app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

**Connecting Drive to Colab**

Mounting Google Drive

```
[146]  1 from google.colab import drive
       2 drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

**Reading Data from Drive**

Unzipping the data

```
[147]  1 !unzip "/content/drive/MyDrive/New-York-Data/train.zip"
```

```
Archive:  /content/drive/MyDrive/New-York-Data/train.zip
replace train.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
```

**Setting up PySpark in Colab**

```
[148]  1 !apt-get install openjdk-8-jdk-headless -qq > /dev/null
```

Installing Apache Spark 3.0.1 with Hadoop 2.7 from the link

```
[149]  1 !wget -q https://downloads.apache.org/spark/spark-3.1.1/spark-3.1.1-bin-hadoop2.7.tgz
```

To unzip that folder

```
[150]  1 !tar xf spark-3.1.1-bin-hadoop2.7.tgz
```

Install findspark library

```
[151]  1 !pip install -q findspark
```

0s completed at 10:01 PM

app.ipynb ☆

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

To set the environment path

```
[152] 1 import os
      2 os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
      3 os.environ["SPARK_HOME"] = "/content/spark-3.1.1-bin-hadoop2.7"
```

To locate Spark in the system

```
[153] 1 import findspark
      2 findspark.init()
```

To know the location where Spark is installed

```
[154] 1 findspark.find()
```

```
'/content/spark-3.1.1-bin-hadoop2.7'
```

To view the Spark UI

```
1 !wget https://bin.equinox.io/c/4VmDzA7iaHb/ngrok-stable-linux-amd64.zip
2 !unzip ngrok-stable-linux-amd64.zip
3 get_ipython().system_raw('./ngrok http 4050 &')
4 !curl -s http://localhost:4040/api/tunnels
```

```
io/c/4VmDzA7iaHb/ngrok-stable-linux-amd64.zip
52.204.190.140, 34.193.189.47, 34.233.212.111, ...
|52.204.190.140|:443... connected.
K
eam]


]  13.19M  6.55MB/s    in 2.0s

le-linux-amd64.zip.1' saved [13832437/13832437]


]ename: n
i/tunnels/command_line","public_url":"https://31c85f4c7aeb.ngrok.io","proto":"https","config":{"addr":"http://localhost:4050","inspect":true},"metrics
```

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

Loading data into PySpark

```
[156]  1 data = spark.read.csv("train.csv", header=True, inferSchema=True)
```

Understanding the Data

```
1 data.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: string (nullable = true)
 |-- dropoff_datetime: string (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
```

To display the informations

```
[158]  1 data.show(5)
```

```
+---------+---------+-------------------+-------------------+---------------+------------------+-----------------+------------------+-----------------+
|       id|vendor_id|    pickup_datetime|   dropoff_datetime|passenger_count|  pickup_longitude|   pickup_latitude| dropoff_longitude|  dropoff_latitu
+---------+---------+-------------------+-------------------+---------------+------------------+-----------------+------------------+-----------------+
|id2875421|        2|2016-03-14 17:24:55|2016-03-14 17:32:30|              1| -73.9821548461914| 40.76793670654297|-73.96463012695312|40.7656021118164
|id2377394|        1|2016-06-12 00:43:35|2016-06-12 00:54:38|              1|-73.98041534423828|40.738563537597656|-73.99948120117188| 40.73115158081
|id3858529|        2|2016-01-19 11:35:24|2016-01-19 12:10:48|              1| -73.9790267944336|40.763938903808594|-74.00533294677734|40.7100868225097
|id3504673|        2|2016-04-06 19:32:31|2016-04-06 19:39:40|              1|-74.01004028320312|   40.719970703125|-74.01226806640625| 40.706718444824
|id2181028|        2|2016-03-26 13:30:55|2016-03-26 13:38:10|              1|-73.97305297851562|40.793209075927734| -73.9729232788086| 40.782520294185
+---------+---------+-------------------+-------------------+---------------+------------------+-----------------+------------------+-----------------+
only showing top 5 rows
```

To count the number of rows

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

To count the number of rows

```
[159] 1 data.count()
```

```
1458644
```

```
[160] 1 data.select("pickup_datetime","dropoff_datetime").show(5)
```

```
+-------------------+-------------------+
|    pickup_datetime|   dropoff_datetime|
+-------------------+-------------------+
|2016-03-14 17:24:55|2016-03-14 17:32:30|
|2016-06-12 00:43:35|2016-06-12 00:54:38|
|2016-01-19 11:35:24|2016-01-19 12:10:48|
|2016-04-06 19:32:31|2016-04-06 19:39:40|
|2016-03-26 13:30:55|2016-03-26 13:38:10|
+-------------------+-------------------+
only showing top 5 rows
```

