Hard Drive Survival

An Indepth & Thorough Analysis

Tyler Beverley - Brian Cao

Why Is Hard Drive Data Important?

- By 2020, it's estimated that 1.7MB of data will be created every second per person.
- That’s 146 GB/Day or 51976 GB/Year per person.
- All that data has to go somewhere.
- Figuring out what Drives have the best resilience can save a lot of $$.

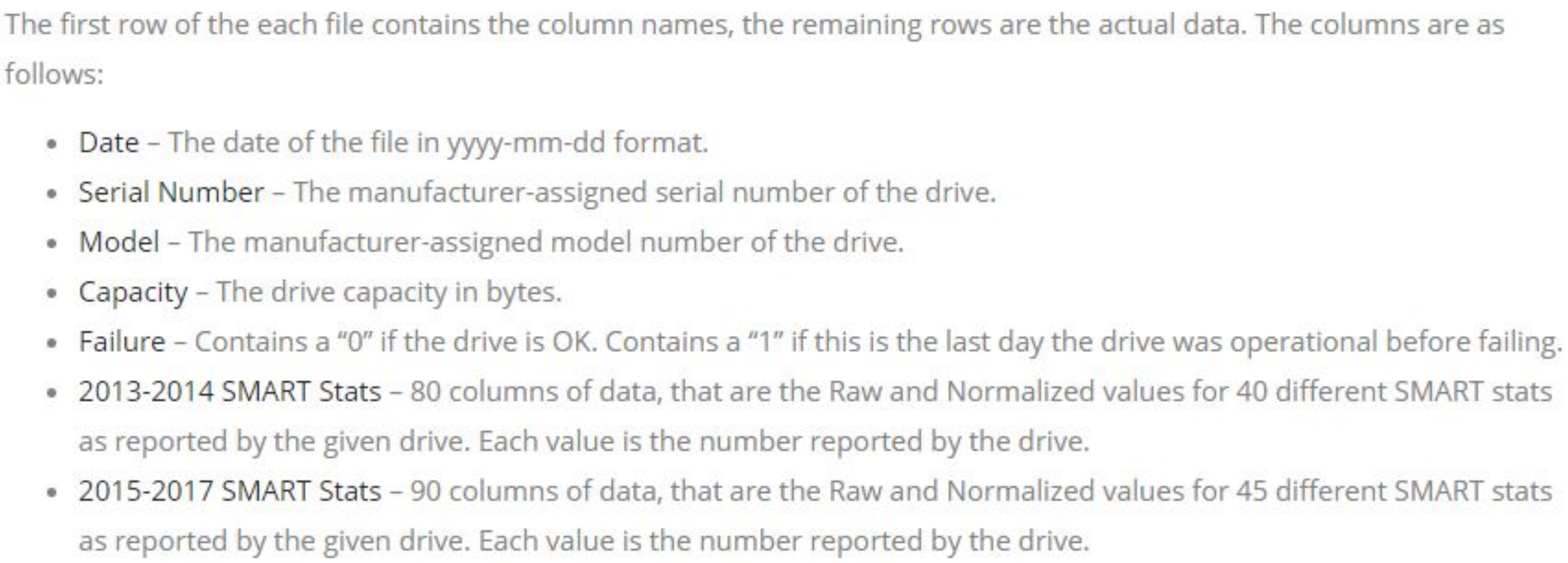
Goals

- Identify which brands/models outperformed other hard drives in terms of Average Survival Tenure.
- Determine which SMART Stats have high correlation with drive failure for the years 2015-2019.
- Create a regression to predict how many days a hard drive might survive.

Overview

Since 2013, Backblaze has published statistics and insights on the hard drives in their data center.

Each day in the Backblaze data center, a snapshot is taken of each operational hard drive. These snapshots include basic drive information and S.M.A.R.T statistics. The daily snapshots of each drive is recorded as one row of data. All of the data, with each row being data for an active hard drive, is put into a csv file.

