Cloud Computing and Big Data Analytics 2022 Fall
Lab 3: PySpark

TA：曾偉倫

Email: wltseng.ee06@nycu.edu.tw

Outline

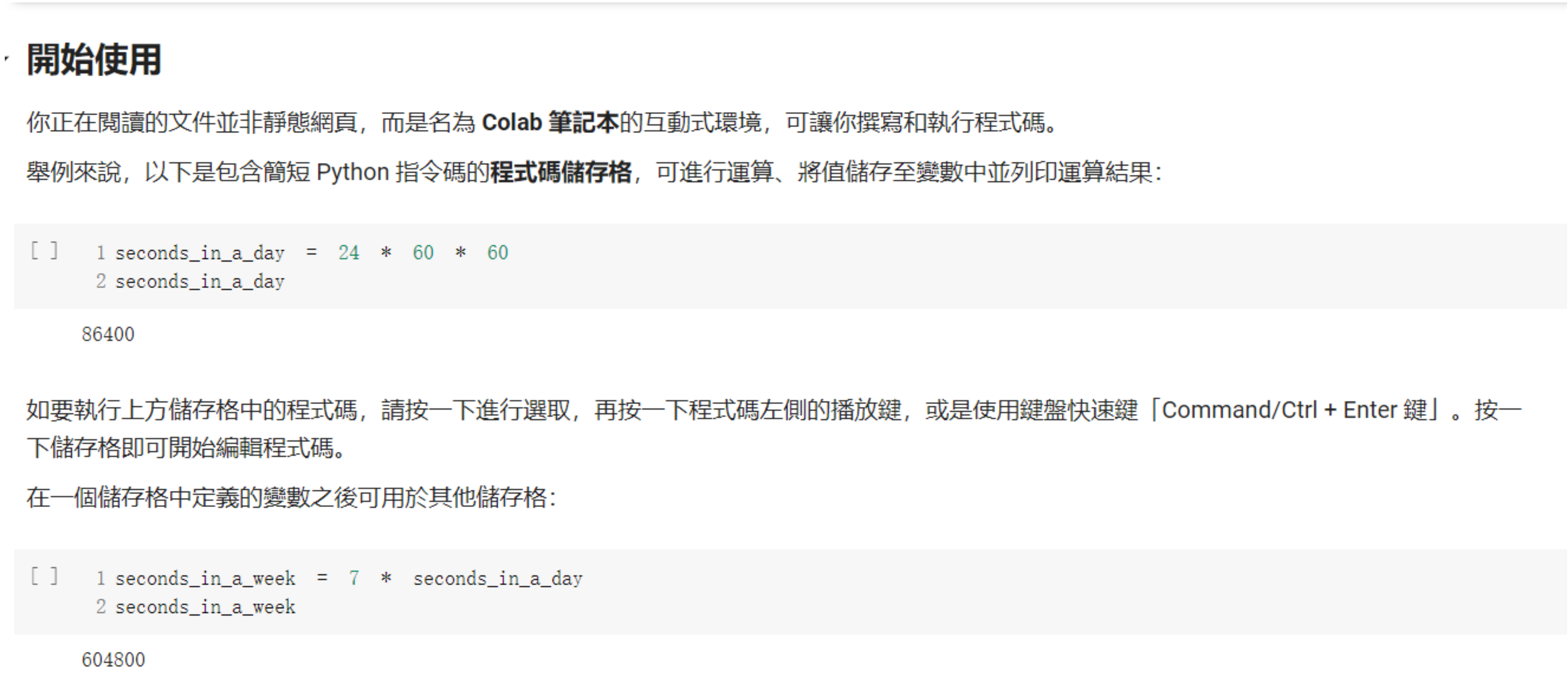
- Introduction
- PySpark Exercise
- Checkpoint
- Grading Policy
- E3 Submission

Introduction

- Python and Spark
  - DataFrame
    - Spark DataFrames hold data in a column and row format.
    - Column: feature or variable
    - Row: individual data point
  - Feature
    - input and output data from a wide variety of sources
    - PySpark contains DataFrame MLlib API for Machine Learning

PySpark Exercise

1. Go to Colab: https://colab.research.google.com/



開始使用

你正在閱讀的文件並非靜態網頁，而是名為 **Colab 筆記本**的互動式環境，可讓你撰寫和執行程式碼。

舉例來說，以下是包含簡短 Python 指令碼的**程式碼儲存格**，可進行運算、將值儲存至變數中並列印運算結果：

```
seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
```

```
86400
```

