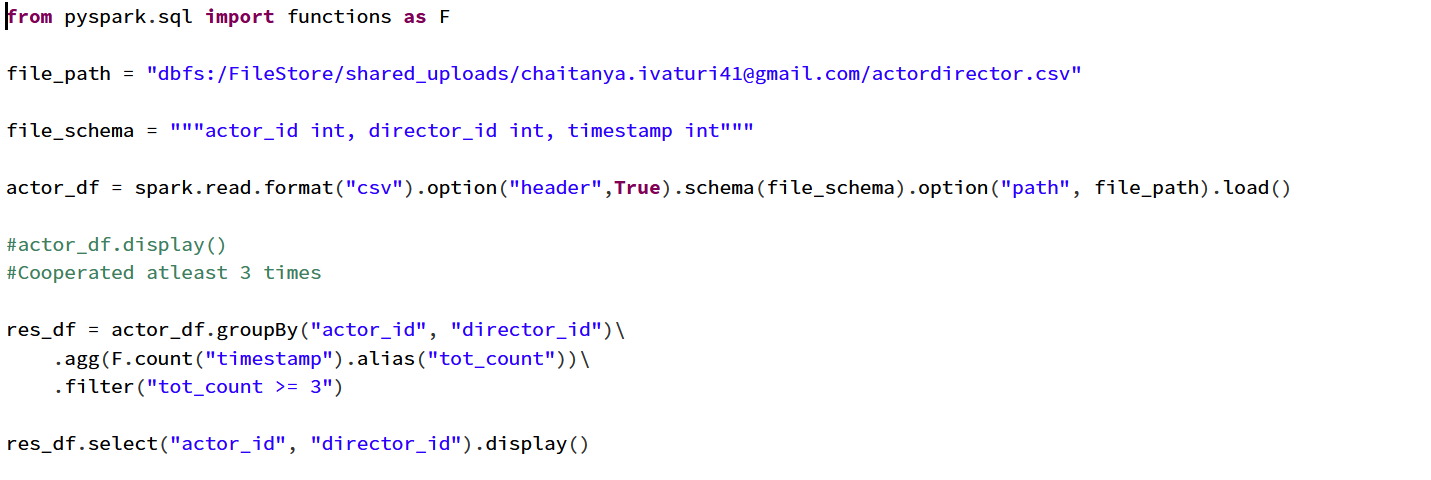
Easy questions

03 November 2022

20:57

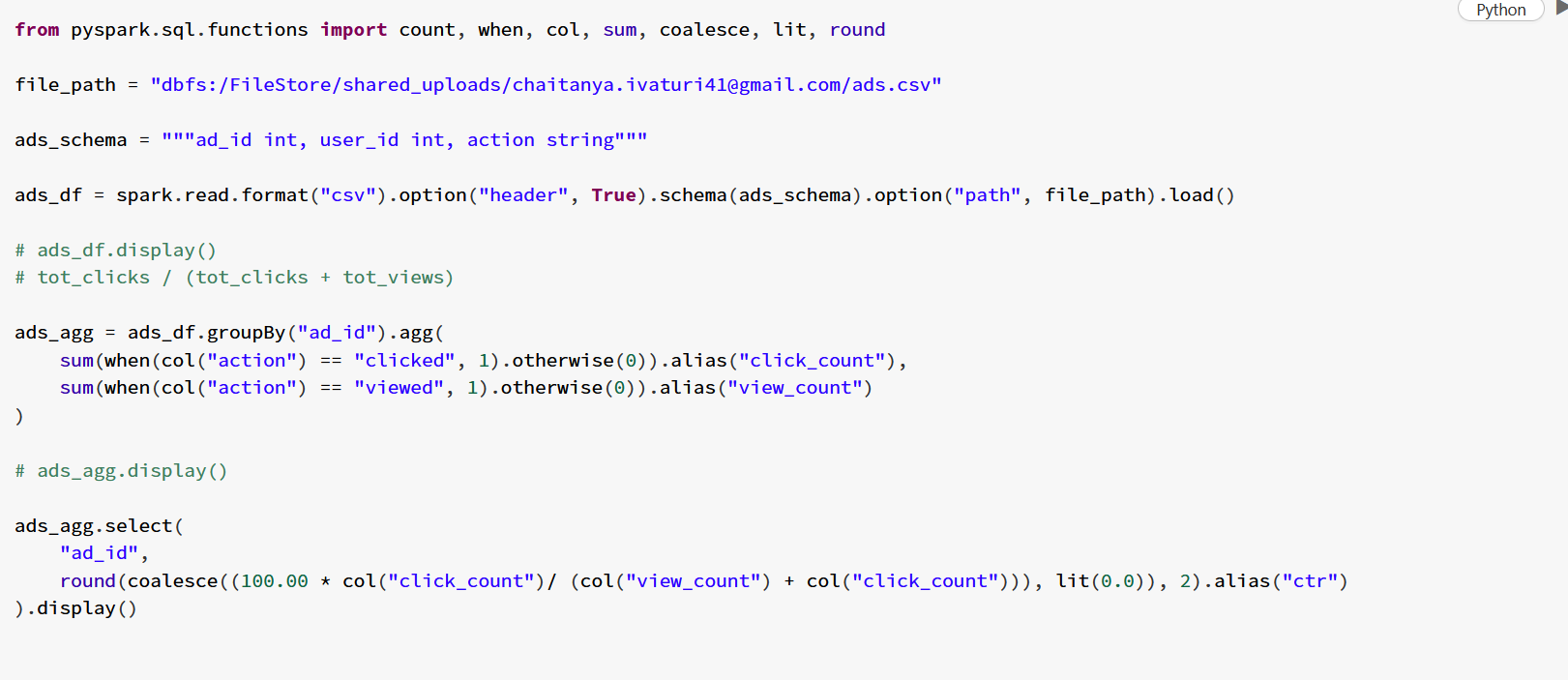
1. **Actors who cooperated with their directors**

