



## **Death Cause Analytics using Spark and Zeppelin**

### **Extra Credit Project**

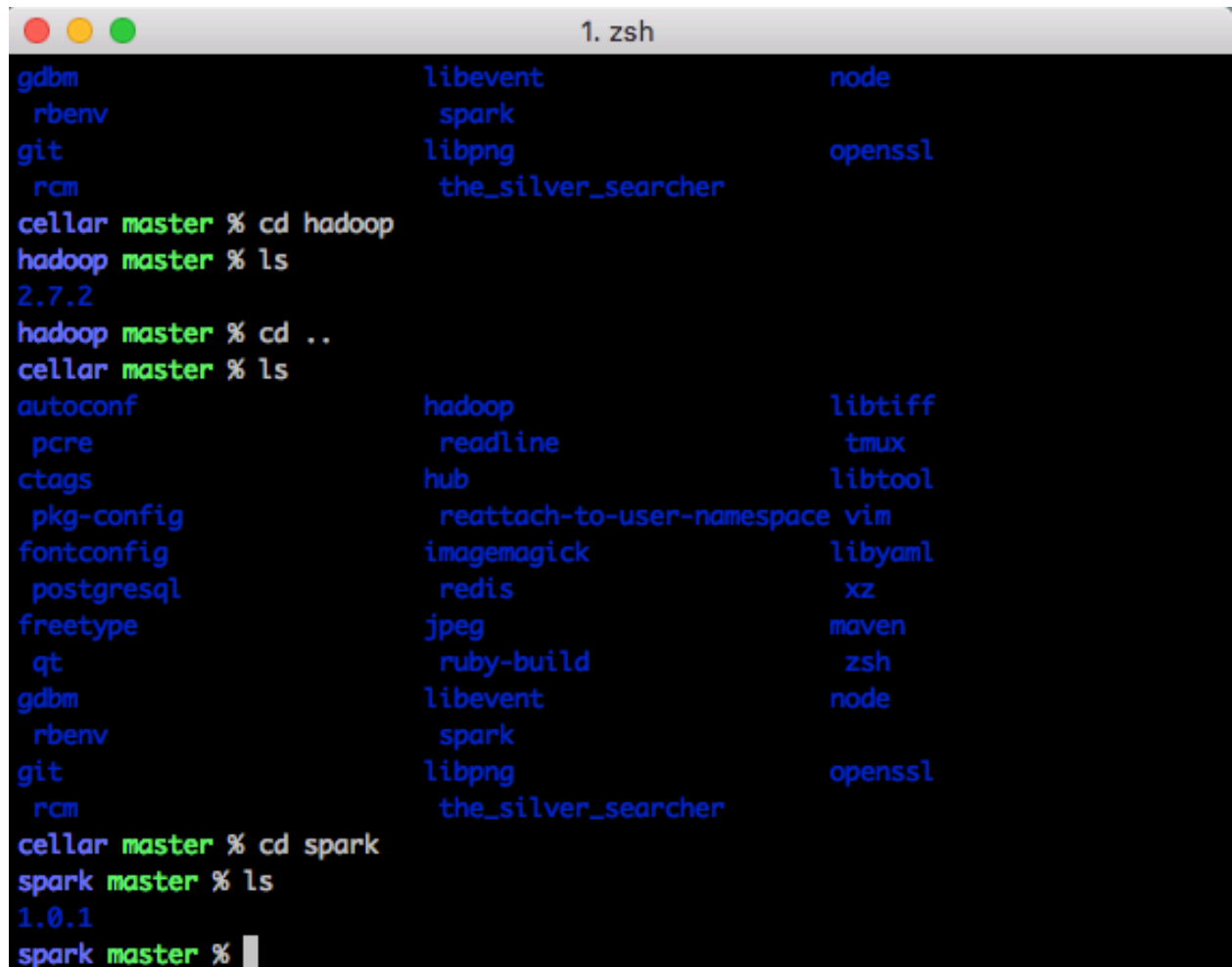
**Submitted By:** Shreyams Jain (010736427)

**CMPE 272 Instructor:** Prof. Rakesh Ranjan

**GitHub link:** <https://github.com/ShreyamsJain/Data-Analytics-using-Apache-Zeppelin-and-Spark>

## Overview of the Steps taken to perform Death Cause Analytics:

- 1) Installed Spark, Hadoop using brew package manager
- 2) Shows spark and Hadoop installed



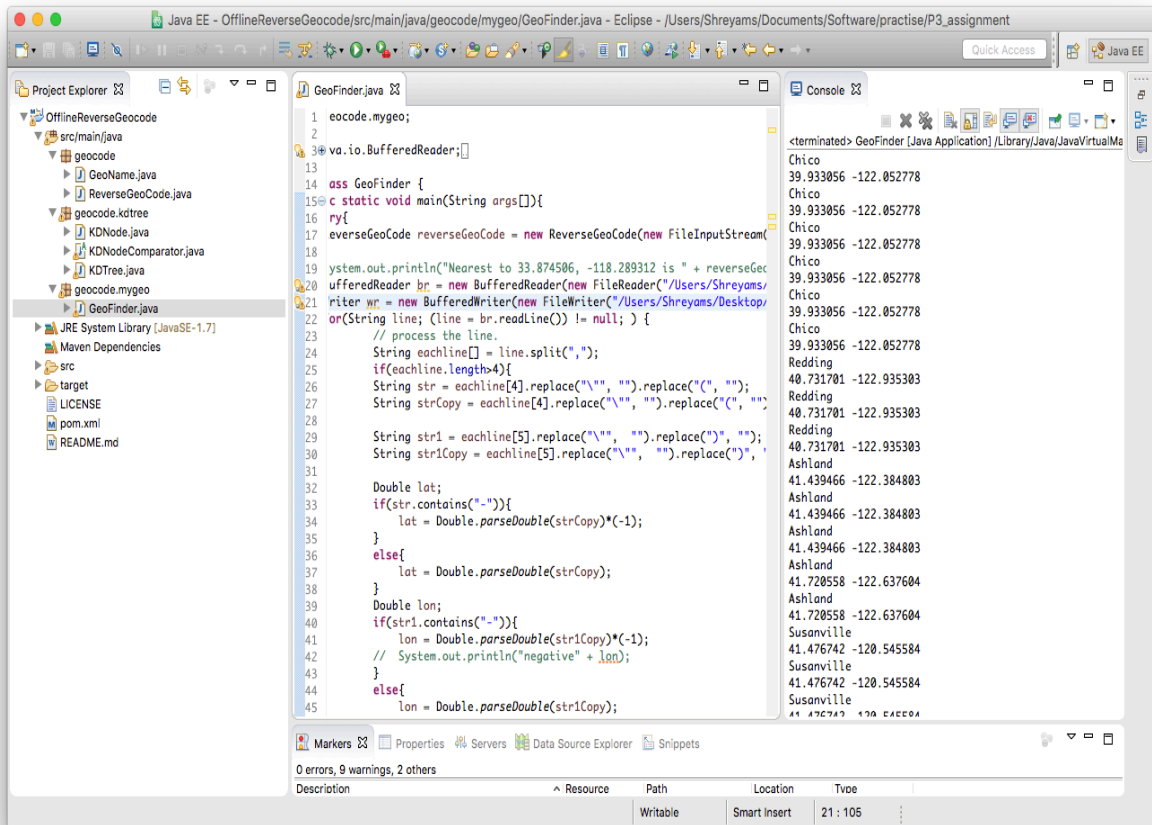
```
1. zsh
gdbm      libevent  node
rbenv     spark
git       libpng   openssl
rcm       the_silver_searcher

cellar master % cd hadoop
hadoop master % ls
2.7.2
hadoop master % cd ..
cellar master % ls
autoconf  hadoop      libtiff
pcre      readline  tmux
ctags     hub        libtool
pkg-config reattach-to-user-namespace vim
fontconfig imagemagick libyaml
postgresql redis      xz
freetype  jpeg       maven
qt        ruby-build zsh
gdbm      libevent  node
rbenv     spark
git       libpng   openssl
rcm       the_silver_searcher

cellar master % cd spark
spark master % ls
1.0.1
spark master %
```

- 3) Cloned Apache Zeppelin from GitHub and configured Apache Zeppelin in local System.
- 4) Downloaded Death Cause Datasets from the link <http://catalog.data.gov/dataset/leading-causes-of-death-by-zip-code-1999-2013>.
- 5) Dataset had location coordinates (longitude, latitude), location was not mentioned.

6) Used reverse geocoder API and wrote a Java Program to convert the location coordinates to the appropriate city in California



7) Used Scala to load the datasets and convert to RDD.

8) Used SparkSQL to perform Death Cause analytics.

9) 16 Questions have been answered with the respective screenshots below.

Started zeppelin

```
1. zsh
651 cd..
652 cd ..
653 s
654 cd\\n
656 pwd
657 kill -9 661
658 top -u
659 kill -9 358
660 exit
661 cd desktop
663 cd cmpe_273
665 cd zeppelin
667 cd in*
669 cd bin
670 ls
bin master % ./bin zeppelin-daemon.sh start
zsh: no such file or directory: ./bin
bin master % cd ..
incubator-zeppelin master % ./bin zeppelin-daemon.sh start
zsh: permission denied: ./bin
incubator-zeppelin master % zeppelin-daemon.sh start
zsh: command not found: zeppelin-daemon.sh
incubator-zeppelin master % bin/zeppelin-daemon.sh start
Zeppelin start [ OK ]
incubator-zeppelin master %
```

Opened zeppelin at localhost:8080 and loaded death Cause csv file using Scala

The screenshot shows the Zeppelin Notebook interface in a web browser. The address bar indicates the URL is localhost:8080/#/notebook/2A94M5J1Z. The Zeppelin logo and navigation tabs (Notebook, Interpreter, Configuration) are visible. A search bar and a 'Connected' status indicator are also present. The main content area displays a Scala script for loading and processing death cause data. The script includes imports for IOUtils, URL, and Charset, followed by comments explaining the context. It then loads a CSV file, defines a case class for DeathCause, and registers a temporary table. The script is marked as 'FINISHED' with a blue play button icon. The execution time is shown as 'Took 9 seconds. Last updated by undefined at time May 19, 2016 12:57:15 AM.'

```
## Welcome to Zeppelin,Shreyams
#### This is a live tutorial, you can run the code yourself. (Shift-Enter to Run)

Welcome to Zeppelin,Shreyams
This is a live tutorial, you can run the code yourself. (Shift-Enter to Run)
Took 2 seconds. Last updated by undefined at time May 12, 2016 7:38:07 PM. (outdated)

Load data into table
import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset

// Zeppelin creates and injects sc (SparkContext) and sqlContext (HiveContext or SqlContext)
// So you don't need create them manually

// load bank data
val bankText = sc.textFile("/Users/Shreyams/desktop/cmpe_273/zeppelin/incubator-zeppelin/datafile/locationOutput1.csv")

case class DeathCause(Year: Integer, ZIP_Code: Integer, Causes_of_Death: String, Count: Integer, Location: String)

val deathCause = bankText.map(s => s.split(",")).filter(s => s(0) != "Year").map(s => DeathCause(s(0).toInt, s(1).toInt, s(2), s(3).toInt, s(4))).toDF()
deathCause.registerTempTable("deathCause")

import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset
bankText: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:26
defined class DeathCause
deathCause: org.apache.spark.sql.DataFrame = [Year: int, ZIP_Code: int, Causes_of_Death: string, Count: int, Location: string]

Took 9 seconds. Last updated by undefined at time May 19, 2016 12:57:15 AM.
```

- 1) Find death count for all years from 1999 to 2013.  
-Below is the list of deaths for each year.