```
1 data.describe().show()
```

```
+-------+---------+------------------+-------------------+-------------------+------------------+-------------------+------------------+--------------
|summary|       id|         vendor_id|    pickup_datetime|   dropoff_datetime|   passenger_count|   pickup_longitude|   pickup_latitude|  dropoff_long
+-------+---------+------------------+-------------------+-------------------+------------------+-------------------+------------------+--------------
|  count|  1458644|           1458644|            1458644|            1458644|           1458644|            1458644|           1458644|             14
|   mean|     null|1.5349502688798637|               null|               null|1.6645295219395548| -73.97348630489282|40.750920908391734|  -73.97341594
| stddev|     null|0.4987771539074042|               null|               null| 1.314242167823114|0.07090185842270283| 0.032881186257633|0.070643268097
|    min|id0000001|                 1|2016-01-01 00:00:17|2016-01-01 00:03:31|                 0|-121.93334197998047| 34.35969543457031|-121.93330383
|    max|id4000000|                 2|2016-06-30 23:59:39|2016-07-01 23:02:03|                 9| -61.33552932739258| 51.88108444213867| -61.335529327
+-------+---------+------------------+-------------------+-------------------+------------------+-------------------+------------------+--------------
```

```
[162] 1 #from pyspark.sql.column import Column as col, _to_java_column, _to_seq
```

```
[163] 1 from pyspark.sql.types import *
```

Converting into timestamp

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

Converting into timestamp

```
[164] 1 data_conv = data.withColumn("pickup_datetime", data["pickup_datetime"].cast(TimestampType()))
```

```
[165] 1 data_conv2 = data_conv.withColumn("dropoff_datetime", data_conv["dropoff_datetime"].cast(TimestampType()))
```

```
[166] 1 data_conv2.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
```

```
[167] 1 import pyspark.sql.functions as f
```

New columns to show the pickup and dropoff days

```
[168] 1 data_conv3 = data_conv2.withColumn("pickup_day", f.date_format("pickup_datetime", "EEEE"))
```

```
1 data_conv3.show(5)
```

```
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------
|       id|vendor_id|    pickup_datetime|   dropoff_datetime|passenger_count|  pickup_longitude|   pickup_latitude| dropoff_longitude|  dropoff_latitu
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------
|id2875421|        2|2016-03-14 17:24:55|2016-03-14 17:32:30|              1| -73.9821548461914| 40.76793670654297|-73.96463012695312|40.7656021118164
|id2377394|        1|2016-06-12 00:43:35|2016-06-12 00:54:38|              1|-73.98041534423828|40.738563537597656|-73.99948120117188| 40.73115158081
|id3858529|        2|2016-01-19 11:35:24|2016-01-19 12:10:48|              1| -73.9790267944336|40.763938903808594|-74.00533294677734|40.7100868225097
|id3504673|        2|2016-04-06 19:32:31|2016-04-06 19:39:40|              1|-74.01004028320312|    40.719970703125|-74.01226806640625| 40.706718444824
|id2181028|        2|2016-03-26 13:30:55|2016-03-26 13:38:10|              1|-73.97305297851562|40.793209075927734| -73.9729232788086| 40.782520294189
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------
only showing top 5 rows
```

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

```
[170] 1 data_conv4 = data_conv3.withColumn("dropoff_day", f.date_format("dropoff_datetime", "EEEE"))
```

