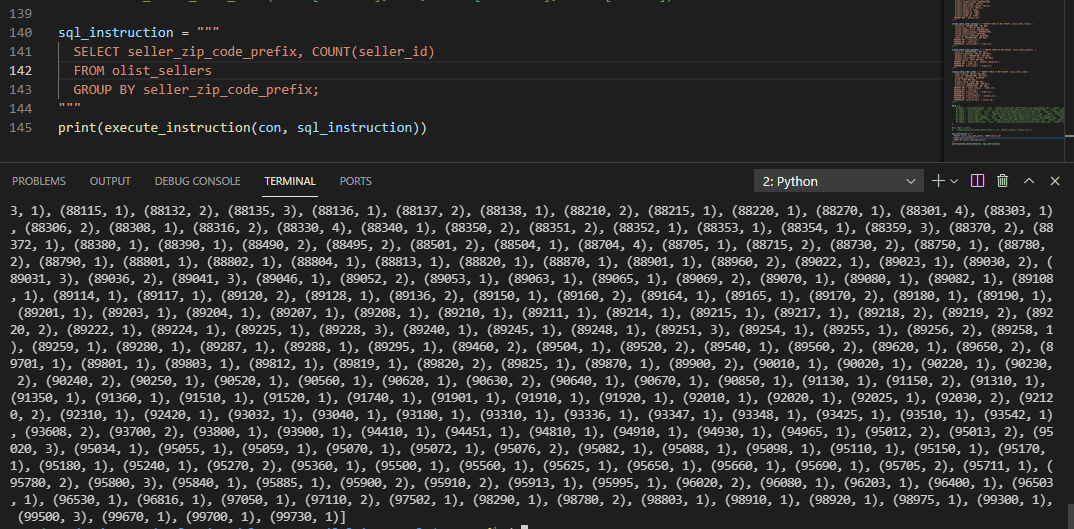


**Tabla 3: olist\_sellers**

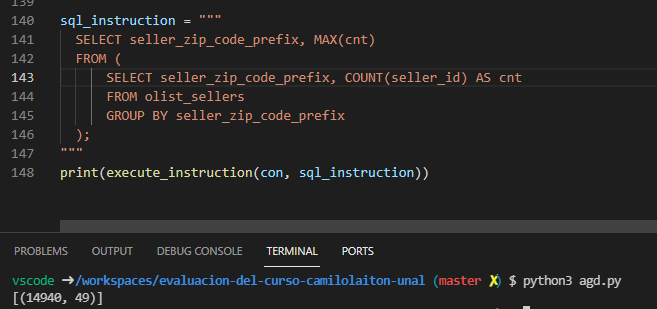
* **Cantidad total de registros**



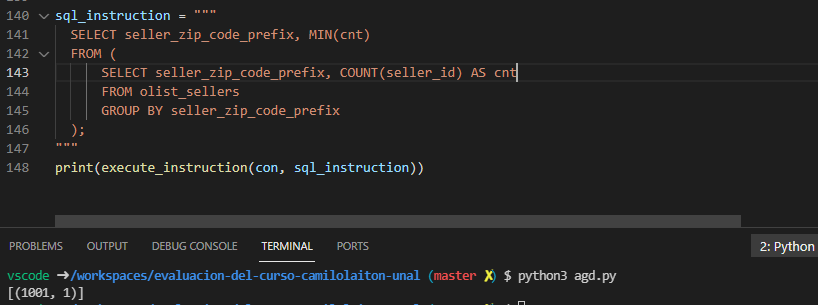
* **Cantidad de vendedores registrados por código postal**



* **Código postal donde hay la mayor cantidad de vendedores registrados**

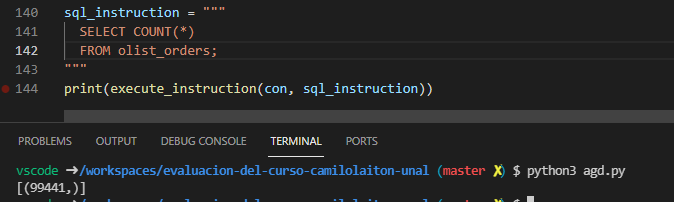


* **Código postal donde hay la menor cantidad de vendedores registrados**

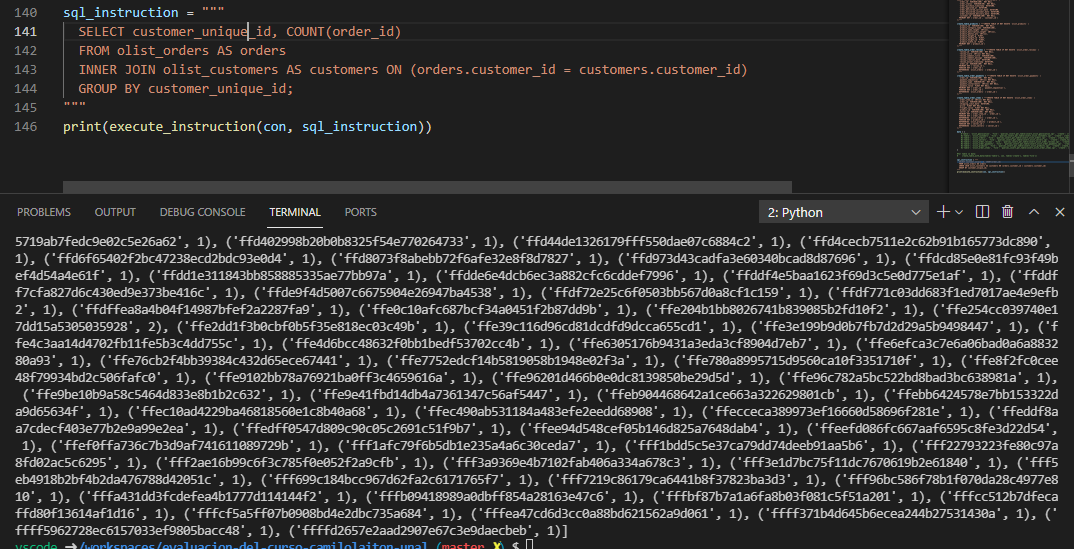


**Tabla 4: olist\_orders**

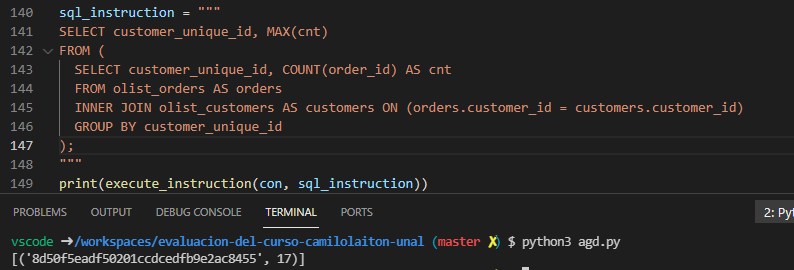
* **Cantidad total de registros**



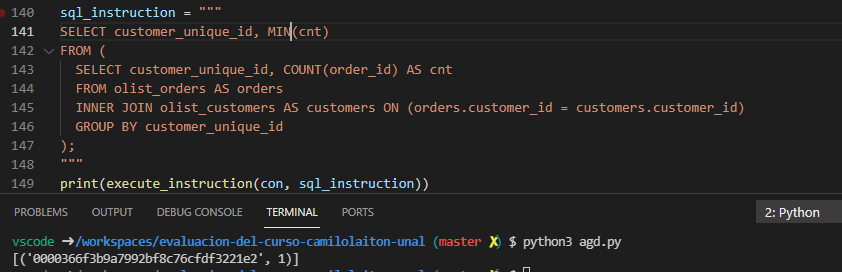
* **Número de pedidos realizados por un cliente**



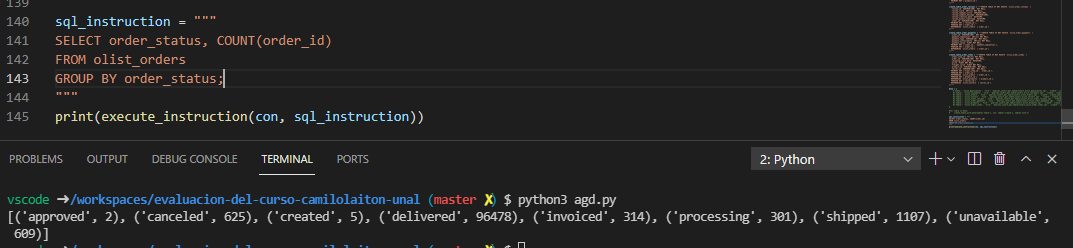
* **Mayor cantidad de pedidos realizados por un cliente**



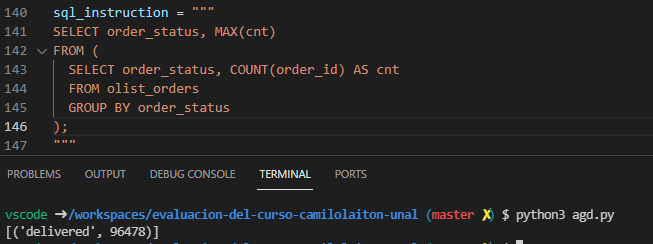
* **Menor cantidad de pedidos realizados por un cliente**



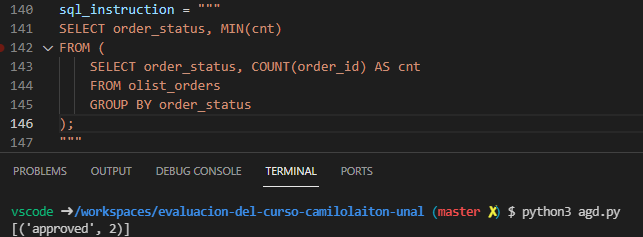
* **Número de pedidos que se encuentran en cada uno de los estados existentes**



* **El estado de pedido que tiene asociado la mayor cantidad de registros**



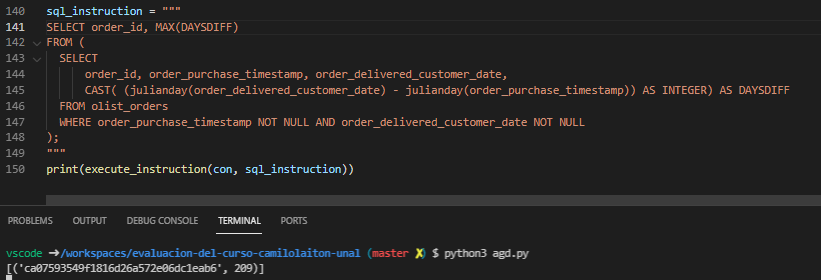
* **El estado de pedido que tiene asociado la menor cantidad de registros**



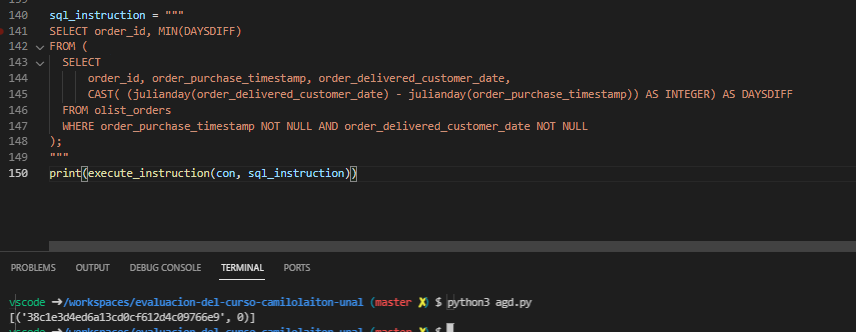
* **Promedio de tiempo (en días) que demora un pedido en ser entregado a un cliente desde su compra por la plataforma**



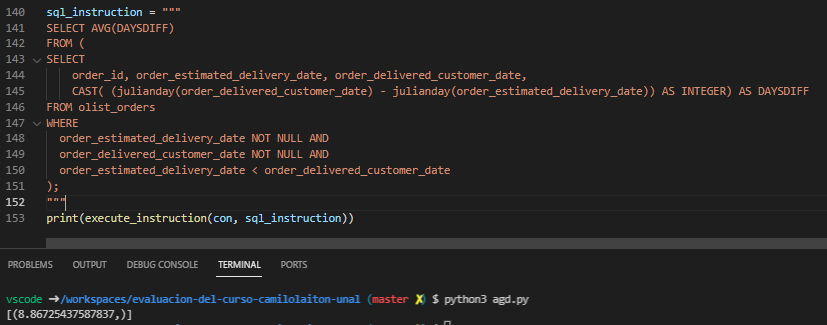
* **Mayor tiempo en días que demoró un pedido en ser entregado a un cliente desde su compra por la plataforma**



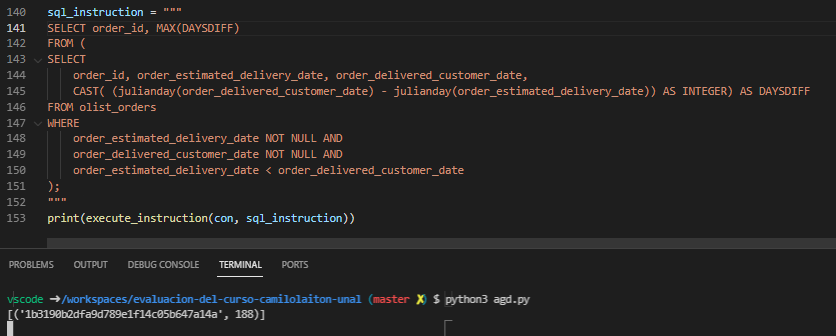
* **Menor tiempo en días que demoró un pedido en ser entregado a un cliente desde su compra por la plataforma**



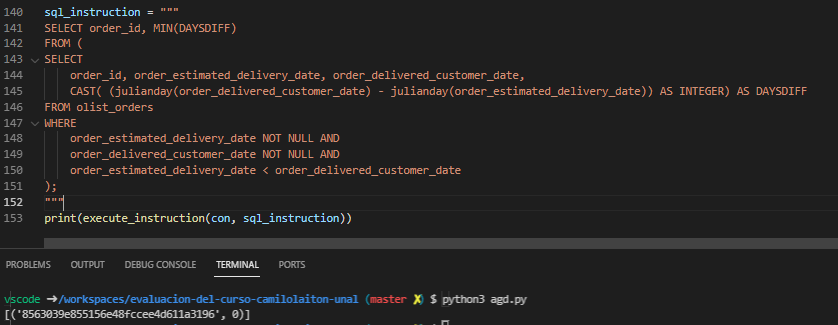
* **Promedio de tiempo en días que demora un pedido en ser entregado después de pasar la fecha estimada de entrega**



