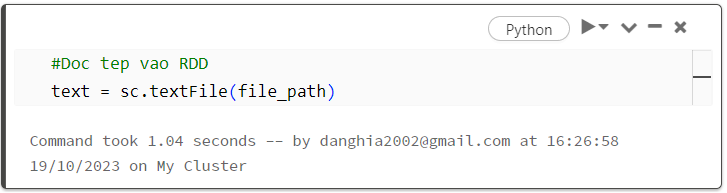
**Bài thực hành số 4. Phân tích dữ liệu điểm thi THPT Quốc gia bằng**

**RDD**





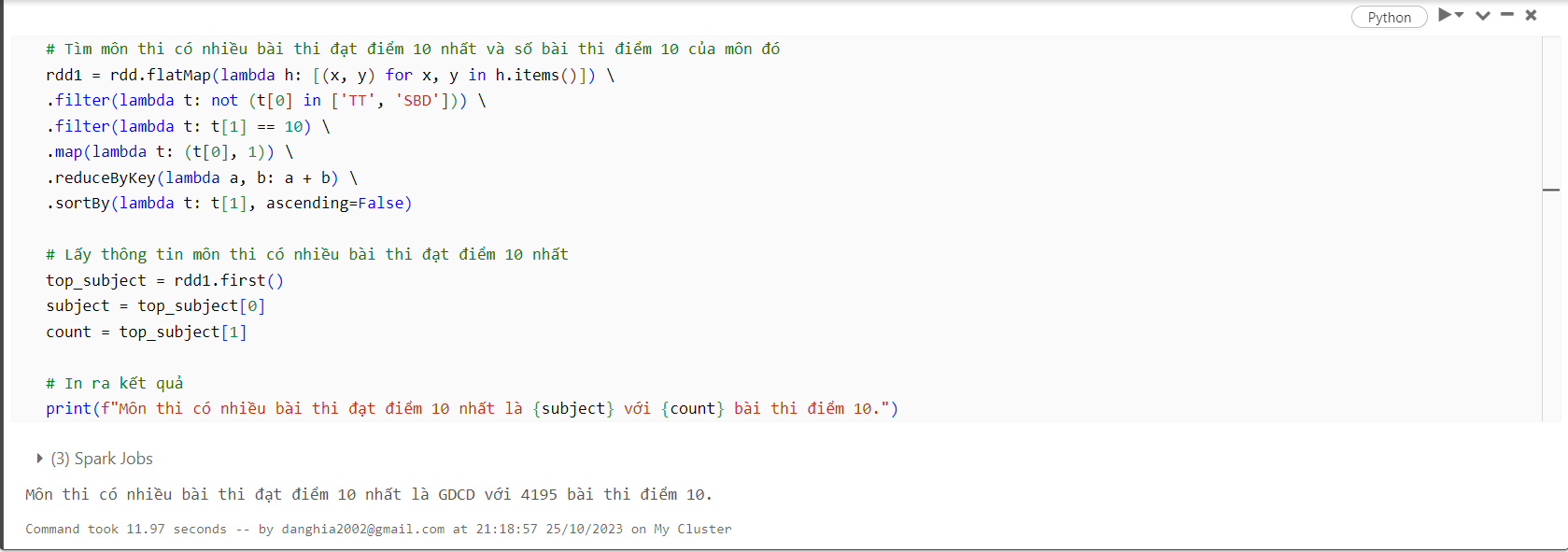


1. Đếm số bài thi điểm 0 từng môn.