```
1 data_conv4.show()
```

```
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+
|       id|vendor_id|    pickup_datetime|   dropoff_datetime|passenger_count|  pickup_longitude|   pickup_latitude| dropoff_longitude|  dropoff_latitu
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+
|id2875421|        2|2016-03-14 17:24:55|2016-03-14 17:32:30|              1| -73.9821548461914| 40.76793670654297|-73.96463012695312|40.7656021118164
|id2377394|        1|2016-06-12 00:43:35|2016-06-12 00:54:38|              1|-73.98041534423828|40.738563537597656|-73.99948120117188| 40.731151580810
|id3858529|        2|2016-01-19 11:35:24|2016-01-19 12:10:48|              1| -73.9790267944336|40.763938903808594|-74.00533294677734|40.7100868225097
|id3504673|        2|2016-04-06 19:32:31|2016-04-06 19:39:40|              1|-74.01004028320312|    40.719970703125|-74.01226806640625| 40.706718444824
|id2181028|        2|2016-03-26 13:30:55|2016-03-26 13:38:10|              1|-73.97305297851562|40.793209075927734| -73.9729232788086| 40.782520294189
|id0801584|        2|2016-01-30 22:01:40|2016-01-30 22:09:03|              6|-73.98285675048828| 40.74219512939453|-73.99208068847656|40.7491836547851
|id1813257|        1|2016-06-17 22:34:59|2016-06-17 22:40:40|              4| -73.9690170288086| 40.75783920288086|-73.95740509033203| 40.765895843505
|id1324603|        2|2016-05-21 07:54:58|2016-05-21 08:20:49|              1|-73.96927642822266| 40.79777908325195|-73.92247009277344| 40.760559082031
|id1301050|        1|2016-05-27 23:12:23|2016-05-27 23:16:38|              1|-73.99948120117188|40.738399505615234|-73.98578643798828| 40.732814788818
|id0012891|        2|2016-03-10 21:45:01|2016-03-10 22:05:26|              1|-73.98104858398438| 40.74433898925781| -73.9729995727539| 40.789989471435
|id1436371|        2|2016-05-10 22:08:41|2016-05-10 22:29:55|              1|-73.98265075683594| 40.76383972167969|-74.00222778320312| 40.732990264892
|id1299289|        2|2016-05-15 11:16:11|2016-05-15 11:34:59|              4|-73.99153137207031| 40.74943923950195|    -73.95654296875|  40.77062988281
|id1187965|        2|2016-02-19 09:52:46|2016-02-19 10:11:20|              2|-73.96298217773438| 40.75667953491211|-73.98440551757812|40.7607192993164
|id0799785|        2|2016-06-01 20:58:29|2016-06-01 21:02:49|              1|-73.95630645751953|40.767940521240234|-73.96611022949219| 40.763000488281
|id2900608|        2|2016-05-27 00:43:36|2016-05-27 01:07:10|              1|-73.99219512939453| 40.72722625732422|-73.97465515136719|  40.78306961059
|id3319787|        1|2016-05-16 15:29:02|2016-05-16 15:32:33|              1|-73.95551300048828|40.768592834472656|-73.94876098632812| 40.771545410156
|id3379579|        2|2016-04-11 17:29:50|2016-04-11 18:08:26|              1|-73.99116516113281| 40.75556182861328| -73.9992904663086|  40.72535324096
|id1154431|        1|2016-04-14 08:48:26|2016-04-14 09:00:37|              1|-73.99425506591797| 40.74580383300781| -73.9996566772461| 40.723342895507
|id3552682|        1|2016-06-27 09:55:13|2016-06-27 10:17:10|              1|-74.00398254394531|  40.7130126953125|-73.97919464111328| 40.749923706054
|id3390316|        2|2016-06-05 13:47:23|2016-06-05 13:51:34|              1|   -73.98388671875|40.738197326660156|-73.99120330810547| 40.727870941162
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+
only showing top 20 rows
```

New columns to show the pickup and dropoff day numbers

```
[172] 1 data_conv5 = data_conv4.withColumn("pickup_day_no", f.date_format("pickup_datetime", "F").cast(IntegerType()))
```

```
[173] 1 data_conv5.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
```

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

```
[173]  1 data_conv5.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
```

```
[174]  1 data_conv6 = data_conv5.withColumn("dropoff_day_no", f.date_format("dropoff_datetime", "F").cast(IntegerType()))
```

```
[175]  1 data_conv6.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
 |-- dropoff_day_no: integer (nullable = true)
```

New columns to show the pickup and dropoff hours

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

New columns to show the pickup and dropoff hours

```
[176] 1 data_conv7 = data_conv6.withColumn("pickup_hour", f.date_format("pickup_datetime", "H").cast(IntegerType()))
```

```
1 data_conv7.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
 |-- dropoff_day_no: integer (nullable = true)
 |-- pickup_hour: integer (nullable = true)
```

```
[178] 1 data_conv8 = data_conv7.withColumn("dropoff_hour", f.date_format("dropoff_datetime", "H").cast(IntegerType()))
```

```
[179] 1 data_conv8.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
```

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

New columns to show the pickup and dropoff months

```
[180]  1 data_conv9 = data_conv8.withColumn("pickup_month", f.date_format("pickup_datetime", "M").cast(IntegerType()))
```

```
[181]  1 data_conv9.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
 |-- dropoff_day_no: integer (nullable = true)
 |-- pickup_hour: integer (nullable = true)
 |-- dropoff_hour: integer (nullable = true)
 |-- pickup_month: integer (nullable = true)
```

```
[182]  1 data_conv10 = data_conv9.withColumn("dropoff_month", f.date_format("dropoff_datetime", "M").cast(IntegerType()))
```