The screenshot shows the Zeppelin Notebook interface in a web browser. The browser address bar shows `localhost:8080/#/notebook/2A94M5J1Z`. The Zeppelin header includes the logo, navigation links (Notebook, Interpreter, Configuration), a search bar, and a 'Connected' status. The notebook title is 'Zeppelin'. Below the title, a status bar indicates 'Took 2 seconds. Last updated by undefined at time May 13, 2016 1:45:11 AM.'

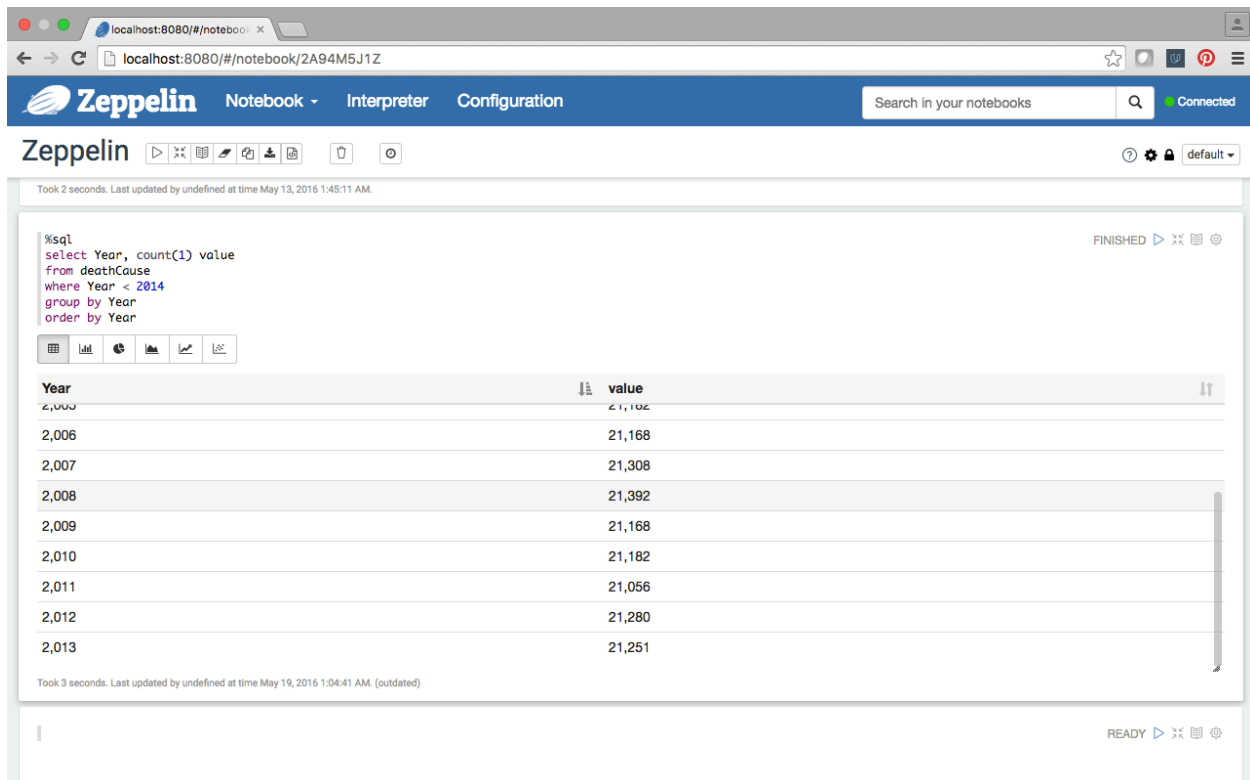
The main area displays an SQL query:

```
%sql
select Year, count(1) value
from deathCause
where Year < 2014
group by Year
order by Year
```

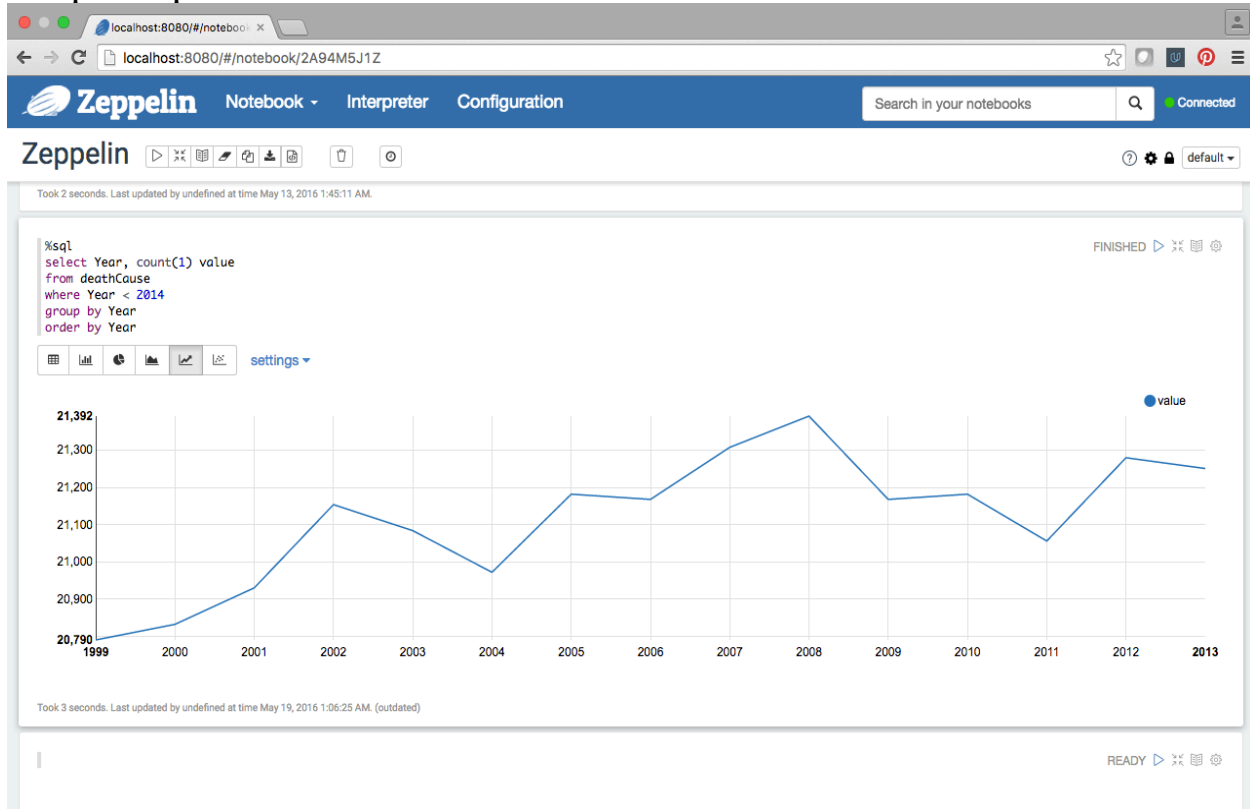
The query has been executed, as indicated by the 'FINISHED' status and a play button icon. Below the query, a table of results is shown:

Year	value
1,999	20,790
2,000	20,832
2,001	20,930
2,002	21,154
2,003	21,084
2,004	20,972
2,005	21,182
2,006	21,168
2,007	21,308

Below the table, a status bar indicates 'Took 3 seconds. Last updated by undefined at time May 19, 2016 1:04:41 AM. (outdated)'. At the bottom of the notebook, a 'READY' status is shown with a play button icon.

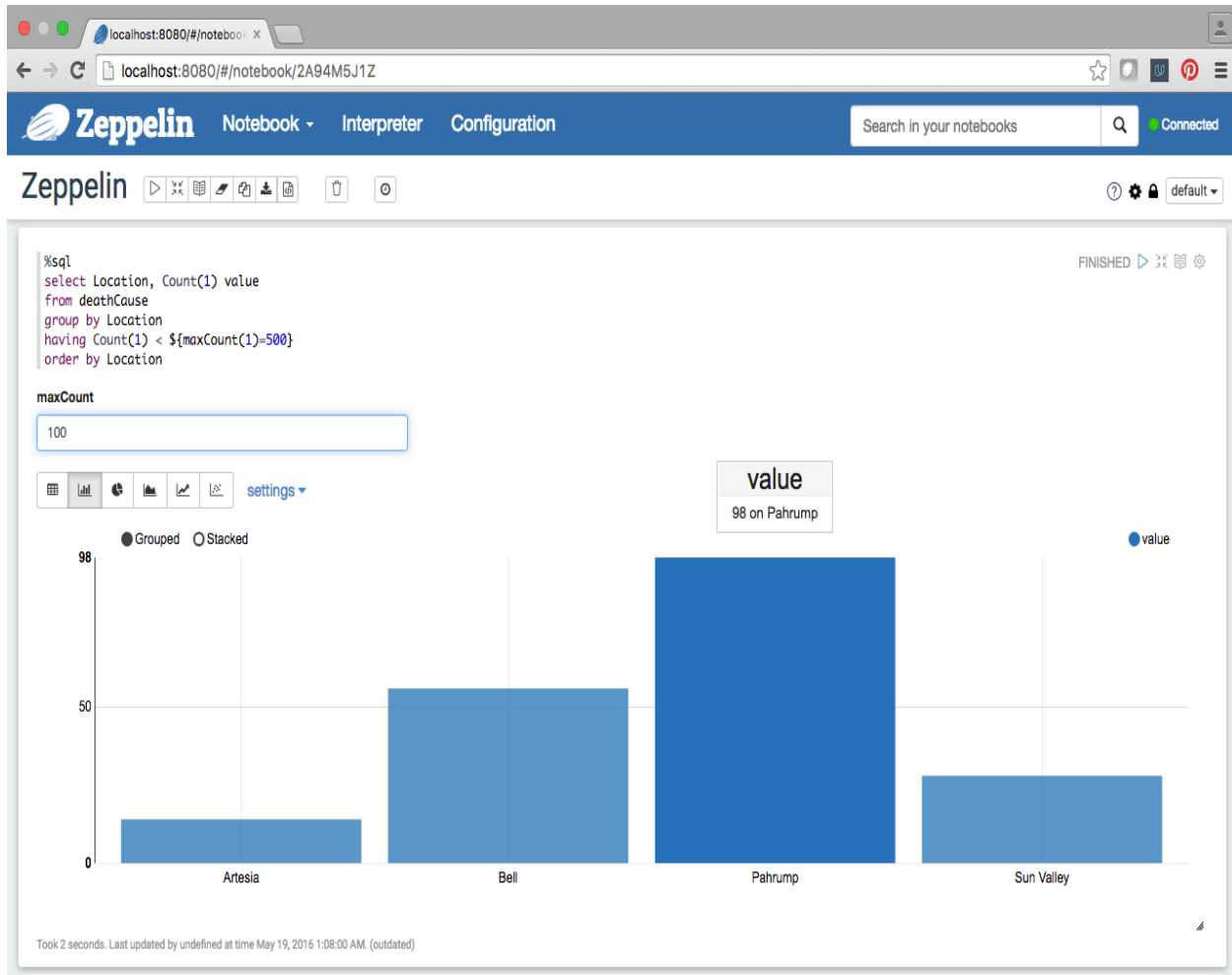


## Graph Representation:



2) Find location where death count is less than hundred.