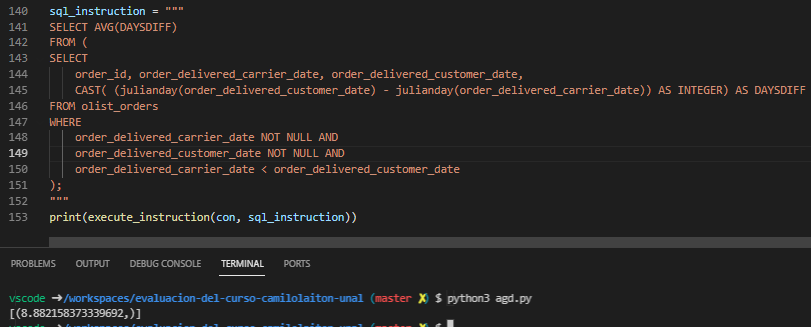
* **Mayor tiempo en días que demoró un pedido en ser entregado después de pasar la fecha estimada de entrega**



* **Menor tiempo en días que demoró un pedido en ser entregado después de pasar la fecha estimada de entrega**

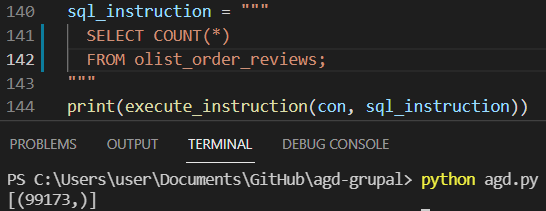


* **Promedio de tiempo en días que demora un pedido en ser entregado después de ser enviado al socio logístico**

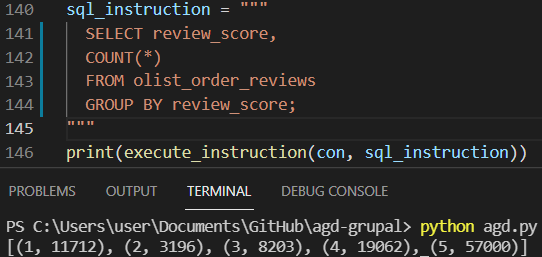


**Tabla 5: olist\_order\_reviews**

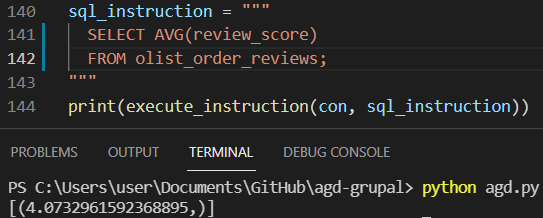
* **Cantidad total de reseñas registradas**

****

* **Cantidad total de registros agrupados según su puntuación en la reseña**

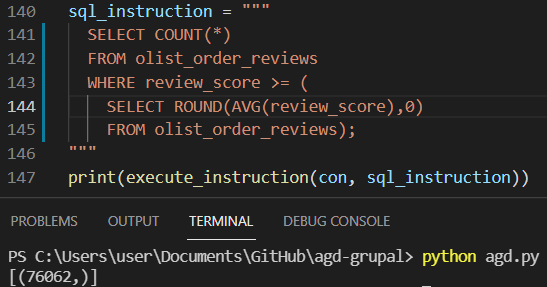
****

* **Promedio de puntuación de todas las reseñas**

****

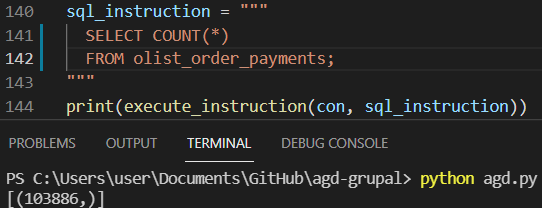
* **Cantidad de registros cuya puntuación es superior al promedio de puntuación de la tabla.**

**Para efectos prácticos, se redondea el promedio al entero más cercano**

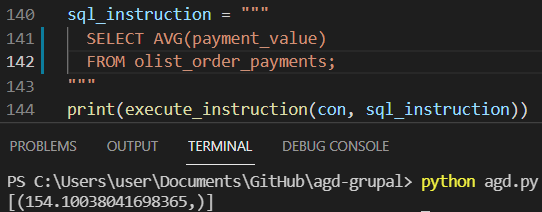
****

**Tabla 6: olist\_order\_payments**

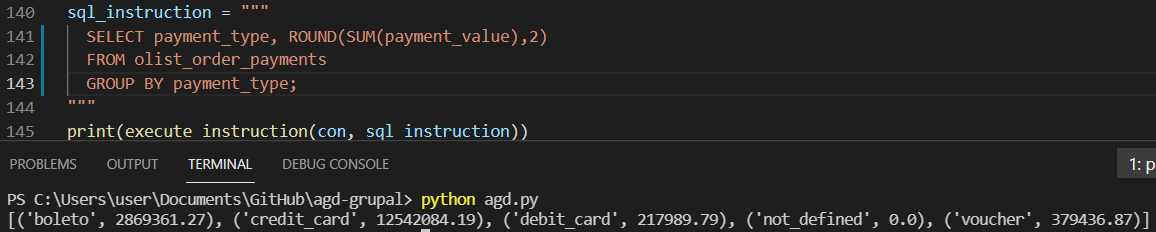
* **Cantidad total de pagos registrados**

****

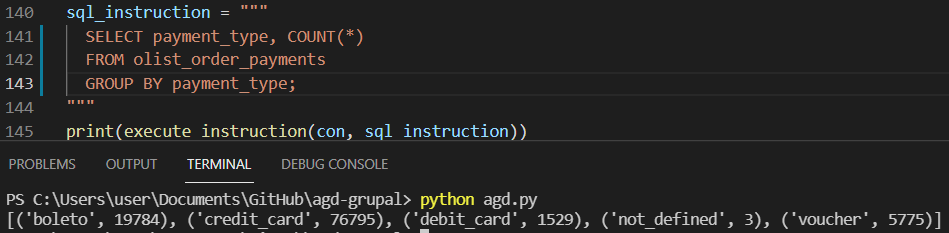
* **Promedio de valor de todos los pagos realizados**

****

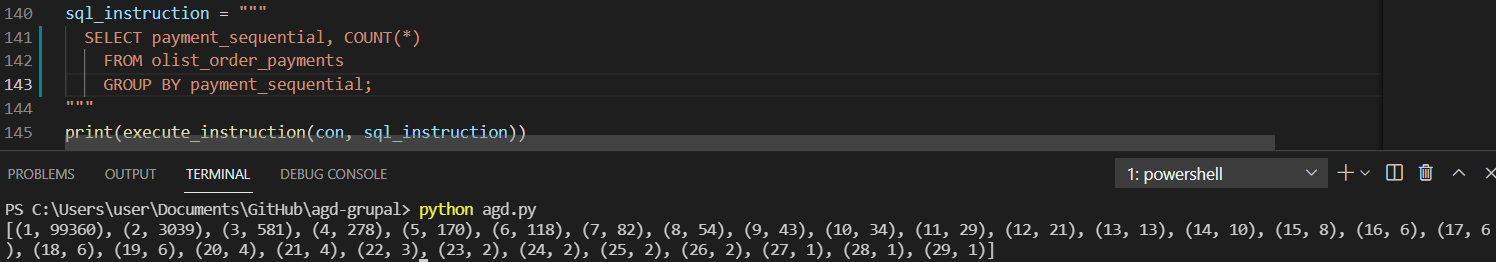
* **Suma de los pagos realizados agrupados por su método de pago**

****

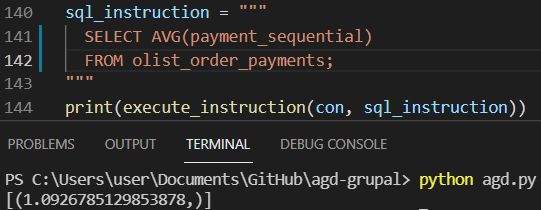
* **Conteo de los pagos realizados agrupados por su método de pago**

****

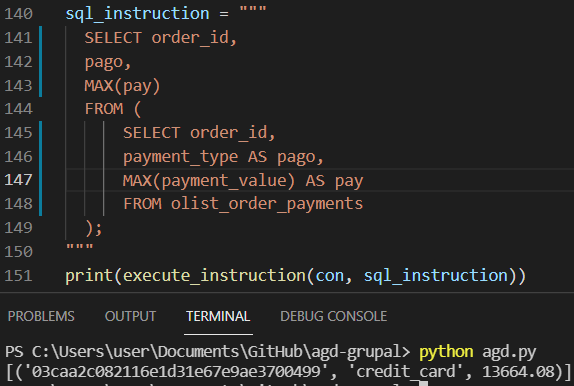
* **Total de registros agrupados por sus secuencias de pago**

****

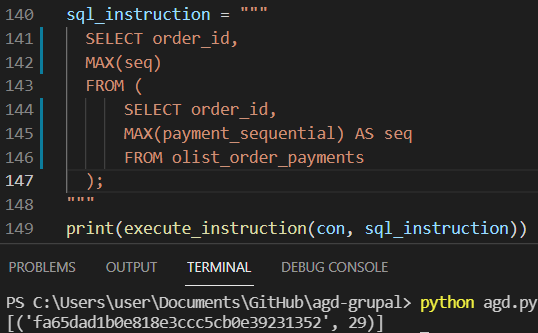
* **Promedio de las secuencias de pago de todos los pagos realizados**

****

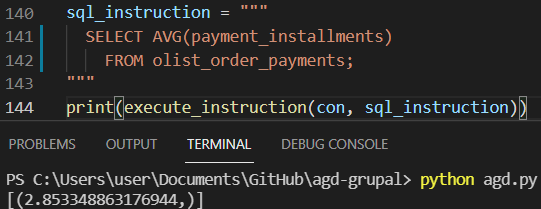
* **Orden y método de pago de la orden que posee el valor máximo**

****

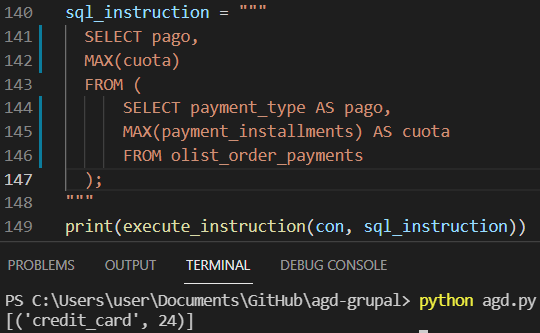
* **Orden que posee el mayor número de secuencias de pago**

****

* **Promedio de cuotas de todos los pagos realizados**

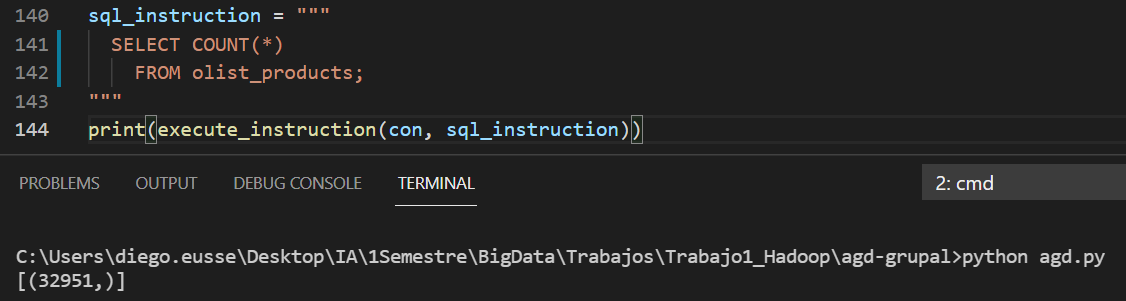
****

* **Método de pago de las órdenes que posee el mayor número de cuotas**

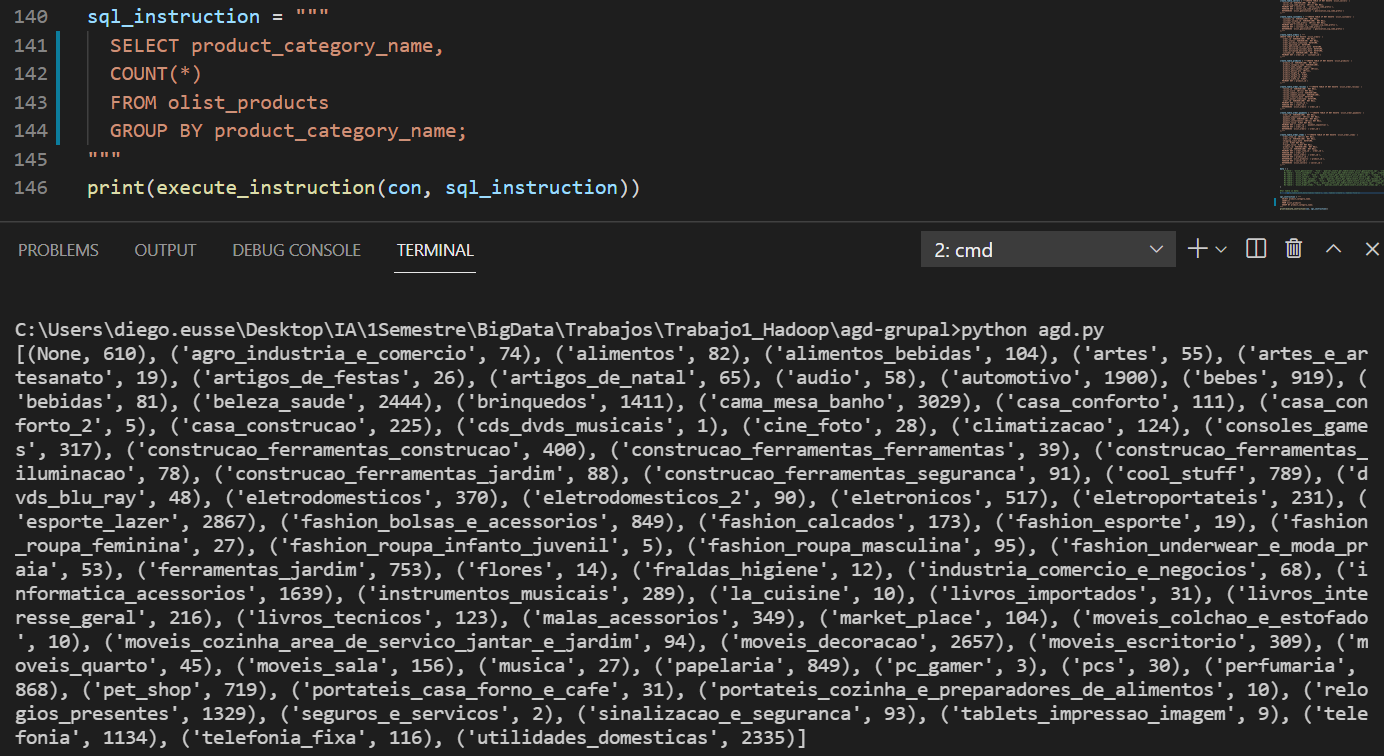
****

**Tabla 7: olist\_products**

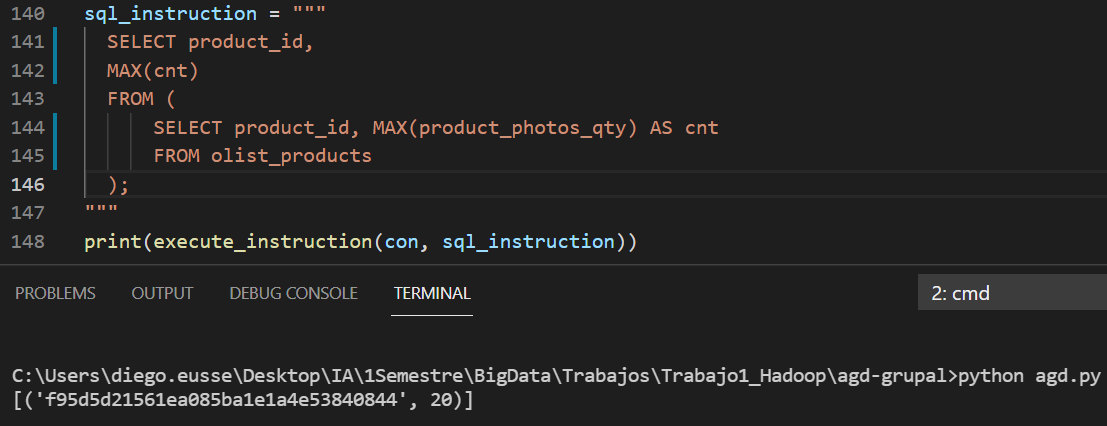
* **Cantidad total de productos registrados**