**File**: dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/actordirector.csv



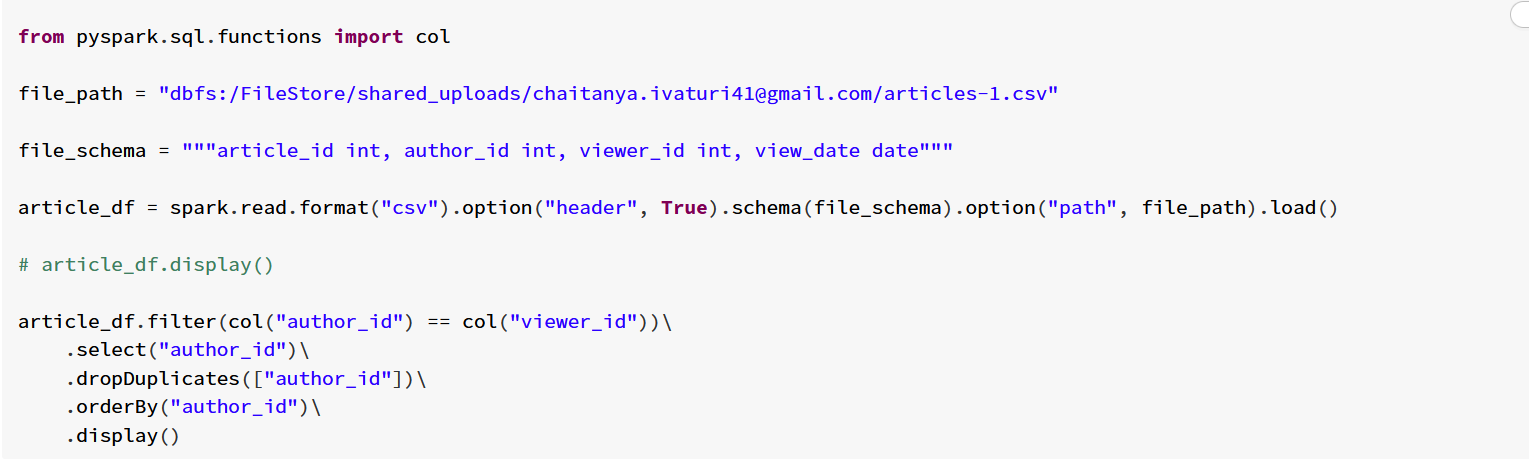
1. **Ads performance**

**File\_path =** dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/ads.csv



1. **Article views:**

**File\_path =** dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/articles-1.csv



* 1. **Average Selling Price**

dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/unitsSold.csv

dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/Prices.csv

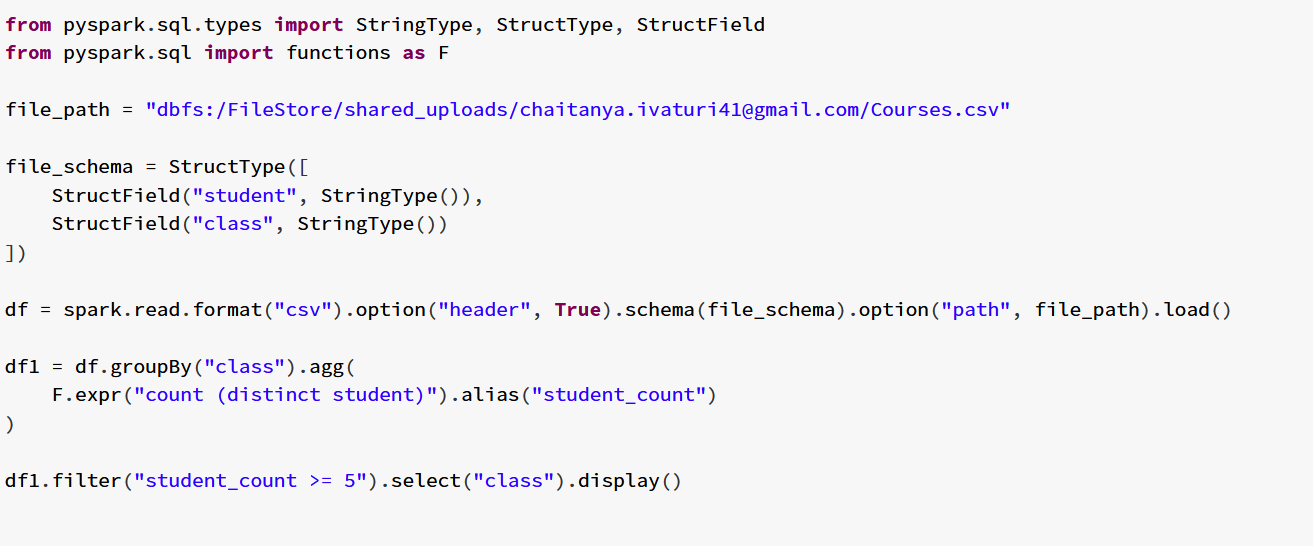


* 1. **Biggest Number**

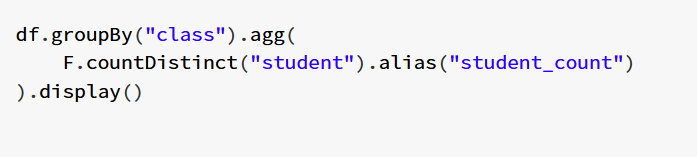


* 1. **Class more than 5 students**

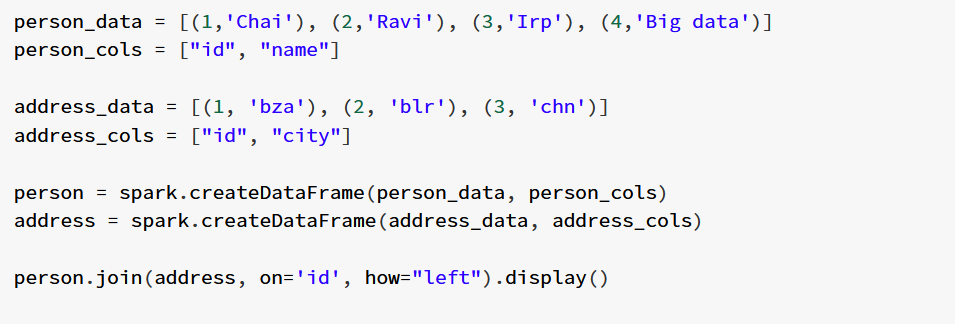
file\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/Courses.csv"



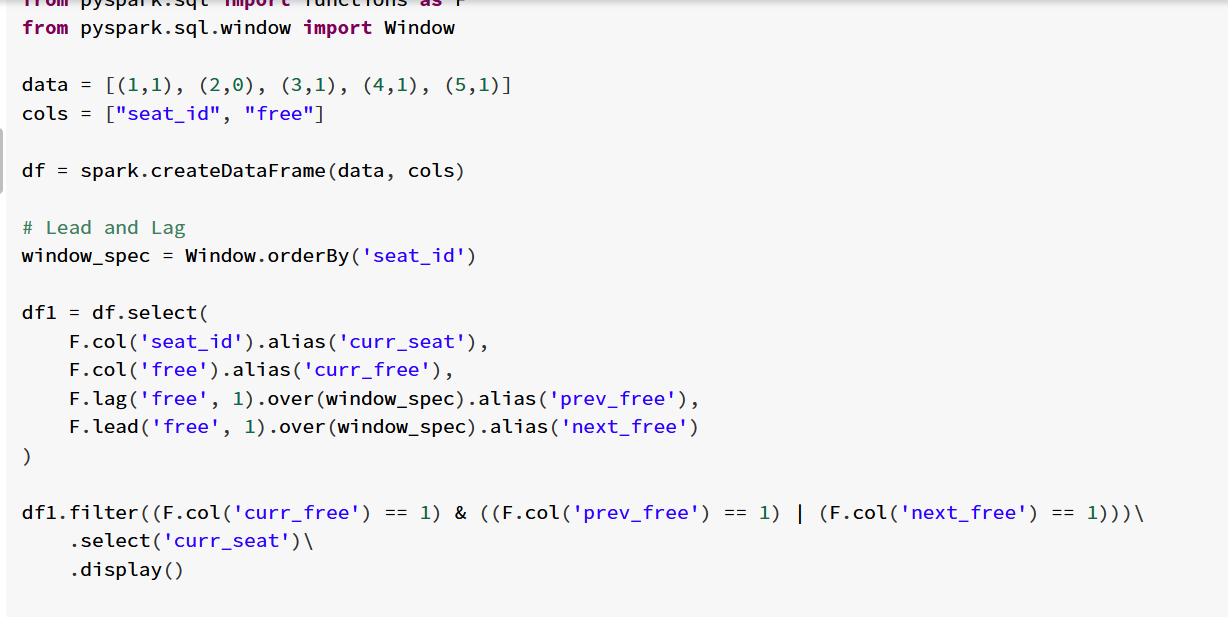
Or you can use countDistinct()