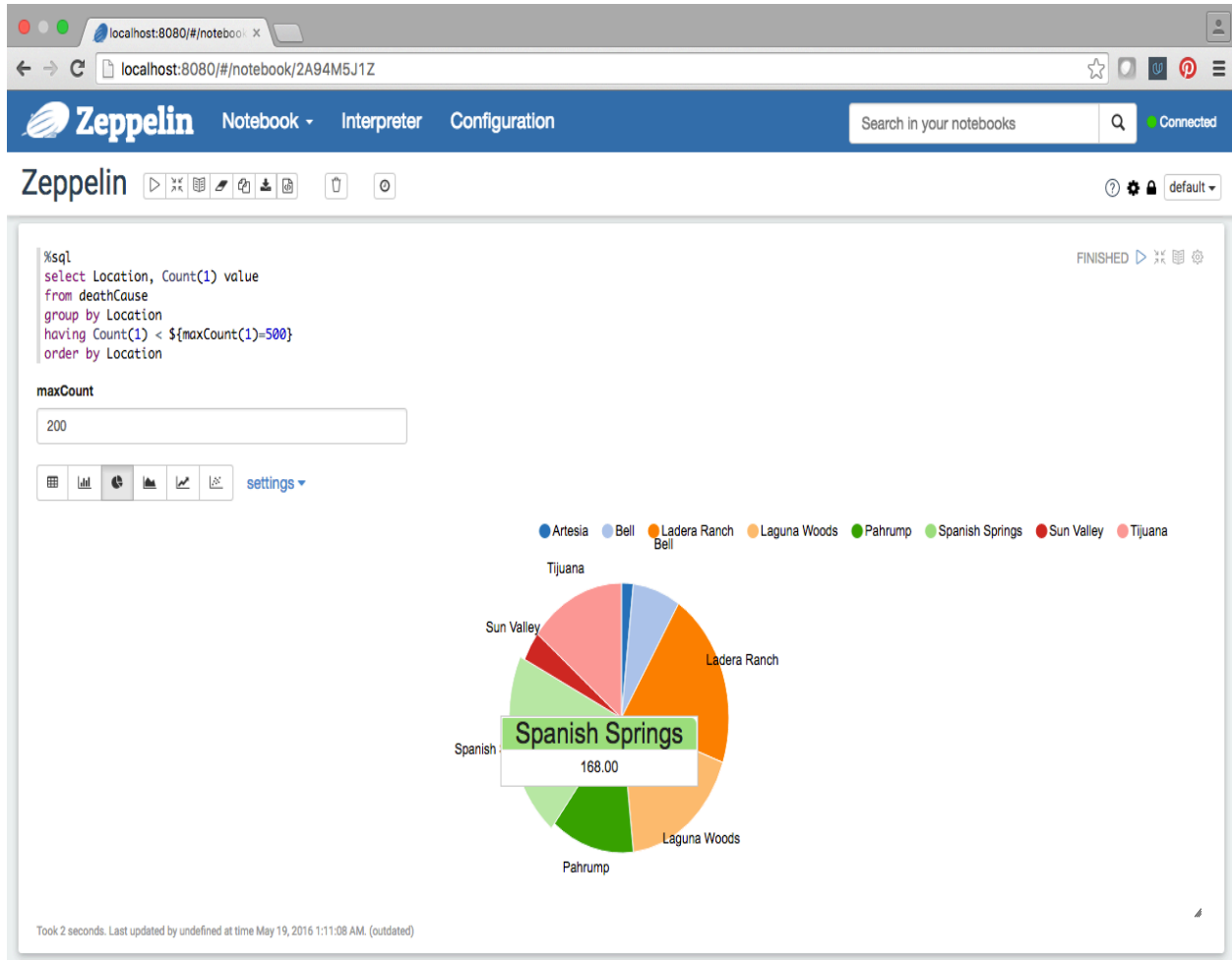
-I have added random value entry 100. Dynamically values can be added to find out respective data



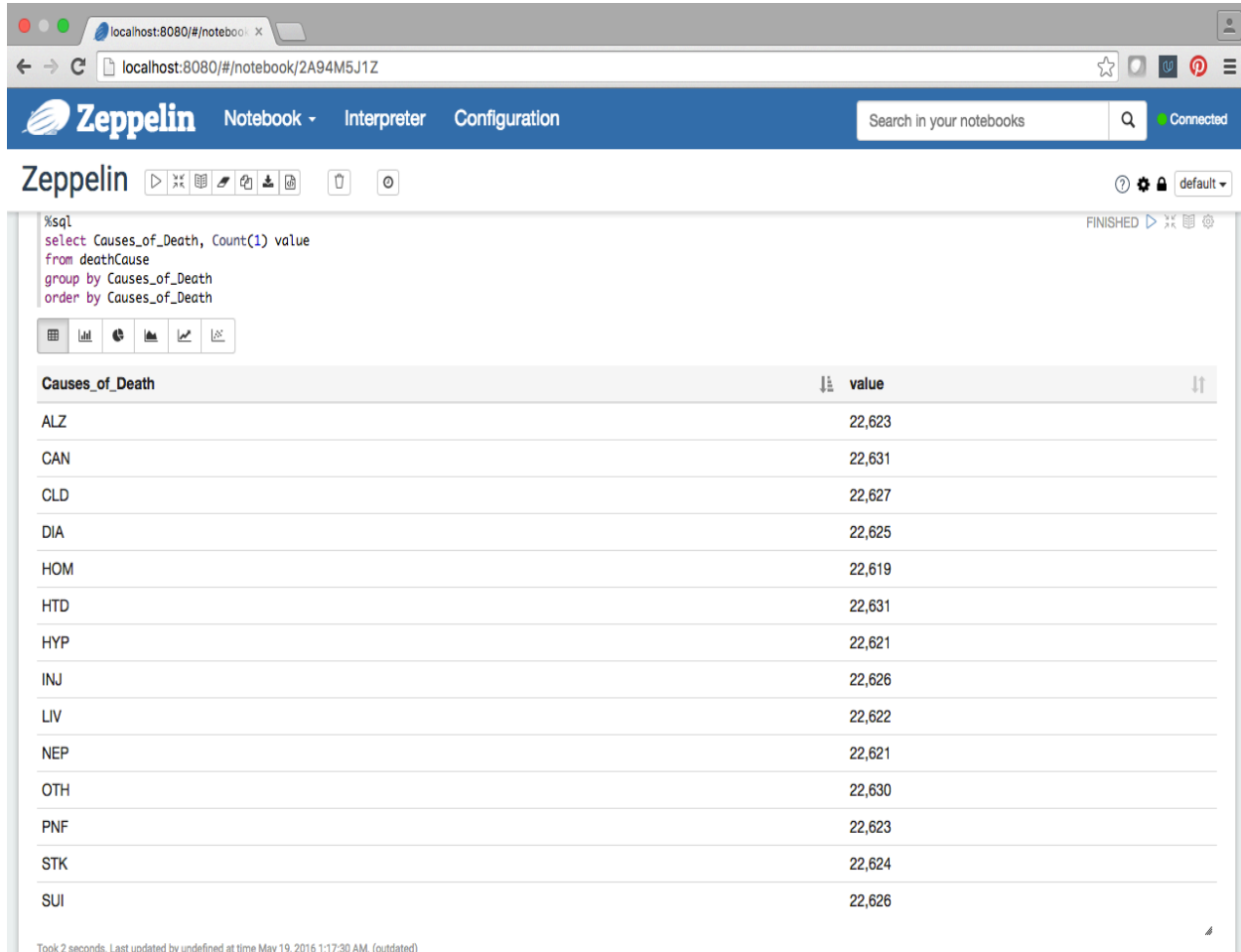


3) Find the list of locations with deaths less than 200 and represent it in a Pie Chart.

-Pie Chart for Locations with deaths less than 200



- 4) Find out which disease caused the highest number of deaths in California during the period 1999-2013?
- Cancer and HTD with number of deaths equal to 22,631 each.



The image shows a Zeppelin Notebook interface. At the top, there's a browser window with the URL `localhost:8080/#/notebook/2A94M5J1Z`. The Zeppelin header includes the logo, 'Notebook', 'Interpreter', and 'Configuration' tabs, along with a search bar and a 'Connected' status. Below the header, the notebook name 'Zeppelin' is displayed. The main area contains a SQL query and its results.