```
1 data_conv10.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
```

✓ 0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

[183]
```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
 |-- dropoff_day_no: integer (nullable = true)
 |-- pickup_hour: integer (nullable = true)
 |-- dropoff_hour: integer (nullable = true)
 |-- pickup_month: integer (nullable = true)
 |-- dropoff_month: integer (nullable = true)
```

[184]
```
from pyspark.sql.types import StructType, StructField, IntegerType, DoubleType, StringType
from pyspark.sql.functions import udf
from pyspark.sql import Row
```

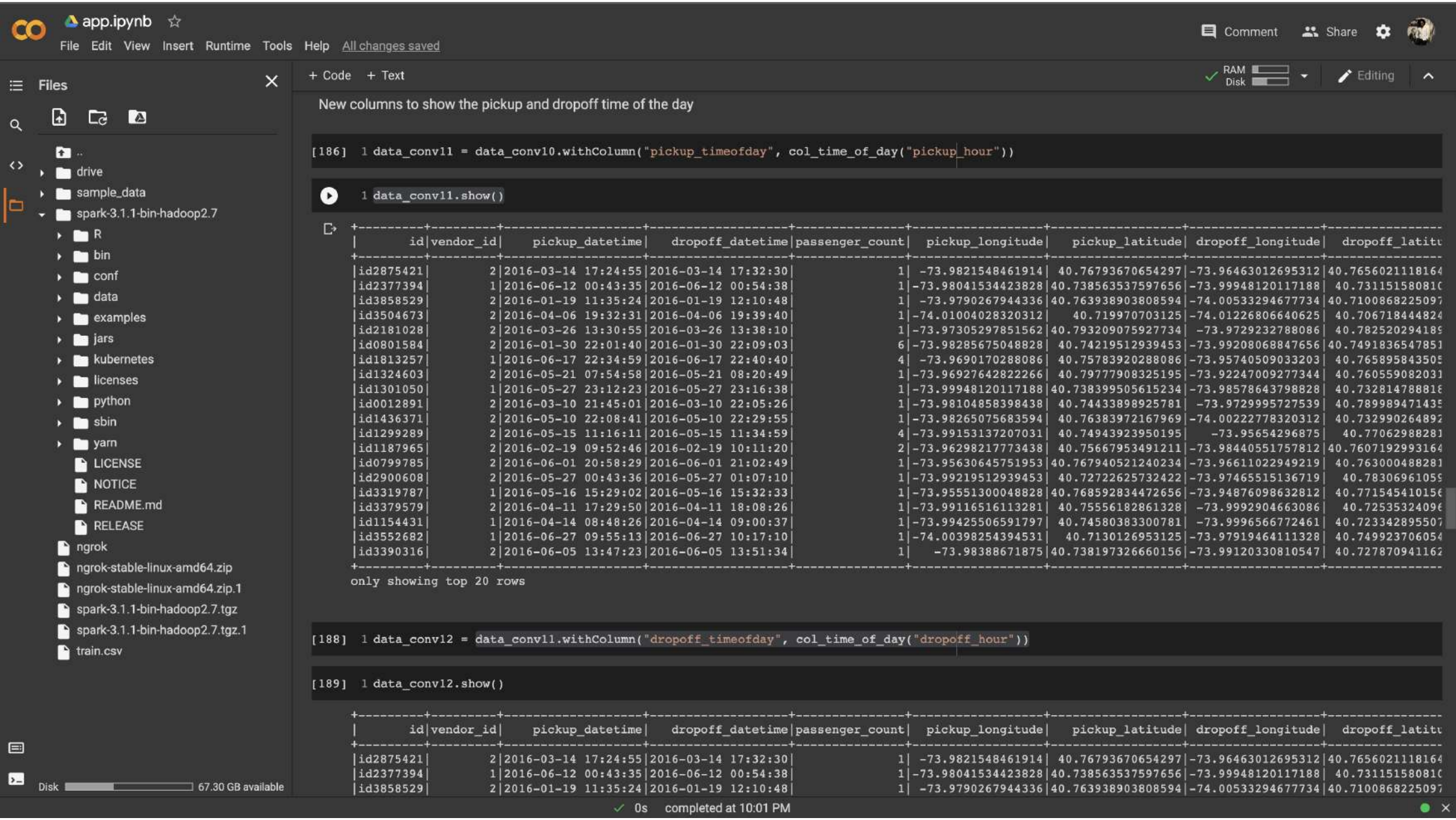
A function to calculate time of day per 4-hour period

```
def time_of_day(x):
    if x in range(6,10):
        return "Morning"
    elif x in range(10,14):
        return "Afternoon"
    elif x in range(14,18):
        return "Evening"
    else:
        return "Night"

col_time_of_day = udf(lambda z: time_of_day(z))
spark.udf.register("col_time_of_day", time_of_day, StringType())
```