如要執行上方儲存格中的程式碼，請按一下進行選取，再按一下程式碼左側的播放鍵，或是使用鍵盤快速鍵「Command/Ctrl + Enter 鍵」。按一下儲存格即可開始編輯程式碼。

在一個儲存格中定義的變數之後可用於其他儲存格：

```
seconds_in_a_week = 7 * seconds_in_a_day
seconds_in_a_week
```

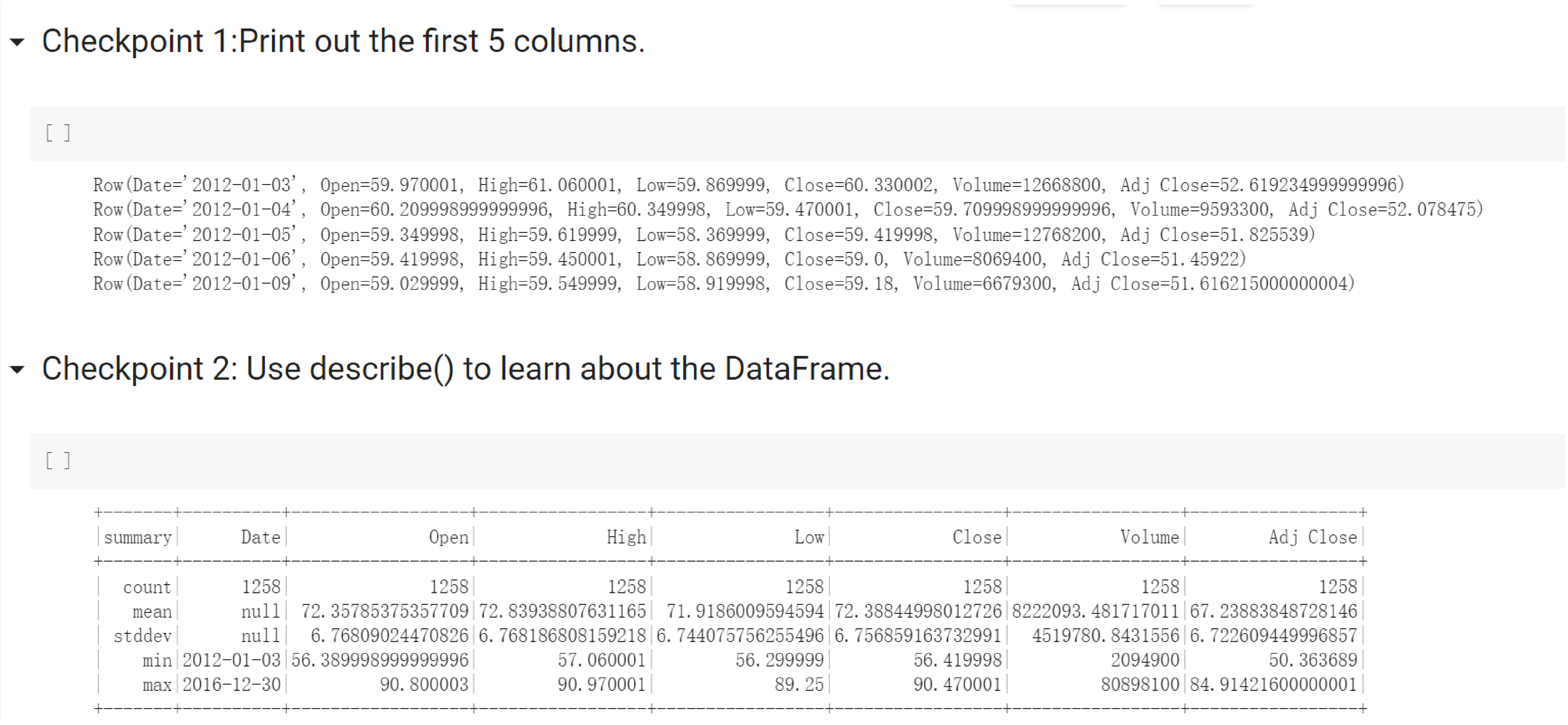
```
604800
```

PySpark Exercise

2. Upload jupyter notebook & data
    - Lab_3_PySpark.ipynb
    - people.json
    - walmart_stock.csv
3. Start !

Checkpoint

- Total: 9
- Use warmup parts and hint to finish 9 checkpoints



Checkpoint 1:Print out the first 5 columns.