****

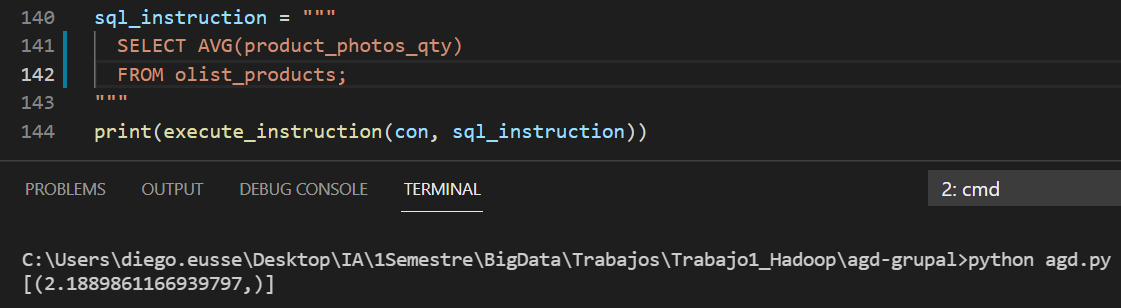
* **Cantidad total de productos por categoría**

****

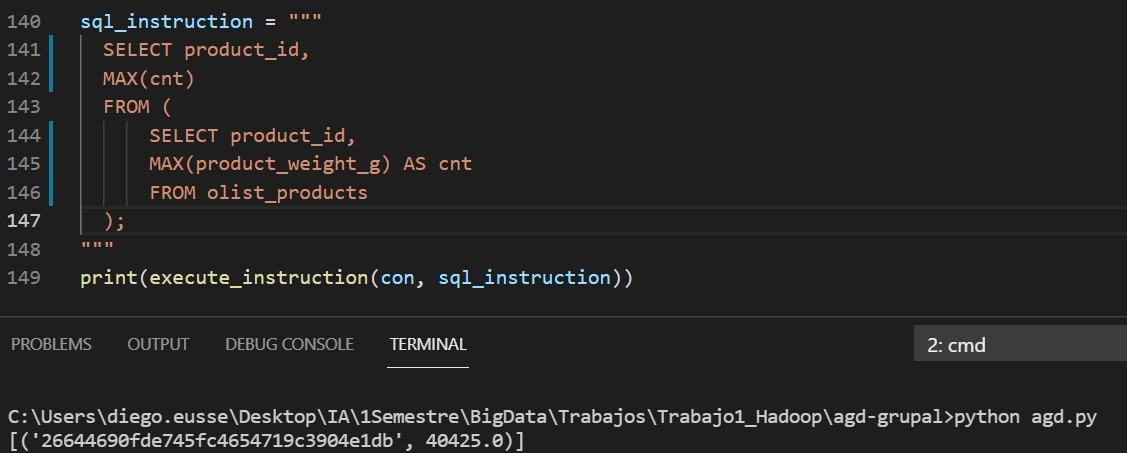
* **Producto con mayor cantidad de fotos asociadas**

****

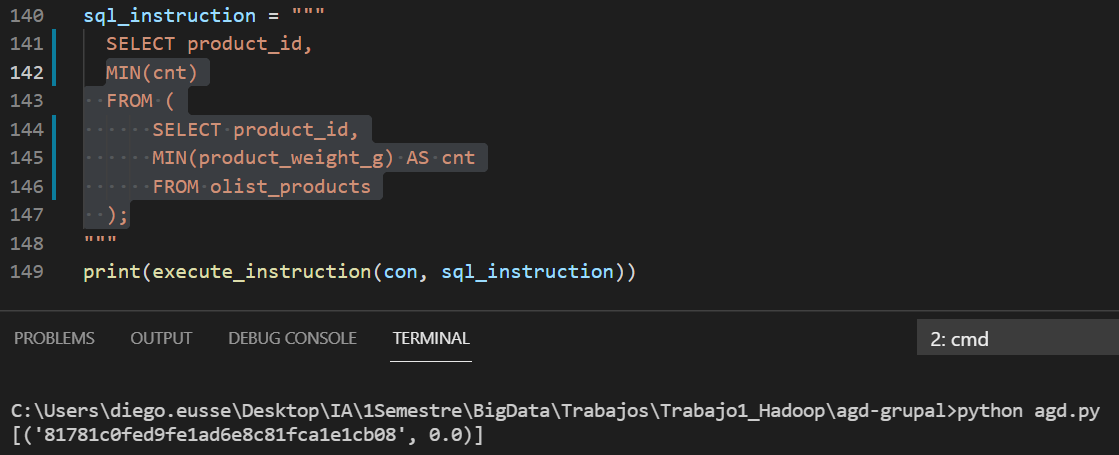
* **Promedio de fotos por producto**

****

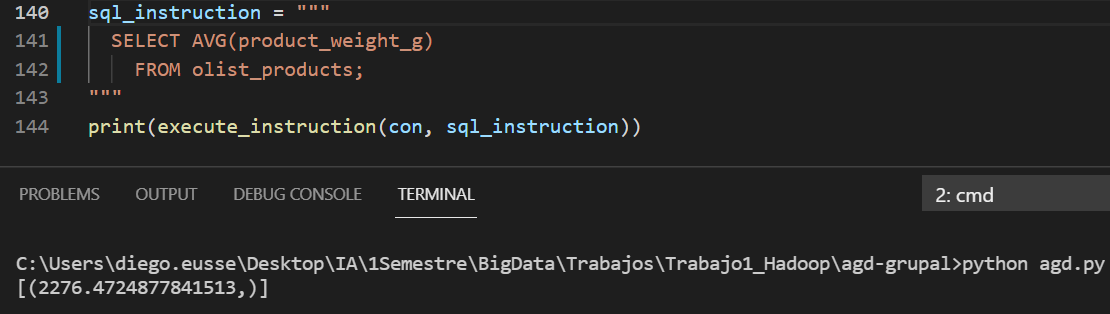
* **Producto más pesado y peso asociado (en gramos)**

****

* **Producto más liviano y peso asociado (en gramos)**

****

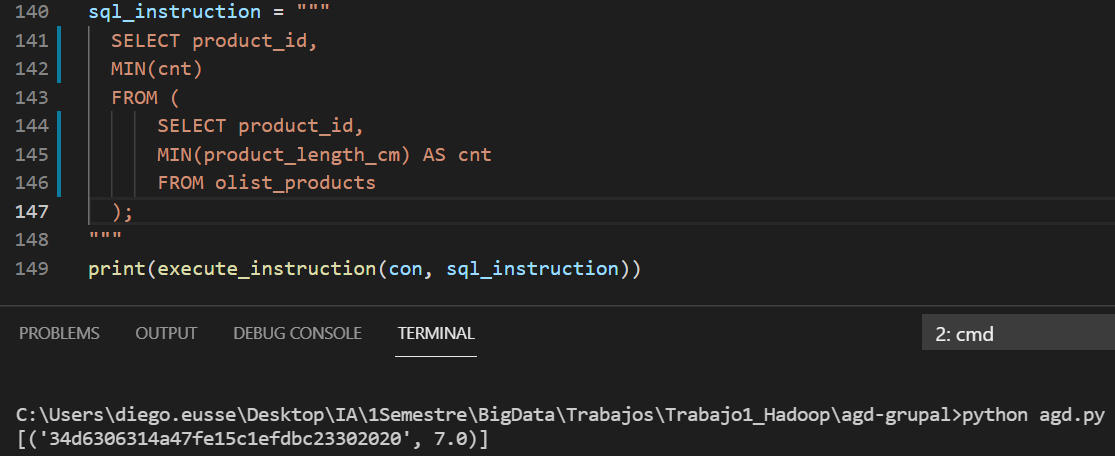
* **Peso promedio de los productos (en gramos)**

****

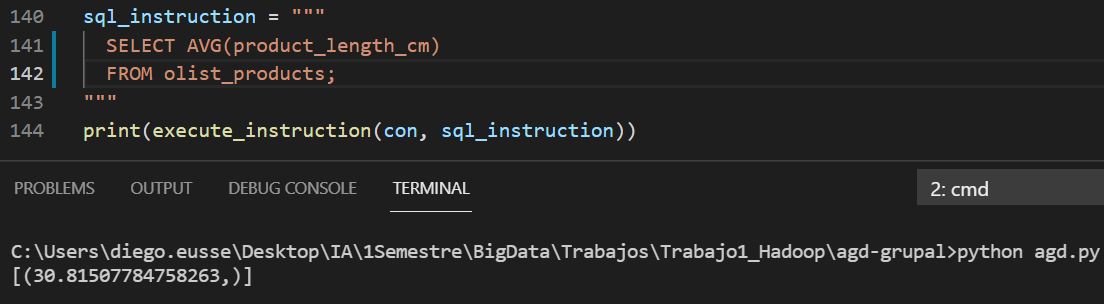
* **Producto más largo y longitud (en cm)**

****

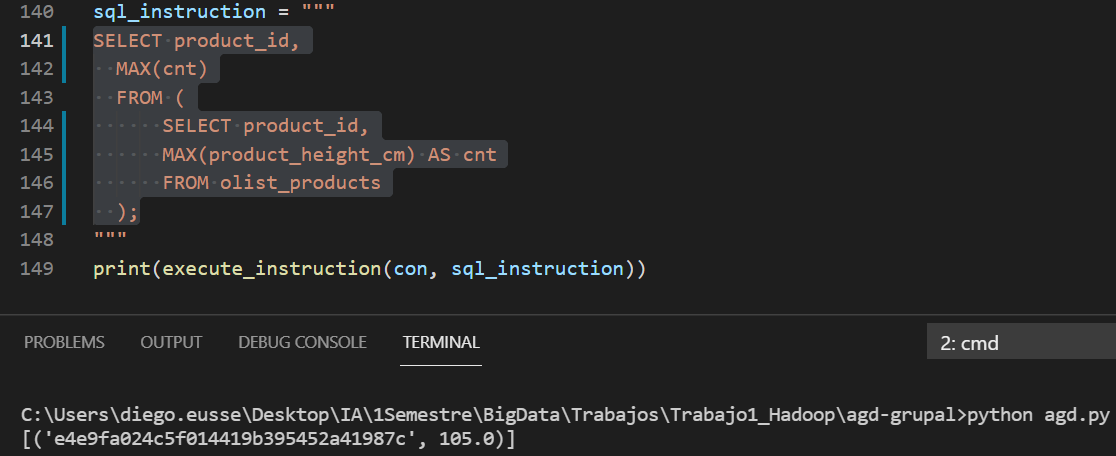
* **Producto menos largo y longitud (en cm)**

****

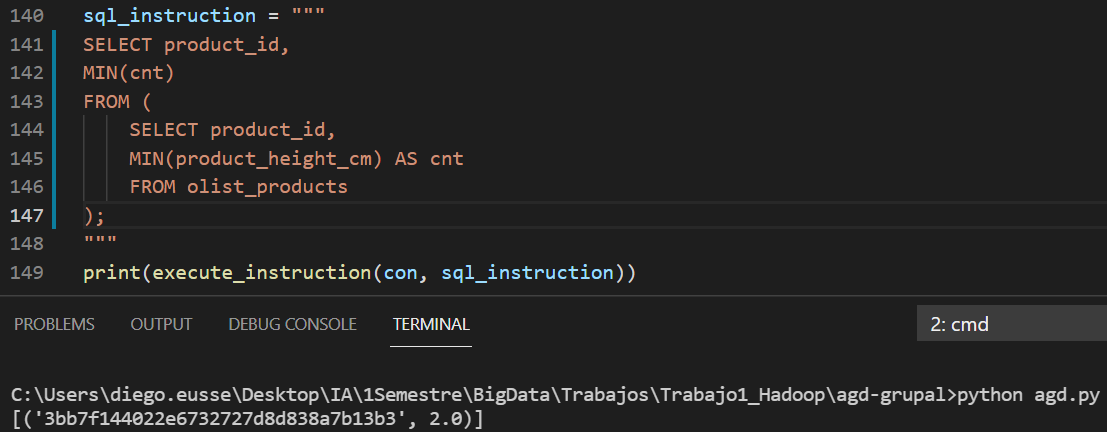
* **Largo promedio de los productos**

****

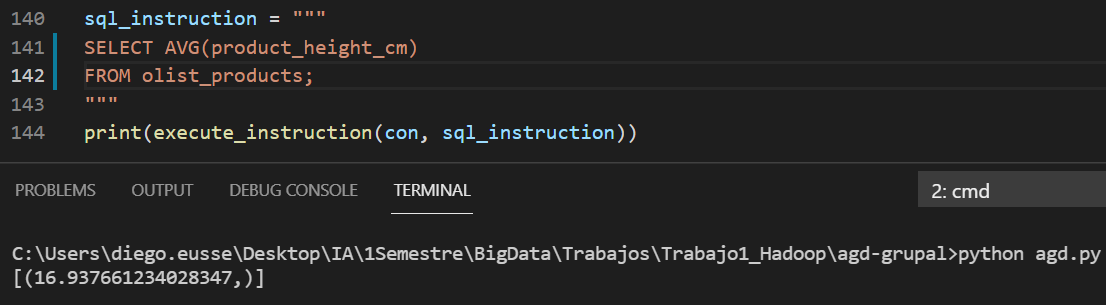
* **Producto más alto y altura (en cm)**