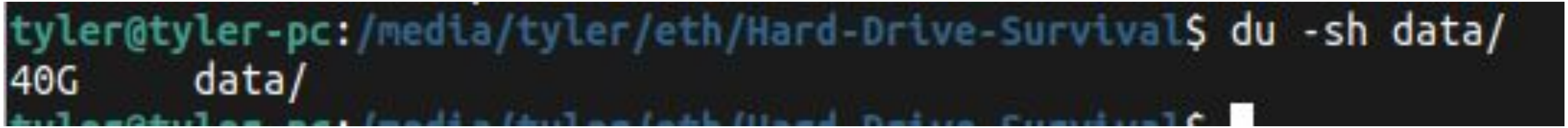


The first row of the each file contains the column names, the remaining rows are the actual data. The columns are as follows:

- Date – The date of the file in yyyy-mm-dd format.
- Serial Number – The manufacturer-assigned serial number of the drive.
- Model – The manufacturer-assigned model number of the drive.
- Capacity – The drive capacity in bytes.
- Failure – Contains a "0" if the drive is OK. Contains a "1" if this is the last day the drive was operational before failing.
- 2013-2014 SMART Stats – 80 columns of data, that are the Raw and Normalized values for 40 different SMART stats as reported by the given drive. Each value is the number reported by the drive.
- 2015-2017 SMART Stats – 90 columns of data, that are the Raw and Normalized values for 45 different SMART stats as reported by the given drive. Each value is the number reported by the drive.

I Have A Problem.

I am a Data Hoarder.



```
tyler@tyler-pc:/media/tyler/eth/Hard-Drive-Survival$ du -sh data/
40G     data/
tyler@tyler-pc:/media/tyler/eth/Hard-Drive-Survival$
```

40GB of text (CSV) data.

Windows OS size

Try opening *that* in minitab.

126,980,000

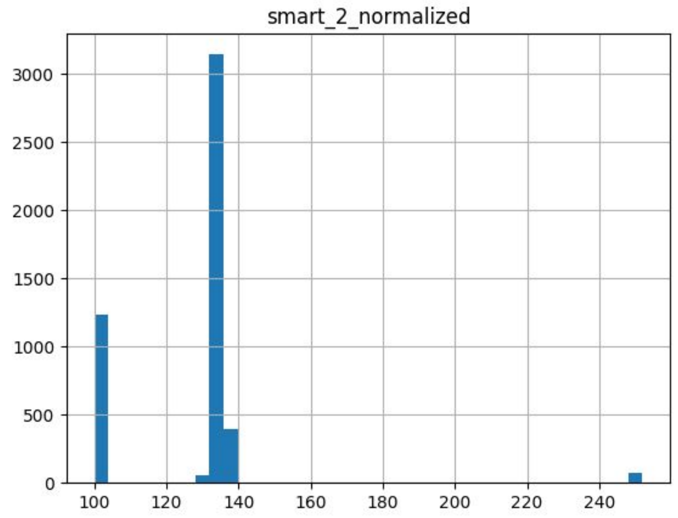
Rows of Data

Data this large required some special tooling.

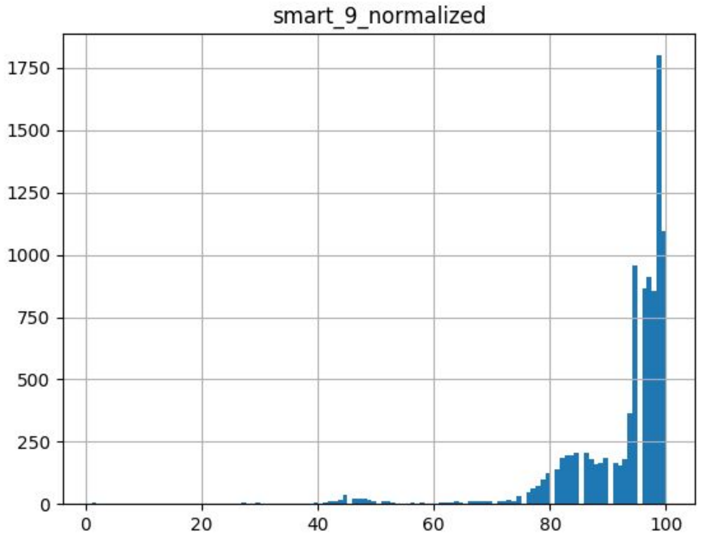
Enter Python.