1. **Combining 2 tables:**



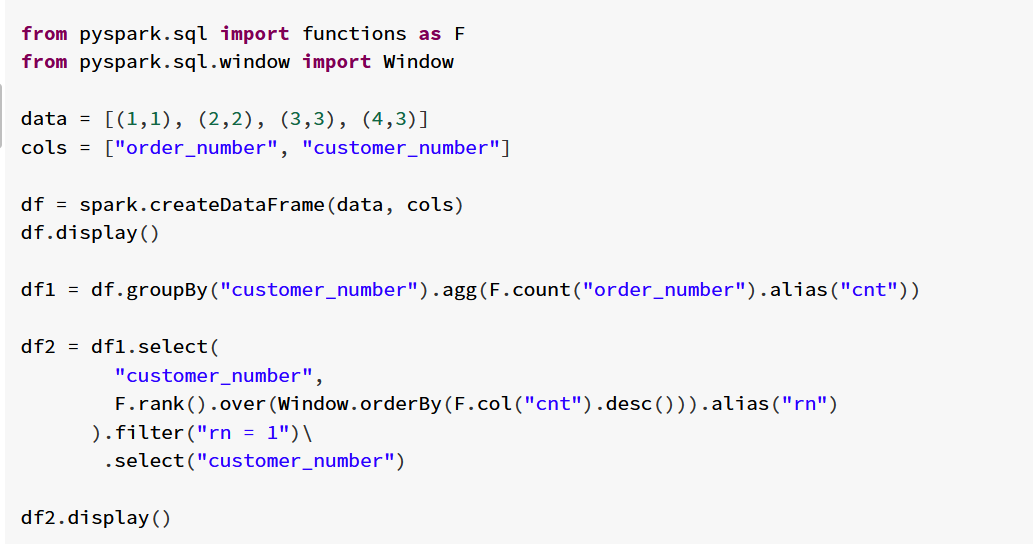
* 1. **Consecutive available seats**



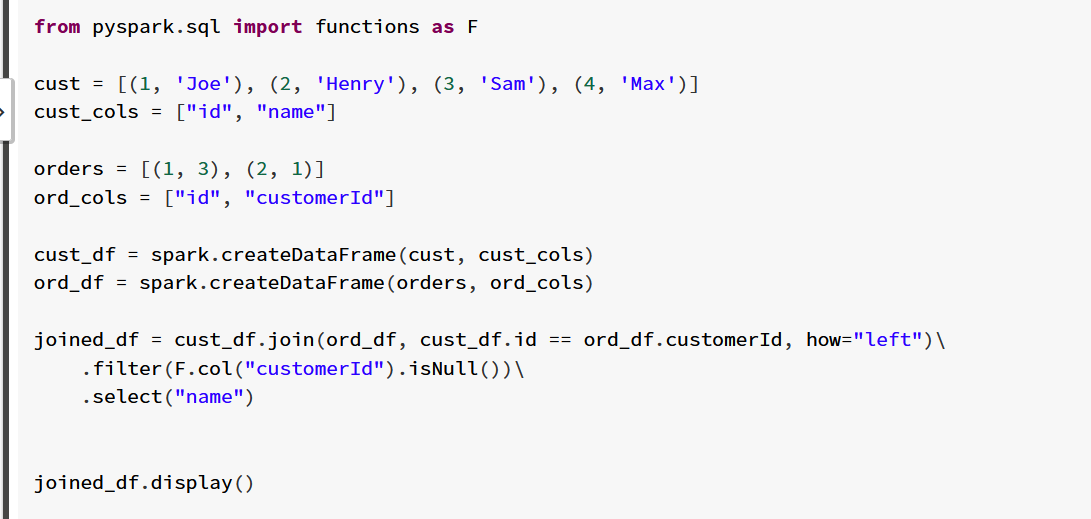
* 1. Session chart



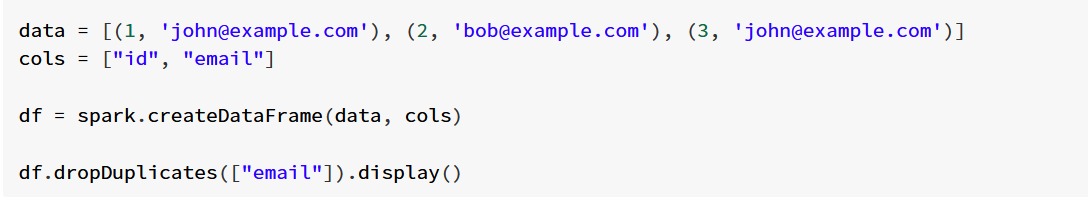
* 1. Largest number of orders



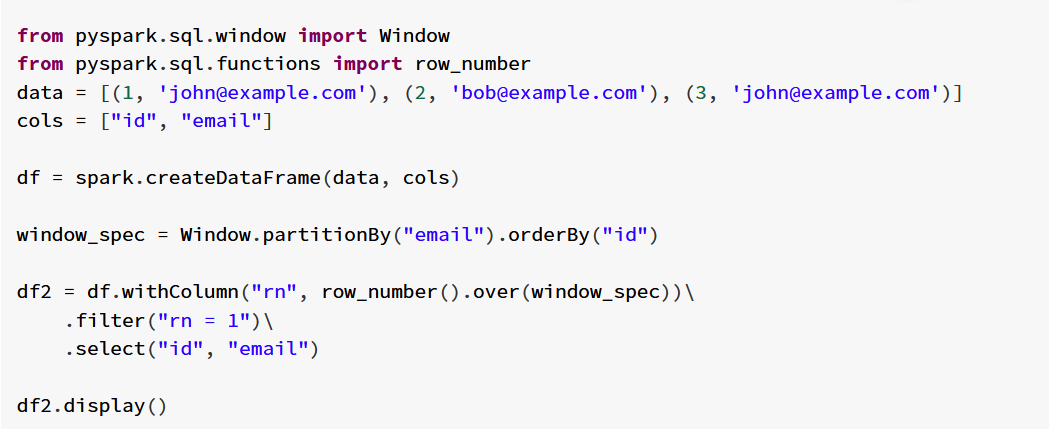
* 1. Customers who never order



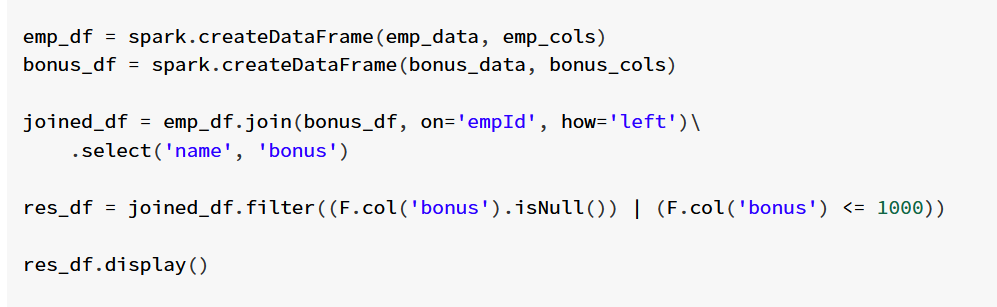
* 1. Drop duplicate emails



Using row\_number:



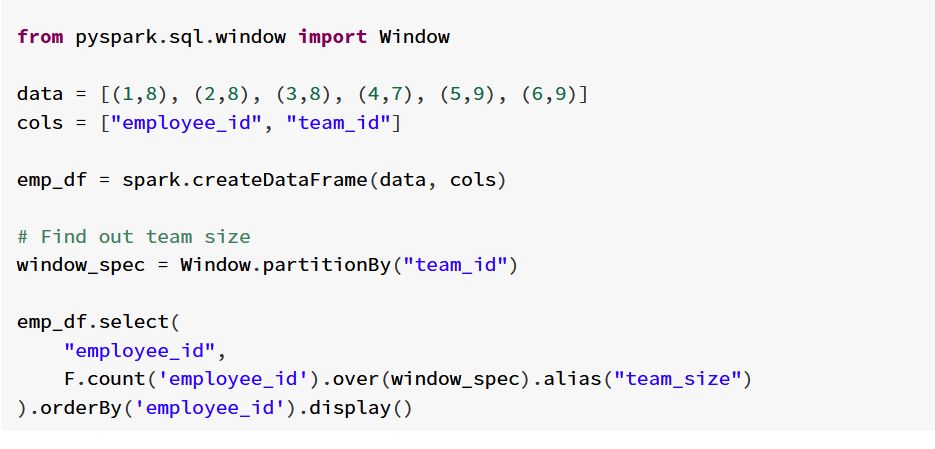
* 1. Employee Bonus



* 1. Employees salary more than managers



* 1. Team size



* 1. Friend Requests 1

friend\_request = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/friend\_request.csv"

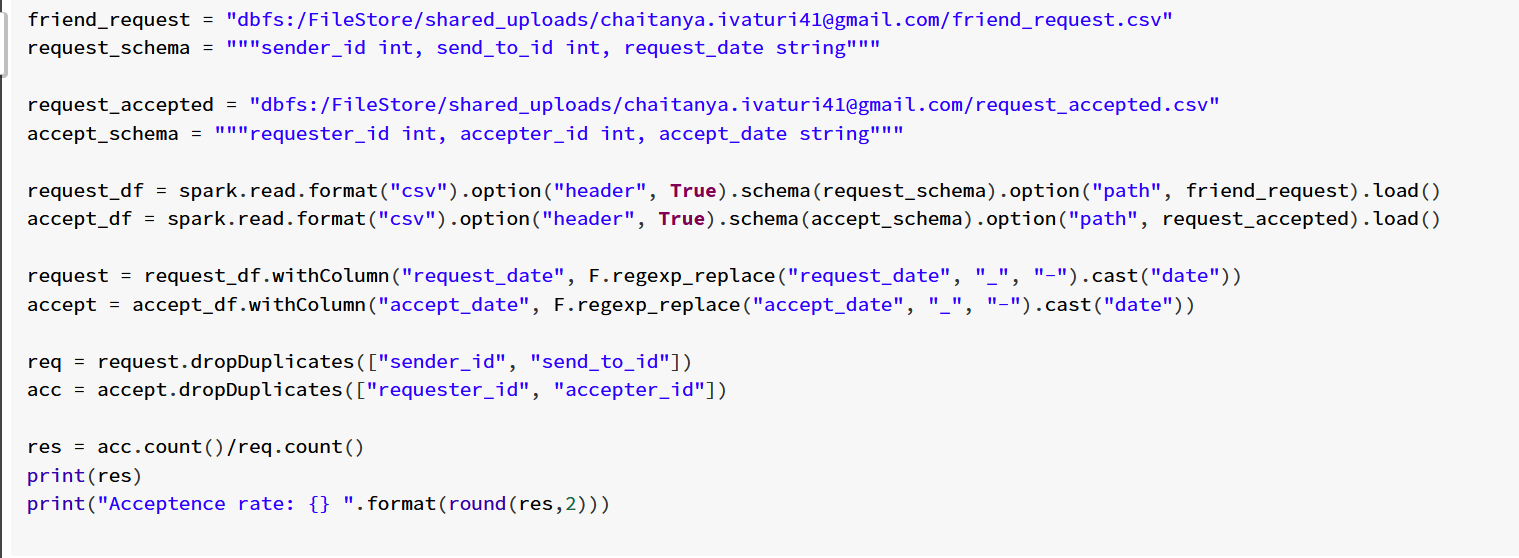
request\_schema = """sender\_id int, send\_to\_id int, request\_date string"""

request\_accepted = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/request\_accepted.csv"

accept\_schema = """requester\_id int, accepter\_id int, accept\_date string"""



Triggering 2 actions:



* 1. **Friendly streamed movies**

tv\_program\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/TVProgram.csv"

tv\_schema = """program\_date string, content\_id int, channel string"""

content\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/content.csv"

content\_schema = """content\_id int, title string, kids\_content string, content\_type string"""



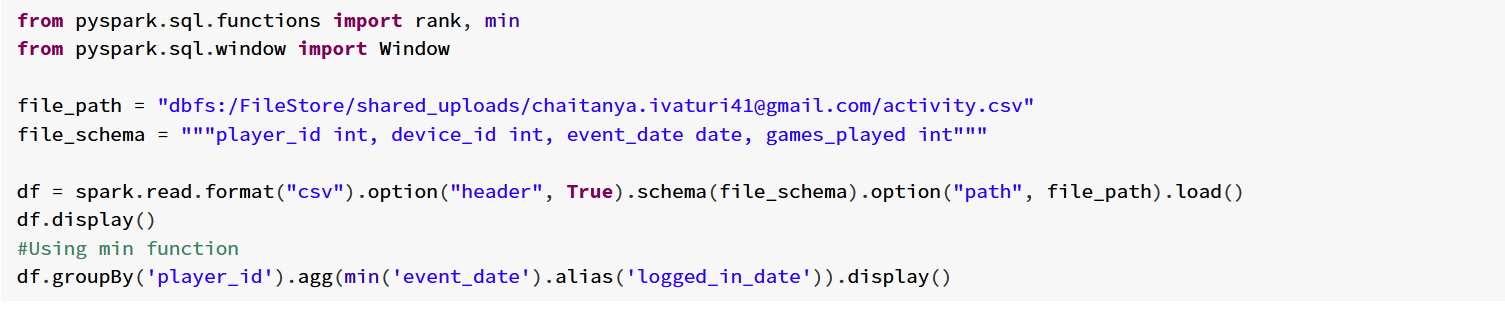
* 1. Game play Analysis 1

file\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/activity.csv"

file\_schema = """player\_id int, device\_id int, event\_date date, games\_played int"""



# Using groupBy



1. Game play analysis 2

Continuation of Game play analysis 1 - Same file



* 1. Group sold product

file\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/Activities.csv"

file\_schema = "sell\_date date, product string"

**Note: Collect list doesn’t guarantee order. It is a non-deterministic function**

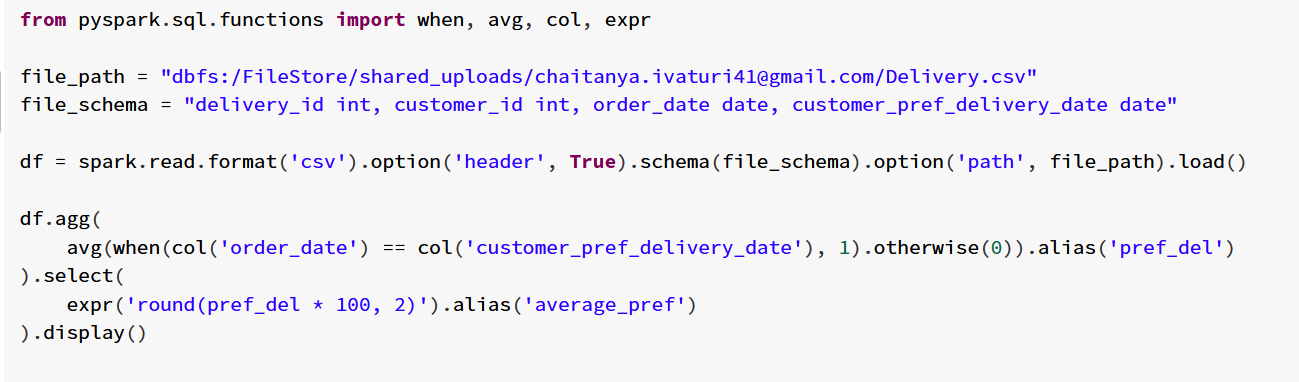
**Need to check how can I achieve order using collect list**