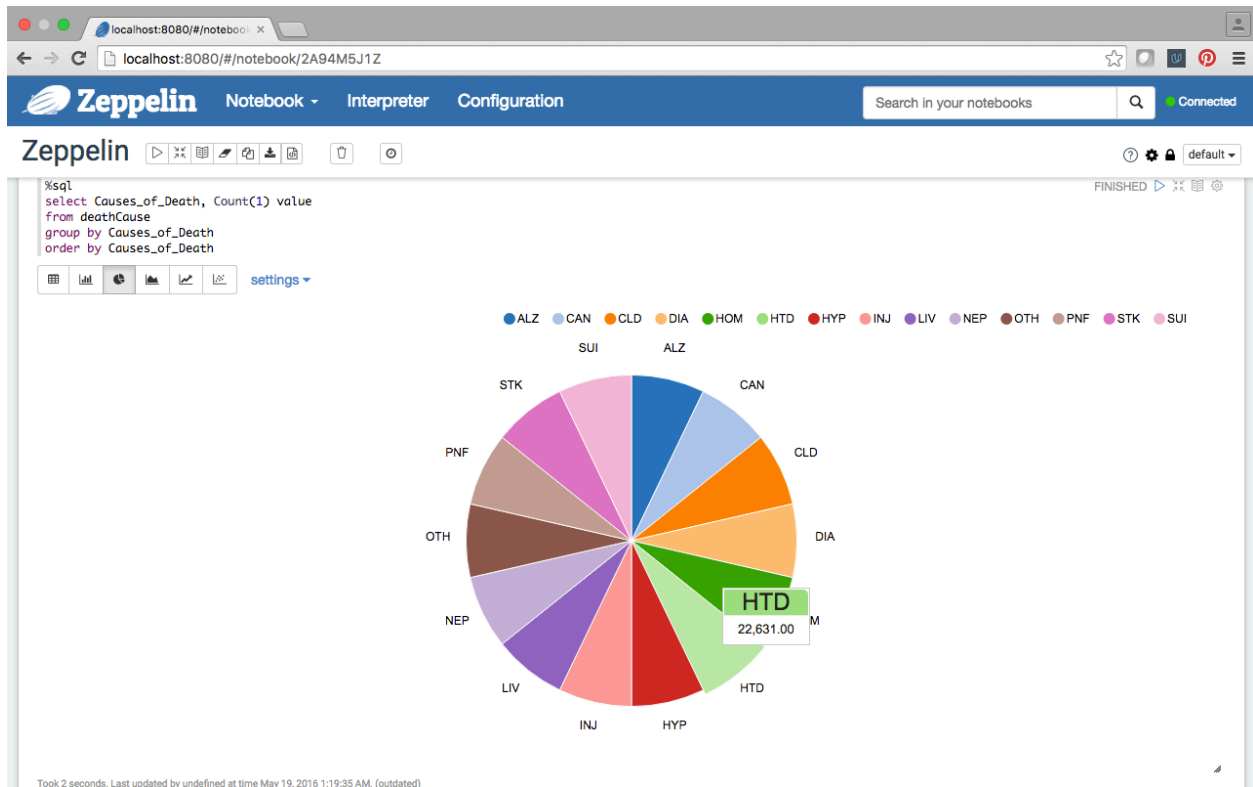
```
%sql
select Causes_of_Death, Count(1) value
from deathCause
group by Causes_of_Death
order by Causes_of_Death
```

The results are displayed in a table with two columns: 'Causes\_of\_Death' and 'value'. The table lists 15 causes of death and their corresponding counts.

Causes_of_Death	value
ALZ	22,623
CAN	22,631
CLD	22,627
DIA	22,625
HOM	22,619
HTD	22,631
HYP	22,621
INJ	22,626
LIV	22,622
NEP	22,621
OTH	22,630
PNF	22,623
STK	22,624
SUI	22,626

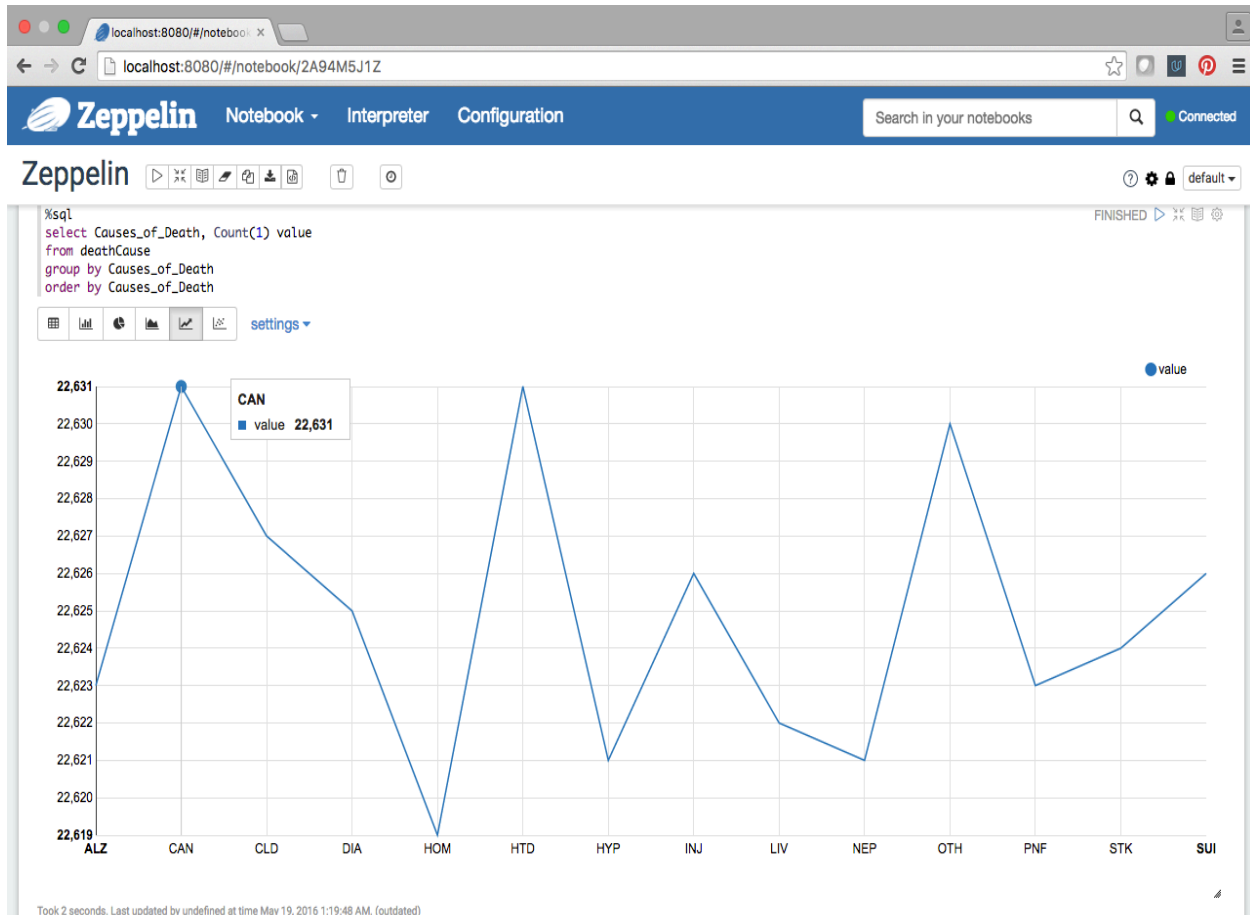
At the bottom, a status bar indicates 'Took 2 seconds. Last updated by undefined at time May 19, 2016 1:17:30 AM. (outdated)'.

# Pie chart



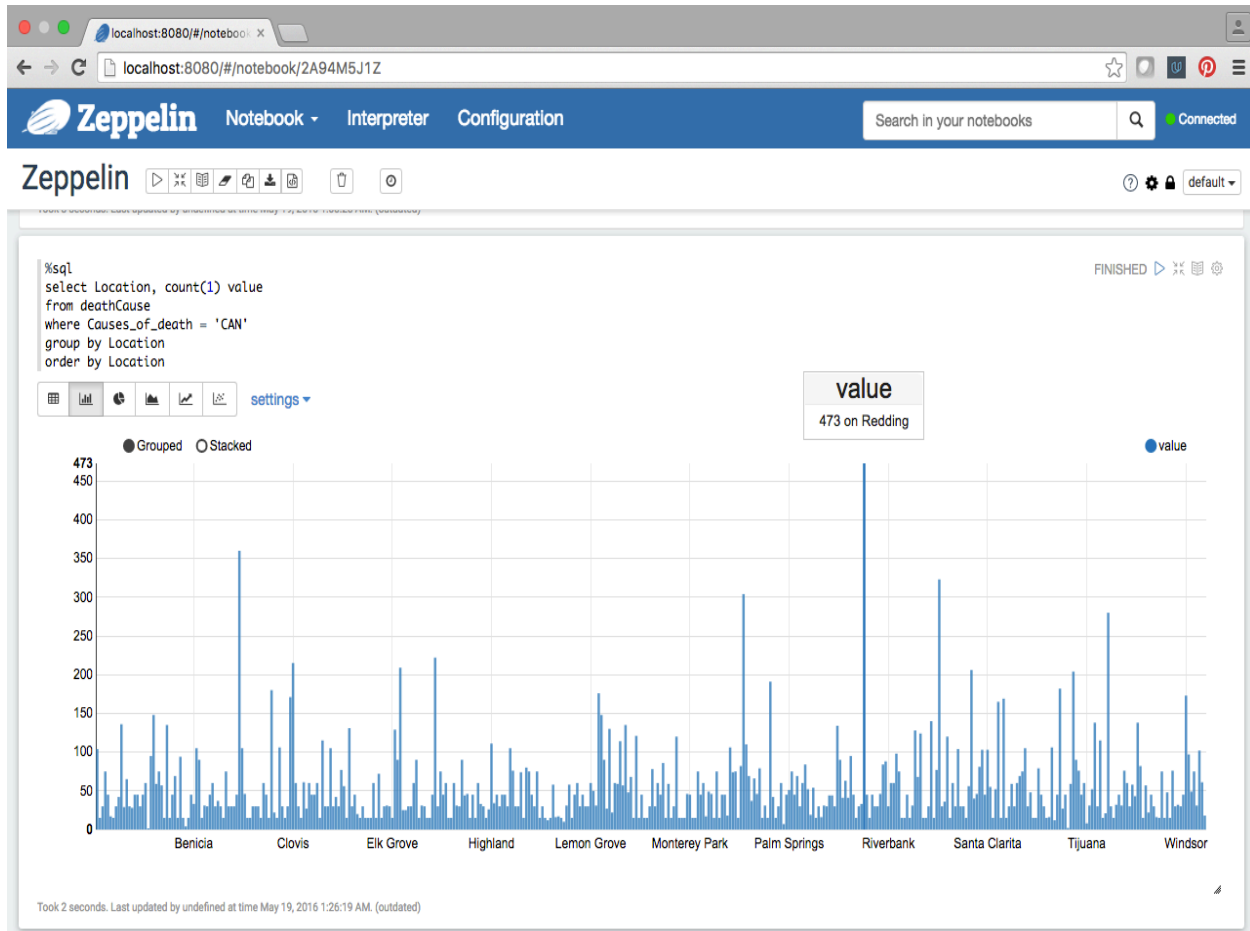
5) Which disease has caused the least number of deaths?