Descriptive Statistics

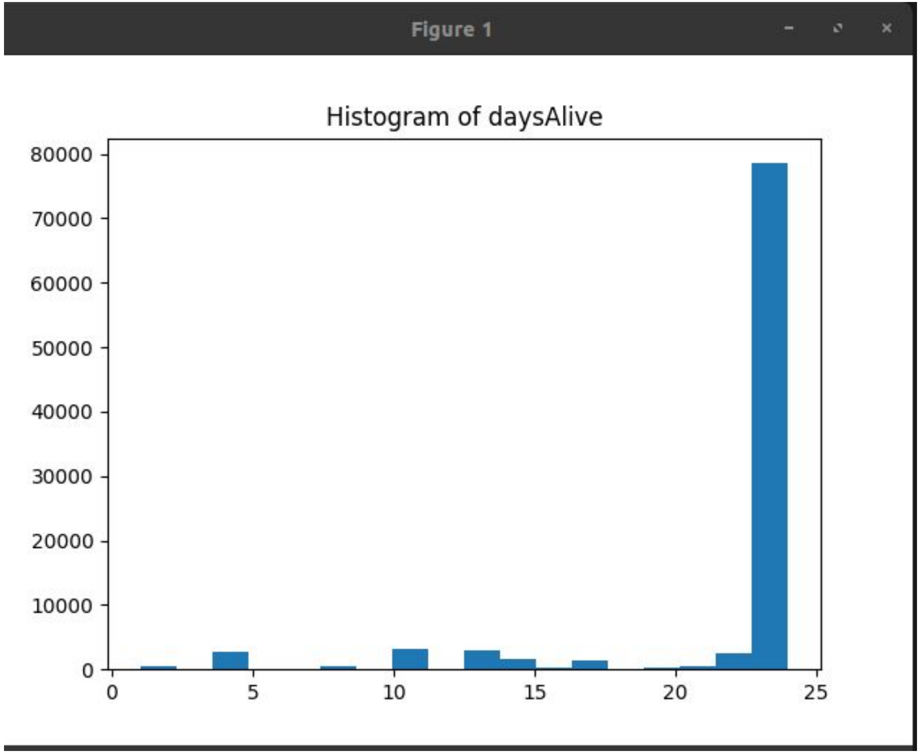
Throughput performance | -65% Correlated