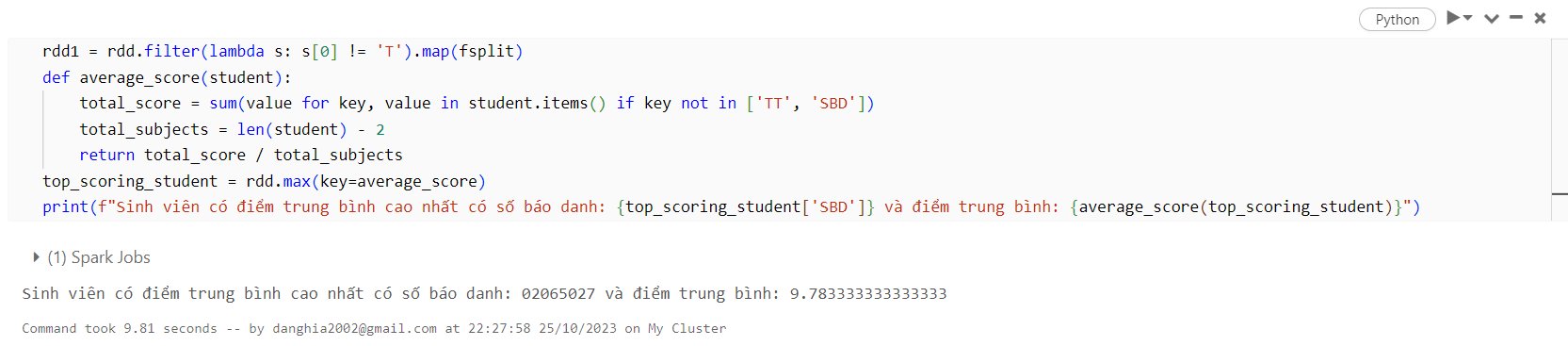


2. Cho biết môn thi có nhiều bài thi đạt điểm 10 nhất và số bài thi điểm 10 môn đó

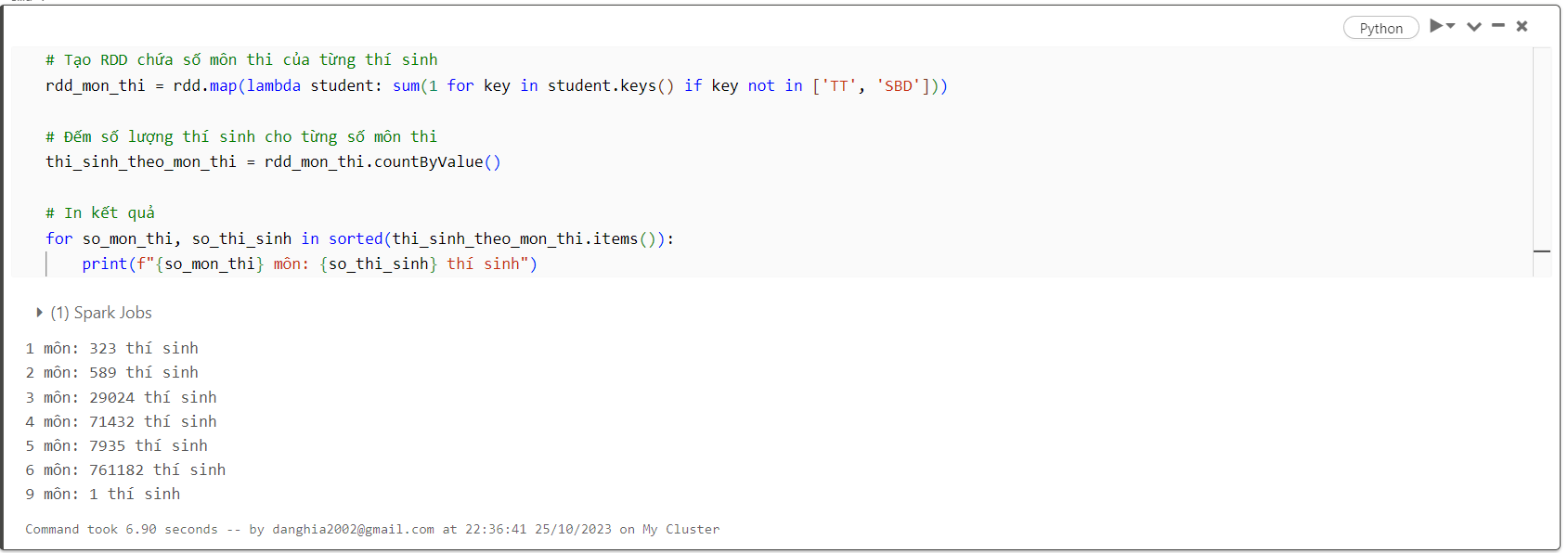
là bao nhiêu.



