(/U4G66226D/spaces)

```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://s3.amazonaws.com/dataviz-curriculum/day_1/bigfoot.csv"
spark.sparkContext.addFile(url)
df = spark.read.csv(SparkFiles.get("bigfoot.csv"), header=True, inferSchema=True, timestampFormat="yyyy/MM/dd HH:mm

# Show DataFrame
df.show()
```

```
+------+-------------------+-------------+-------------------+--------+---------+
|number|              title|classification|          timestamp|latitude|longitude|
+------+-------------------+-------------+-------------------+--------+---------+
|   637|Report 637: Campe...|       Class A|2000-06-16T12:00:00Z|    61.5|   -142.9|
|  2917|Report 2917: Fami...|       Class A|1995-05-15T12:00:00Z| 55.1872|-132.7982|
|  7963|Report 7963: Sasq...|       Class A|2004-02-09T12:00:00Z| 55.2035|-132.8202|
|  9317|Report 9317: Driv...|       Class A|2004-06-18T12:00:00Z| 62.9375|-141.5667|
| 13038|Report 13038: Sno...|       Class A|2004-02-15T12:00:00Z| 61.0595|-149.7853|
| 23666|Report 23666: Pas...|       Class A|2008-04-23T12:00:00Z|62.77335|-141.3165|
| 26604|Report 26604: Day...|       Class A|2009-07-15T12:00:00Z|64.89139|-147.8142|
|   179|Report 179: Man a...|       Class A|1981-09-15T12:00:00Z|32.31435|-85.16235|
|   245|Report 245: Two o...|       Class A|1999-07-15T12:00:00Z|33.28375|-87.32655|
|   416|Report 416: A res...|       Class A|1983-11-15T12:00:00Z|34.95605| -86.4559|
|   435|Report 435: Dayli...|       Class A|2000-10-10T12:00:00Z| 34.5422|-86.66465|
|   451|Report 451: Young...|       Class A|1993-08-20T12:00:00Z| 34.9263|-87.02025|
|   577|Report 577: Man h...|       Class A|1999-11-15T12:00:00Z|34.80405|-87.50905|
|   799|Report 799: Perso...|       Class A|1978-04-15T12:00:00Z|34.92855| -87.1105|
|   832|Report 832: Witne...|       Class A|1980-11-15T12:00:00Z|33.13195|-88.17885|
|   961|Report 961: Motor...|       Class A|1997-01-06T12:00:00Z| 31.4515|-88.08305|
|  1022|Report 1022: Hunt...|       Class A|1990-09-15T12:00:00Z|33.97575|-87.45876|
|  1907|Report 1907: Moto...|       Class A|1996-12-05T12:00:00Z|31.58255|-87.96095|
|  3028|Report 3028: Dayl...|       Class A|2000-06-01T12:00:00Z| 34.4881| -86.6333|
|  3296|Report 3296: Man ...|       Class A|2001-10-15T12:00:00Z| 34.6802|-87.00665|
+------+-------------------+-------------+-------------------+--------+---------+
only showing top 20 rows
```

Run Started Juno

Interpreter: spark.pyspark. **FINISHED** Took 29 sec 854 millisec. Updated by ars0107 on February 01 2019, 2:24:50 PM (CST)

```
%pyspark (/U4G66226D/spaces)
# Import date time functions
from pyspark.sql.functions import month, year
```