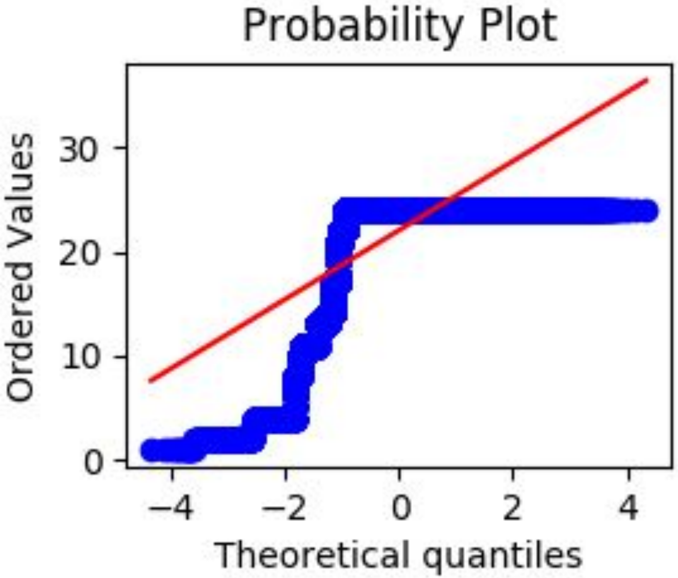


Power On Hours | -64% Correlated

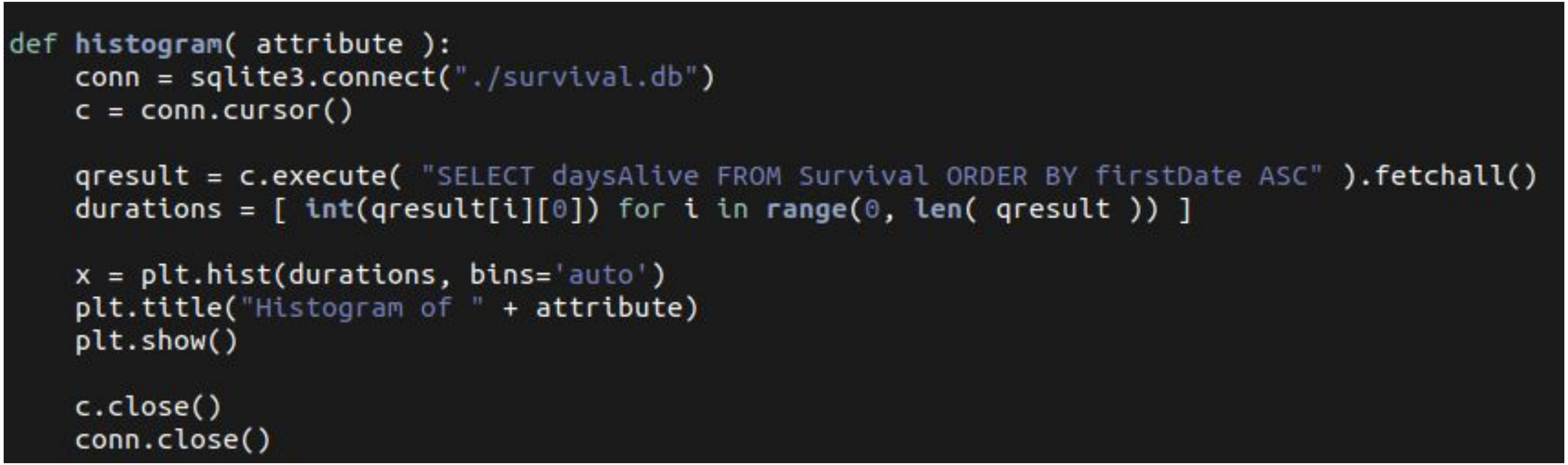


Descriptive Statistics





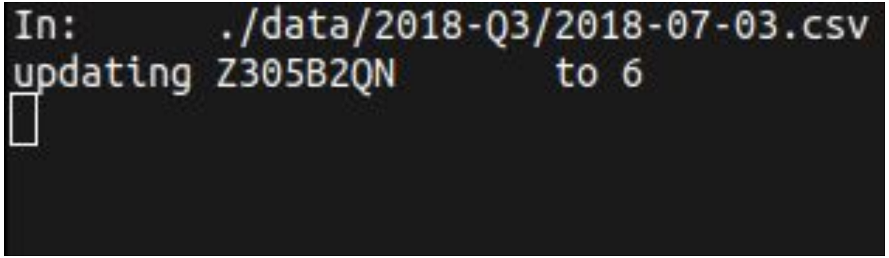
Histograms -- In Python!



```python
def histogram( attribute ):
    conn = sqlite3.connect("./survival.db")
    c = conn.cursor()

    qresult = c.execute( "SELECT daysAlive FROM Survival ORDER BY firstDate ASC" ).fetchall()
    durations = [ int(qresult[i][0]) for i in range(0, len( qresult )) ]

    x = plt.hist(durations, bins='auto')
    plt.title("Histogram of " + attribute)
    plt.show()

    c.close()
    conn.close()
```

Data Aggregation And Cleaning

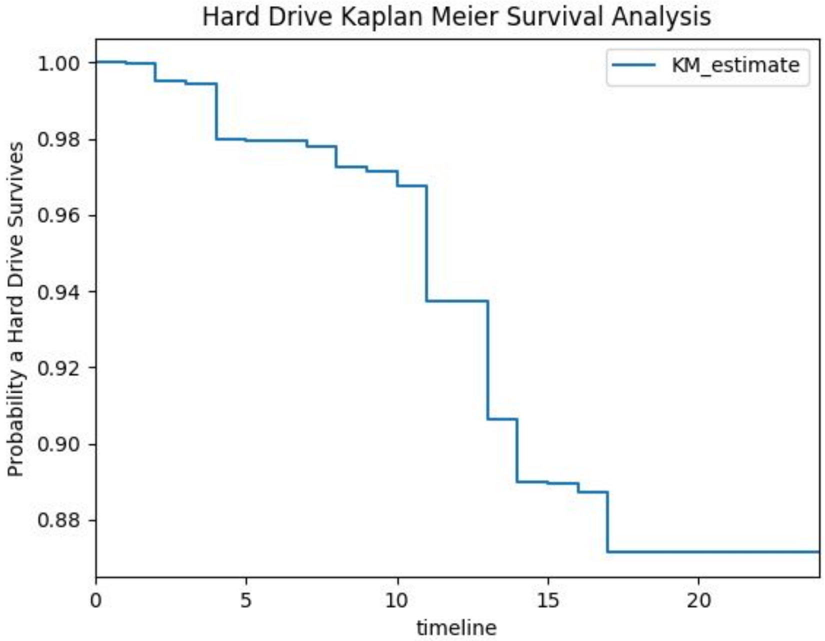
- Problem 1 - Hard Drives are inserted as time goes on.
- Solution - Aggregate the number of days the hard drive was alive.