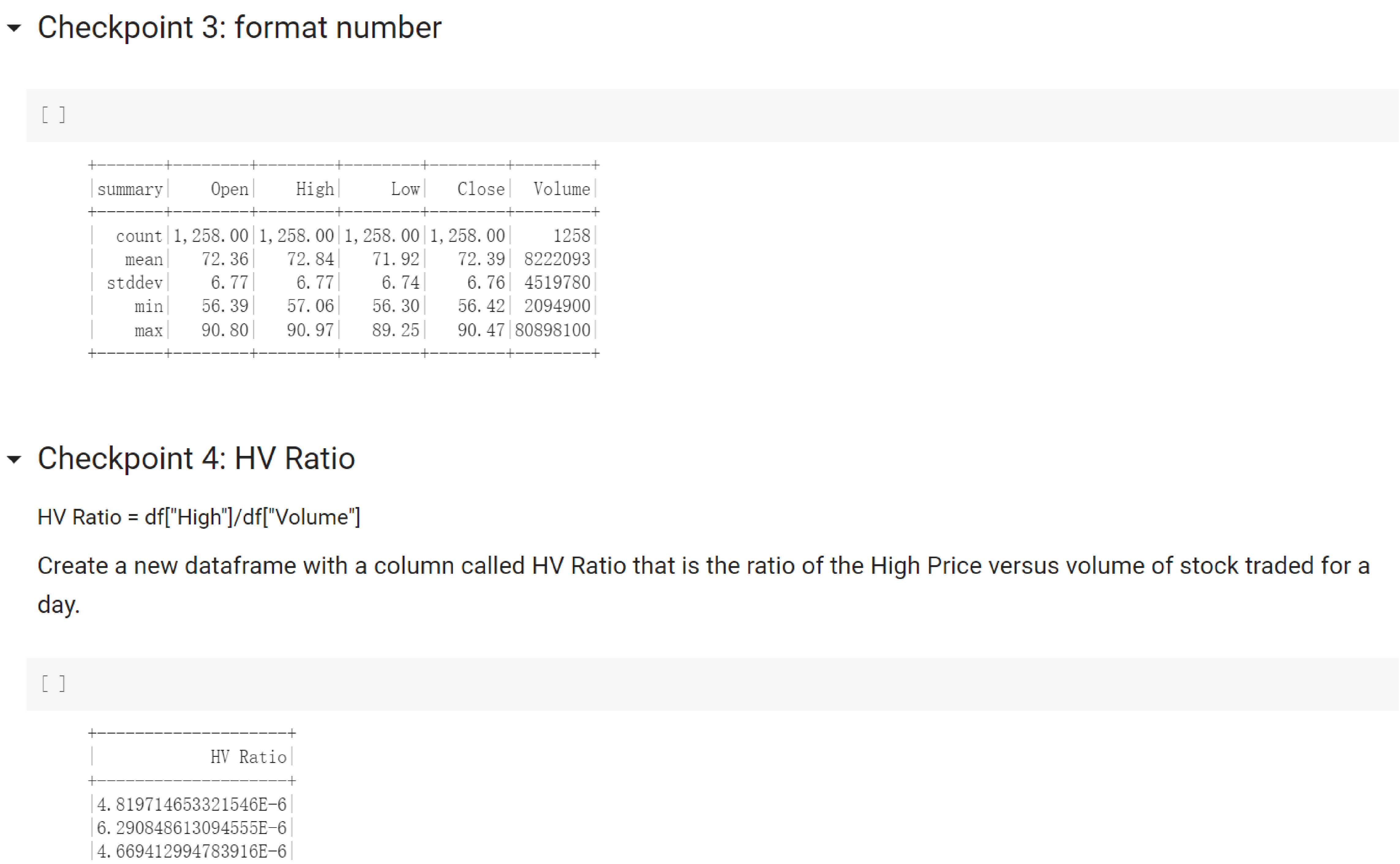
[ ]

```
Row(Date='2012-01-03', Open=59.970001, High=61.060001, Low=59.869999, Close=60.330002, Volume=12668800, Adj Close=52.619234999999996)
Row(Date='2012-01-04', Open=60.209998999999996, High=60.349998, Low=59.470001, Close=59.709998999999996, Volume=9593300, Adj Close=52.078475)
Row(Date='2012-01-05', Open=59.349998, High=59.619999, Low=58.369999, Close=59.419998, Volume=12768200, Adj Close=51.825539)
Row(Date='2012-01-06', Open=59.419998, High=59.450001, Low=58.869999, Close=59.0, Volume=8069400, Adj Close=51.45922)
Row(Date='2012-01-09', Open=59.029999, High=59.549999, Low=58.919998, Close=59.18, Volume=6679300, Adj Close=51.616215000000004)
```