****

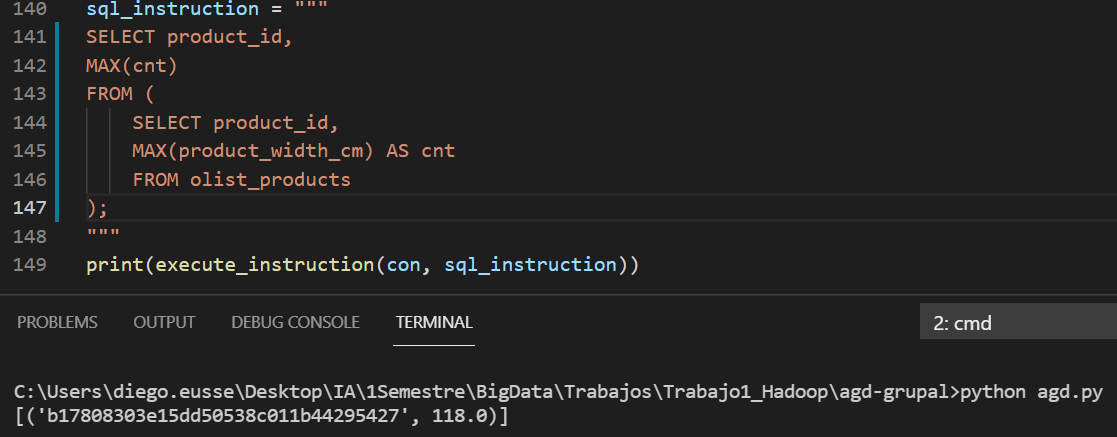
* **Producto menos alto y altura (en cm)**

****

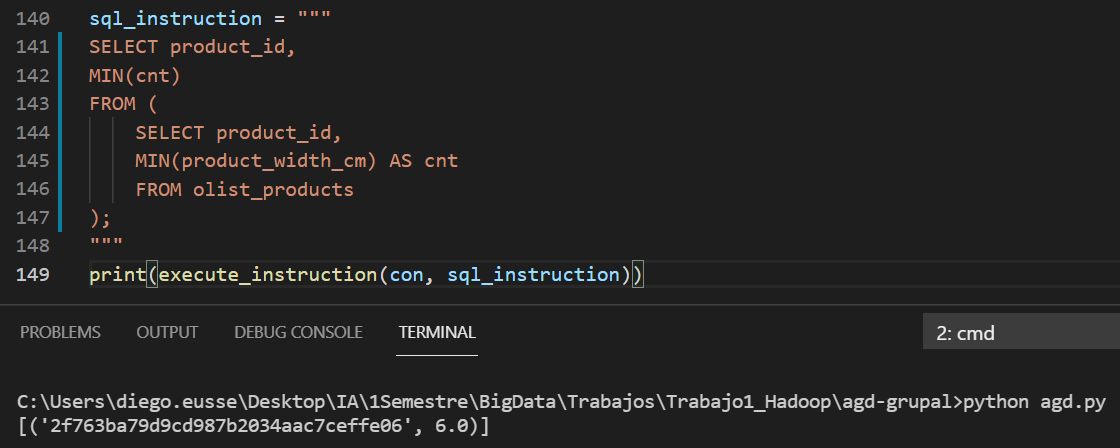
* **Altura promedio de los productos (en cm)**

****

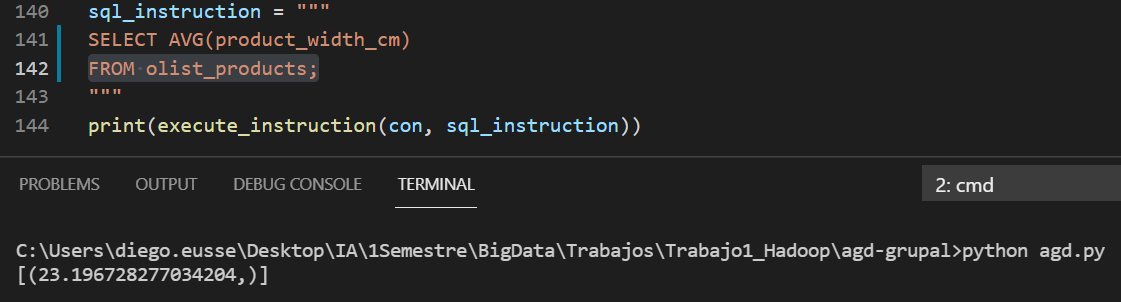
* **Producto más ancho y anchura (en cm)**

****

* **Producto menos ancho y anchura (en cm)**

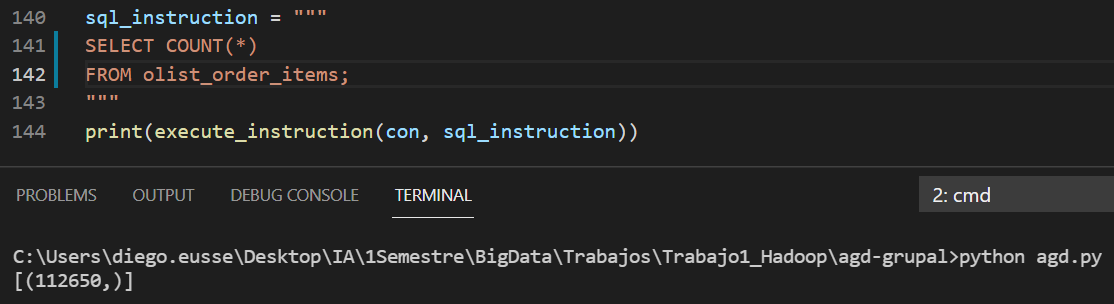
****

* **Anchura promedio de los productos (en cm)**

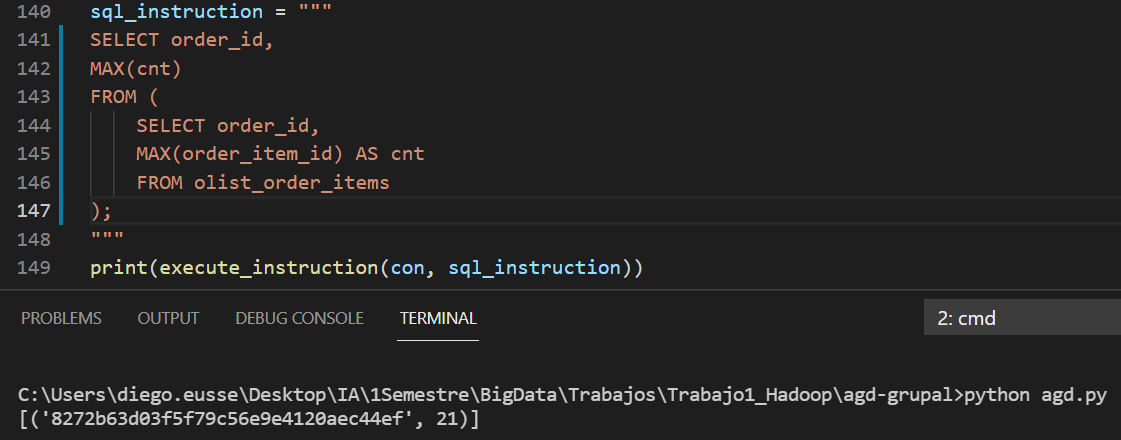
****

**Tabla 8: olist\_order\_items**

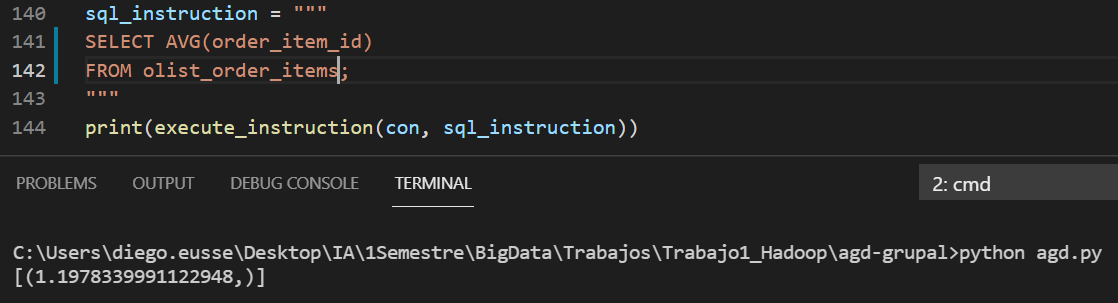
* **Cantidad total de order\_items registrados**

****

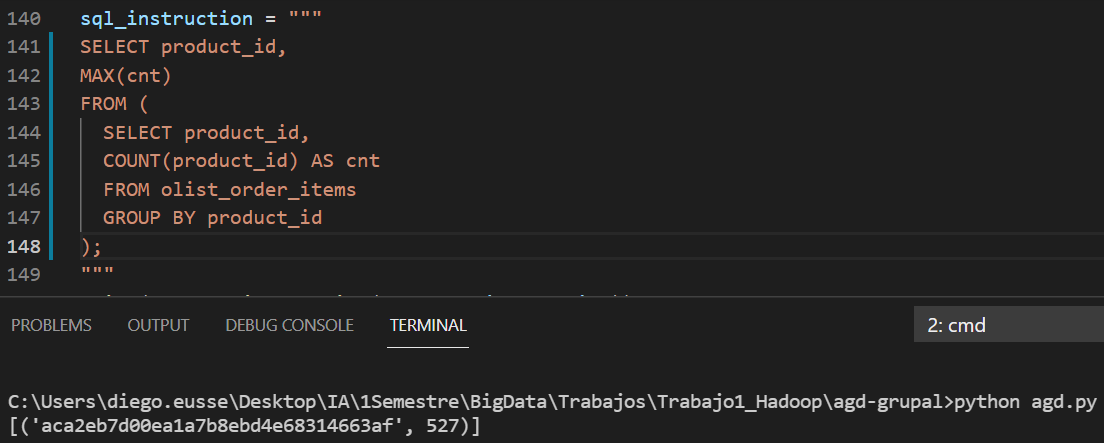
* **Orden con mayor cantidad de productos y cantidad de productos asociada**

****

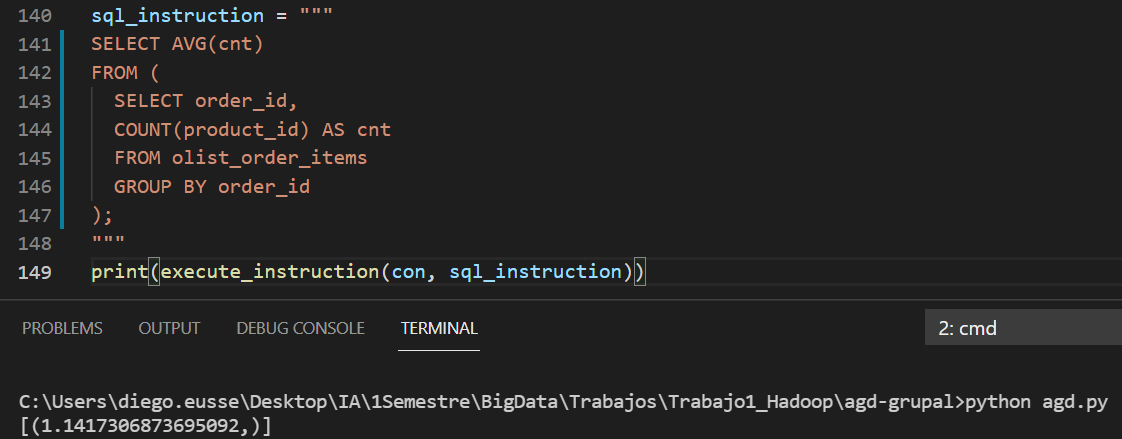
* **Número promedio de productos por orden**

****

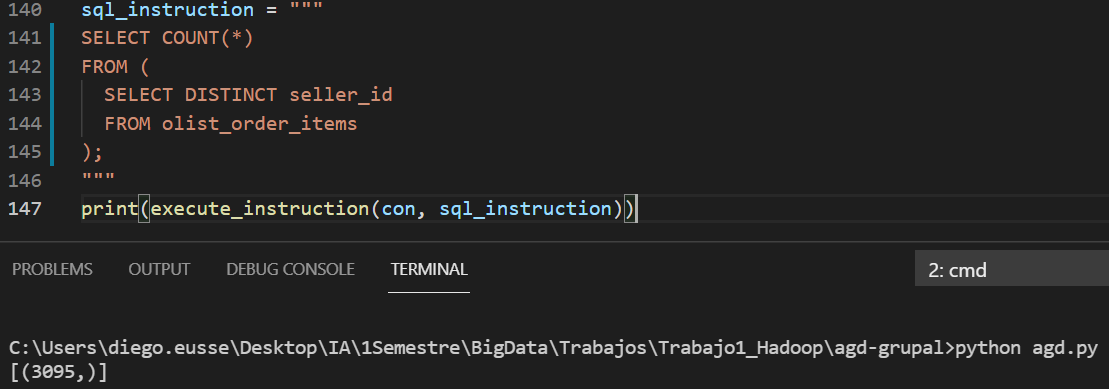
* **Producto más vendido dentro de las órdenes y cantidad de órdenes en las que aparece**

****

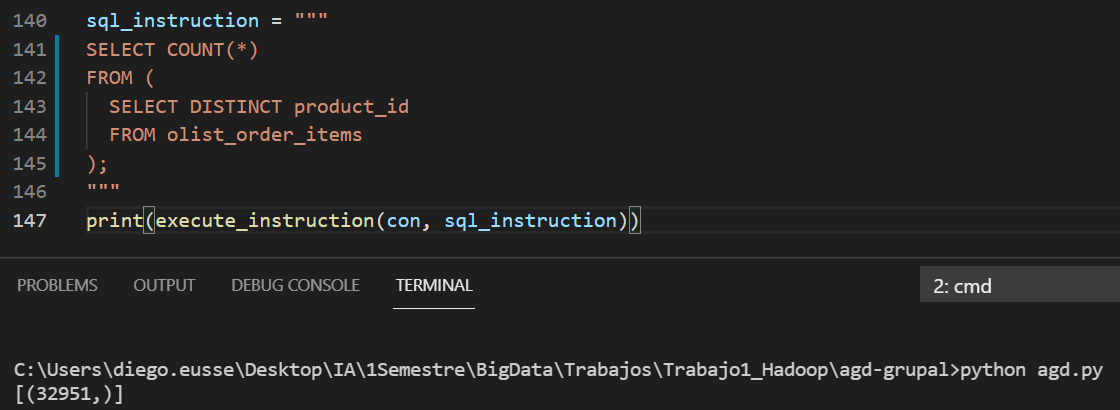
* **Número promedio de productos por orden**