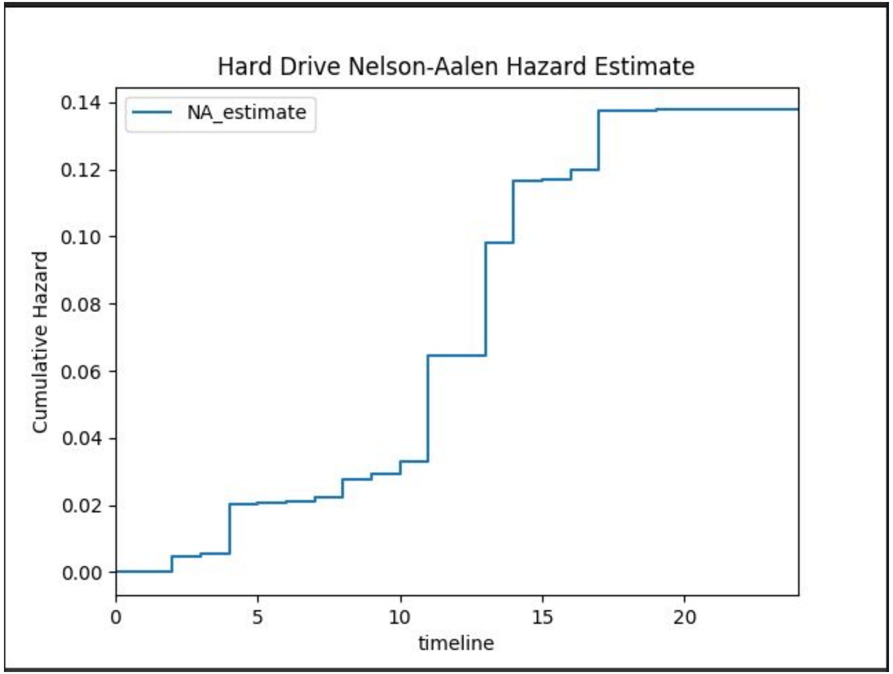


```
In:      ./data/2018-Q3/2018-07-03.csv
updating Z305B2QN       to 6
```

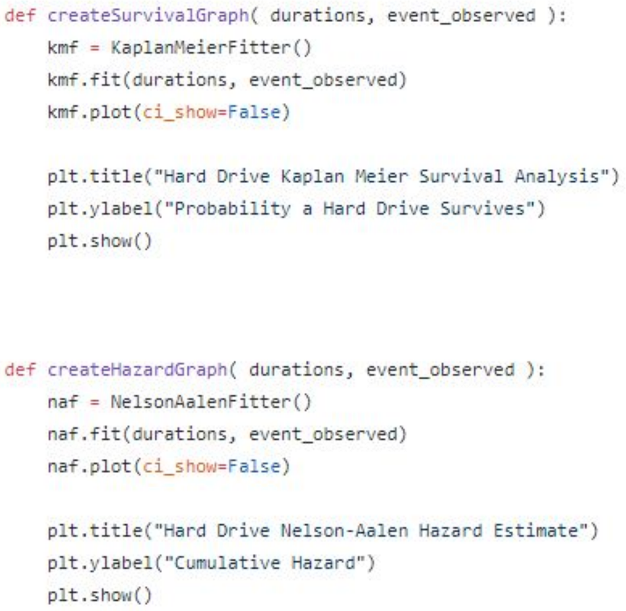
- Problem 2 - Failed dates were not a part of the data set.
- Solution - Calculate
  - First Date + Days Alive == Last Date?
  - Assuming that drives are only pulled once they die