```
<function __main__.time_of_day>
```

0s completed at 10:01 PM


app.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
+ Code + Text
RAM Disk
Editing
New columns to show the pickup and dropoff time of the day
[186] 1 data_conv11 = data_conv10.withColumn("pickup_timeofday", col_time_of_day("pickup_hour"))
1 data_conv11.show()
+---------+---------+-------------------+-------------------+---------------+-------------------+-------------------+-------------------+-----------------
| id|vendor_id| pickup_datetime| dropoff_datetime|passenger_count| pickup_longitude| pickup_latitude| dropoff_longitude| dropoff_latitu
+---------+---------+-------------------+-------------------+---------------+-------------------+-------------------+-------------------+-----------------
|id2875421| 2|2016-03-14 17:24:55|2016-03-14 17:32:30| 1| -73.9821548461914| 40.76793670654297|-73.96463012695312|40.7656021118164
|id2377394| 1|2016-06-12 00:43:35|2016-06-12 00:54:38| 1|-73.98041534423828|40.738563537597656|-73.99948120117188| 40.73115158081
|id3858529| 2|2016-01-19 11:35:24|2016-01-19 12:10:48| 1| -73.9790267944336|40.763938903808594| -74.00533294677734|40.7100868225097
|id3504673| 2|2016-04-06 19:32:31|2016-04-06 19:39:40| 1|-74.01004028320312| 40.719970703125|-74.01226806640625| 40.706718444824
|id2181028| 2|2016-03-26 13:30:55|2016-03-26 13:38:10| 1|-73.97305297851562|40.793209075927734| -73.9729232788086| 40.782520294189
|id0801584| 2|2016-01-30 22:01:40|2016-01-30 22:09:03| 6|-73.98285675048828| 40.74219512939453|-73.99208068847656|40.7491836547851
|id1813257| 1|2016-06-17 22:34:59|2016-06-17 22:40:40| 4| -73.9690170288086| 40.75783920288086|-73.95740509033203| 40.765895843505
|id1324603| 2|2016-05-21 07:54:58|2016-05-21 08:20:49| 1|-73.96927642822266| 40.79777908325195|-73.92247009277344| 40.760559082031
|id1301050| 1|2016-05-27 23:12:23|2016-05-27 23:16:38| 1|-73.99948120117188|40.738399505615234|-73.98578643798828| 40.732814788818
|id0012891| 2|2016-03-10 21:45:01|2016-03-10 22:05:26| 1|-73.98104858398438| 40.74433898925781| -73.9729995727539| 40.789989471435
|id1436371| 2|2016-05-10 22:08:41|2016-05-10 22:29:55| 1|-73.98265075683594| 40.76383972167969|-74.00222778320312| 40.732990264892
|id1299289| 2|2016-05-15 11:16:11|2016-05-15 11:34:59| 4|-73.99153137207031| 40.74943923950195| -73.95654296875| 40.77062988281
|id1187965| 2|2016-02-19 09:52:46|2016-02-19 10:11:20| 2|-73.96298217773438| 40.75667953491211|-73.98440551757812|40.7607192993164
|id0799785| 2|2016-06-01 20:58:29|2016-06-01 21:02:49| 1|-73.95630645751953|40.767940521240234|-73.96611022949219| 40.763000488281
|id2900608| 2|2016-05-27 00:43:36|2016-05-27 01:07:10| 1|-73.99219512939453| 40.72722625732422|-73.97465515136719| 40.78306961059
|id3319787| 1|2016-05-16 15:29:02|2016-05-16 15:32:33| 1|-73.95551300048828|40.768592834472656|-73.94876098632812| 40.771545410156
|id3379579| 2|2016-04-11 17:29:50|2016-04-11 18:08:26| 1|-73.99116516113281| 40.75556182861328| -73.9992904663086| 40.72535324096
|id1154431| 1|2016-04-14 08:48:26|2016-04-14 09:00:37| 1|-73.99425506591797| 40.74580383300781| -73.9996566772461| 40.723342895507
|id3552682| 1|2016-06-27 09:55:13|2016-06-27 10:17:10| 1|-74.00398254394531| 40.7130126953125|-73.97919464111328| 40.749923706054
|id3390316| 2|2016-06-05 13:47:23|2016-06-05 13:51:34| 1| -73.98388671875|40.738197326660156|-73.99120330810547| 40.727870941162
+---------+---------+-------------------+-------------------+---------------+-------------------+-------------------+-------------------+-----------------
only showing top 20 rows
[188] 1 data_conv12 = data_conv11.withColumn("dropoff_timeofday", col_time_of_day("dropoff_hour"))
[189] 1 data_conv12.show()
+---------+---------+-------------------+-------------------+---------------+-------------------+-------------------+-------------------+-----------------
| id|vendor_id| pickup_datetime| dropoff_datetime|passenger_count| pickup_longitude| pickup_latitude| dropoff_longitude| dropoff_latitu
+---------+---------+-------------------+-------------------+---------------+-------------------+-------------------+-------------------+-----------------
|id2875421| 2|2016-03-14 17:24:55|2016-03-14 17:32:30| 1| -73.9821548461914| 40.76793670654297|-73.96463012695312|40.7656021118164
|id2377394| 1|2016-06-12 00:43:35|2016-06-12 00:54:38| 1|-73.98041534423828|40.738563537597656|-73.99948120117188| 40.73115158081
|id3858529| 2|2016-01-19 11:35:24|2016-01-19 12:10:48| 1| -73.9790267944336|40.763938903808594| -74.00533294677734|40.7100868225097
Files
..
drive
sample_data
spark-3.1.1-bin-hadoop2.7
R
bin
conf
data
examples
jars
kubernetes
licenses
python
sbin
yarn
LICENSE
NOTICE
README.md
RELEASE
ngrok
ngrok-stable-linux-amd64.zip
ngrok-stable-linux-amd64.zip.1
spark-3.1.1-bin-hadoop2.7.tgz
spark-3.1.1-bin-hadoop2.7.tgz.1
train.csv
Disk 67.30 GB available
0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