****

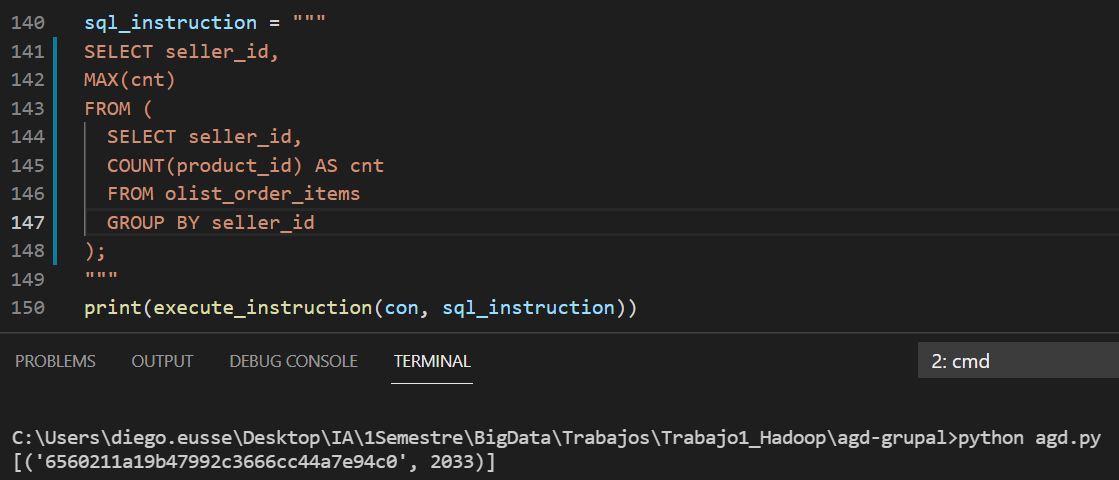
* **Número total de vendedores con al menos una orden**

****

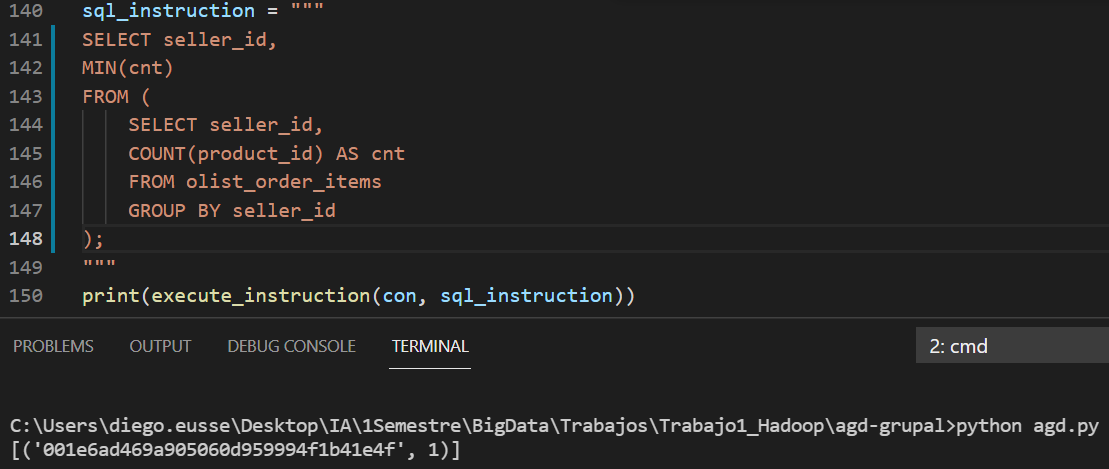
* **Número total de productos distintos**

****

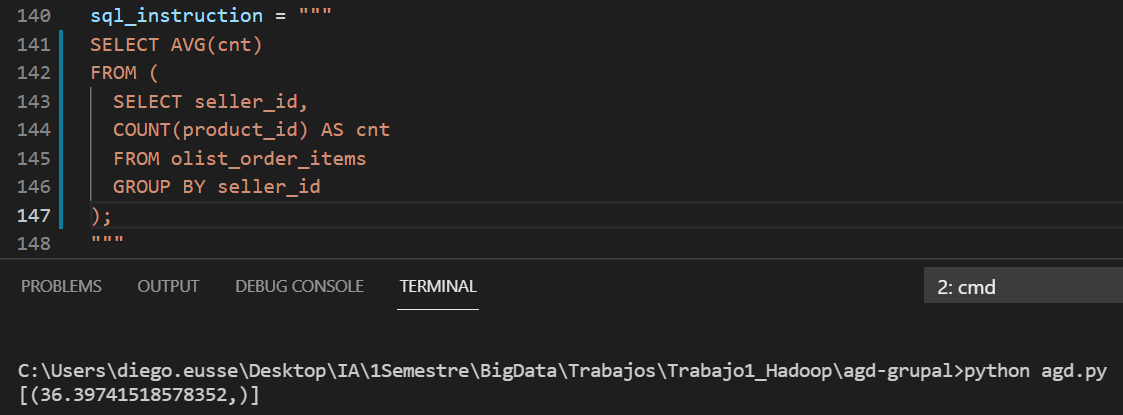
* **Vendedor con mayor número de productos vendidos y cantidad de productos vendidos**

****

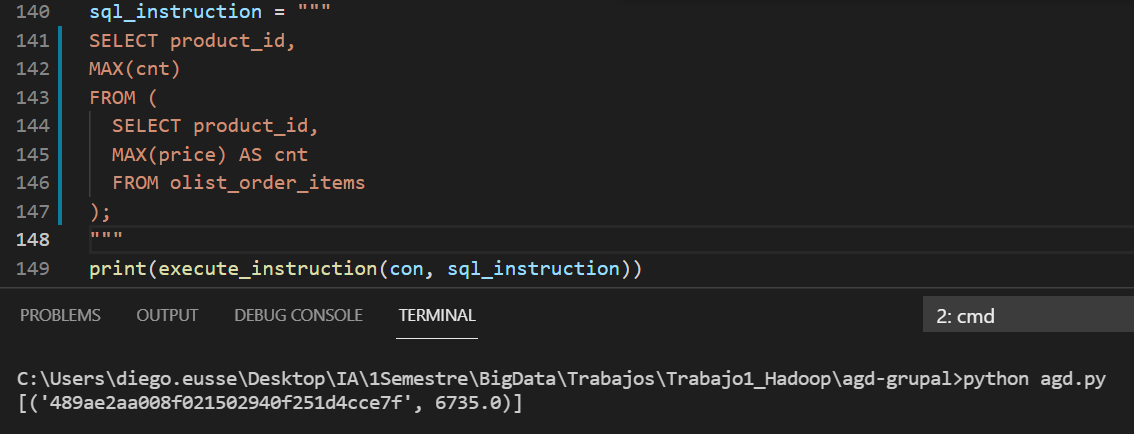
* **Vendedor con menor número de órdenes y cantidad de órdenes asociadas**

****

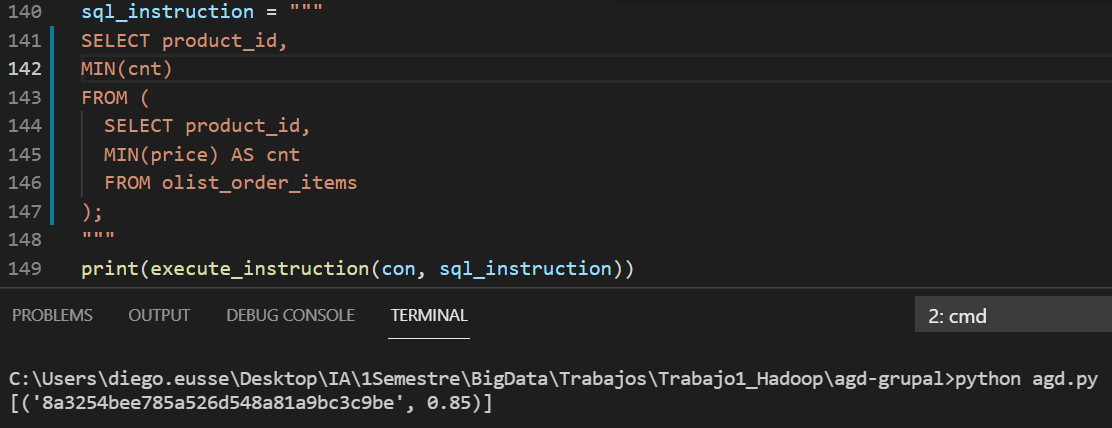
* **Número promedio de productos que vende cada vendedor**

****

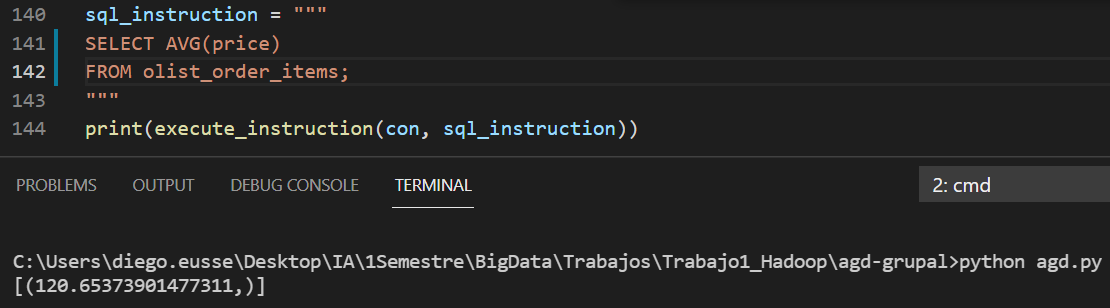
* **Producto más costoso y id asociado**

****

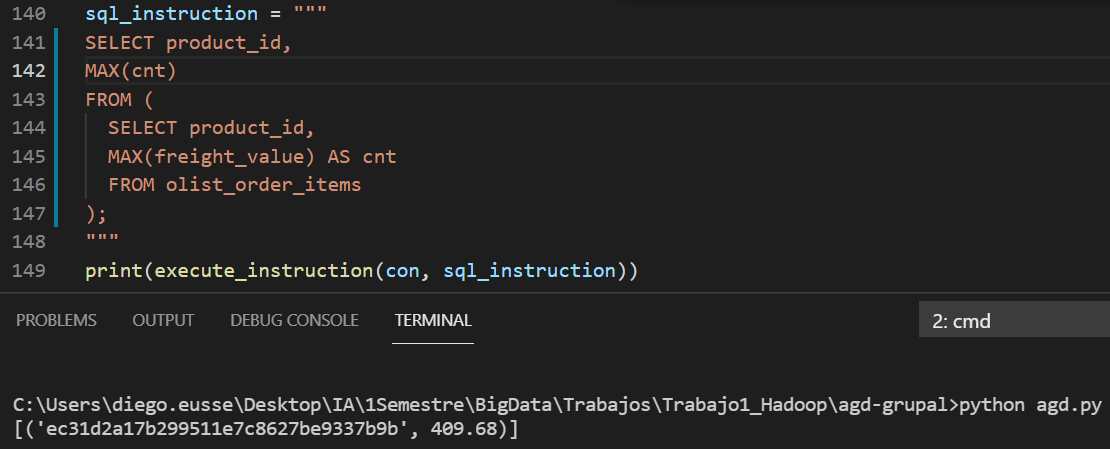
* **Producto más barato y id asociado**

****

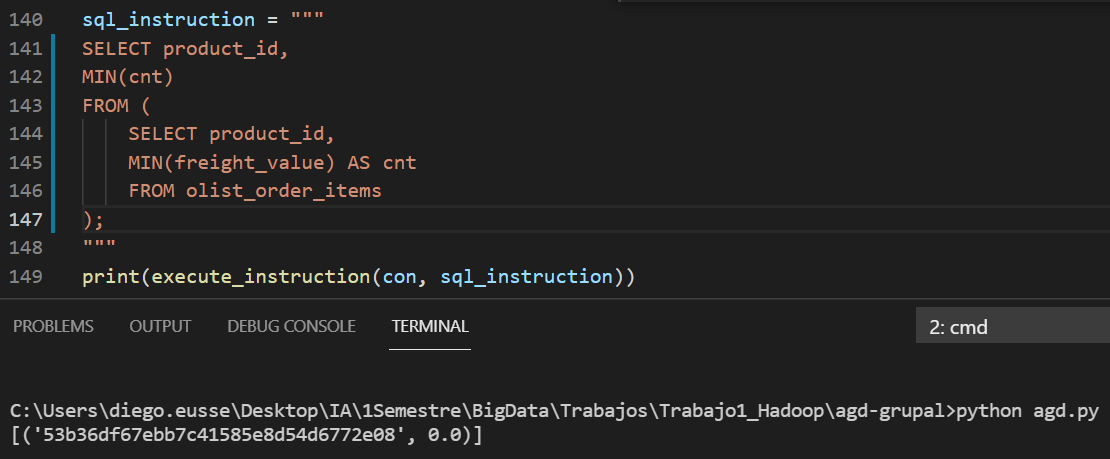
* **Precio promedio de los productos**

****

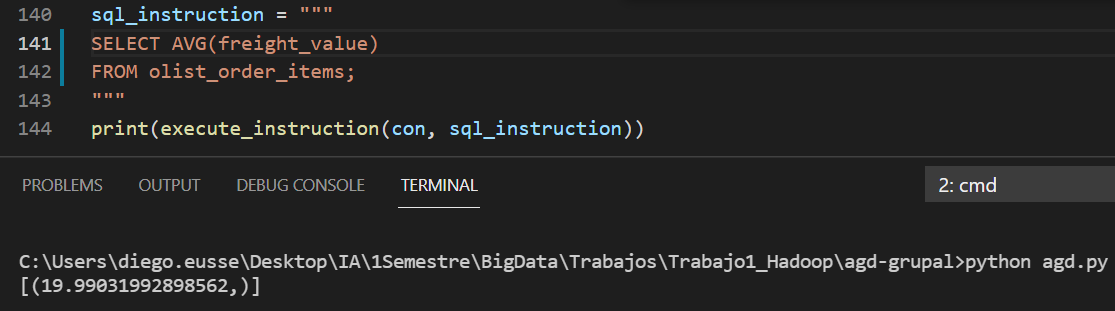
* **Costo de envío más alto y orden asociada**