3. Cho biết số báo danh của những thí sinh có điểm trung bình các môn thi cao nhất.

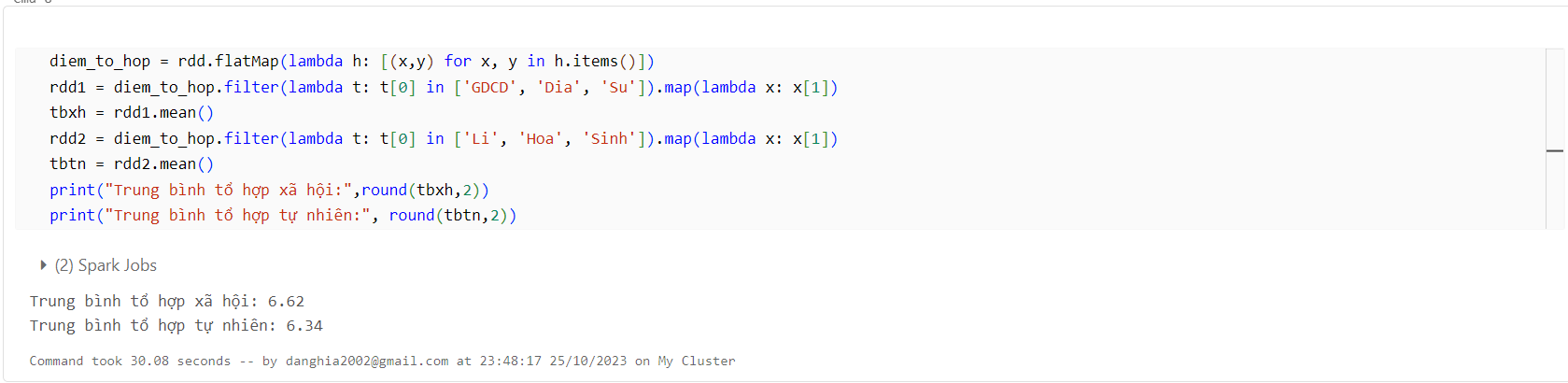


4. Thống kê số lượng thí sinh theo số môn thi. Ví dụ: 1 môn có bao nhiêu thí sinh, 2 môn .., ..., 9 môn có bao nhiêu thí sinh.



5. Tính điểm trung bình bài thi Tự nhiên, bài thi Xã hội. Biết rằng điểm bài thi tổ

hợp là tổng điểm thi 3 môn thành phần của bài thi đó.



6. Vẽ biểu đồ biểu diễn số lượng bài thi Tự nhiên và Xã hội theo từng mức điểm.

7. Vẽ biểu đồ boxplot thể hiện thống kê điểm các môn: Toán, Văn, Ngoại ngữ.

8. Cho biết mã tỉnh có nhiều thí sinh có điểm thi 10. Biết rằng mỗi tỉnh khác nhau

ở 2 ký tự đầu trong số báo danh của thí sinh.

9. Thống kê điểm trung bình theo từng tỉnh, vẽ biểu đồ cho thống kê này.

10.Thống kê mức điểm theo khối A (Toán, Lí, Hóa), B (Toán, Hóa, Sinh), C(Văn,

Sử, Địa), D(Toán, Văn, NN). Mỗi thí sinh chỉ lấy 1 khối có tổng điểm cao nhất.

11. Vẽ biểu đồ cho thống kê mức điểm tổng theo khối.

NẠP DỮ LIỆU BÀI 5



df\_u\_item = spark.read.option('inferSchema','true').option("delimiter", "|").csv("/content/u.item")