```
[190]  1 data_conv12.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
 |-- dropoff_day_no: integer (nullable = true)
 |-- pickup_hour: integer (nullable = true)
 |-- dropoff_hour: integer (nullable = true)
 |-- pickup_month: integer (nullable = true)
 |-- dropoff_month: integer (nullable = true)
 |-- pickup_timeofday: string (nullable = true)
 |-- dropoff_timeofday: string (nullable = true)
```

```
1 from geopy.distance import great_circle
```

A function to calculate the distance between two coordinates

```
[192]  1 def cal_distance(pickup_lat , pickup_long , dropoff_lat, dropoff_long):
       2     start_coordinates = pickup_lat, pickup_long
       3     stop_coordinates = dropoff_lat, dropoff_long
       4
       5     return great_circle(start_coordinates, stop_coordinates).km
       6
       7 cal_distance_udf = udf(lambda x1,x2,y1,y2: cal_distance(x1,x2,y1,y2))
       8 spark.udf.register("cal_distance_udf", cal_distance, DoubleType())
```

```
<function __main__.cal_distance>
```

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

+ Code + Text

RAM Disk Editing

New column to show the distances in km of trips

```
[193] 1 _conv13 = data_conv12.withColumn("distance", cal_distance_udf("pickup_latitude", "pickup_longitude", "dropoff_latitude", "dropoff_longitude"))
```

```
1 data_conv13.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- vendor_id: integer (nullable = true)
 |-- pickup_datetime: timestamp (nullable = true)
 |-- dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: integer (nullable = true)
 |-- pickup_longitude: double (nullable = true)
 |-- pickup_latitude: double (nullable = true)
 |-- dropoff_longitude: double (nullable = true)
 |-- dropoff_latitude: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- trip_duration: integer (nullable = true)
 |-- pickup_day: string (nullable = true)
 |-- dropoff_day: string (nullable = true)
 |-- pickup_day_no: integer (nullable = true)
 |-- dropoff_day_no: integer (nullable = true)
 |-- pickup_hour: integer (nullable = true)
 |-- dropoff_hour: integer (nullable = true)
 |-- pickup_month: integer (nullable = true)
 |-- dropoff_month: integer (nullable = true)
 |-- pickup_timeofday: string (nullable = true)
 |-- dropoff_timeofday: string (nullable = true)
 |-- distance: string (nullable = true)
```

```
[195] 1 data_conv13.show()
```

| id | vendor_id | pickup_datetime | dropoff_datetime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitu |
|---|---|---|---|---|---|---|---|---|
| id2875421 | 2 | 2016-03-14 17:24:55 | 2016-03-14 17:32:30 | 1 | -73.9821548461914 | 40.76793670654297 | -73.96463012695312 | 40.7656021118164 |
| id2377394 | 1 | 2016-06-12 00:43:35 | 2016-06-12 00:54:38 | 1 | -73.98041534423828 | 40.738563537597656 | -73.99948120117188 | 40.731151580810 |
| id3858529 | 2 | 2016-01-19 11:35:24 | 2016-01-19 12:10:48 | 1 | -73.9790267944336 | 40.763938903808594 | -74.00533294677734 | 40.7100868225097 |
| id3504673 | 2 | 2016-04-06 19:32:31 | 2016-04-06 19:39:40 | 1 | -74.01004028320312 | 40.719970703125 | -74.0122806640625 | 40.706718444824 |
| id2181028 | 2 | 2016-03-26 13:30:55 | 2016-03-26 13:38:10 | 1 | -73.97305297851562 | 40.793209075927734 | -73.9729232788086 | 40.782520294185 |
| id0801584 | 2 | 2016-01-30 22:01:40 | 2016-01-30 22:09:03 | 6 | -73.98285675048828 | 40.74219512939453 | -73.99208068847656 | 40.7491836547851 |
| id1813257 | 1 | 2016-06-17 22:34:59 | 2016-06-17 22:40:40 | 4 | -73.9690170288086 | 40.75783920288086 | -73.95740509033203 | 40.765895843505 |
| id1324603 | 2 | 2016-05-21 07:54:58 | 2016-05-21 08:20:49 | 1 | -73.96927642822266 | 40.79777908325195 | -73.92247009277344 | 40.760559082031 |

0s completed at 10:01 PM

app.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Files

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

[196]
```
from pyspark.sql import SparkSession