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

```
+---------------+
|year(timestamp)|
+---------------+
|           2000|
|           1995|
|           2004|
|           2004|
|           2004|
|           2008|
|           2009|
|           1981|
|           1999|
|           1983|
|           2000|
|           1993|
|           1999|
|           1978|
|           1980|
|           1997|
|           1990|
|           1996|
|           2000|
|           2001|
+---------------+
only showing top 20 rows
```

```
%pyspark (/U4G66226D/spaces)
# Save the year as a new column
df = df.withColumn("year", year(df['timestamp']))
df.show()
```

```
+------+--------------------+--------------+--------------------+--------+---------+----+
|number|               title|classification|           timestamp|latitude|longitude|year|
+------+--------------------+--------------+--------------------+--------+---------+----+
|   637|Report 637: Campe...|       Class A|2000-06-16T12:00:00Z|    61.5|   -142.9|2000|
|  2917|Report 2917: Fami...|       Class A|1995-05-15T12:00:00Z| 55.1872|-132.7982|1995|
|  7963|Report 7963: Sasq...|       Class A|2004-02-09T12:00:00Z| 55.2035|-132.8202|2004|
|  9317|Report 9317: Driv...|       Class A|2004-06-18T12:00:00Z| 62.9375|-141.5667|2004|
| 13038|Report 13038: Sno...|       Class A|2004-02-15T12:00:00Z| 61.0595|-149.7853|2004|
| 23666|Report 23666: Pas...|       Class A|2008-04-23T12:00:00Z|62.77335|-141.3165|2008|
| 26604|Report 26604: Day...|       Class A|2009-07-15T12:00:00Z|64.89139|-147.8142|2009|
|   179|Report 179: Man a...|       Class A|1981-09-15T12:00:00Z|32.31435|-85.16235|1981|
|   245|Report 245: Two o...|       Class A|1999-07-15T12:00:00Z|33.28375|-87.32655|1999|
|   416|Report 416: A res...|       Class A|1983-11-15T12:00:00Z|34.95605| -86.4559|1983|
|   435|Report 435: Dayli...|       Class A|2000-10-10T12:00:00Z| 34.5422|-86.66465|2000|
|   451|Report 451: Young...|       Class A|1993-08-20T12:00:00Z| 34.9263|-87.02025|1993|
|   577|Report 577: Man h...|       Class A|1999-11-15T12:00:00Z|34.80405|-87.50905|1999|
|   799|Report 799: Perso...|       Class A|1978-04-15T12:00:00Z|34.92855| -87.1105|1978|
|   832|Report 832: Witne...|       Class A|1980-11-15T12:00:00Z|33.13195|-88.17885|1980|
|   961|Report 961: Motor...|       Class A|1997-01-06T12:00:00Z| 31.4515|-88.08305|1997|
|  1022|Report 1022: Hunt...|       Class A|1990-09-15T12:00:00Z|33.97575|-87.45876|1990|
|  1907|Report 1907: Moto...|       Class A|1996-12-05T12:00:00Z|31.58255|-87.96095|1996|
|  3028|Report 3028: Dayl...|       Class A|2000-06-01T12:00:00Z| 34.4881| -86.6333|2000|
|  3296|Report 3296: Man ...|       Class A|2001-10-15T12:00:00Z| 34.6802|-87.00665|2001|
+------+--------------------+--------------+--------------------+--------+---------+----+
only showing top 20 rows
```

```
%pyspark (/U4G66226D/spaces)
# Find the total bigfoot sightings per year
averages = df.groupBy("year").count()
averages.orderBy("year").select("year", "count").show()
```

```
+----+-----+
|year|count|
+----+-----+
|null|    6|
|1869|    1|
|1921|    1|
|1925|    1|
|1930|    1|
|1932|    1|
|1934|    1|
|1937|    1|
|1938|    1|
|1941|    1|
|1942|    1|
|1944|    2|
|1947|    1|
|1948|    1|
|1949|    2|
|1950|    3|
|1952|    1|
|1953|    2|
|1954|    1|
|1955|    3|
+----+-----+
only showing top 20 rows
```

```
%pyspark (/U4G66226D/spaces)
# 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()
```

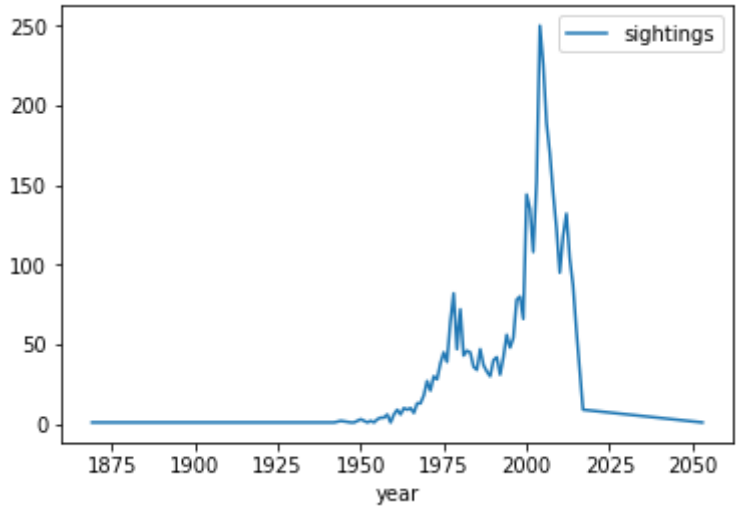
```
     year  count
0     NaN      6
1  1869.0      1
2  1921.0      1
3  1925.0      1
4  1930.0      1
```

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

```
     year  sightings
1  1869.0          1
2  1921.0          1
3  1925.0          1
4  1930.0          1
5  1932.0          1
```

```
%pyspark (/U4G66226D/spaces)
# Plot the year and sightings
%matplotlib inline
pandas_df.plot("year", "sightings")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f42032b8b50>
```




```
<Figure size 432x288 with 1 Axes>
```