Checkpoint 2: Use describe() to learn about the DataFrame.

[ ]

```
+-------+----------+------------------+-----------------+-----------------+-----------------+-----------------+-----------------+
|summary|      Date|              Open|             High|              Low|            Close|           Volume|        Adj Close|
+-------+----------+------------------+-----------------+-----------------+-----------------+-----------------+-----------------+
|  count|      1258|              1258|             1258|             1258|             1258|             1258|             1258|
|   mean|      null| 72.35785375357709|72.83938807631165| 71.9186009594594|72.38844998012726|8222093.481717011|67.23883848728146|
| stddev|      null|  6.76809024470826|6.768186808159218|6.744075756255496|6.756859163732991|  4519780.8431556|6.722609449996857|
|    min|2012-01-03|56.389998999999996|        57.060001|        56.299999|        56.419998|          2094900|        50.363689|
|    max|2016-12-30|         90.800003|        90.970001|            89.25|        90.470001|         80898100|84.91421600000001|
+-------+----------+------------------+-----------------+-----------------+-----------------+-----------------+-----------------+
```



Checkpoint 3: format number

```
[ ]
```

```
+-------+--------+--------+--------+--------+--------+
|summary|    Open|    High|     Low|   Close|  Volume|
+-------+--------+--------+--------+--------+--------+
|  count|1,258.00|1,258.00|1,258.00|1,258.00|    1258|
|   mean|   72.36|   72.84|   71.92|   72.39| 8222093|
| stddev|    6.77|    6.77|    6.74|    6.76| 4519780|
|    min|   56.39|   57.06|   56.30|   56.42| 2094900|
|    max|   90.80|   90.97|   89.25|   90.47|80898100|
+-------+--------+--------+--------+--------+--------+
```