spark = SparkSession.builder\
        .master("local[*]")\
        .appName("New York Cab Info")\
        .config('spark.ui.port', '4050')\
        .getOrCreate()
print("A Technical Case Study of New York Cab Trips")
```

```
A Technical Case Study of New York Cab Trips
```

[197]
```
# os.environ['PYSPARK_SUBMIT_ARGS']= spark.sparkContext.addPyFile('../')
```

```
spark
```

**SparkSession - in-memory**

**SparkContext**

Spark UI

Version
v3.1.1
Master
local
AppName
New York Taxi Trip Information

Sql query to display thw total trips

[209]
```
data_conv13.createOrReplaceTempView("data_conv13")
spark.sql("SELECT COUNT(id) AS total_trip from data_conv13").show()

```

```
+----------+
|total_trip|
+----------+
|   1458644|
+----------+
```

2s completed at 10:10 PM

app.ipynb

File Edit View Insert Runtime Tools Help

Comment Share

Files

+ Code + Text

RAM Disk

Editing

- ..
- drive
- sample_data
- spark-3.1.1-bin-hadoop2.7
  - R
  - bin
  - conf
  - data
  - examples
  - jars
  - kubernetes
  - licenses
  - python
  - sbin
  - yarn
  - LICENSE
  - NOTICE
  - README.md
  - RELEASE
- ngrok
- ngrok-stable-linux-amd64.zip
- ngrok-stable-linux-amd64.zip.1
- spark-3.1.1-bin-hadoop2.7.tgz
- spark-3.1.1-bin-hadoop2.7.tgz.1
- train.csv

Disk 67.30 GB available

To show the number of trips made according to the day of the week

```
[208]  1 spark.sql("SELECT pickup_day, COUNT(id) AS total_trips FROM data_conv13 GROUP BY pickup_day ORDER BY total_trips DESC").show()
       2
```

```
+----------+-----------+
|pickup_day|total_trips|
+----------+-----------+
|    Friday|     223533|
|  Saturday|     220868|
|  Thursday|     218574|
| Wednesday|     210136|
|   Tuesday|     202749|
|    Sunday|     195366|
|    Monday|     187418|
+----------+-----------+
```

number of trips made according to the day of the week