* 1. Immediate food delivery

file\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/Delivery.csv"

file\_schema = "delivery\_id int, customer\_id int, order\_date date, customer\_pref\_delivery\_date date"



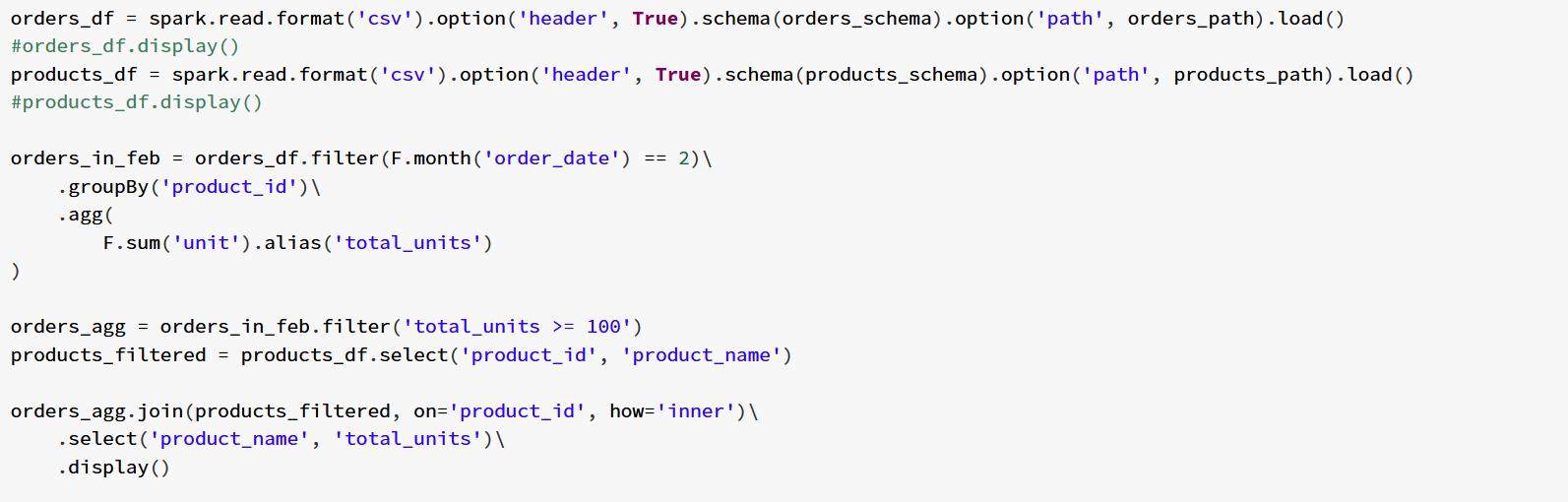
* 1. Products ordered in a period

orders\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/orders-3.csv"

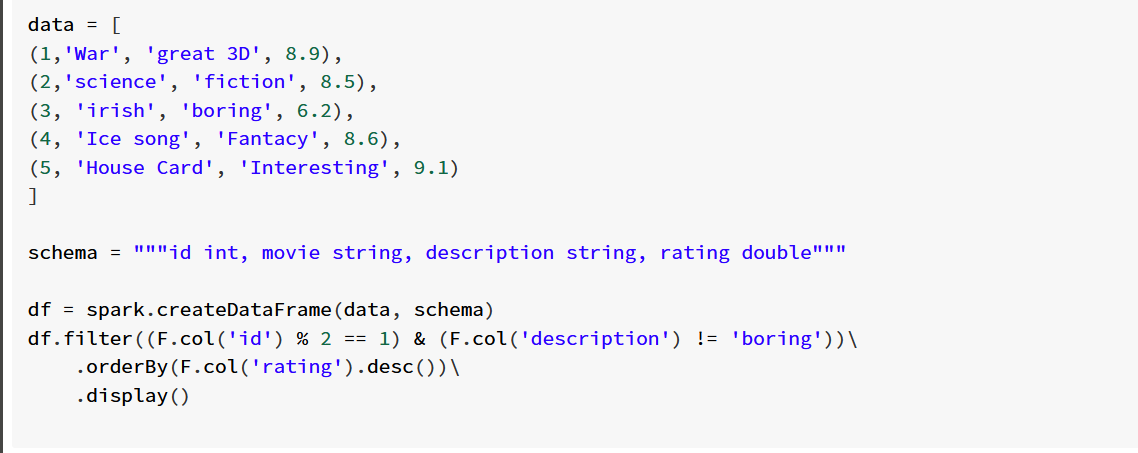
orders\_schema = """product\_id int, order\_date date, unit int"""

products\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/products.csv"

products\_schema = """product\_id int, product\_name string, product\_category string"""



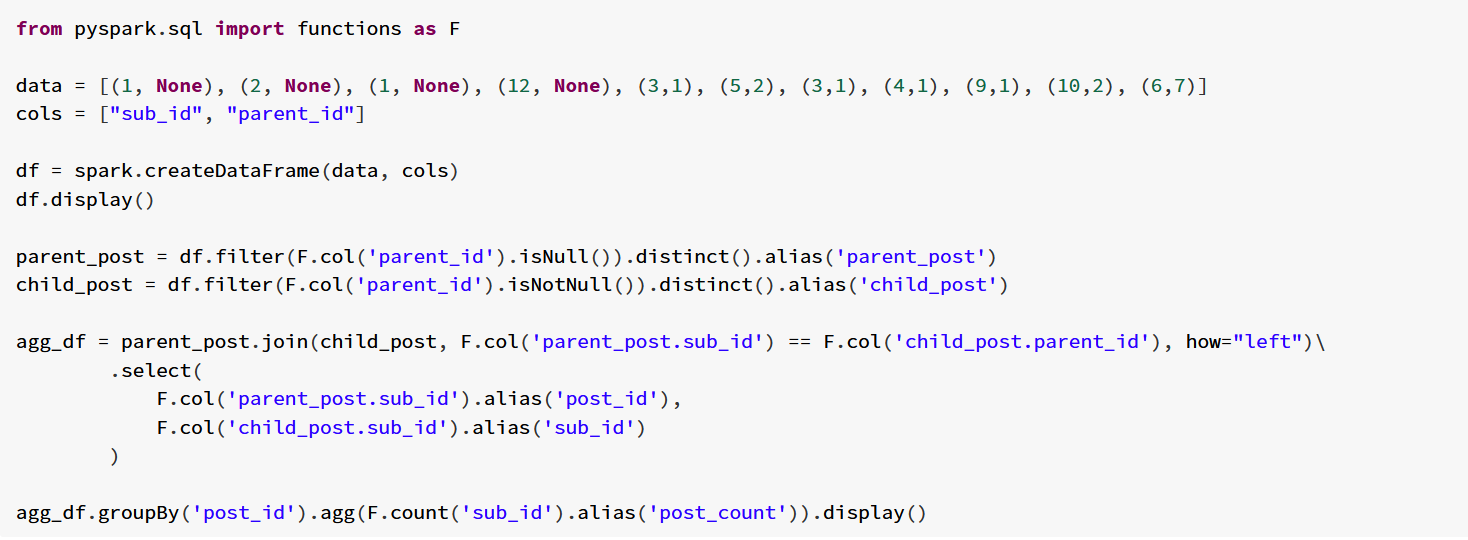
* 1. Not boring movies

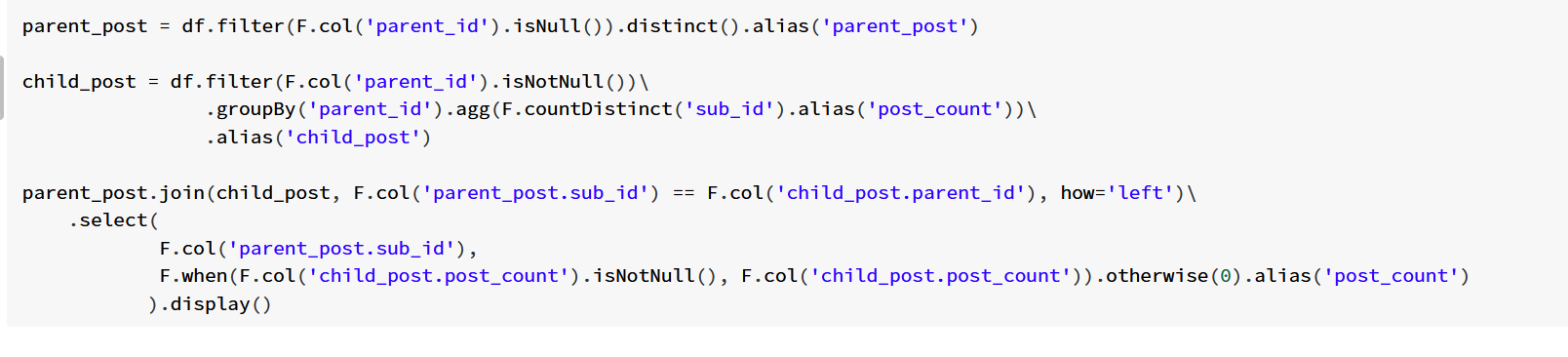


* 1. Number of comments

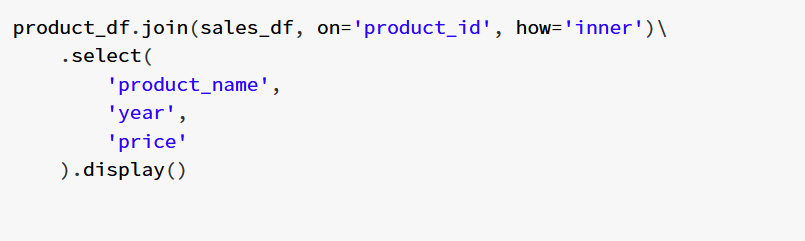
data = [(1, None), (2, None), (1, None), (12, None), (3,1), (5,2), (3,1), (4,1), (9,1), (10,2), (6,7)]

cols = ["sub\_id", "parent\_id"]