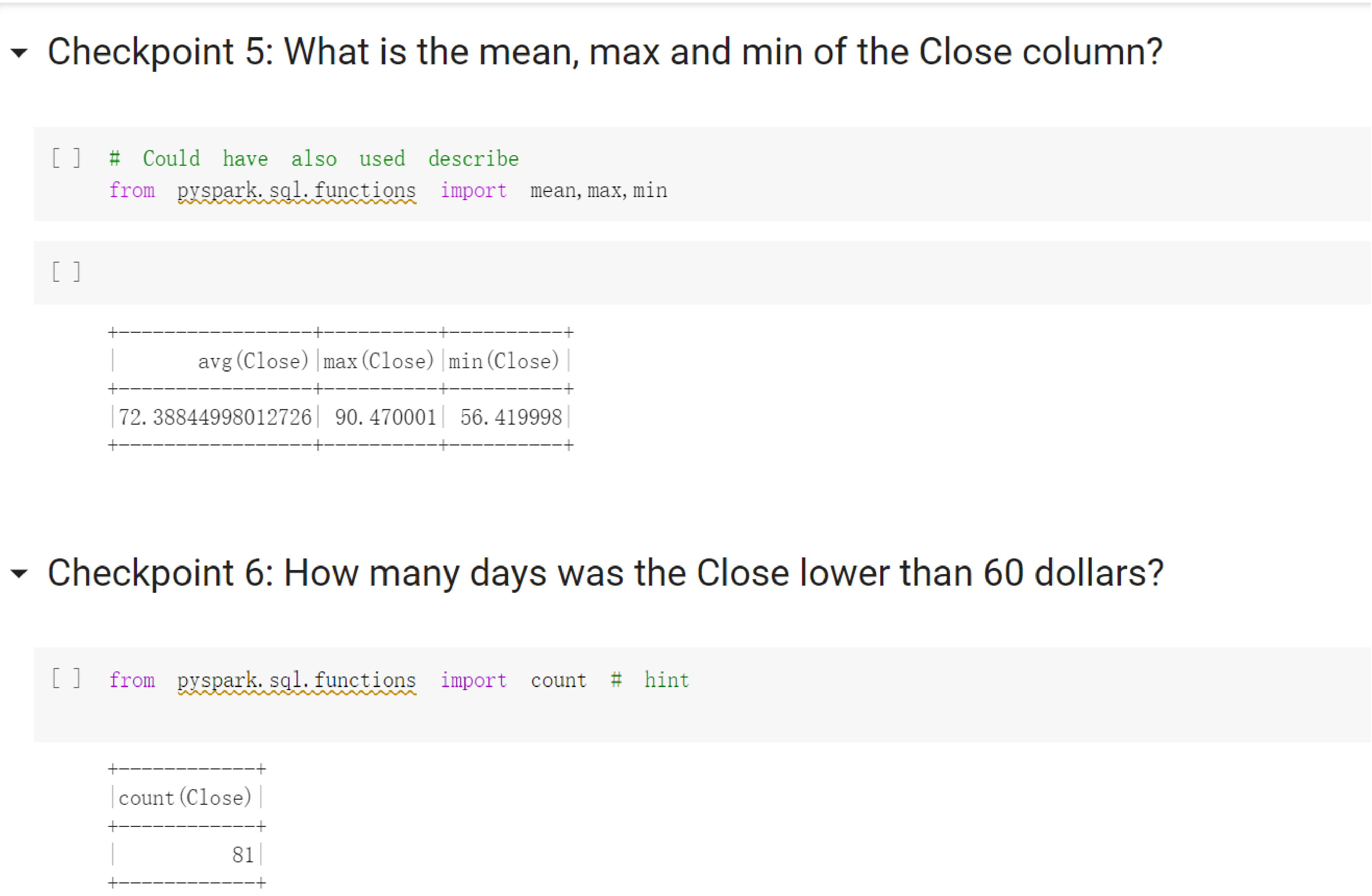
Checkpoint 4: HV Ratio

HV Ratio = df["High"]/df["Volume"]

Create a new dataframe with a column called HV Ratio that is the ratio of the High Price versus volume of stock traded for a day.

```
[ ]
```

```
+--------------------+
|            HV Ratio|
+--------------------+
|4.819714653321546E-6|
|6.290848613094555E-6|
|4.669412994783916E-6|
```



Checkpoint 5: What is the mean, max and min of the Close column?

```
[ ] # Could have also used describe
    from pyspark.sql.functions import mean,max,min
```