****

* **Costo de envío más bajo y orden asociada**

****

* **Costo promedio de envío de los productos**

****

1. **MongoDB**
   1. **Sentencia o consulta de creación del documento(s)**

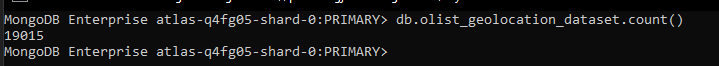
***Observación****: Escriba el código en MongoDB para crear al menos 20 documentos que correspondan a su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.1.Creacion\_Documentos.sql*

* 1. **Sentencia de consulta**

***Observación****: Realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.2.Consultar\_Datos.sql*

**Tabla 1. olist\_geolocation**

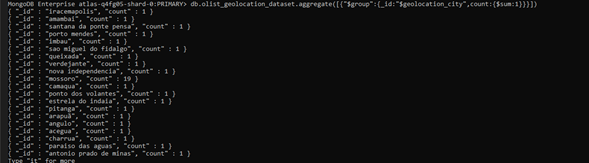
* **Conteo cantidad total de registros**



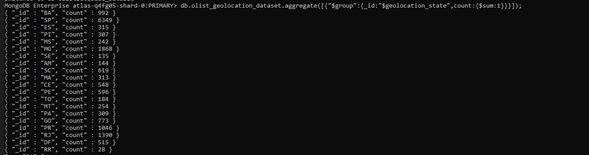
* **Conteo cantidad total de estados diferentes**

****

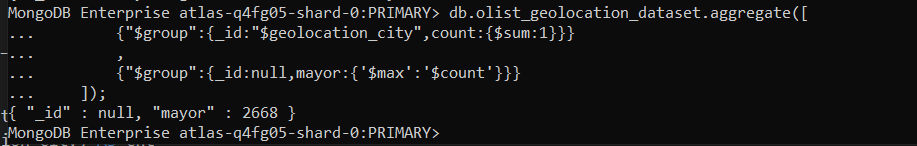
* **Conteo cantidad total de registros por ciudad**

****

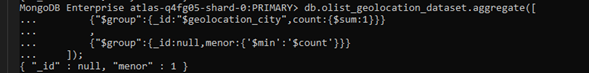
* **Conteo cantidad total de registros por estado**

****

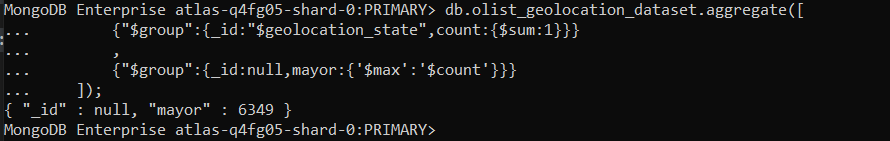
* **Ciudad con mayor número de registros**

****

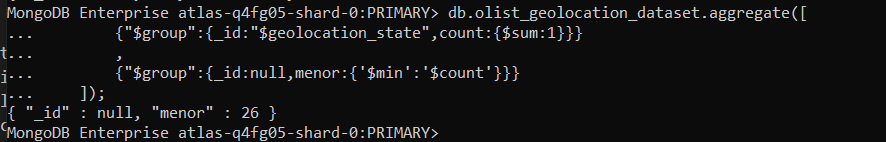
* **Ciudad con menor número de registros**

****

* **Estado con mayor número de registros**

****

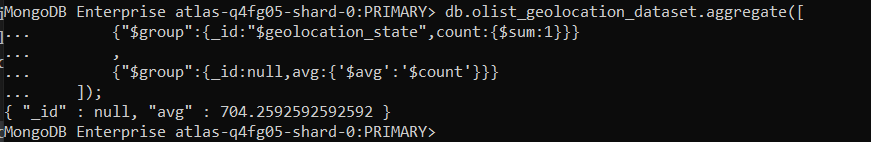
* **Ciudad con menor número de registros**

****

* **Número promedio de registros por ciudad**

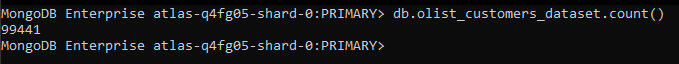
****

* **Número promedio de registros por estado**

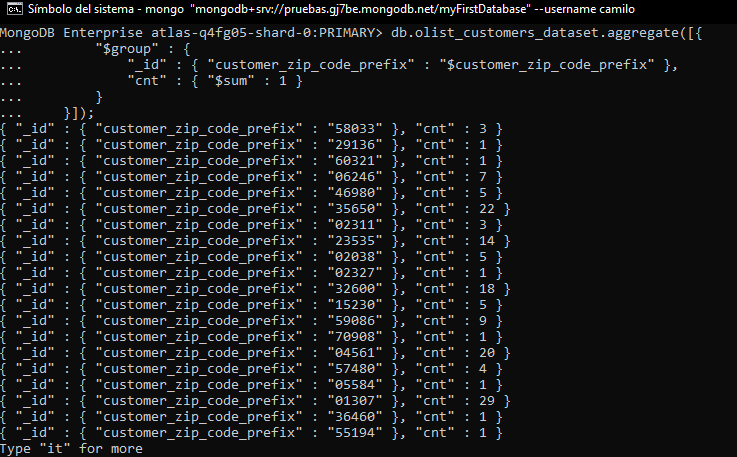
****

**Tabla 2: olist\_customers**

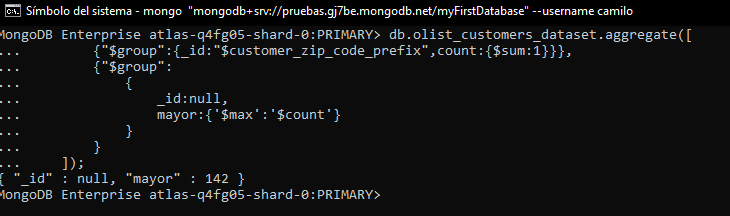
* **Cantidad total de registros**



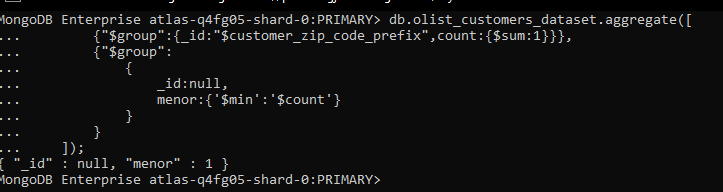
* **Cantidad de clientes registrados por código postal**



* **Código postal en donde hay la mayor cantidad de clientes registrados**

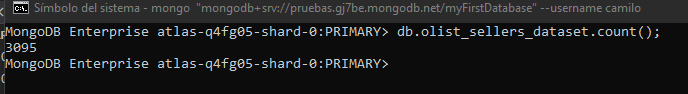


* **Código postal en donde hay la menor cantidad de clientes registrados**

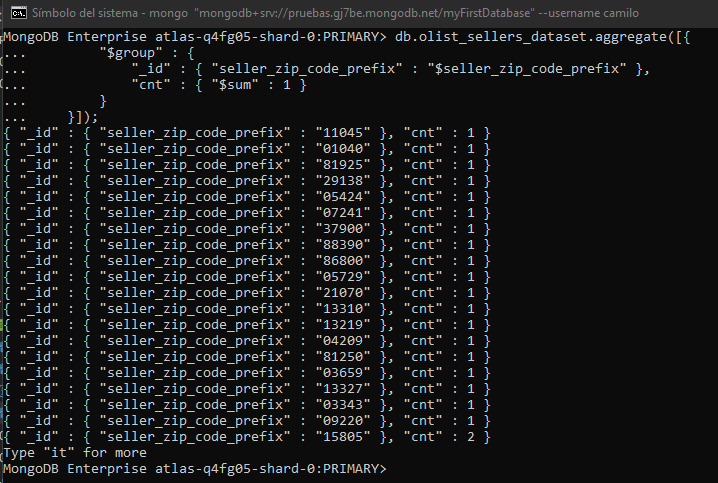


**Tabla 3: olist\_sellers**

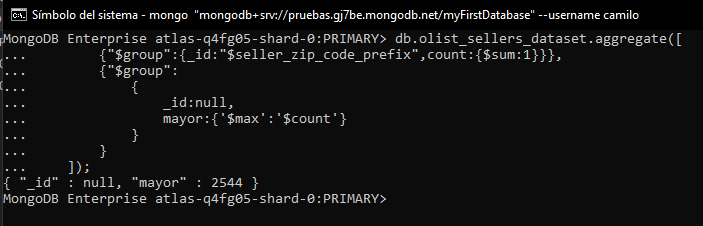
* **Cantidad total de registros**

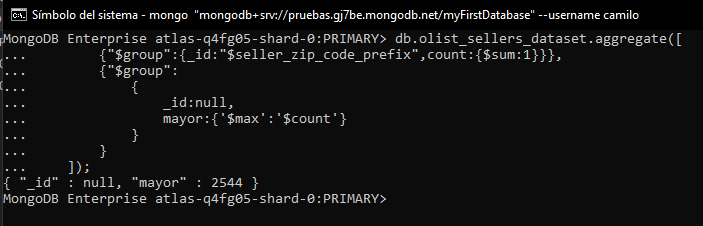


* **Cantidad de vendedores registrados por código postal**

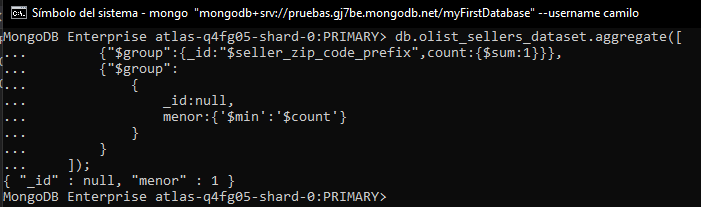


* **Código postal donde hay la mayor cantidad de vendedores registrados**

****



* **Código postal donde hay la menor cantidad de vendedores registrados**

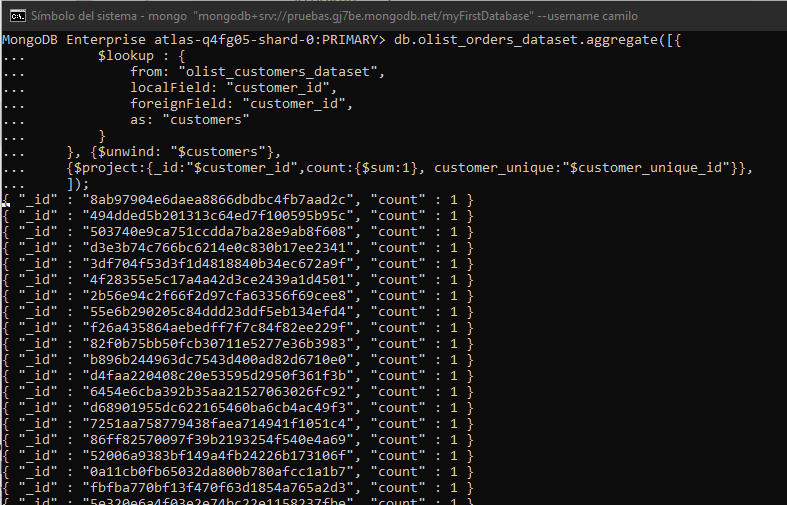


**Tabla 4: olist\_orders**

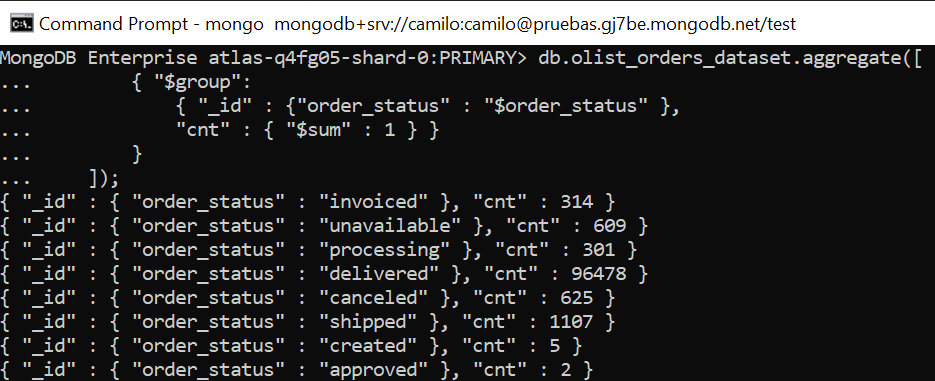
* **Cantidad total de registros**



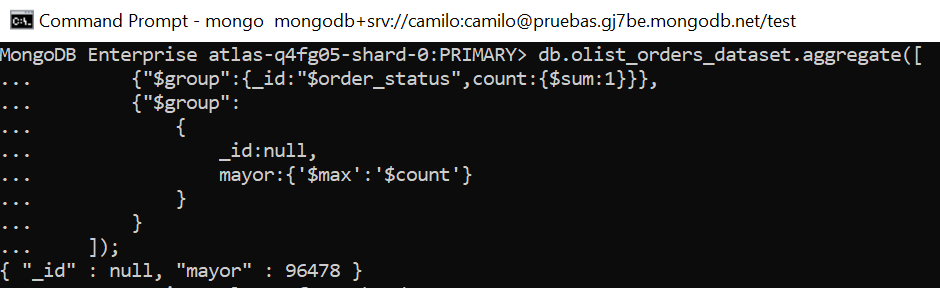
* **Número de pedidos realizados por un cliente**



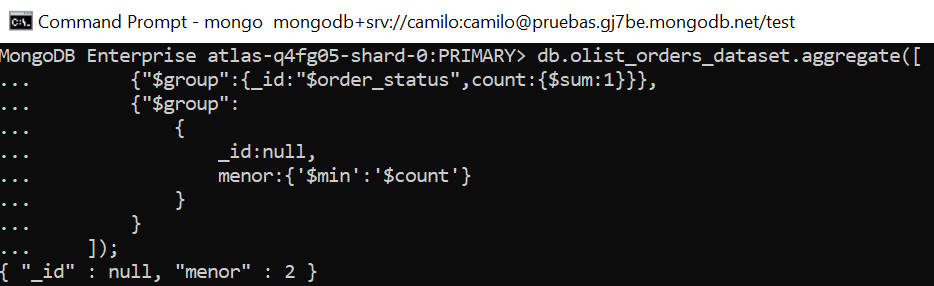
* **Número de pedidos que se encuentran en cada uno de los estados existentes**