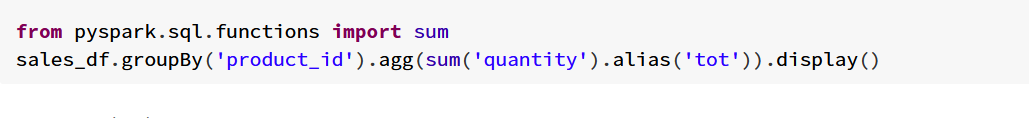




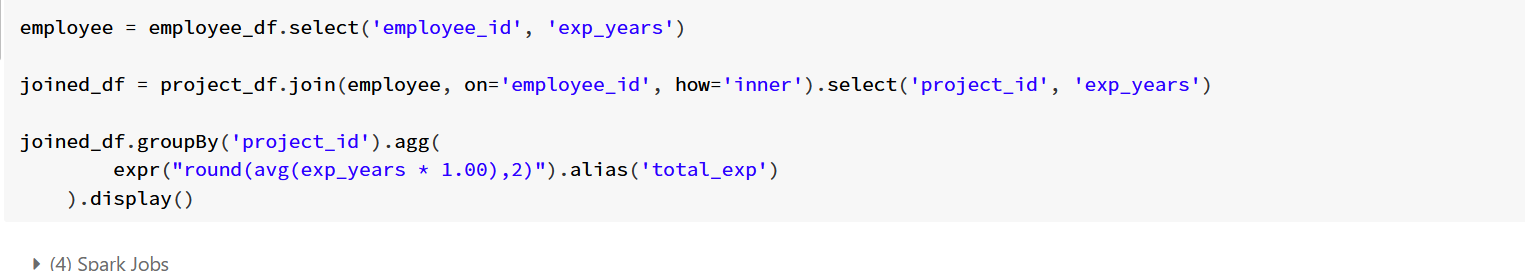
* 1. Product sales analysis 1



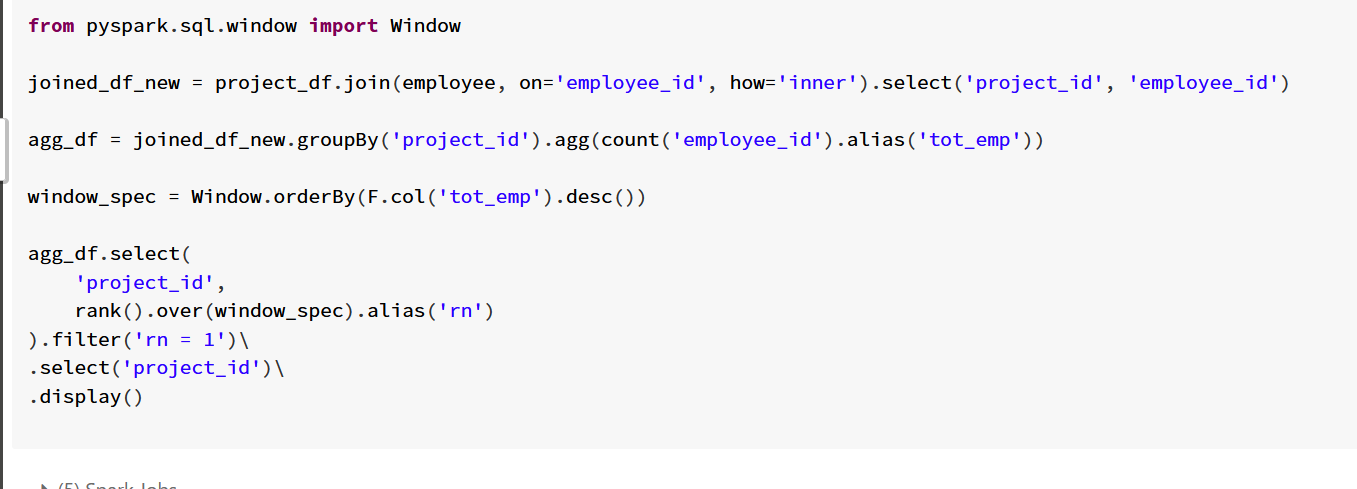
* 1. Product sales analysis 2



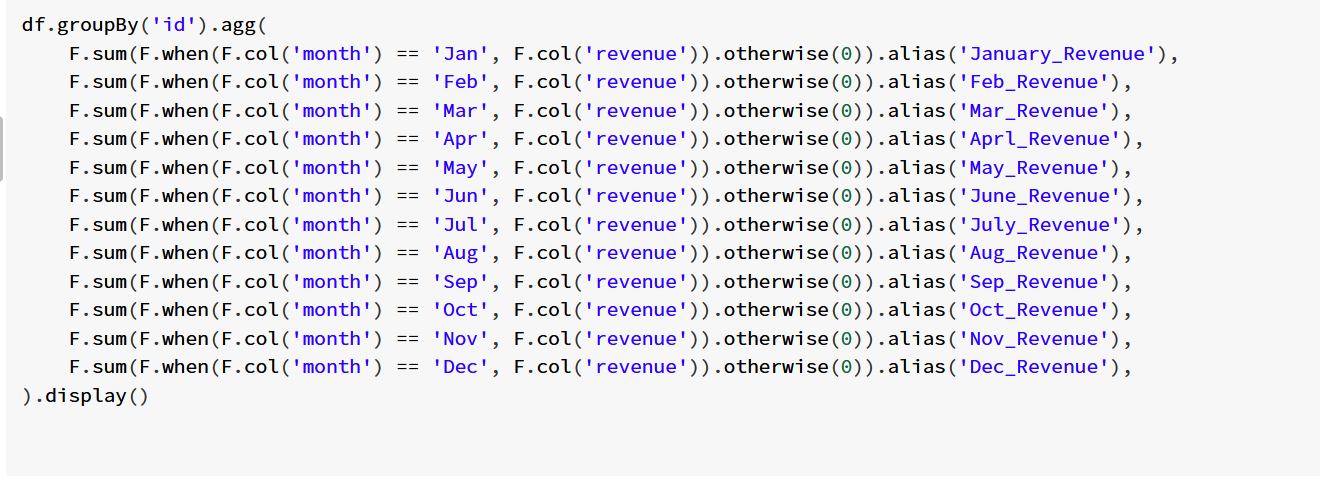
* 1. Project Employees 1



* 1. Project Employees 2



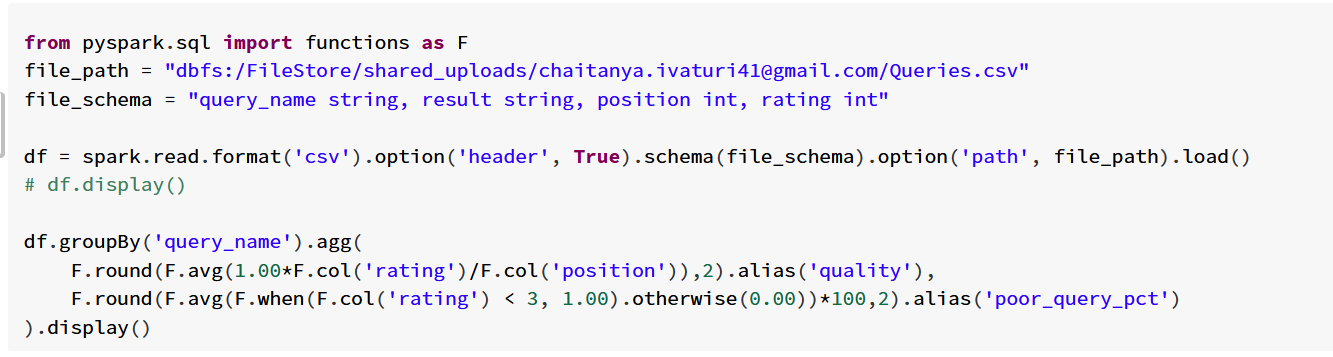
* 1. Reformat Department



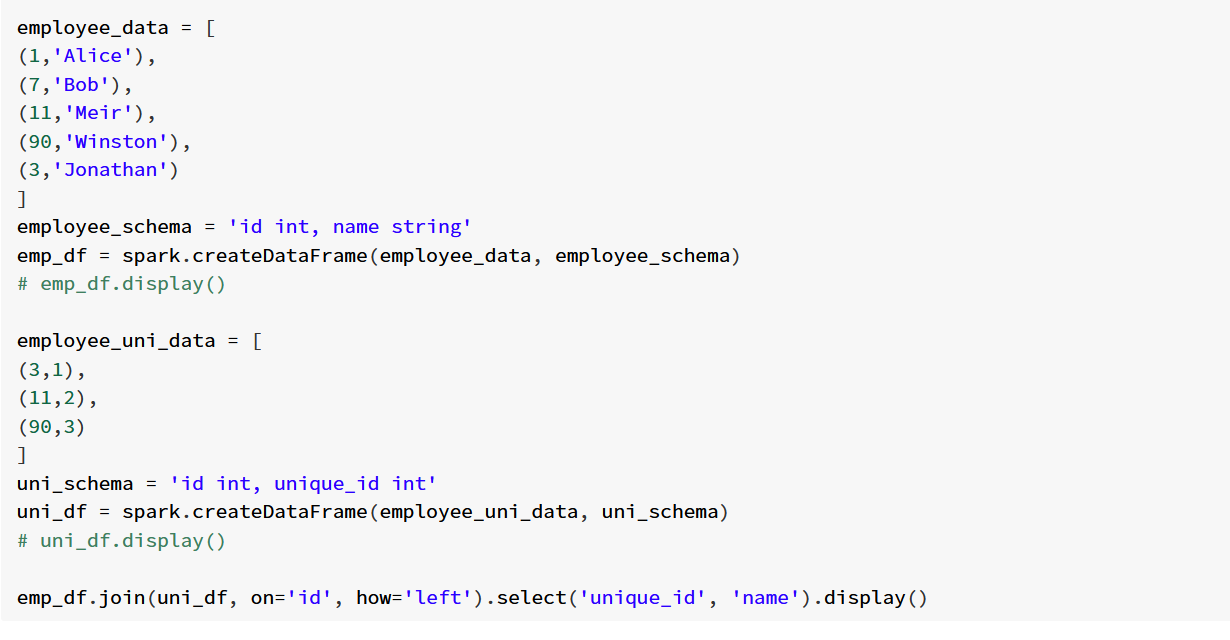
* 1. Query quality and percentage

file\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/Queries.csv"

file\_schema = "query\_name string, result string, position int, rating int"



