```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://bootcampms.s3.amazonaws.com/earthquakesclean.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("earthquakesclean.csv"), header=True, inferSchema=True, timestampFormat="yyyy/MM/dd HH:mm:ss")

# Show DataFrame
df.show()
```

```
+--------------------+--------+-------------------+-------------------+----------+-----+---------+--------------------+
|                Date|    Time|           Latitude|          Longitude|      Type|Depth|Magnitude|                  ID|
+--------------------+--------+-------------------+-------------------+----------+-----+---------+--------------------+
|1965-01-02 00:00:...|13:44:18|             19.246|            145.616|Earthquake|131.6|      6.0|        ISCGEM860706|
|1965-01-04 00:00:...|11:29:49| 1.8630000000000002|            127.352|Earthquake| 80.0|      5.8|        ISCGEM860737|
|1965-01-05 00:00:...|18:05:58|            -20.579|           -173.972|Earthquake| 20.0|      6.2|        ISCGEM860762|
|1965-01-08 00:00:...|18:49:43|            -59.076|            -23.557|Earthquake| 15.0|      5.8|        ISCGEM860856|
|1965-01-09 00:00:...|13:32:50| 11.937999999999999|            126.427|Earthquake| 15.0|      5.8|        ISCGEM860890|
|1965-01-10 00:00:...|13:36:32|            -13.405| 166.62900000000002|Earthquake| 35.0|      6.7|        ISCGEM860922|
|1965-01-12 00:00:...|13:32:25|             27.357|             87.867|Earthquake| 20.0|      5.9|        ISCGEM861007|
|1965-01-15 00:00:...|23:17:42|-13.309000000000001| 166.21200000000002|Earthquake| 35.0|      6.0|        ISCGEM861111|
|1965-01-16 00:00:...|11:32:37|            -56.452|-27.043000000000003|Earthquake| 95.0|      6.0|     ISCGEMSUP861125|
|1965-01-17 00:00:...|10:43:17|-24.563000000000002|            178.487|Earthquake|565.0|      5.8|        ISCGEM861148|
|1965-01-17 00:00:...|20:57:41|             -6.807|            108.988|Earthquake|227.9|      5.9|        ISCGEM861155|
|1965-01-24 00:00:...| 0:11:17|             -2.608| 125.95200000000001|Earthquake| 20.0|      8.2|        ISCGEM861299|
|1965-01-29 00:00:...| 9:35:30|             54.636|            161.703|Earthquake| 55.0|      5.5|        ISCGEM861461|
|1965-02-01 00:00:...| 5:27:06|            -18.697|           -177.864|Earthquake|482.9|      5.6|        ISCGEM859136|
|1965-02-02 00:00:...|15:56:51|             37.523|             73.251|Earthquake| 15.0|      6.0|        ISCGEM859164|
|1965-02-04 00:00:...| 3:25:00|             -51.84|            139.741|Earthquake| 10.0|      6.1|        ISCGEM859200|
|1965-02-04 00:00:...| 5:01:22| 51.251000000000005|            178.715|Earthquake| 30.3|      8.7|OFFICIAL196502040...|
|1965-02-04 00:00:...| 6:04:59| 51.638999999999996|            175.055|Earthquake| 30.0|      6.0|     ISCGEMSUP859215|
|1965-02-04 00:00:...| 6:37:06|             52.528|            172.007|Earthquake| 25.0|      5.7|        ISCGEM859221|
|1965-02-04 00:00:...| 6:39:32| 51.626000000000005| 175.74599999999998|Earthquake| 25.0|      5.8|        ISCGEM859222|
+--------------------+--------+-------------------+-------------------+----------+-----+---------+--------------------+
only showing top 20 rows
```

```
%pyspark
# Import date time functions
from pyspark.sql.functions import month, year
```

```
%pyspark
df.printSchema()
```

```
root
 |-- Date: string (nullable = true)
 |-- Time: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Type: string (nullable = true)
 |-- Depth: double (nullable = true)
 |-- Magnitude: double (nullable = true)
 |-- ID: string (nullable = true)
```

```
%pyspark
# Create a new DataFrame with the column Month
df.select(year(df["Date"])).show()
```

```
+----------+
|year(Date)|
+----------+
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
|      1965|
+----------+
only showing top 20 rows
```

```
%pyspark
# Save the year as a new column
df = df.withColumn("year", year(df['Date']))
df.show()
```

```
+--------------------+--------+------------------+------------------+----------+-----+---------+--------------------+----+
|                Date|    Time|          Latitude|         Longitude|      Type|Depth|Magnitude|                  ID|year|