* **Mayor cantidad de registros en un estado de pedido**



* **Menor cantidad de registros en un estado de pedidos**

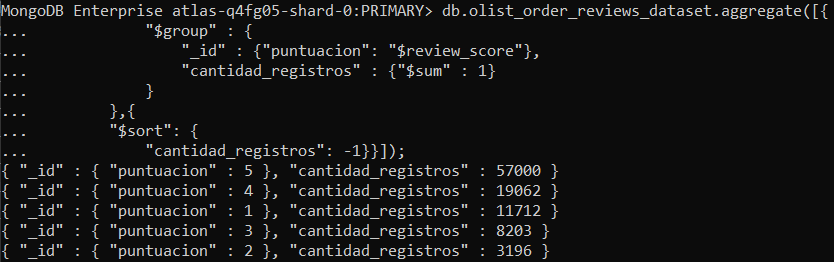


**Tabla 5: olist\_order\_reviews**

* **Cantidad total de reseñas registradas**

****

* **Cantidad total de registros agrupados según su puntuación en la reseña**

****

* **Promedio de puntuación de todas las reseñas**

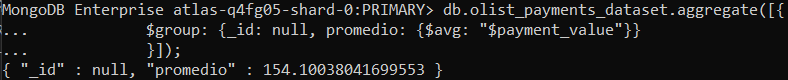
****

**Tabla 6: olist\_order\_payments**

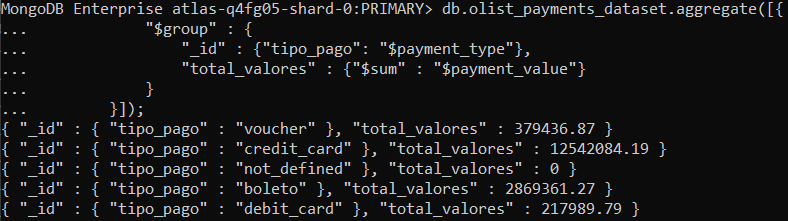
* **Cantidad total de pagos registrados**

****

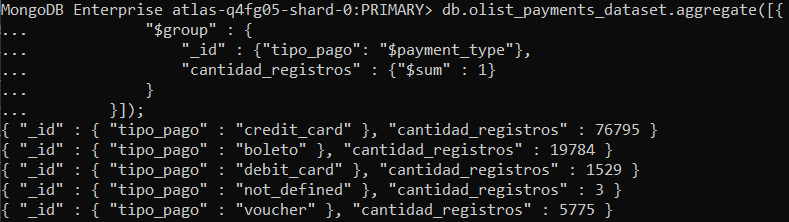
* **Promedio de valor de todos los pagos realizados**

****

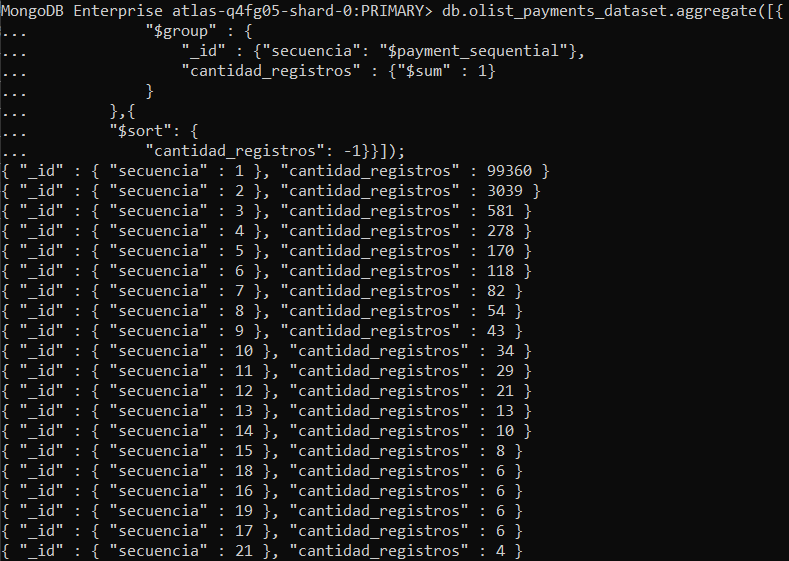
* **Suma de los pagos realizados agrupados por su método de pago**

****

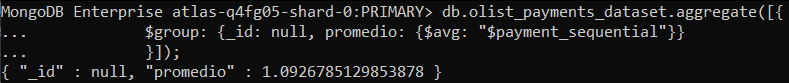
* **Conteo de los pagos realizados agrupados por su método de pago**

****

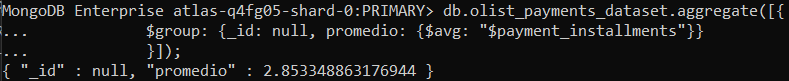
* **Total de registros agrupados por sus secuencias de pago**

****

* **Promedio de las secuencias de pago de todos los pagos realizados**

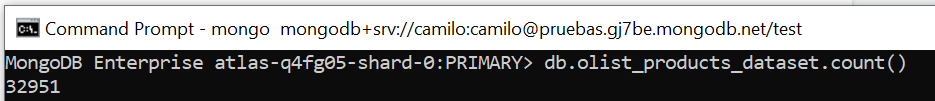
****

* **Promedio de cuotas de todos los pagos realizados**

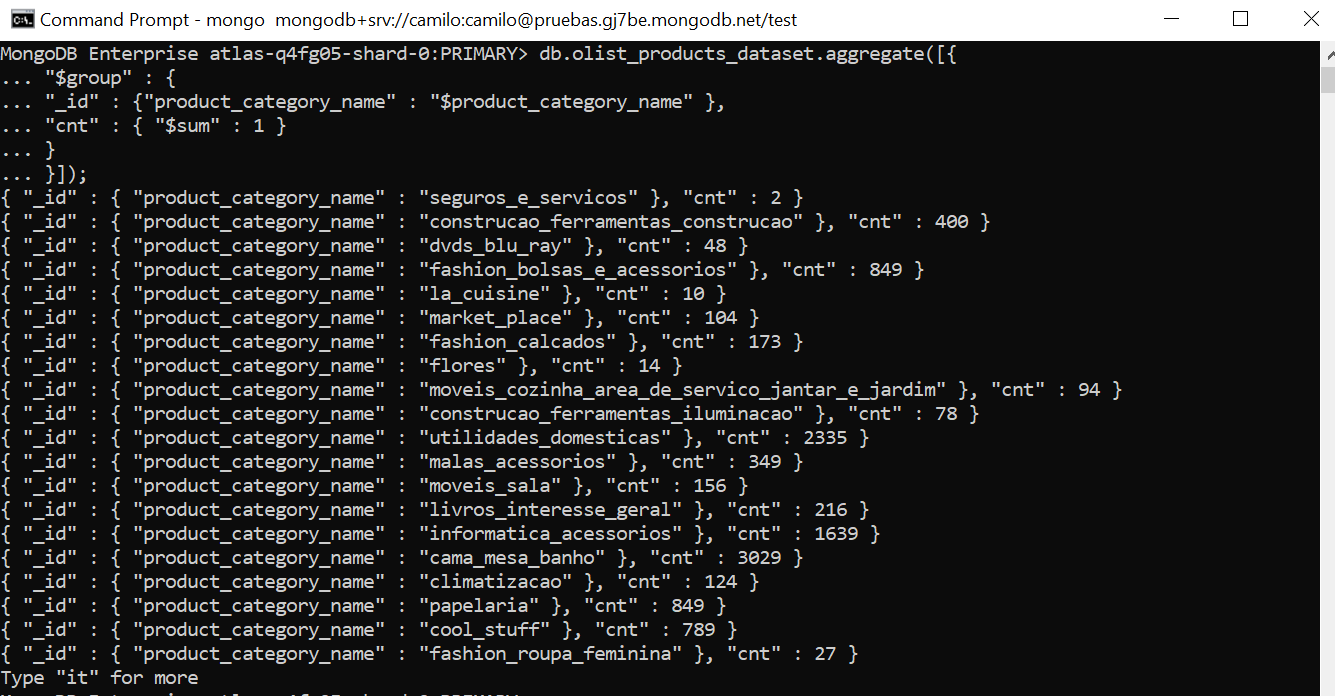
****

**Tabla 7: olist\_products**

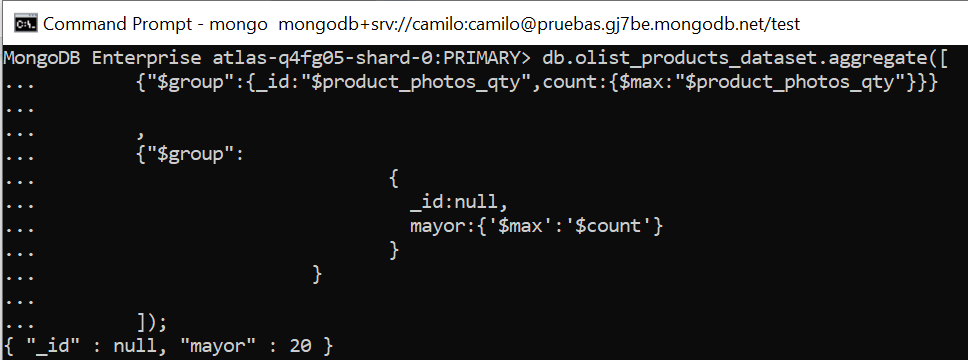
* **Cantidad total de productos registrados**

****

* **Cantidad total de productos por categoría**

****

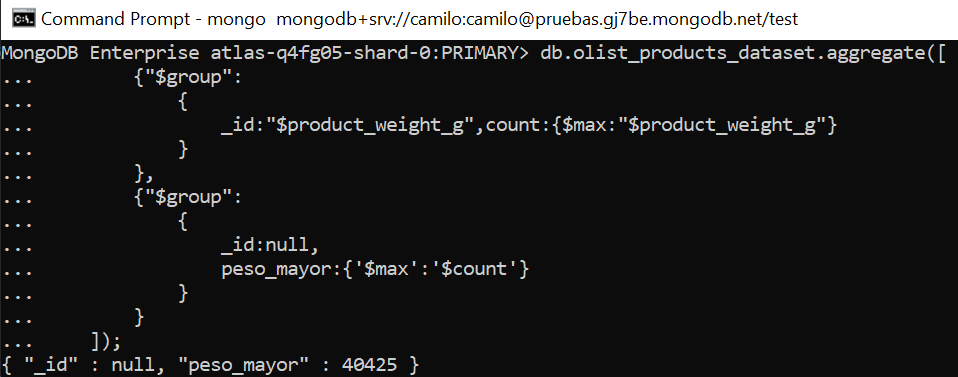
* **Mayor cantidad de fotos en un producto**

****

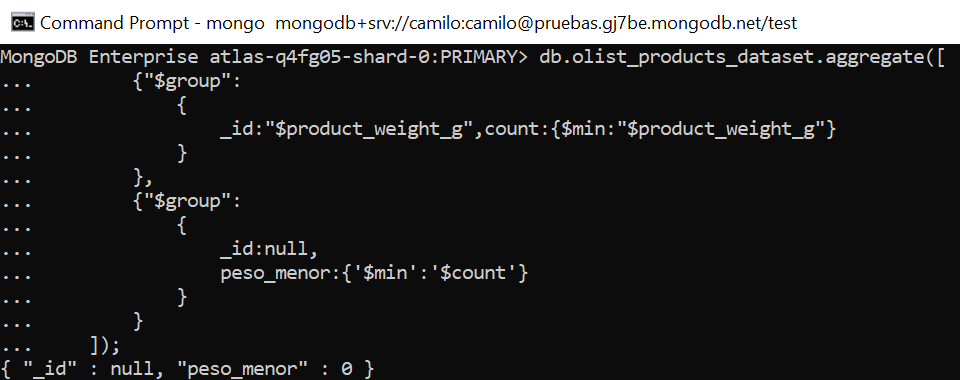
* **Promedio de fotos por producto**

****

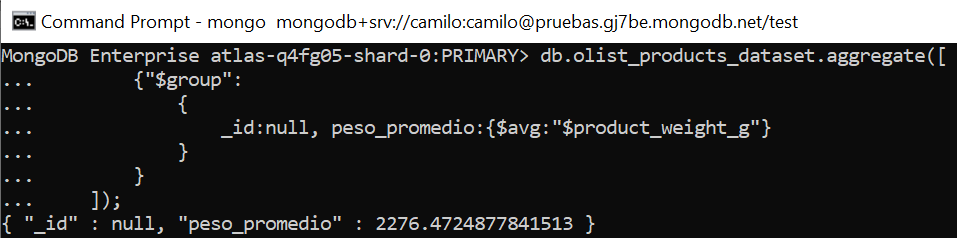
* **Peso mayor de un producto**

****

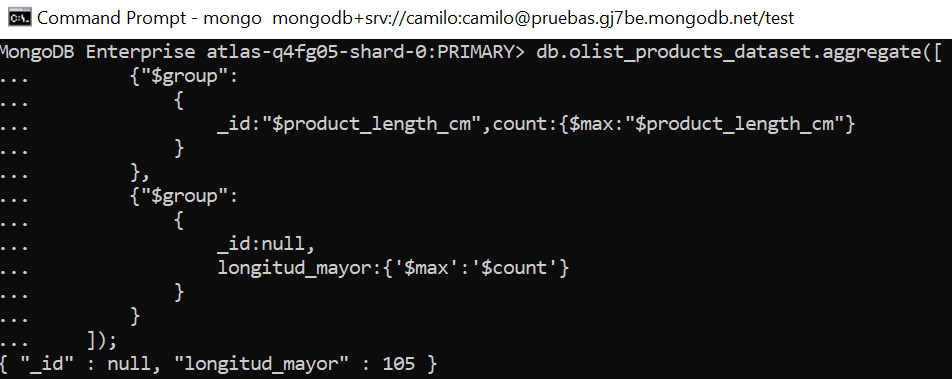
* **Peso menor de un producto (en gramos)**

****

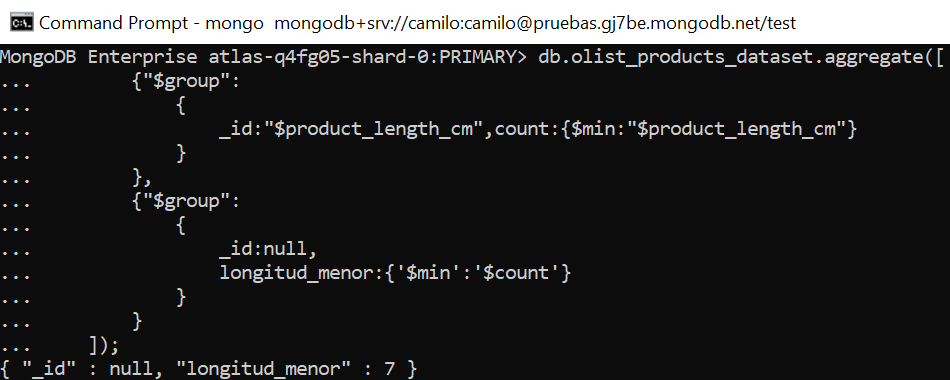
* **Peso promedio de los productos (en gramos)**