Survival Curves -- All



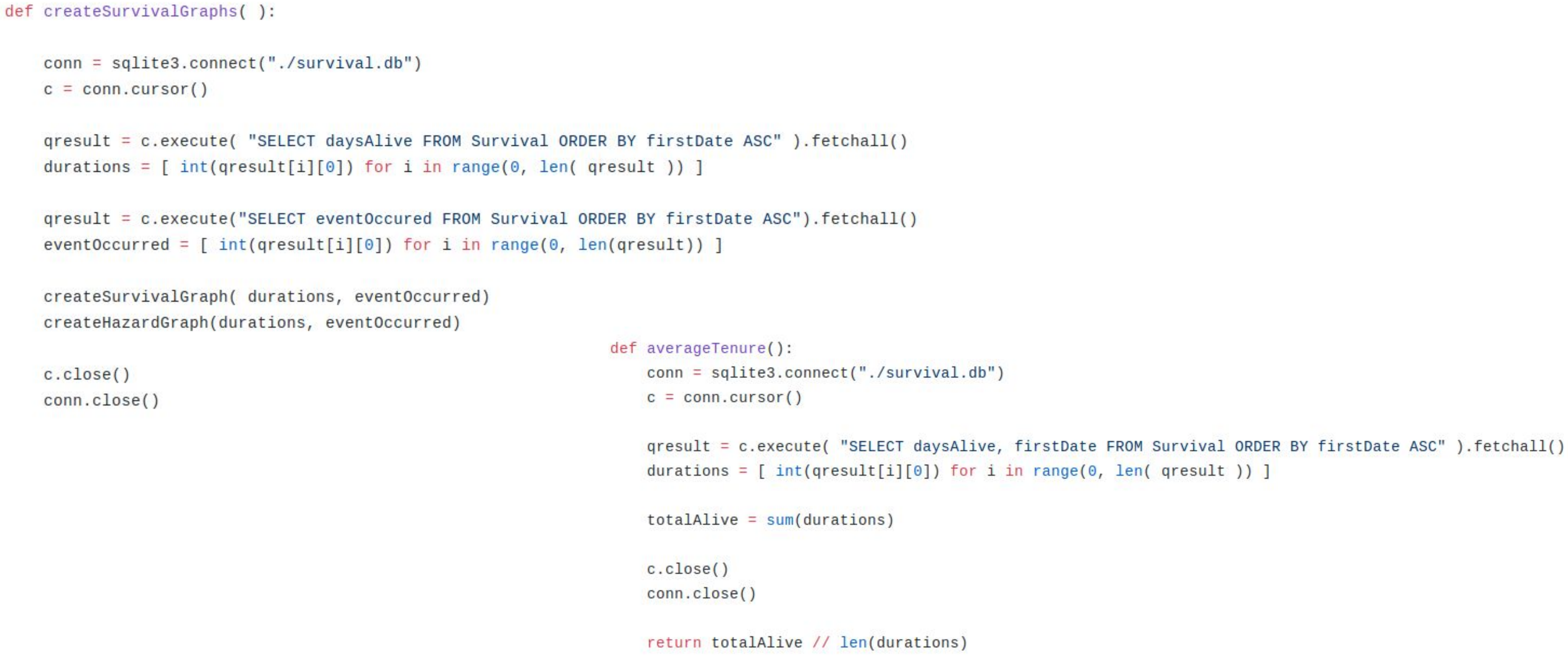


Survival Curves in Python



```python
def createSurvivalGraph( durations, event_observed ):
    kmf = KaplanMeierFitter()
    kmf.fit(durations, event_observed)
    kmf.plot(ci_show=False)

    plt.title("Hard Drive Kaplan Meier Survival Analysis")
    plt.ylabel("Probability a Hard Drive Survives")
    plt.show()



def createHazardGraph( durations, event_observed ):
    naf = NelsonAalenFitter()
    naf.fit(durations, event_observed)
    naf.plot(ci_show=False)

    plt.title("Hard Drive Nelson-Aalen Hazard Estimate")
    plt.ylabel("Cumulative Hazard")
    plt.show()
```

Survival Curves -- In Python!



```
def createSurvivalGraphs( ):

    conn = sqlite3.connect("./survival.db")
    c = conn.cursor()

    qresult = c.execute( "SELECT daysAlive FROM Survival ORDER BY firstDate ASC" ).fetchall()
    durations = [ int(qresult[i][0]) for i in range(0, len( qresult )) ]

    qresult = c.execute("SELECT eventOccured FROM Survival ORDER BY firstDate ASC").fetchall()
    eventOccurred = [ int(qresult[i][0]) for i in range(0, len(qresult)) ]

    createSurvivalGraph( durations, eventOccurred)
    createHazardGraph(durations, eventOccurred)

    c.close()
    conn.close()
```

```
def averageTenure():
    conn = sqlite3.connect("./survival.db")
    c = conn.cursor()

    qresult = c.execute( "SELECT daysAlive, firstDate FROM Survival ORDER BY firstDate ASC" ).fetchall()
    durations = [ int(qresult[i][0]) for i in range(0, len( qresult )) ]

    totalAlive = sum(durations)

    c.close()
    conn.close()

    return totalAlive // len(durations)
```

Where is the rest?

- We are currently running the survival analysis, for populations based on Capacity, and Brand.
- Ran out of time, but will be in final paper.

Questions?