df\_u\_item = df\_u\_item.withColumnRenamed('\_c0','movieId')\

.withColumnRenamed('\_c1','movieTitle')\

.withColumnRenamed('\_c2','releaseDate')\

.withColumnRenamed('\_c3','videoReleaseDate')\

.withColumnRenamed('\_c4','IMDbURL')\

.withColumnRenamed('\_c5','unknown')\

.withColumnRenamed('\_c6','Action')\

.withColumnRenamed('\_c7','Adventure')\

.withColumnRenamed('\_c8','Animation')\

.withColumnRenamed('\_c9','Childrens')\

.withColumnRenamed('\_c10','Comedy')\

.withColumnRenamed('\_c11','Crime')\

.withColumnRenamed('\_c12','Documentary')\

.withColumnRenamed('\_c13','Drama')\

.withColumnRenamed('\_c14','Fantasy')\

.withColumnRenamed('\_c15','Film-Noir')\

.withColumnRenamed('\_c16','Horror')\

.withColumnRenamed('\_c17','Musical')\

.withColumnRenamed('\_c18','Mystery')\

.withColumnRenamed('\_c19','Romance')\

.withColumnRenamed('\_c20','Sci-Fi')\

.withColumnRenamed('\_c21','Thriller')\

.withColumnRenamed('\_c22','War')\

.withColumnRenamed('\_c23','Western')

df\_u\_item.show()

////////////////////////////////////////////////

!sudo apt update

!apt-get install openjdk-8-jdk-headless -qq> /dev/null

!wget -q https://archive.apache.org/dist/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz

!tar xf spark-3.0.1-bin-hadoop2.7.tgz

from google.colab import drive

drive.mount('/content/drive')

import os

os.environ["JAVA\_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"

os.environ["SPARK\_HOME"] = "/content/spark-3.0.1-bin-hadoop2.7"

!pip install -q findspark

import findspark

findspark.init()

from pyspark.context import SparkContext

#create SparkContext

sc = SparkContext("local", "My First Spark Application")

print("SparkContext :",sc)

from pyspark.sql import SparkSession

# Create a SparkSession

spark = SparkSession.builder.getOrCreate()

# Thống kê số lượng từng mức điểm của môn Toán

df.groupBy('Toan').count().show()

from pyspark.sql.functions import year

df\_2015 = df.filter(year(df['booking\_date']) == 2015)

total\_bookings = df\_2015.count()

successful\_bookings = df\_2015.filter(df['booking\_status'] == 'successful').count()

booking\_rate = successful\_bookings / total\_bookings

import matplotlib.pyplot as plt

from pyspark.sql import SparkSession

from pyspark.sql.functions import month, count, sum

# Khởi tạo SparkSession

spark = SparkSession.builder.getOrCreate()

# Đọc dữ liệu từ tệp CSV vào DataFrame

data = spark.read.csv('hotel.csv', header=True)

# Chuyển đổi kiểu dữ liệu của các cột liên quan đến số thành Integer

int\_columns = ['is\_canceled', 'lead\_time', 'arrival\_date\_year', 'arrival\_date\_week\_number',

'arrival\_date\_day\_of\_month', 'stays\_in\_weekend\_nights', 'stays\_in\_week\_nights',

'adults', 'children', 'babies', 'previous\_cancellations', 'previous\_bookings\_not\_canceled',

'booking\_changes', 'days\_in\_waiting\_list', 'adr', 'required\_car\_parking\_spaces',

'total\_of\_special\_requests']

for column in int\_columns:

data = data.withColumn(column, data[column].cast('int'))

# Thống kê số lượt đặt phòng theo tháng trong năm

bookings\_by\_month = data.groupBy('arrival\_date\_month').agg(count('\*').alias('booking\_count'))

bookings\_by\_month = bookings\_by\_month.sort(months\_dict['arrival\_date\_month'])

months = bookings\_by\_month.select('arrival\_date\_month').rdd.flatMap(lambda x: x).collect()

booking\_counts = bookings\_by\_month.select('booking\_count').rdd.flatMap(lambda x: x).collect()

# Vẽ biểu đồ số lượt đặt phòng theo tháng

plt.figure(figsize=(10, 6))

plt.bar(months, booking\_counts, color='blue')

plt.xlabel('Tháng')

plt.ylabel('Số lượt đặt phòng')

plt.title('Thống kê số lượt đặt phòng theo tháng trong năm')

plt.xticks(rotation=45)

plt.show()

# Thống kê số lượt đặt phòng theo số ngày ở và quốc gia

bookings\_by