```
%pyspark
# Show schema to confirm date type
df.printSchema()
```

(/U4G66226D/spaces)

```
root
 |-- station: string (nullable = true)
 |-- date: timestamp (nullable = true)
 |-- prcp: double (nullable = true)
 |-- tobs: integer (nullable = true)
```

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

Run

Started Juno

```
%pyspark (/U4G66226D/spaces)
# Show the year for the date column
df.select(year(df["date"])).show()
```

```
+----------+
|year(date)|
+----------+
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
|      2010|
+----------+
only showing top 20 rows
```

Run

Started Juno

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

```
+-----------+-------------------+----+----+----+
|    station|               date|prcp|tobs|year|
+-----------+-------------------+----+----+----+
|USC00519397|2010-01-01 00:00:...|0.08|  65|2010|
|USC00519397|2010-01-02 00:00:...| 0.0|  63|2010|
|USC00519397|2010-01-03 00:00:...| 0.0|  74|2010|
|USC00519397|2010-01-04 00:00:...| 0.0|  76|2010|
|USC00519397|2010-01-07 00:00:...|0.06|  70|2010|
|USC00519397|2010-01-08 00:00:...| 0.0|  64|2010|
|USC00519397|2010-01-09 00:00:...| 0.0|  68|2010|
|USC00519397|2010-01-10 00:00:...| 0.0|  73|2010|
|USC00519397|2010-01-11 00:00:...|0.01|  64|2010|
|USC00519397|2010-01-12 00:00:...| 0.0|  61|2010|
|USC00519397|2010-01-14 00:00:...| 0.0|  66|2010|
|USC00519397|2010-01-15 00:00:...| 0.0|  65|2010|
|USC00519397|2010-01-16 00:00:...| 0.0|  68|2010|
|USC00519397|2010-01-17 00:00:...| 0.0|  64|2010|
|USC00519397|2010-01-18 00:00:...| 0.0|  72|2010|
|USC00519397|2010-01-19 00:00:...| 0.0|  66|2010|
|USC00519397|2010-01-20 00:00:...| 0.0|  66|2010|
|USC00519397|2010-01-21 00:00:...| 0.0|  69|2010|
|USC00519397|2010-01-22 00:00:...| 0.0|  67|2010|
|USC00519397|2010-01-23 00:00:...| 0.0|  67|2010|
+-----------+-------------------+----+----+----+
only showing top 20 rows
```

Run

```
%pyspark (/U4G66226D/spaces)
# Find the average precipitation per year
averages = df.groupBy("year").avg()
averages.orderBy("year").select("year", "avg(prcp)").show()
```

```
+----+------------------+
|year|          avg(prcp)|
+----+------------------+
|2010|0.13852293920179035|
|2011| 0.1637348927875241|
|2012| 0.1163805668016194|
|2013|0.15554567502020986|
|2014|0.17855953372189803|
|2015|0.19919999999999985|
|2016|0.17984533591106822|
|2017|0.16592738752959774|
+----+------------------+
```

```
%pyspark
from pyspark.sql.functions import month
df.select(month(df['Date'])).show()
```

```
+-----------+
|month(Date)|
+-----------+
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
|          1|
+-----------+
only showing top 20 rows
```

```
%pyspark
df = df.withColumn("month", month(df['date']))
df.head()
```

```
Row(station=u'USC00519397', date=datetime.datetime(2010, 1, 1, 0, 0), prcp=0.08, tobs=65, year=2010, month=1)
```

```
%pyspark (/U4G66226D/spaces)
averages = df.groupBy("month").max()
averages.orderBy("month").select("month", "max(prcp)").show()
```

```
+-----+---------+
|month|max(prcp)|
+-----+---------+
|    1|     8.81|
|    2|     5.04|
|    3|     6.38|
|    4|     6.25|
|    5|     4.07|
|    6|     4.43|
|    7|    11.53|
|    8|     4.81|
|    9|     6.83|
|   10|     4.47|
|   11|     8.06|
|   12|     6.42|
+-----+---------+
```

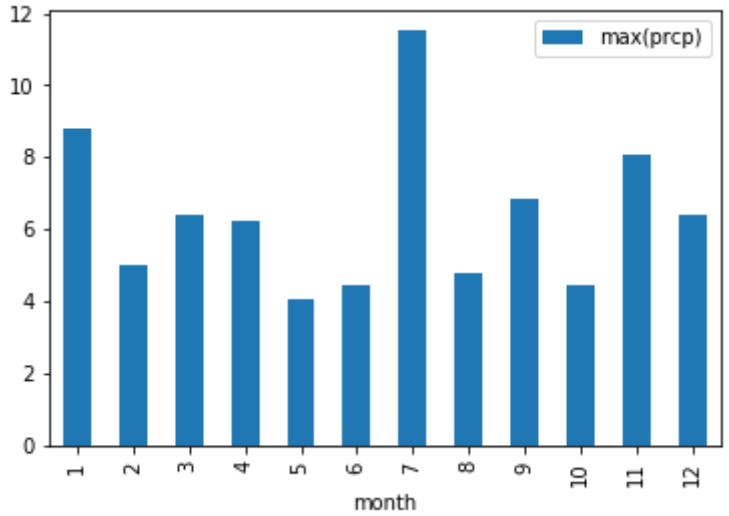
```
%pyspark
# Import the summarized data to a pandas dataframe for plotting
# Note: If your summarized data is still too big for your local memory then your notebook may crash

pandas_df = averages.orderBy("month").select("month", "max(prcp)").toPandas()
pandas_df.head()
```

```
   month  max(prcp)
0      1       8.81
1      2       5.04
2      3       6.38
3      4       6.25
4      5       4.07
```

```
%pyspark
import matplotlib.pyplot as plt
pandas_df.set_index("month", inplace=True)
pandas_df.plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3cc9a60450>
```




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

Interpreter: spark.pyspark. **FINISHED** Took 1 sec 137 millisec. Updated by ars0107 on February 01 2019, 2:18:17 PM (CST)

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

```
root
 |-- station: string (nullable = true)
 |-- date: timestamp (nullable = true)
 |-- prcp: double (nullable = true)
 |-- tobs: integer (nullable = true)
 |-- year: integer (nullable = true)
 |-- month: integer (nullable = true)
```

Started Juno

Interpreter: unknown.