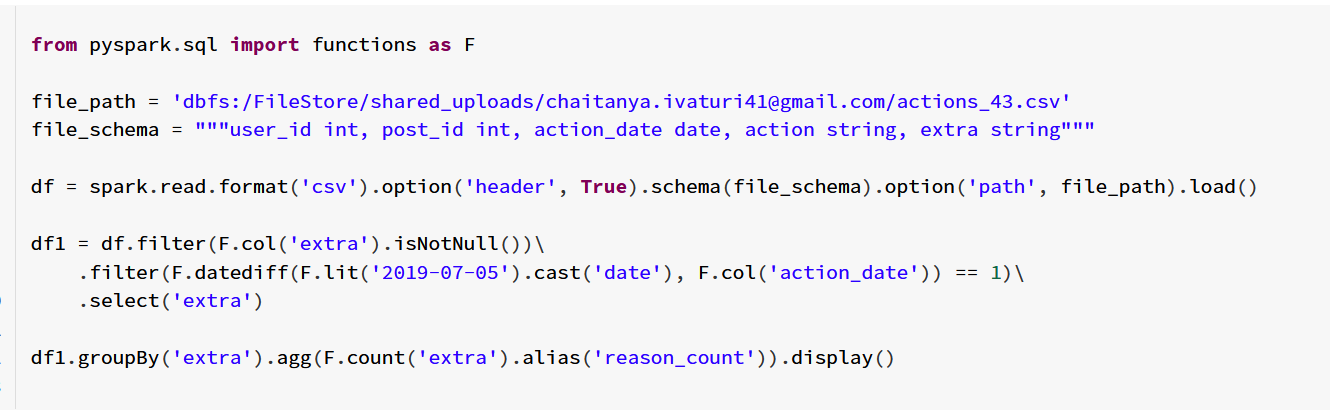
* 1. Employee unique identifier



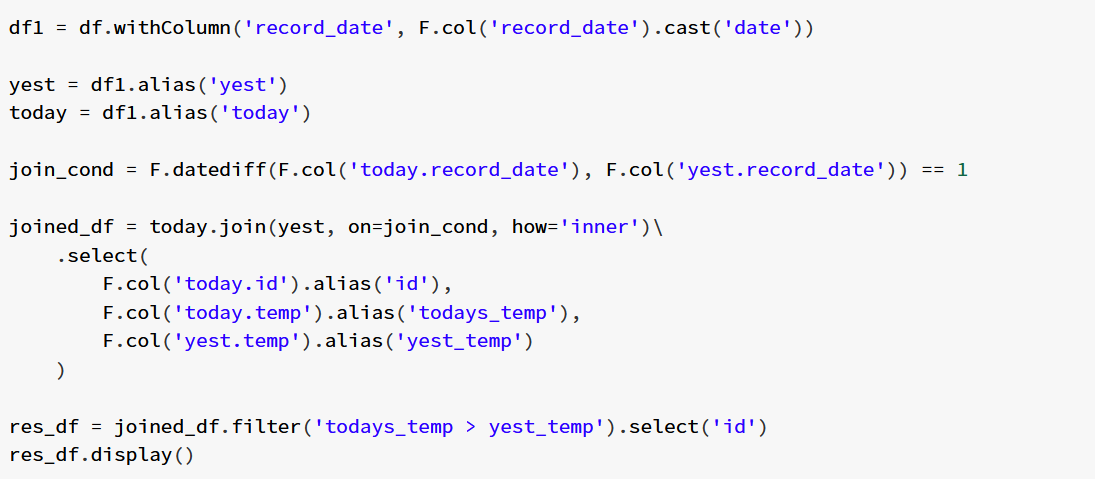
* 1. Reported posts

file\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/actions\_43.csv'

file\_schema = """user\_id int, post\_id int, action\_date date, action string, extra string"""



* 1. Rising Temperature



* 1. Sales analysis 1 - **Applying window over entire dataframe will cause data to move to single partition. How to tackle this?**

product\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/product.csv'

product\_schema = '''product\_id int, product\_name string, unitprice int'''

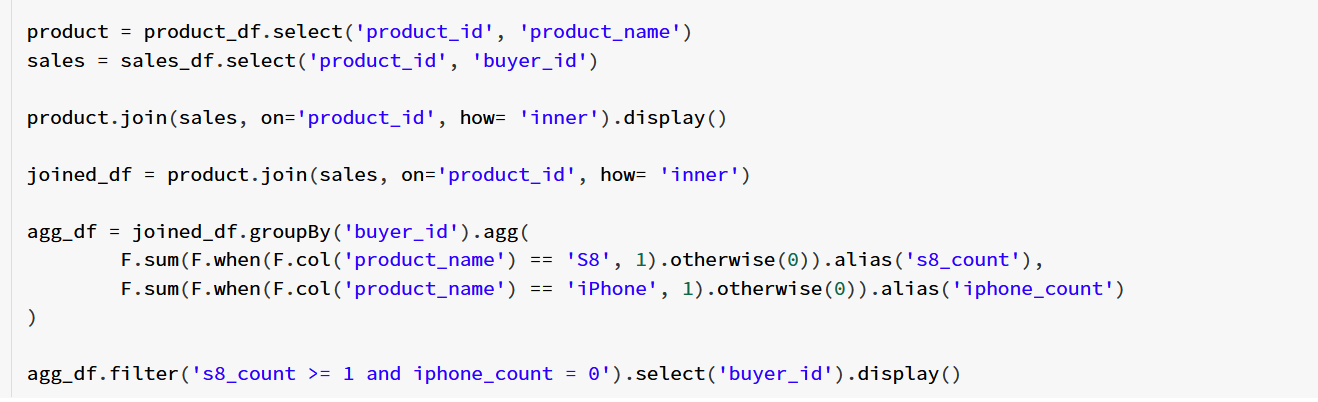
sales\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/sales.csv'

sales\_schema = '''seller\_id int, product\_id int, buyer\_id int, sale\_date date, quantity int, price int'''

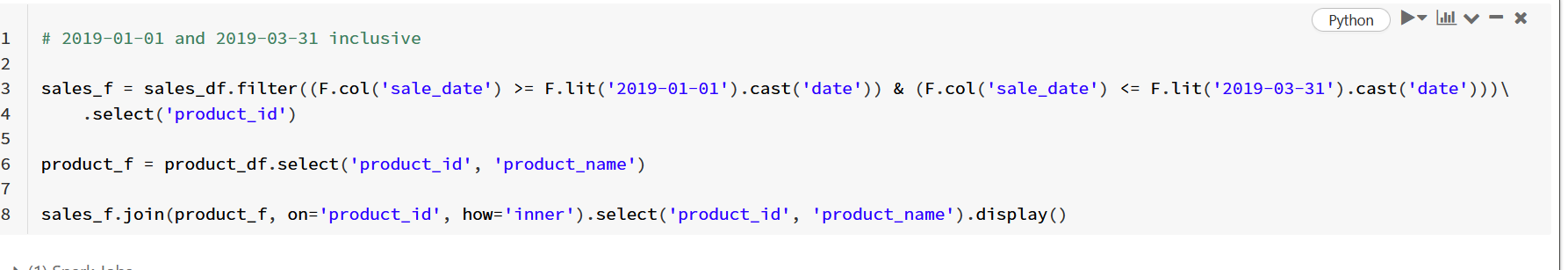


* 1. Sales Analysis 2

Dataset - sales-3



* 1. Sales Analysis 3



* 1. Sales person

sales\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/salesperson-1'

sales\_schema = '''sales\_id int, name string, salary int, commission\_rate int, hire\_date string'''

sales\_df = spark.read.format('csv').option('header', True).schema(sales\_schema).option('path', sales\_path).load()

company\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/company'

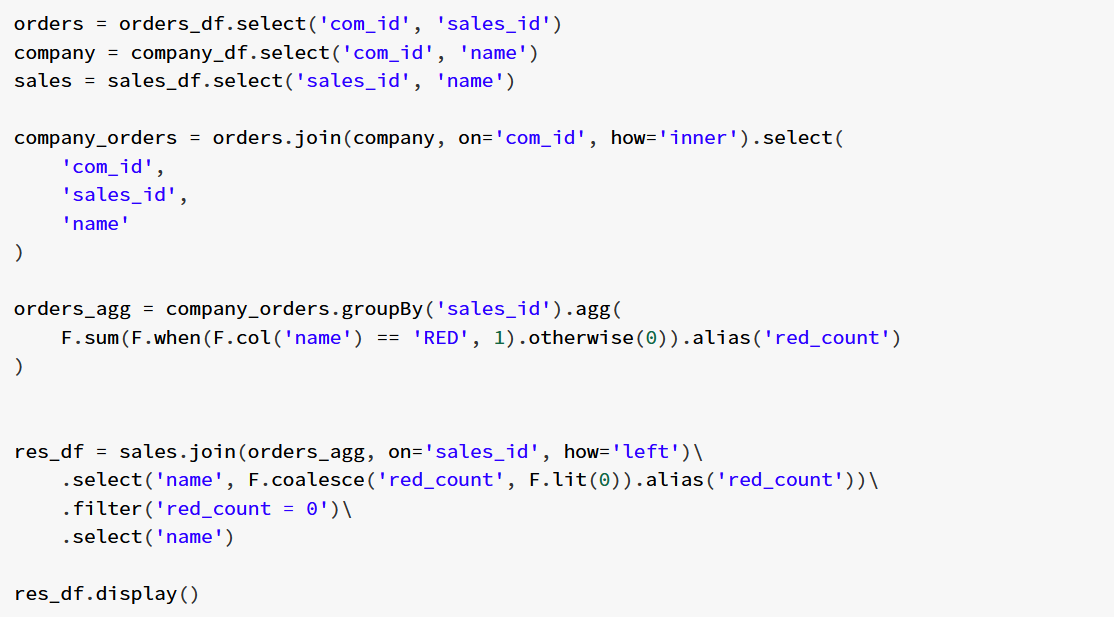
company\_schema = '''com\_id int, name string, city string'''

company\_df = spark.read.format('csv').option('header', True).schema(company\_schema).option('path', company\_path).load()

orders\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/orders\_12.csv'

orders\_schema = '''order\_id int, order\_date string, com\_id int, sales\_id int, amount int'''

orders\_df = spark.read.format('csv').option('header', True).schema(orders\_schema).option('path', orders\_path).load()