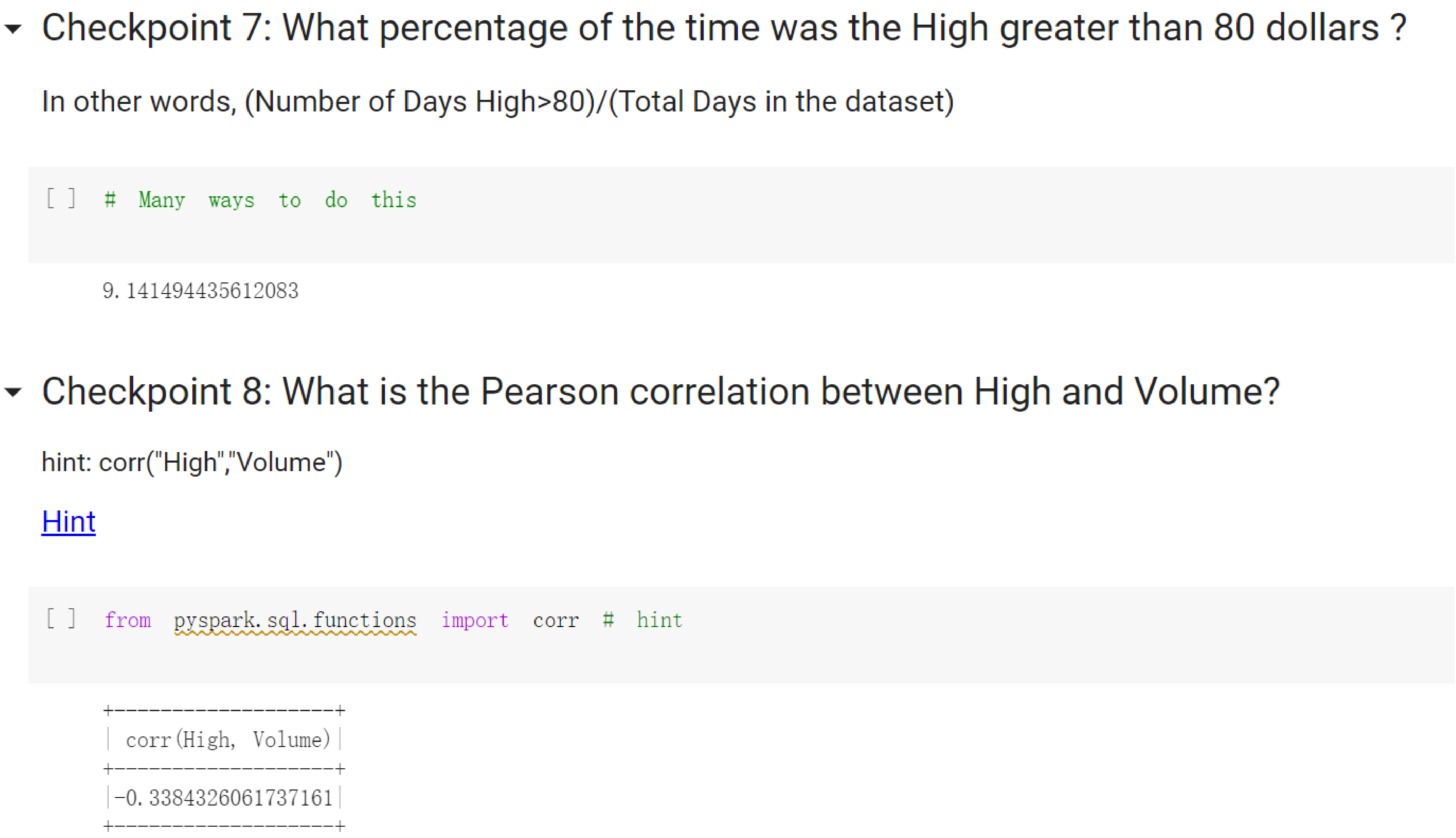
```
[ ]
```

```
+-----------------+----------+----------+
|       avg(Close)|max(Close)|min(Close)|
+-----------------+----------+----------+
|72.38844998012726| 90.470001| 56.419998|
+-----------------+----------+----------+
```

Checkpoint 6: How many days was the Close lower than 60 dollars?

```
[ ] from pyspark.sql.functions import count # hint
```

```
+------------+
|count(Close)|
+------------+
|          81|
+------------+
```



Checkpoint 7: What percentage of the time was the High greater than 80 dollars ?

In other words, (Number of Days High>80)/(Total Days in the dataset)

```
[ ] # Many ways to do this
```

```
9.141494435612083
```