****

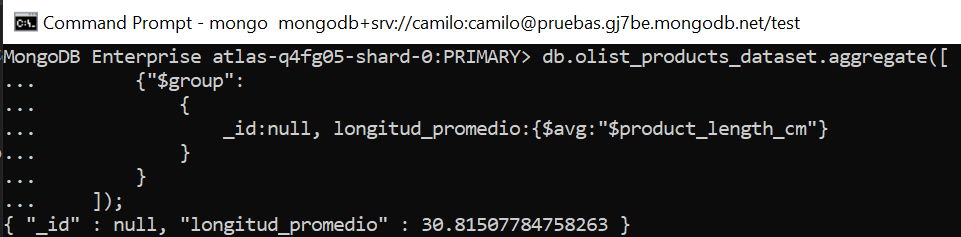
* **Longitud mayor de un producto (en cm)**

****

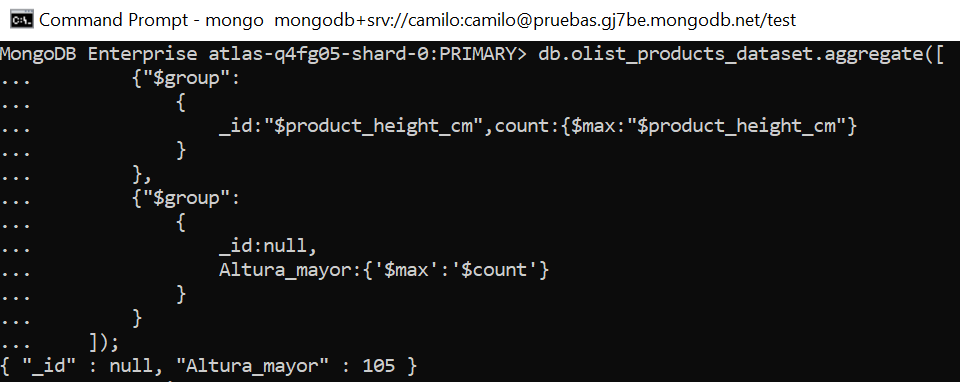
* **Longitud menor de un producto (en cm)**

****

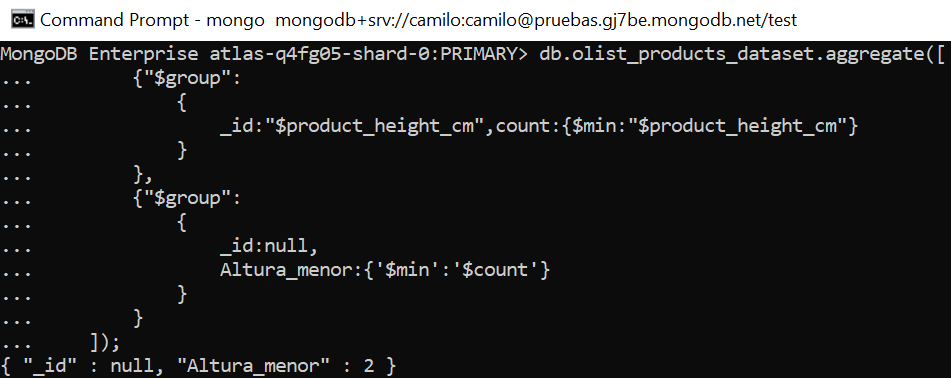
* **Longitud promedio de los productos**

****

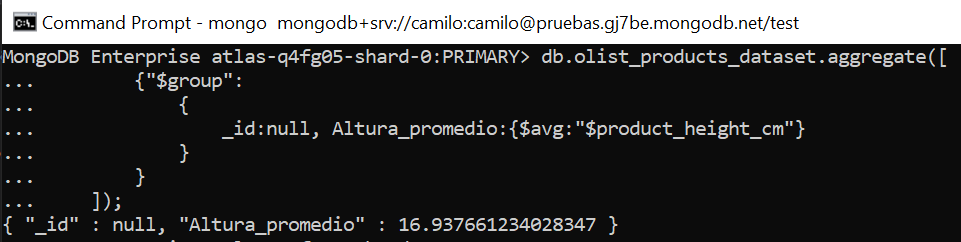
* **Altura máxima de un producto (en cm)**

****

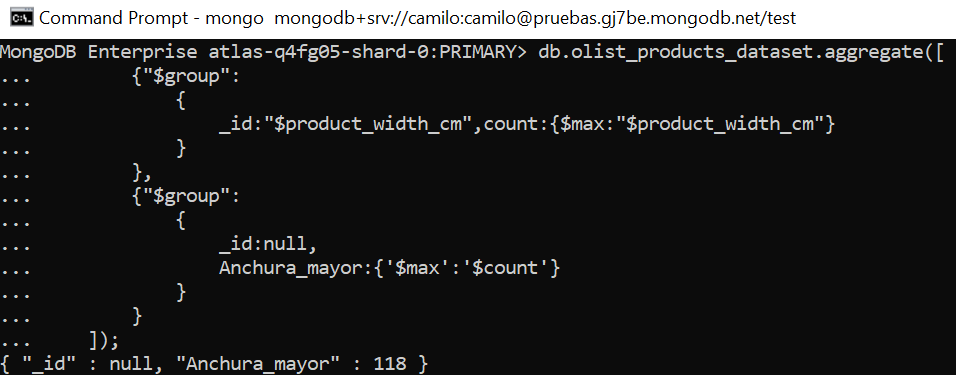
* **Altura mínima de un producto (en cm)**

****

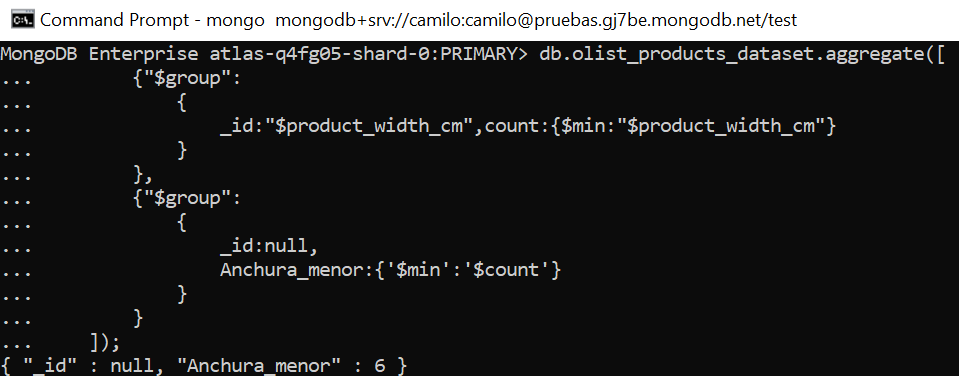
* **Altura promedio de los productos (en cm)**

****

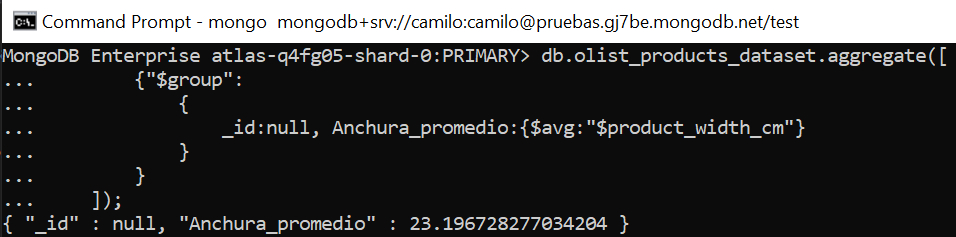
* **Ancho máximo de un producto (en cm)**

****

* **Ancho mínimo de un producto (en cm)**

****

* **Ancho promedio de los productos (en cm)**

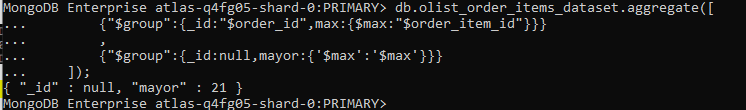
****

**Tabla 8: olist\_order\_items**

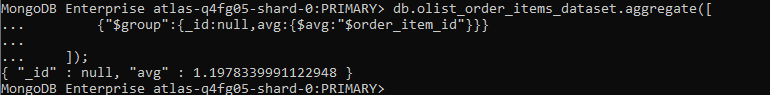
* **Cantidad total de order\_items registrados**

****

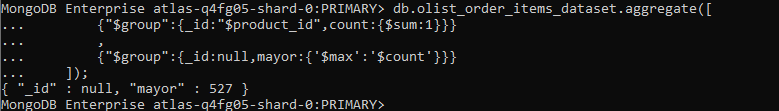
* **Orden con mayor cantidad de productos y cantidad de productos asociada**

****

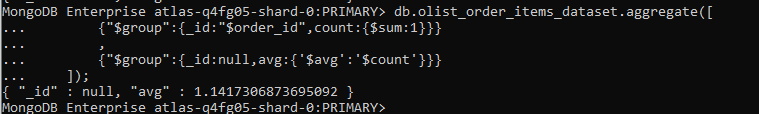
* **Número promedio de productos por orden**

****

* **Producto más vendido dentro de las órdenes y cantidad de órdenes en las que aparece**

****

* **Número promedio de productos por orden**

****

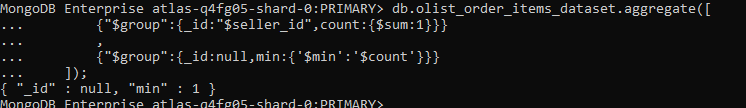
* **Número total de vendedores con al menos una orden**

****

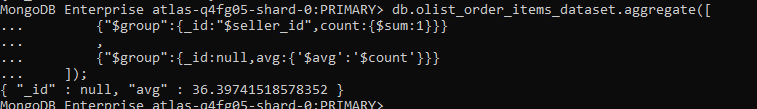
* **Número total de productos distintos**

****

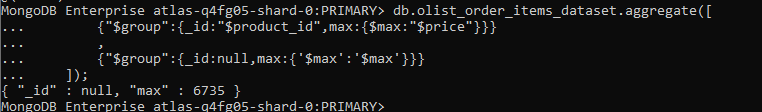
* **Vendedor con menor número de órdenes y cantidad de órdenes asociadas**

****

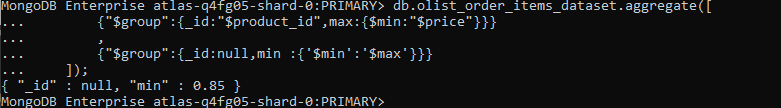
* **Número promedio de productos que vende cada vendedor**

****

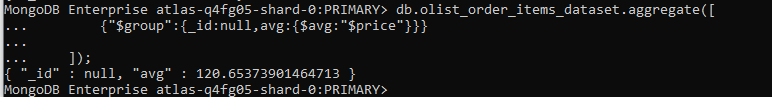
* **Producto más costoso y id asociado**

****

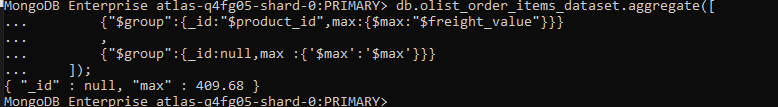
* **Producto más barato y id asociado**

****

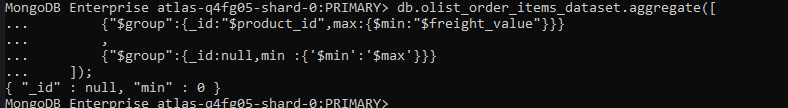
* **Precio promedio de los productos**



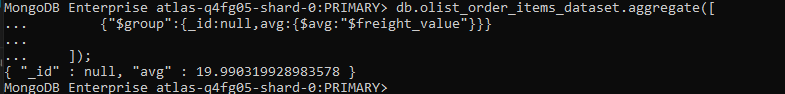
* **Costo de envío más alto y orden asociada**

****

* **Costo de envío más bajo y orden asociada**

****

* **Costo promedio de envío de los productos**

****

1. **Análisis de lectura**

***Observación****: Considerando el artículo: “*The Definitive Guide to Graph Databases for the RDBMS Developer*” de Neo4J. Compartido en las carpeta de lecturas recomendadas. Analice y responda cada pregunta en máximo 150 palabras:*

1. **¿Cuáles son las limitaciones, que se pueden inferir de la lectura, para migrar los conjuntos de datos relacionales a NoSQL?**

Las aplicaciones y los requisitos de los usuarios de hoy en día se han incrementado con el tiempo en cuanto a funcionalidades, volúmenes de información, performance y relaciones entre la información. Al validar estos requisitos en las bases de datos relacionales tradicionales, se pueden evidenciar inconvenientes relacionados con su tiempo de respuesta, rendimiento y cantidad de relaciones entre los datos, ya que las bases de datos relacionales no fueron diseñadas en un inicio para capturar y analizar las relaciones entre los datos, logrando que se pierdan conexiones críticas esenciales para las decisiones de hoy en día, basadas en datos. Por el contrario, las bases de datos tipo *graph* permiten almacenar billones de relaciones, acoplarse a las demandantes y cambiantes necesidades de los negocios, brindar ventaja competitiva a negocios que requieran datos altamente conectados, y donde el rendimiento es una prioridad, ya que se obtienen tiempos de respuesta en milisegundos, consiguiendo responder las preguntas que se acomodan a sus datos.

1. **¿Cuáles limitaciones adicionales se deben considerar, a parte de las mencionadas en el artículo?**

Se deben tener en cuenta algunas necesidades del mercado del mañana como lo pueden ser: La obtención de respuestas a partir de los datos en tiempo real, la diversidad en cuanto a los tipos de datos que se recolectan tales como: audios, videos o fotos, los cuales se requieren almacenar y que exigen que las tecnologías en las empresas se modernicen y capaciten a sus trabajadores en estos temas. A su vez, la mantenibilidad que se debe realizar a los sistemas de almacenamiento es un tema que se debe considerar para así tener disponibilidad 24/7 de estos sistemas y poder recopilar la mayor cantidad de datos posibles, así como también la búsqueda de algoritmos eficientes que sean capaces de brindar soluciones rápidas ante grandes volúmenes de información.

1. **¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?**

Para las empresas de hoy, los datos son el nuevo oro. Por lo cual, los negocios buscan aprovechar al máximo la consulta de estos para así obtener resultados que le brinden ventaja competitiva frente al mercado. Las bases de datos relacionales no sostienen las demandas y requerimientos de los negocios debido a su esquema rígido, y su nula capacidad de adaptarse a las metodologías ágiles que exigen constante cambio. Otro criterio es su inhabilidad de contemplar íntegra y completamente las relaciones que posee los datos. Asimismo, la complejidad al momento de realizar consultas SQL debido a la cantidad de líneas, lo que puede llevar a errores humanos, el rendimiento que se ve afectado por la cantidad de sentencias JOIN, y las consultas de ejecución lenta las cuales no respaldan las necesidades de las aplicaciones hacen que las empresas busquen migrar sus datos a mejores opciones que cumplan sus necesidades.