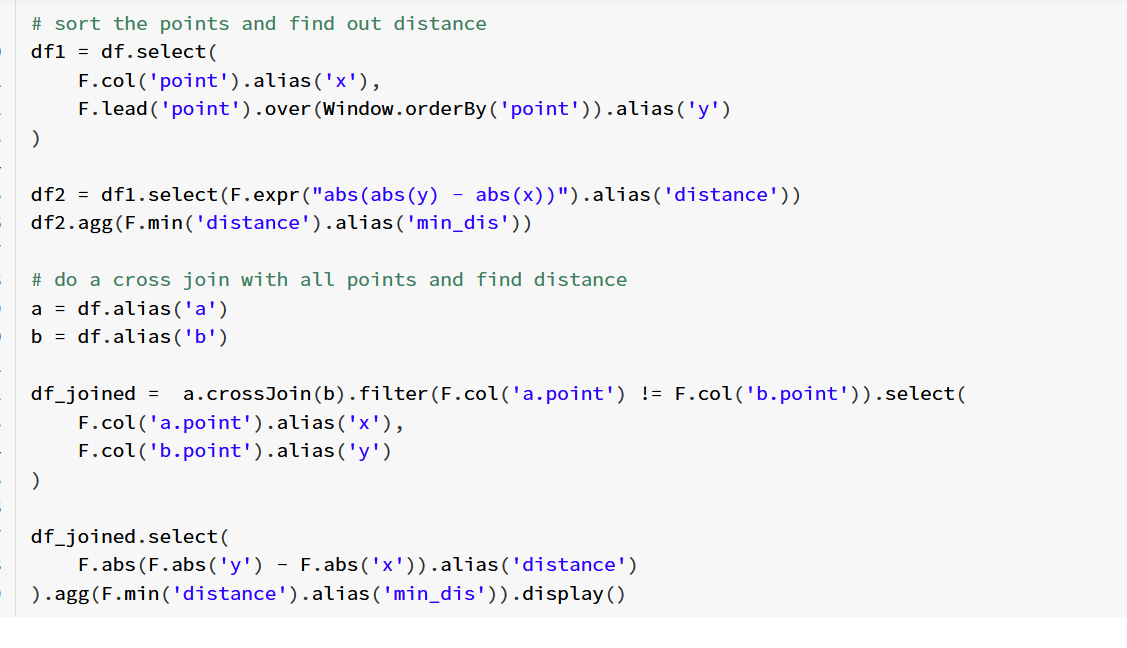


* 1. Second highest salary

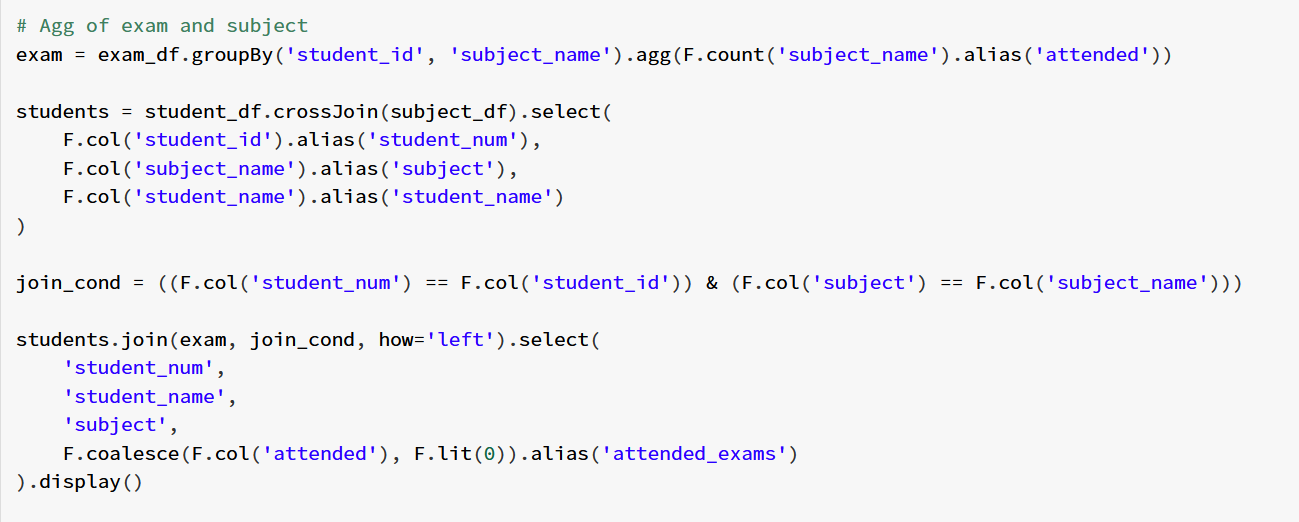


* 1. Shortest Distance

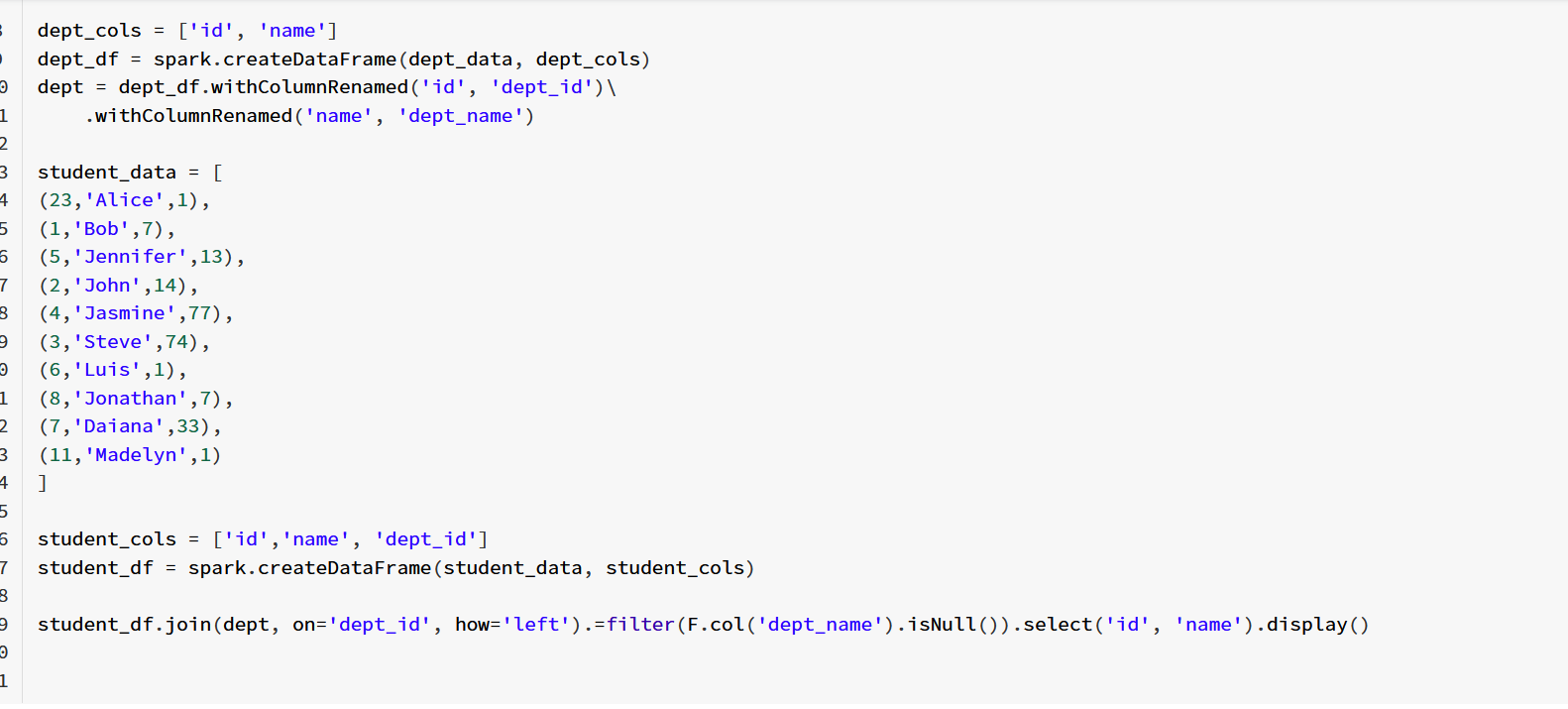
Which is best ? Sort or cross join



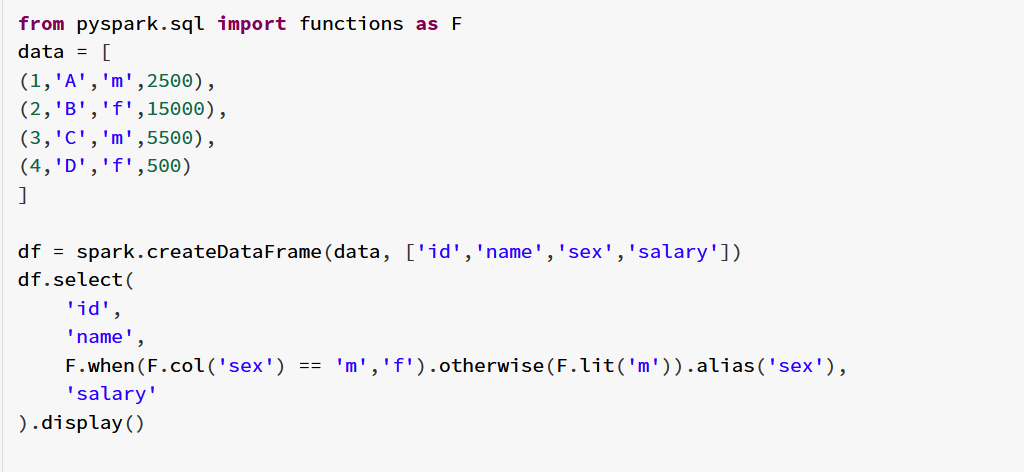
* 1. Student and Examinations



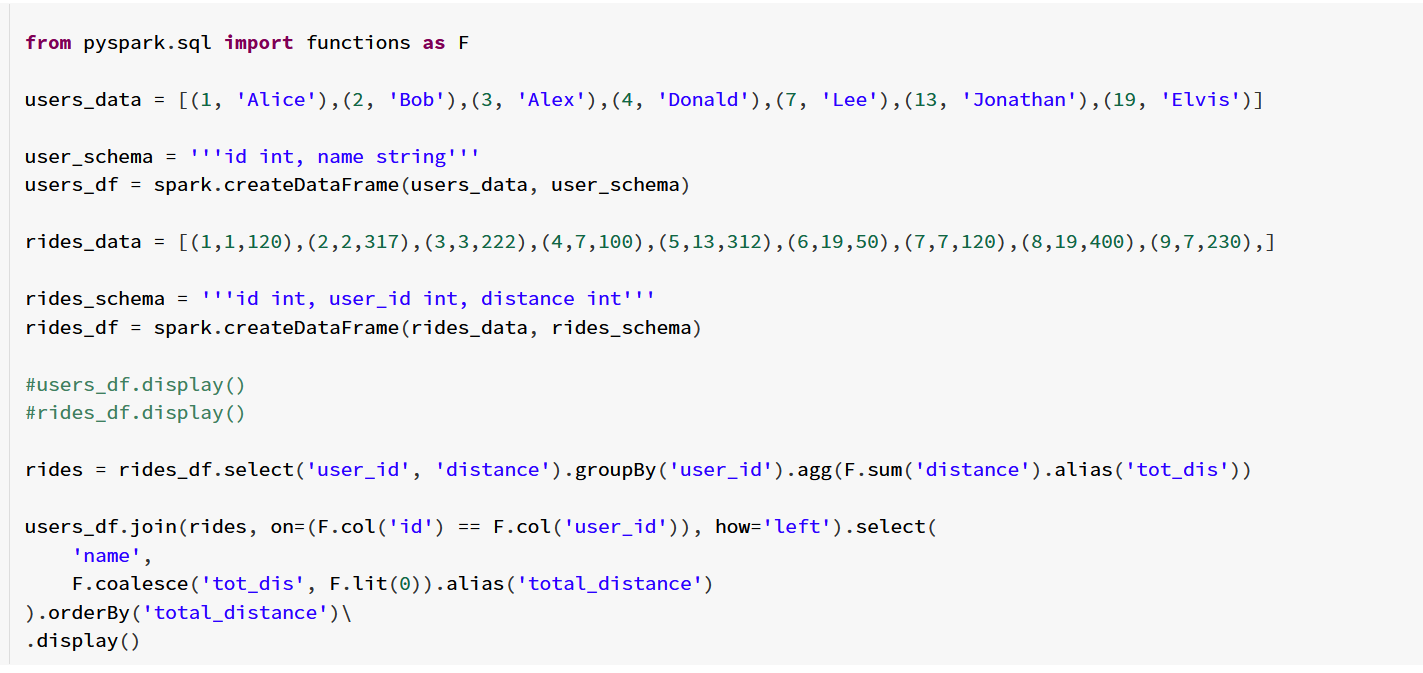
* 1. Invalid Departments



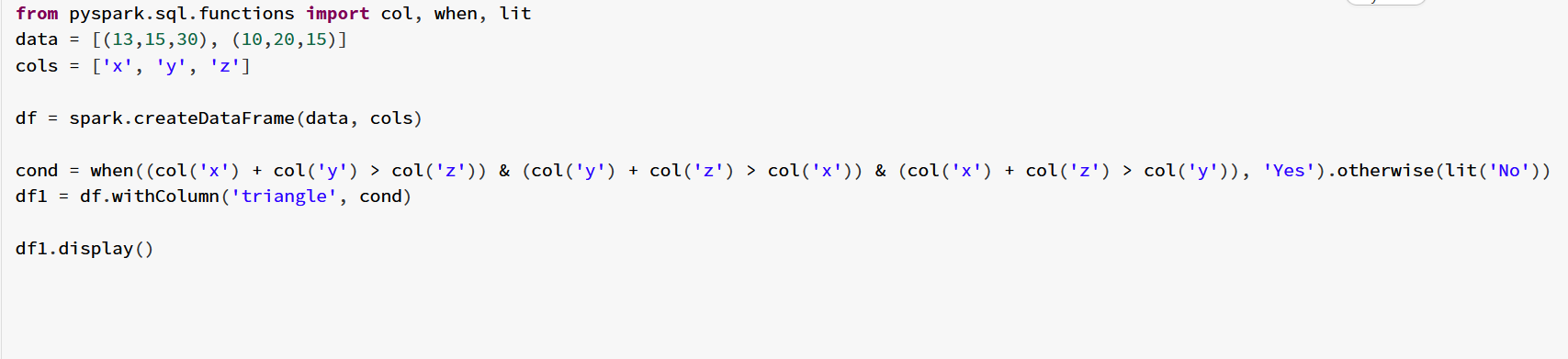
* 1. Swap salary



* 1. Top Travellers



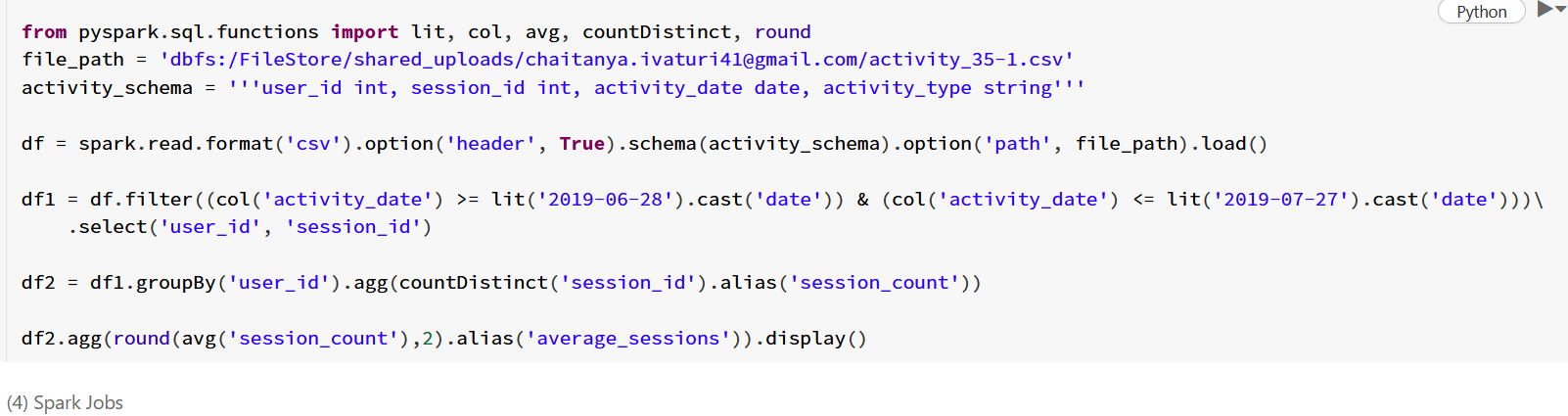
* 1. Triangle judgement



* 1. User activity in 30 days 2 - **Any other way to optimize this? Window or groupBy?**