- HOM



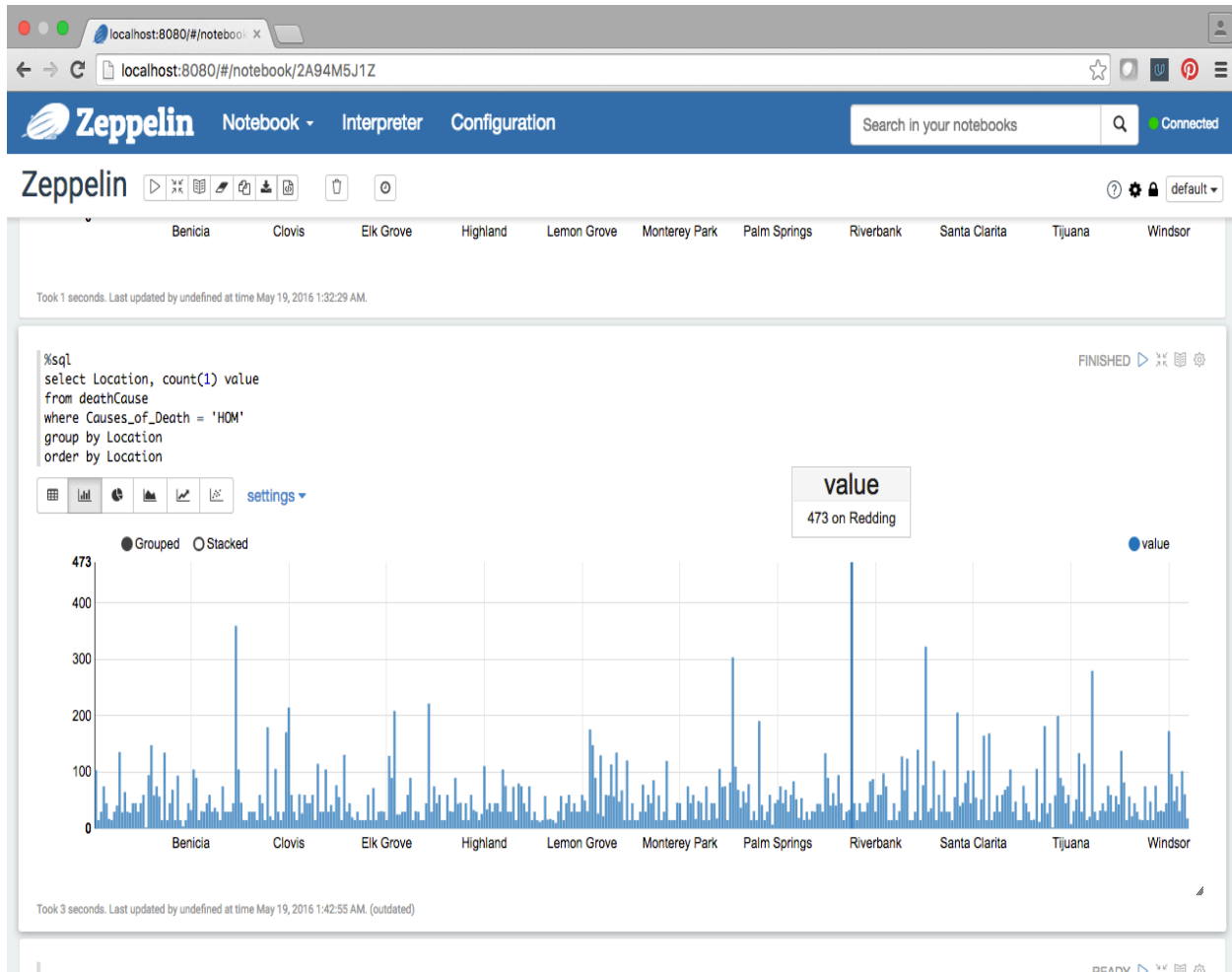
6) Which location has highest number of Death due to Cancer ?

- Redding, CA with 473 deaths

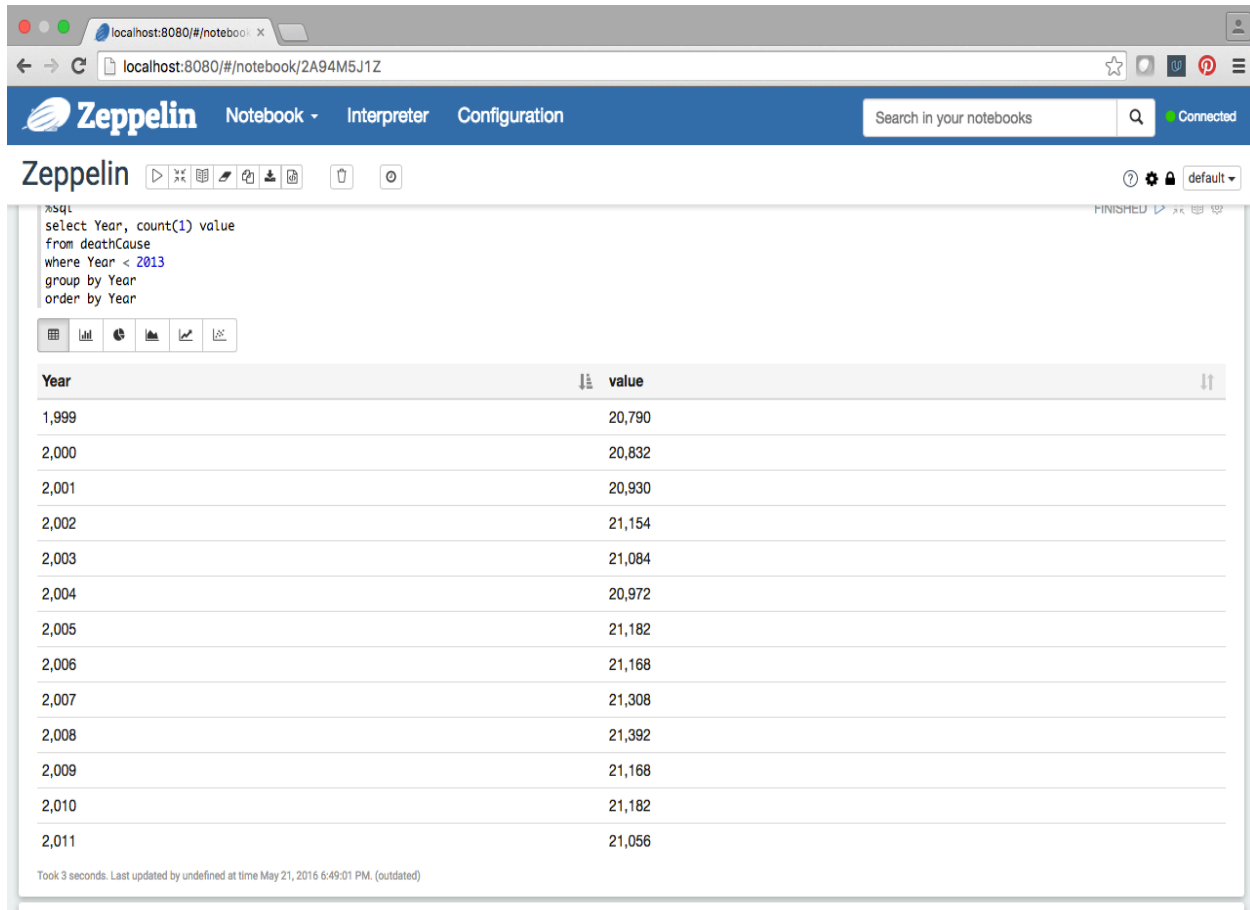


7) Which location has highest number of Death due to HOM ?

- Redding, CA – 473 deaths



- 8) In which year has the number of deaths been the least?
- Year 1999, 20790 deaths



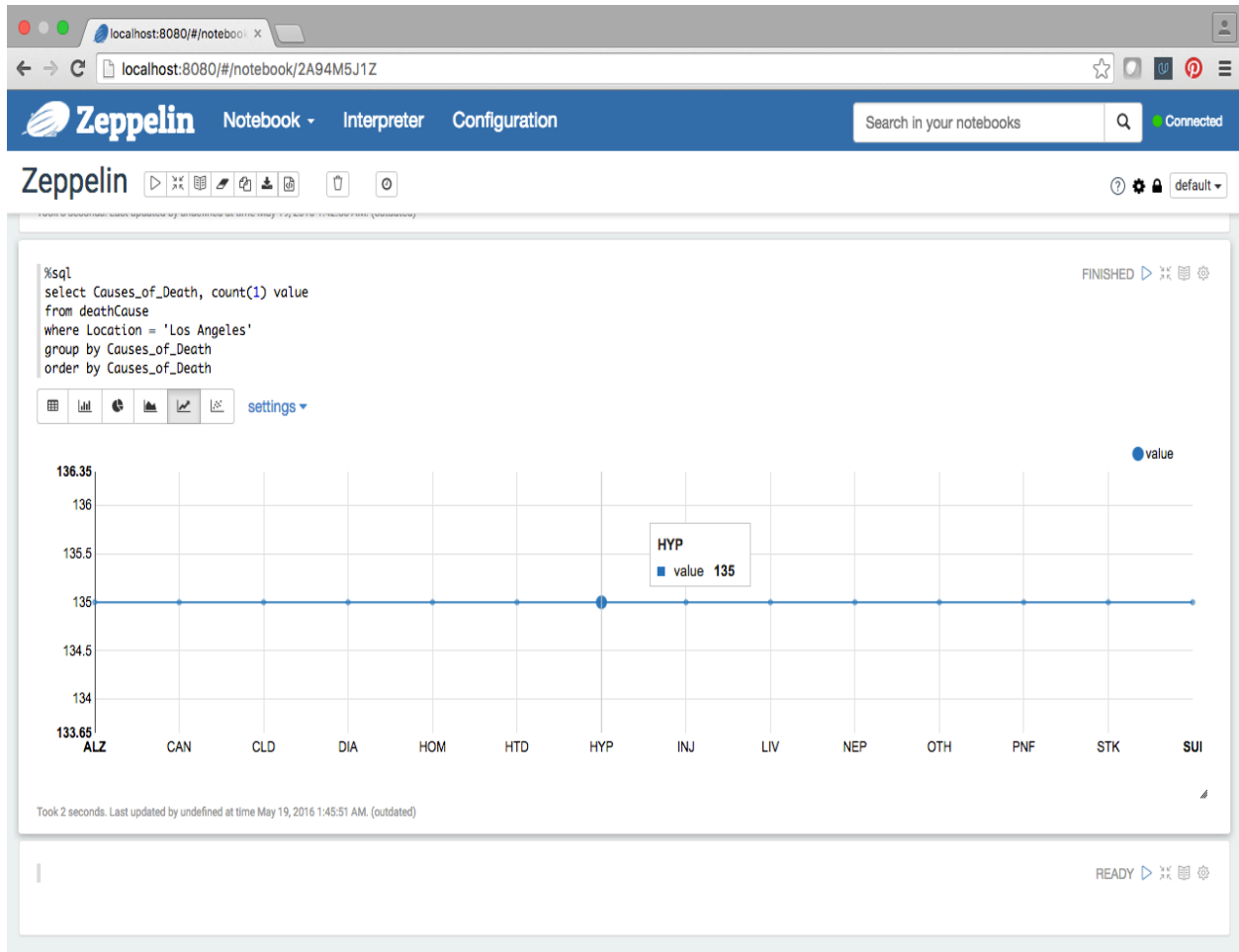
The screenshot shows the Zeppelin Notebook interface. The top navigation bar includes the Zeppelin logo, 'Notebook', 'Interpreter', and 'Configuration' tabs. A search bar and a 'Connected' status indicator are on the right. Below the navigation bar, the 'Zeppelin' title is followed by a toolbar with icons for running, saving, and other actions. The main area displays an SQL query and its results in a table.

```
select Year, count(1) value
from deathCause
where Year < 2013
group by Year
order by Year
```