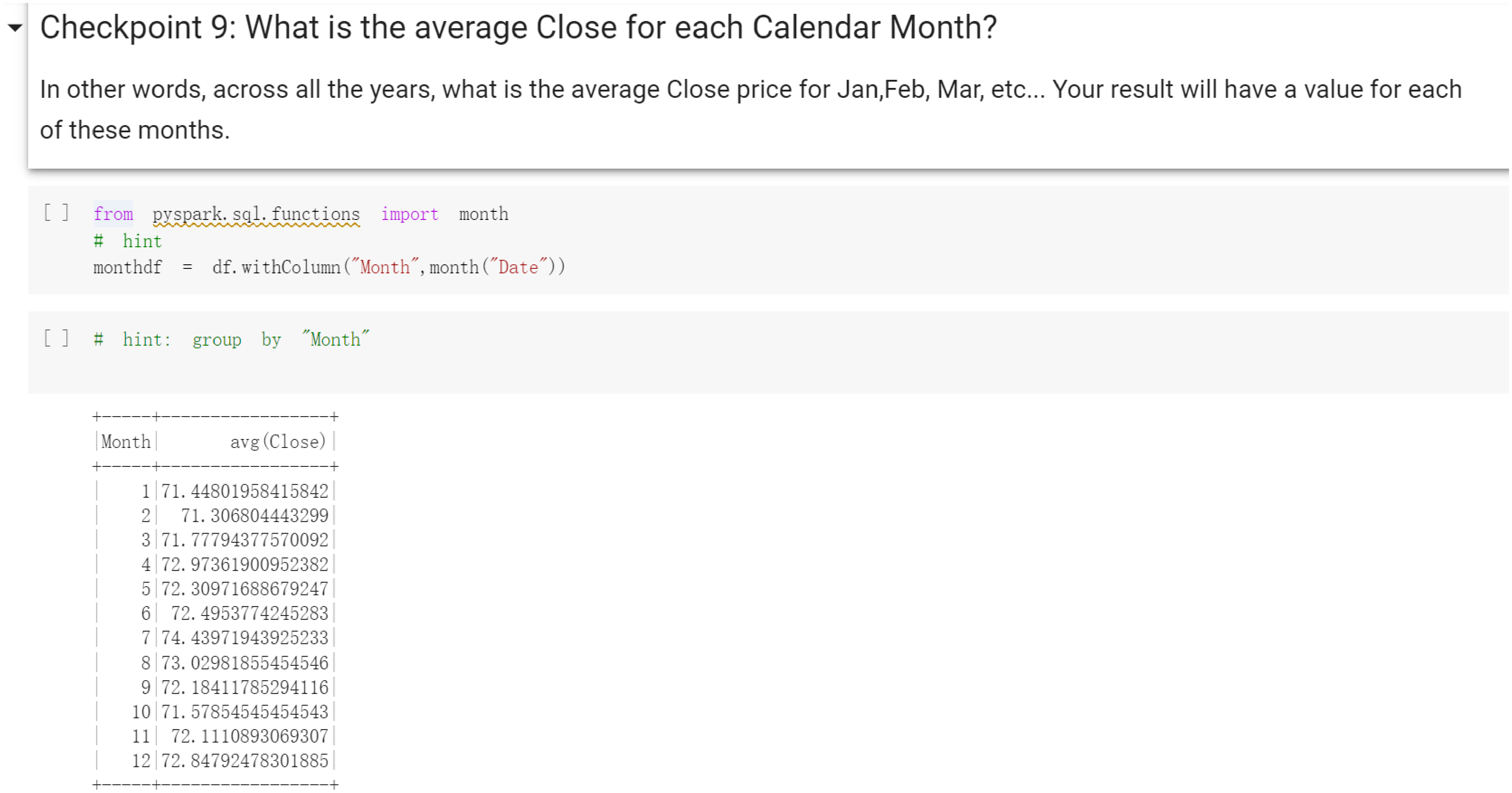
Checkpoint 8: What is the Pearson correlation between High and Volume?

hint: corr("High","Volume")

Hint

```
[ ] from pyspark.sql.functions import corr # hint
```

```
+------------------+
| corr(High, Volume)|
+------------------+
|-0.3384326061737161|
+------------------+
```



Checkpoint 9: What is the average Close for each Calendar Month?

In other words, across all the years, what is the average Close price for Jan,Feb, Mar, etc... Your result will have a value for each of these months.

```
[ ]  from pyspark.sql.functions import month
     # hint
     monthdf = df.withColumn("Month",month("Date"))
```

```
[ ]  # hint: group by "Month"
```

```
+-----+-----------------+
|Month|       avg(Close)|
+-----+-----------------+
|    1|71.44801958415842|
|    2|  71.306804443299|
|    3|71.77794377570092|
|    4|72.97361900952382|
|    5|72.30971688679247|
|    6| 72.4953774245283|
|    7|74.43971943925233|
|    8|73.02981855454546|
|    9|72.18411785294116|
|   10|71.57854545454543|
|   11| 72.1110893069307|
|   12|72.84792478301885|
+-----+-----------------+
```

Grading Policy

Total: 100 pts

- Checkpoint 1~9: 10 pts *9
- E3 submission: 10 pts

E3 Submission

- 1 PDF file
  - File name: <student_ID>.pdf
  - Paste screenshot contain each checkpoint
  - Check sample_submission.pdf
- **Deadline: 12/13 23:59**

- TA：曾偉倫
- Email: wltseng.ee06@nycu.edu.tw

THANK YOU FOR LISTENING

Big data Analytics and Social Intelligent Computing LABoratory

BASIC LAB

NYCU