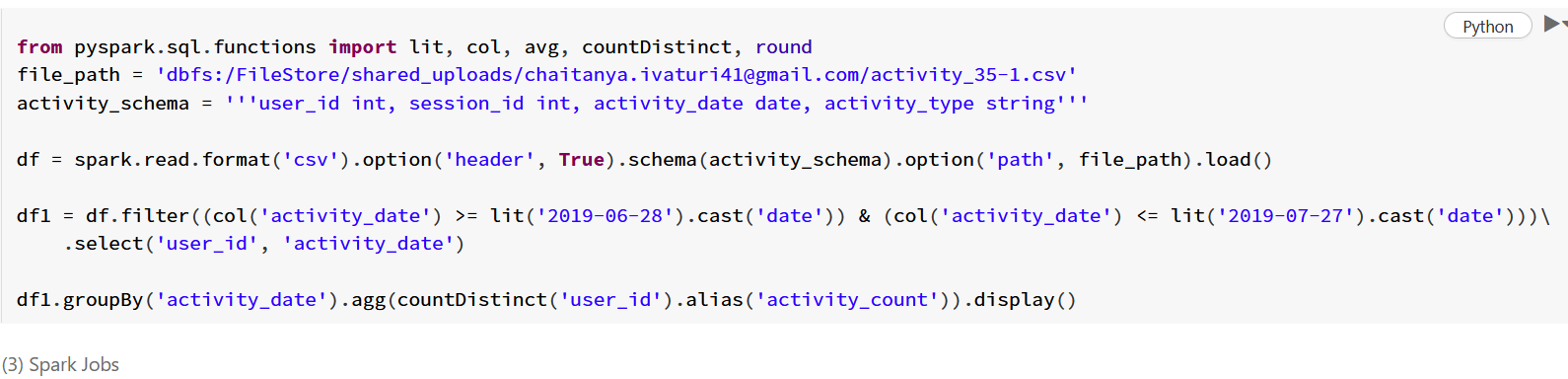
file\_path = 'dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/activity\_35-1.csv'

activity\_schema = '''user\_id int, session\_id int, activity\_date date, activity\_type string'''



* 1. User activity in 30 days 1

Same data as above



* 1. Weather type

country\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/countries.csv"

country\_schema = """country\_id int, country\_name string"""

weather\_path = "dbfs:/FileStore/shared\_uploads/chaitanya.ivaturi41@gmail.com/weather.csv"

weather\_schema = """country\_id int, weather\_state int, day date"""