Year	value
1,999	20,790
2,000	20,832
2,001	20,930
2,002	21,154
2,003	21,084
2,004	20,972
2,005	21,182
2,006	21,168
2,007	21,308
2,008	21,392
2,009	21,168
2,010	21,182
2,011	21,056

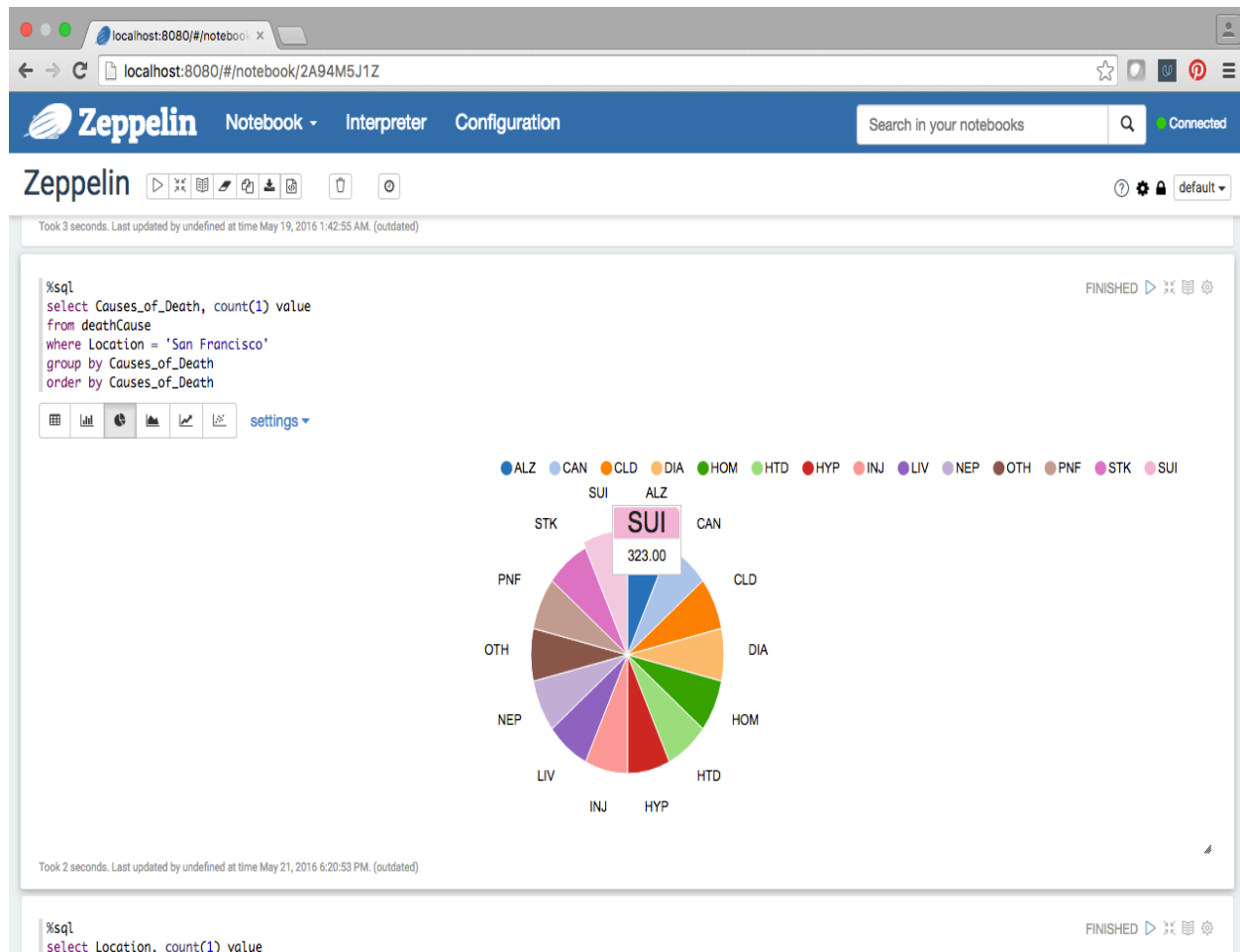
Took 3 seconds. Last updated by undefined at time May 21, 2016 6:49:01 PM. (outdated)

9) Find out the No. of deaths in Los Angeles for all the diseases?

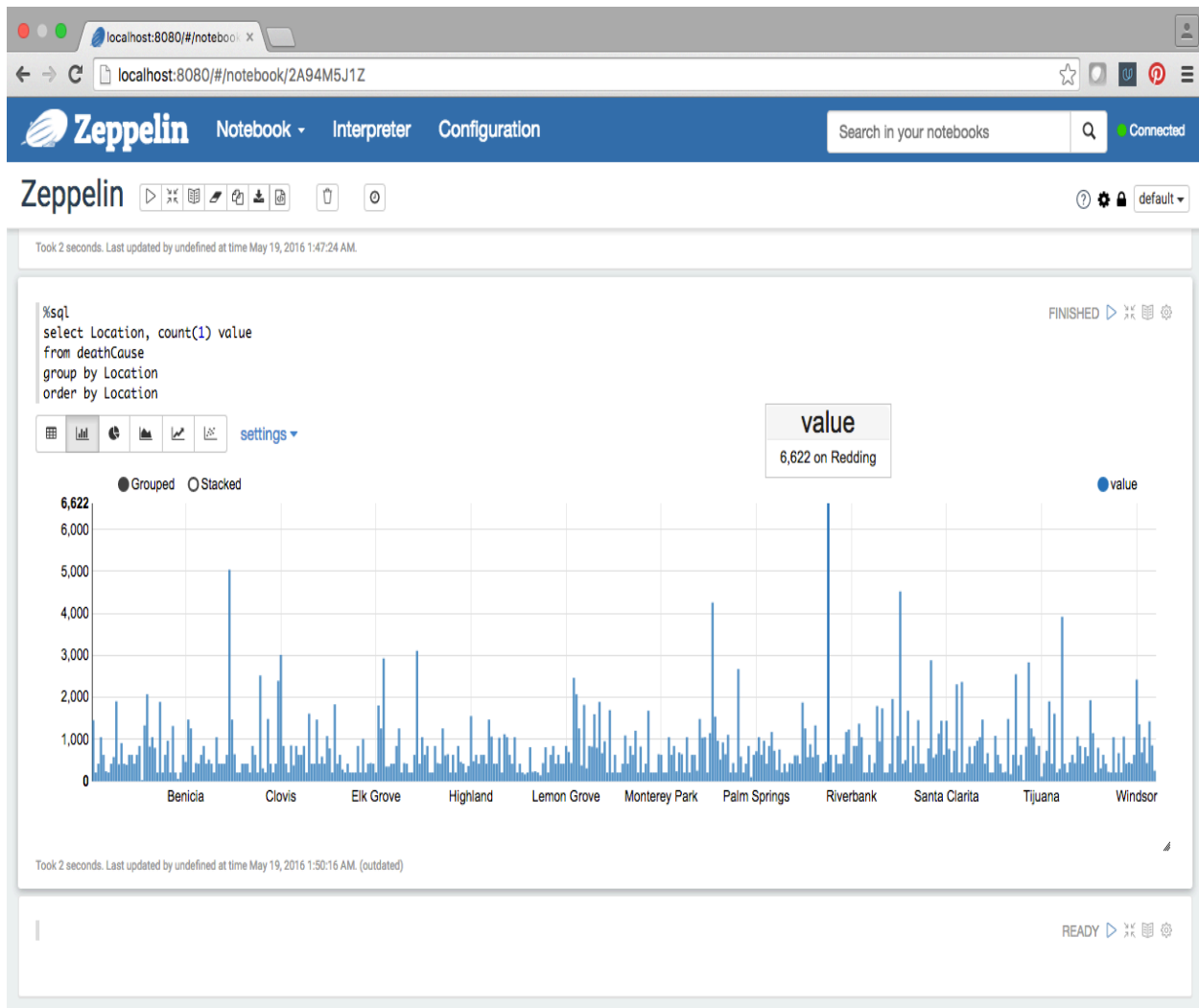




10) What is number of deaths in San Francisco due to Suicides?  
-323 Suicides count from 1999-2013