+--------------------+--------+------------------+------------------+----------+-----+---------+--------------------+----+
|1965-01-02 00:00:...|13:44:18|            19.246|           145.616|Earthquake|131.6|      6.0|       ISCGEM860706|1965|
|1965-01-04 00:00:...|11:29:49|1.8630000000000002|           127.352|Earthquake| 80.0|      5.8|       ISCGEM860737|1965|
|1965-01-05 00:00:...|18:05:58|           -20.579|          -173.972|Earthquake| 20.0|      6.2|       ISCGEM860762|1965|
|1965-01-08 00:00:...|18:49:43|           -59.076|           -23.557|Earthquake| 15.0|      5.8|       ISCGEM860856|1965|
|1965-01-09 00:00:...|13:32:50|11.937999999999999|           126.427|Earthquake| 15.0|      5.8|       ISCGEM860890|1965|
|1965-01-10 00:00:...|13:36:32|           -13.405|166.62900000000002|Earthquake| 35.0|      6.7|       ISCGEM860922|1965|
|1965-01-12 00:00:...|13:32:25|            27.357|            87.867|Earthquake| 20.0|      5.9|       ISCGEM861007|1965|
|1965-01-15 00:00:...|23:17:42|-13.309000000000001|166.21200000000002|Earthquake| 35.0|      6.0|       ISCGEM861111|1965|
|1965-01-16 00:00:...|11:32:37|           -56.452|-27.043000000000003|Earthquake| 95.0|      6.0|    ISCGEMSUP861125|1965|
|1965-01-17 00:00:...|10:43:17|-24.563000000000002|           178.487|Earthquake|565.0|      5.8|       ISCGEM861148|1965|
|1965-01-17 00:00:...|20:57:41|            -6.807|           108.988|Earthquake|227.9|      5.9|       ISCGEM861155|1965|
|1965-01-24 00:00:...| 0:11:17|            -2.608|125.95200000000001|Earthquake| 20.0|      8.2|       ISCGEM861299|1965|
|1965-01-29 00:00:...| 9:35:30|            54.636|           161.703|Earthquake| 55.0|      5.5|       ISCGEM861461|1965|
|1965-02-01 00:00:...| 5:27:06|           -18.697|          -177.864|Earthquake|482.9|      5.6|       ISCGEM859136|1965|
|1965-02-02 00:00:...|15:56:51|            37.523|            73.251|Earthquake| 15.0|      6.0|       ISCGEM859164|1965|
|1965-02-04 00:00:...| 3:25:00|            -51.84|           139.741|Earthquake| 10.0|      6.1|       ISCGEM859200|1965|
|1965-02-04 00:00:...| 5:01:22| 51.251000000000005|           178.715|Earthquake| 30.3|      8.7|OFFICIAL196502040...|1965|
|1965-02-04 00:00:...| 6:04:59| 51.638999999999996|           175.055|Earthquake| 30.0|      6.0|    ISCGEMSUP859215|1965|
|1965-02-04 00:00:...| 6:37:06|            52.528|           172.007|Earthquake| 25.0|      5.7|       ISCGEM859221|1965|
|1965-02-04 00:00:...| 6:39:32| 51.626000000000005| 175.74599999999998|Earthquake| 25.0|      5.8|       ISCGEM859222|1965|
+--------------------+--------+------------------+------------------+----------+-----+---------+--------------------+----+
only showing top 20 rows
```

```
%pyspark
# Find the total earthquakes per year
averages = df.groupBy("year").count()
averages.orderBy("year").select("year", "count").show()
```

```
+----+-----+
|year|count|
+----+-----+
|1965|  339|
|1966|  234|
|1967|  255|
|1968|  305|
|1969|  323|
|1970|  345|
|1971|  386|
|1972|  388|
|1973|  401|
|1974|  361|
|1975|  412|
|1976|  457|
|1977|  425|
|1978|  410|
|1979|  356|
|1980|  348|
|1981|  321|
|1982|  346|
|1983|  453|
|1984|  482|
+----+-----+
only showing top 20 rows
```

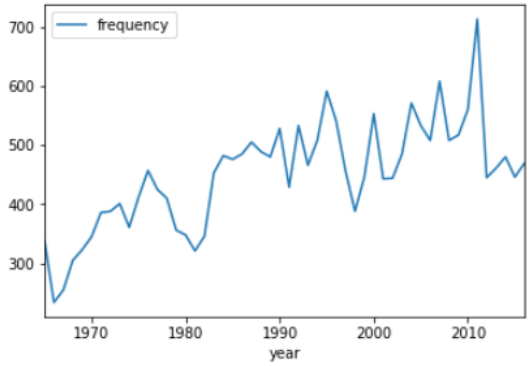
```
%pyspark
# Import the summarized data to a pandas DataFrame for plotting
# Note: If your summarized data is still too big for your local memory then your notebook may crash
import pandas as pd
pandas_df = averages.orderBy("year").select("year", "count").toPandas()
pandas_df.head()
```

```
   year  count
0  1965    339
1  1966    234
2  1967    255
3  1968    305
4  1969    323
```

```
%pyspark
# Clean the data and rename the columns to "year" and "count"
pandas_df = pandas_df.dropna()
pandas_df = pandas_df.rename(columns={"count": "frequency"})
pandas_df.head()
```

```
   year  frequency
0  1965        339
1  1966        234
2  1967        255
3  1968        305
4  1969        323
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa765952250><Figure size 432x288 with 1 Axes>
```




```
%pyspark
diff = pandas_df.diff(periods=1)
diff=diff[1:]
diff.head()
```

```
   year  frequency
1   1.0     -105.0
2   1.0       21.0
3   1.0       50.0
4   1.0       18.0
5   1.0       22.0
```

Interpreter: .