```
[210]  1 spark.sql("SELECT pickup_day, COUNT(vendor_id) AS total_trips FROM data_conv13 GROUP BY pickup_day ORDER BY total_trips DESC").show()
```

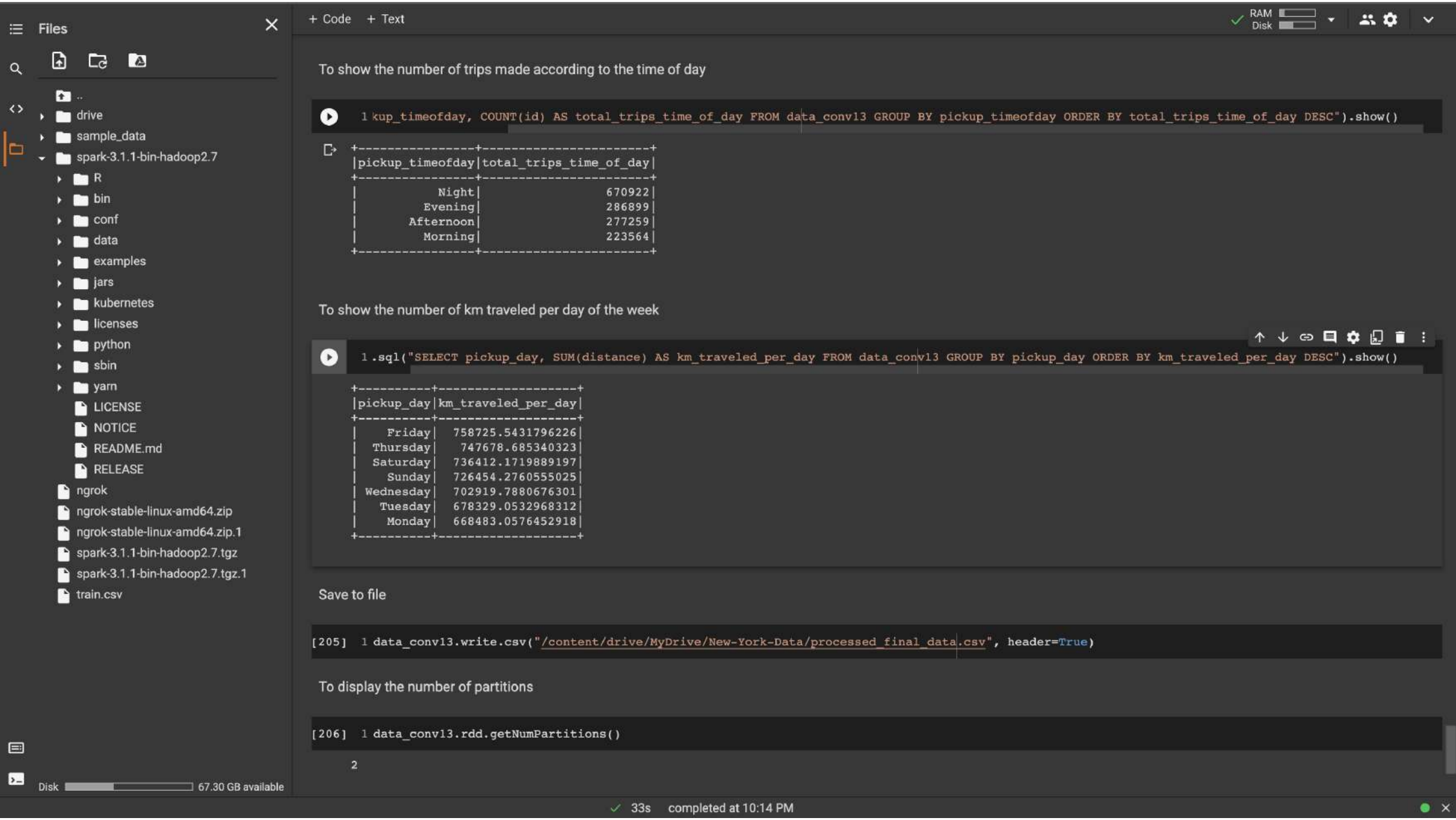
```
+----------+-----------+
|pickup_day|total_trips|
+----------+-----------+
|    Friday|     223533|
|  Saturday|     220868|
|  Thursday|     218574|
| Wednesday|     210136|
|   Tuesday|     202749|
|    Sunday|     195366|
|    Monday|     187418|
+----------+-----------+
```

To show the number of trips made according to the time of day

```
1 spark.sql("SELECT pickup_timeofday, COUNT(vendor_id) AS total_trips_time_of_day FROM data_conv13 GROUP BY pickup_timeofday ORDER BY total_trips_ti
```

```
+----------------+----------------------+
|pickup_timeofday|total_trips_time_of_day|
```

10s completed at 10:12 PM


Files
drive
sample_data
spark-3.1.1-bin-hadoop2.7
R
bin
conf
data
examples
jars
kubernetes
licenses
python
sbin
yarn
LICENSE
NOTICE
README.md
RELEASE
ngrok
ngrok-stable-linux-amd64.zip
ngrok-stable-linux-amd64.zip.1
spark-3.1.1-bin-hadoop2.7.tgz
spark-3.1.1-bin-hadoop2.7.tgz.1
train.csv
Disk 67.30 GB available
+ Code + Text
RAM
Disk
To show the number of trips made according to the time of day
kup_timeofday, COUNT(id) AS total_trips_time_of_day FROM data_conv13 GROUP BY pickup_timeofday ORDER BY total_trips_time_of_day DESC").show()
|pickup_timeofday|total_trips_time_of_day|
| Night| 670922|
| Evening| 286899|
| Afternoon| 277259|
| Morning| 223564|
To show the number of km traveled per day of the week
.sql("SELECT pickup_day, SUM(distance) AS km_traveled_per_day FROM data_conv13 GROUP BY pickup_day ORDER BY km_traveled_per_day DESC").show()
|pickup_day|km_traveled_per_day|
| Friday| 758725.5431796226|
| Thursday| 747678.685340323|
| Saturday| 736412.1719889197|
| Sunday| 726454.2760555025|
| Wednesday| 702919.7880676301|
| Tuesday| 678329.0532968312|
| Monday| 668483.0576452918|
Save to file
[205] data_conv13.write.csv("/content/drive/MyDrive/New-York-Data/processed_final_data.csv", header=True)
To display the number of partitions
[206] data_conv13.rdd.getNumPartitions()
2
33s completed at 10:14 PM