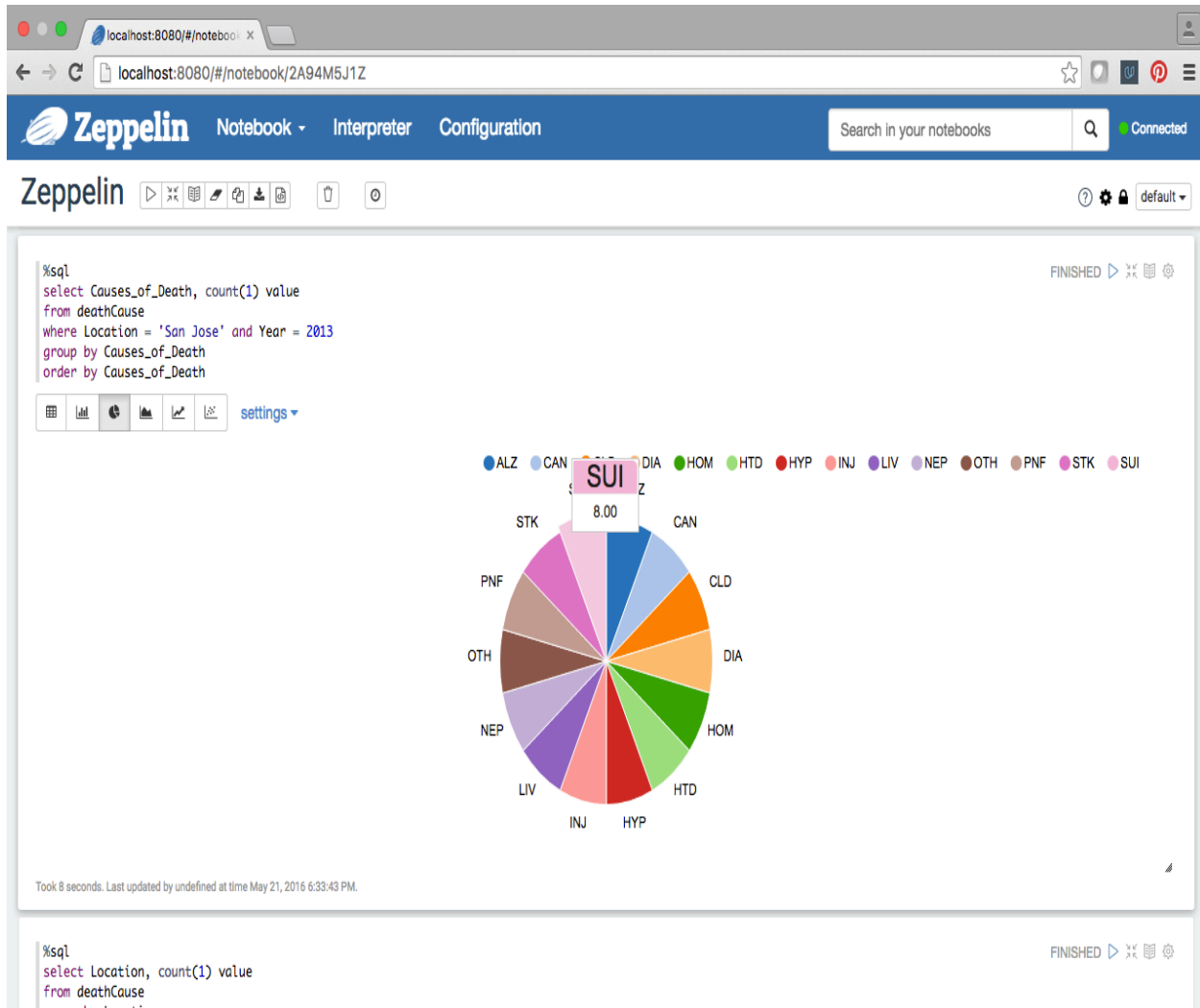


11) Which city has highest number of deaths in California?  
- Redding with 6622 deaths from 1999-2013

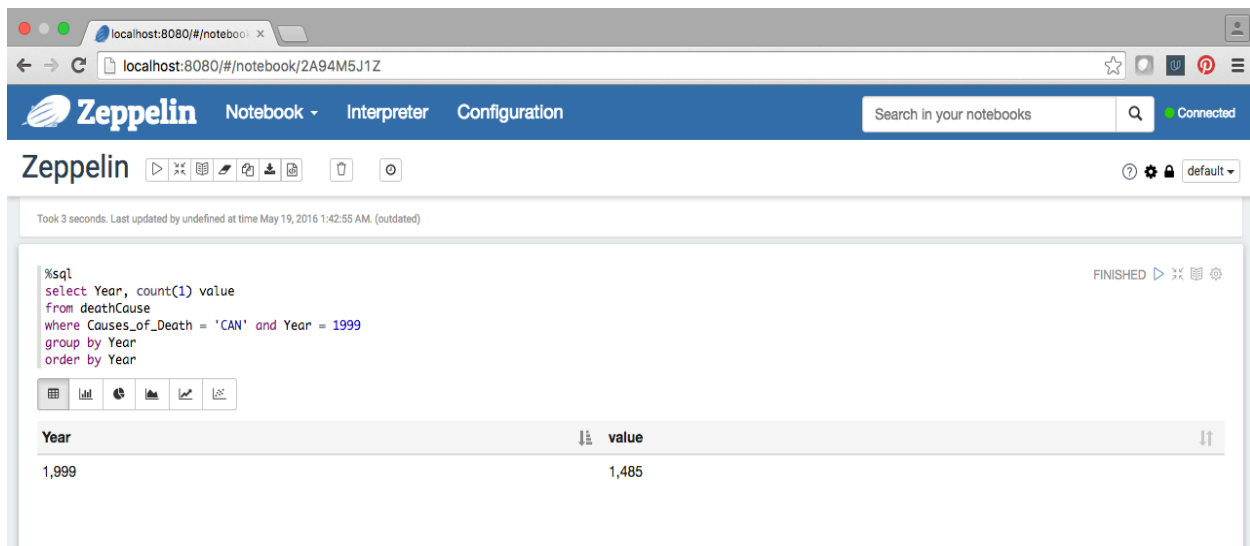


12) Find out the number of deaths due to Suicide in San Jose, CA in the year 2013?

- No. of Suicides in San Jose in 2013 is 8.



- 13) One in how many people died due to Cancer in the year 1999?
- California Population in 1999, 33.5 million
  - $33500000 / 1485 = 22,559$ .
  - One out of 22,559 people dies of Cancer in California in the Year 1999



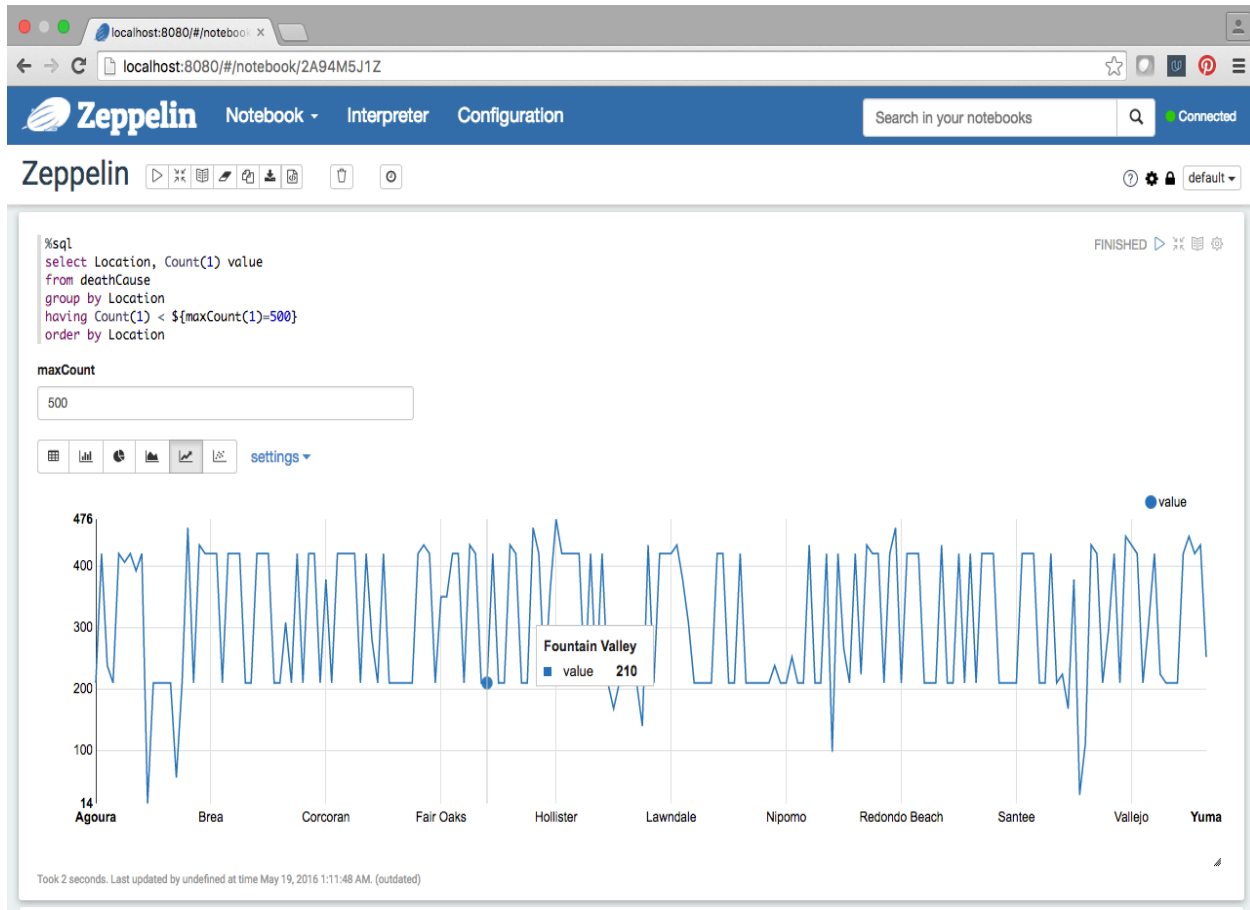
The screenshot shows the Zeppelin Notebook interface in a web browser. The URL is `localhost:8080/#/notebook/2A94M5J1Z`. The notebook has tabs for "Notebook", "Interpreter", and "Configuration". A search bar is present with the text "Search in your notebooks". The notebook content area shows a SQL query:

```
%sql
select Year, count(1) value
from deathCause
where Causes_of_Death = 'CAN' and Year = 1999
group by Year
order by Year
```

The query status is "FINISHED". Below the query, there is a table with the following data:

Year	value
1,999	1,485

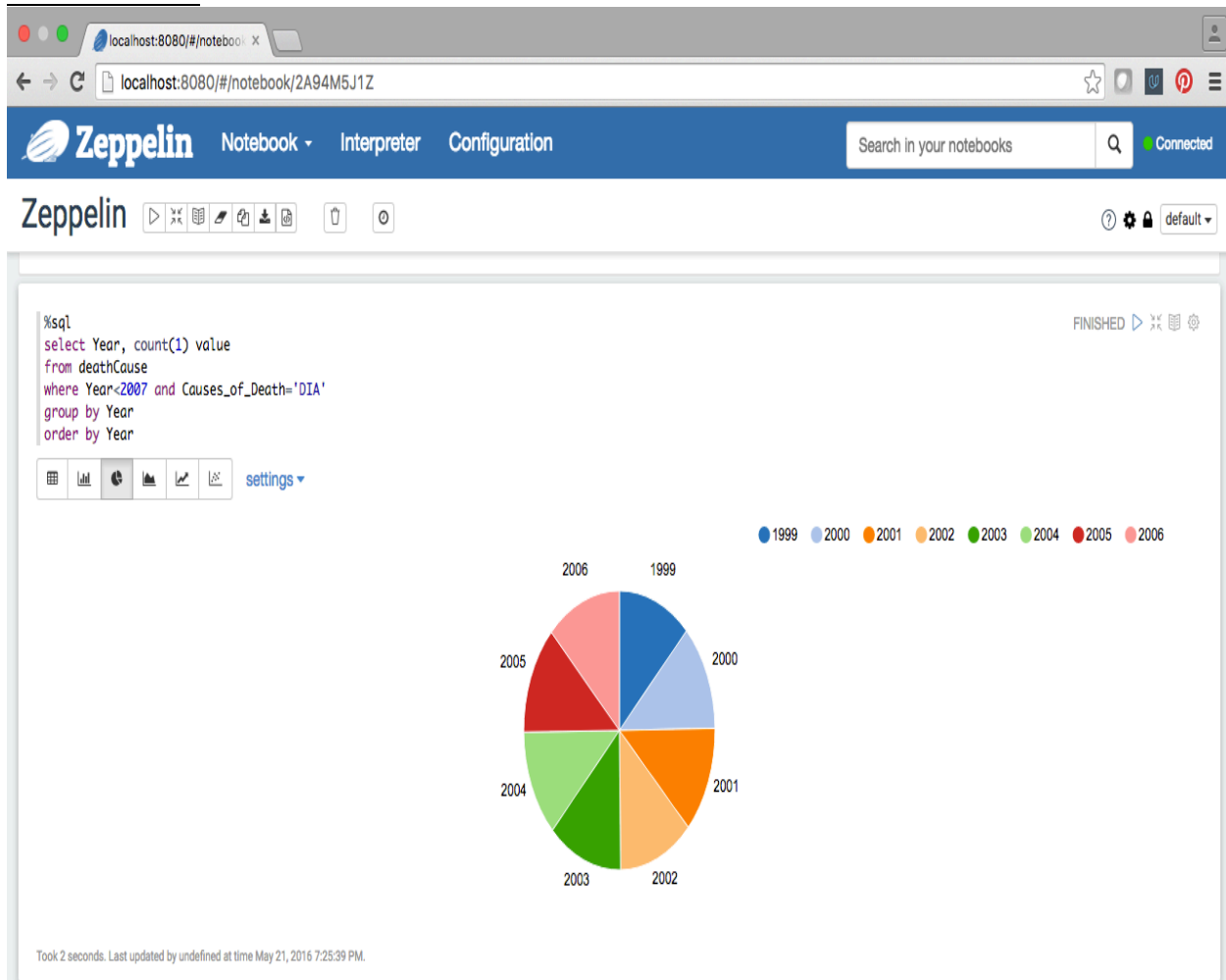
- 14) Find the locations with deaths less than 500
- Below is the representation of data in a graph.



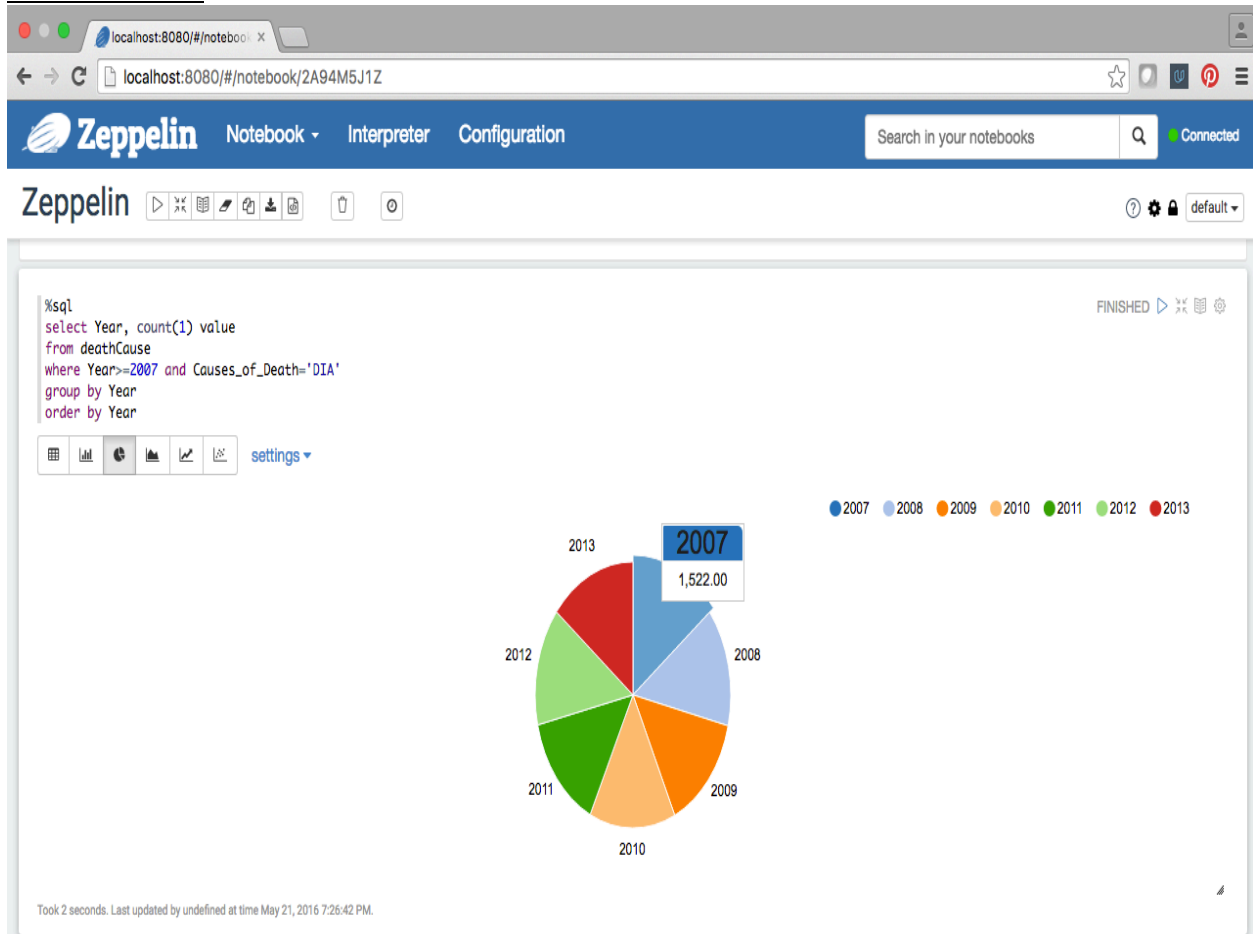
15) When has the number of Deaths due to Diabetes been more? From 1999-2006 or 2007-2013?

- The average number deaths due to Diabetes has been more every year for the range 2007-2013

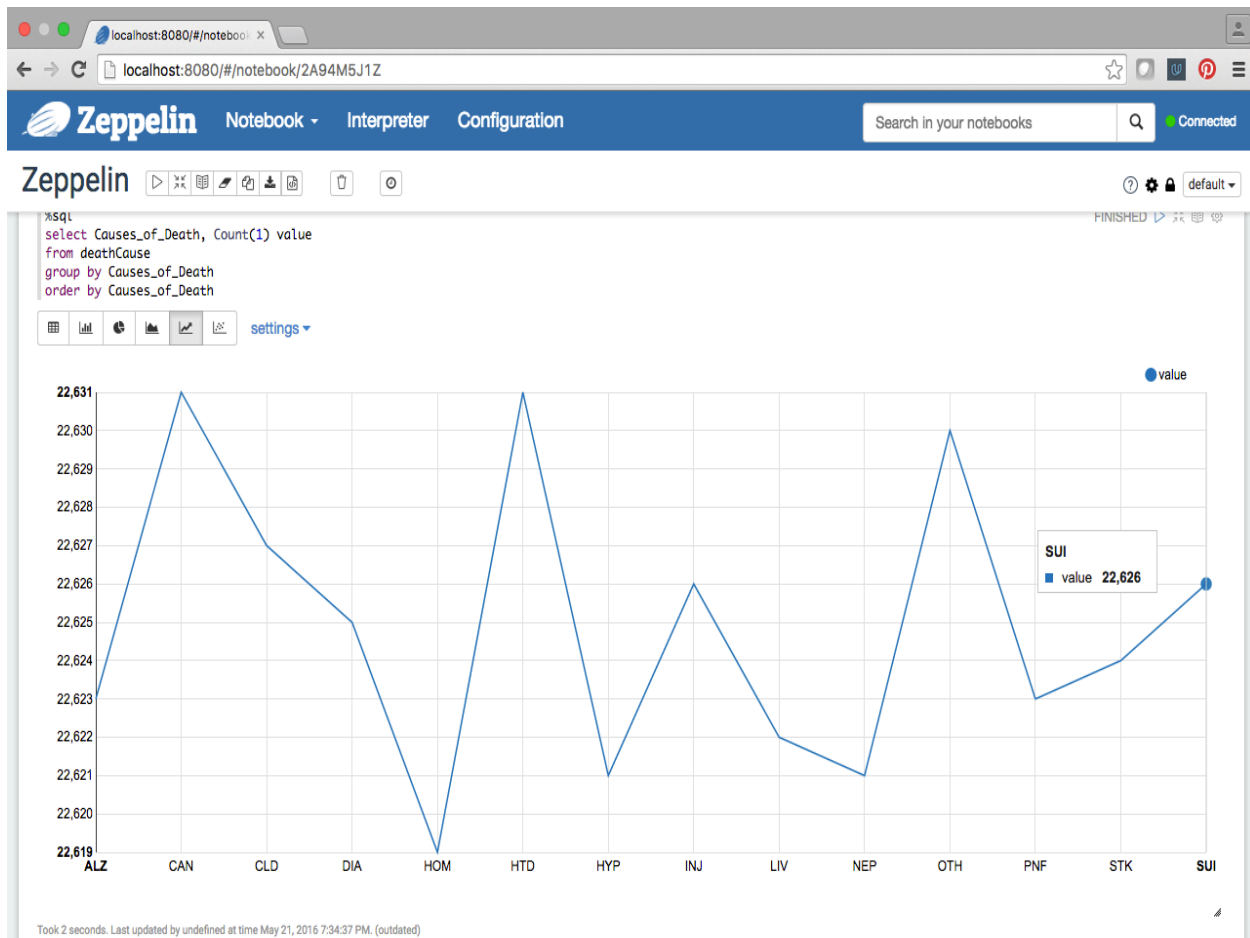
1999-2006:



2007-2013:



16) What is the no. of deaths by Suicide from 1999-2013?  
- No. of deaths by Suicides 22,626 from 1999-2013





## Stopped Zeppelin from terminal

```
1. zsh
653 s
654 cd\\n
656 pwd
657 kill -9 661
658 top -u
659 kill -9 358
660 exit
661 cd desktop
663 cd cmpe_273
665 cd zeppelin
667 cd in*
669 cd bin
670 ls
bin master % ./bin zeppelin-daemon.sh start
zsh: no such file or directory: ./bin
bin master % cd ..
incubator-zeppelin master % ./bin zeppelin-daemon.sh start
zsh: permission denied: ./bin
incubator-zeppelin master % zeppelin-daemon.sh start
zsh: command not found: zeppelin-daemon.sh
incubator-zeppelin master % bin/zeppelin-daemon.sh start
Zeppelin start [ OK ]
incubator-zeppelin master % bin/zeppelin-daemon.sh stop
Zeppelin stop [ OK ]
incubator-zeppelin master %
```