Springboard Foundation of Data Science Capstone Project - Hard Drive Reliability - Survival Analysis

Chinpei Tang

November 11, 2016

Introduction

Hard drive reliability details are important not only for data centers to provide high uptime service, but also to hard drive manufacturer to ensure the production of high quality hard drives to support the competitive and growing business. The project aims at analyzing the trends of the survival period of hard drives of variety of hard drives available in the market. These trends along with some physical data, its major causes that results in hard drive failures can possibly be deduced.

Value

The project will explore and analyze the trends in hard drive failures. Three major areas that can benefit these analysis are:

1. The hard drive manufacturers can use such information to focus on improving the key weaknesses.
2. The data centers can also use the information to predict the risk and decide which types of hard drives they should choose to run data centers to ensure high uptime and low customer dissatisfaction.
3. The consumers can also use such data to make decision on which types of hard drives they should be purchasing to backup their valuable files and reducing the risk of losing them.

Hard Drive Failures

There are two types of hard drive failures:

- Predictable failures: resulting from slow processes such as mechanical wear and gradual degradation of storage surfaces. Monitoring can determine when such failures are becoming more likely.
- Unpredictable failures: happening without warning and ranging from electronic components becoming defective to a sudden mechanical failure (which may be related to improper handling).

Goals

The main goals of the projects are:

- Understanding the failure rates of the hard drives based on its different manufacturer, models, and sizes. These may lead to the list of "high risk" hard drives that may be avoided for critical operations.
- Estimate and predict the survival rates of the hard drives based on its different manufacturer and sizes. The data center can then proactively plan for the hard drive procurement and replacement to minimize inventory and maximize uptime.

Data

Backblaze is an online personal/business backup and cloud storage service provider, which consumes about 1000 hard drive per month. The company wrote scripts to track the hard drive health information since 2013. See more information here. The dataset is made open-source here.

The data contains key properties of the hard drives (serial number, manufacturer/model and capacity), whether or not it failed, and 80 to 90 SMART (Self-Monitoring, Analysis and Reporting Technology) parameters (or 40 to 45 normalized values). This can leads to failure mode identification.

More specifically:

- Date - The date of the file in yyyy-mm-dd.csv 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.
- 2015-2016 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.

Data Organization

The data is organized into directories of different time period. For instance, the data available in 2013 is named data_2013. Each folder contains csv file of the hard drive details on the shelf on each day. The file names are organized in the format of yyyy-mm-dd.csv.

In each of the csv file, each row of data is the details of each hard drive that is operational on the shelf each day. It is identified by the serial numbers of the individual hard drives. The model and capacity of the hard drive is reported. If the hard drive is failed on the particular day, it is labeled 1 in the failure column; otherwise it is labeled 0. The row with the failed hard drive will be removed in the csv file of the next day, and, potentially a new hard drive will be added. The SMART data of each of the hard drive is also reported in columns.

Data Import

Due to large amount of data logged in the dataset, we can only import certain key columns in the data to perform various functions. We found that RStudio was unable to import and create a data frame of larger than 90 files, since it will be larger than the memory that can hold the data (about 8 GB RAM). Hence, as the first step, we will look into only the date, serial number, model number, capacity and if the harddrive failed, and ignore the SMART data. We believe that by looking at the above data, we can deduce some failure conclusion.

It is impossible to import all the available data into a huge data frame to perform analysis. This is even impossible for just a year of data (the author spent quite some times to make the attempts but failed due to the limited available memory). Also, importing the data also takes significant amount of time. Hence, the approach adopted here is to write a script to explicitly import the data into RStudio, then save them to the RData file. Then the RData files will be loaded to perform the corresponding analysis.

The "dplyr" package is used extensively to write the scripts such that it extracts only the required columns of data and/or calculating the statistics required to perform the corresponding plotting and/or analysis. They are well documented in the scripts. Also, the scripts are first tested on a small manageable set of csv files, i.e. only 15 days of csv data in a folder, then extended to the full "production" to take the full year of data.

While this project is not utilizing the SMART data, the author suggests in the later stage if SMART data study is involved, good strategy during the data import stage will be required. It will need to import only relevant information (or just enough details) for study as the data size can quickly grows in size. Some of the information may need to be processed as soon as a csv file is imported to build the data frame. For instance, if one of the SMART data is required to study with respect to brand, the serial number may not need to be imported, and the model number may need to be converted immediately to brand, and the model number is then disgarded since the model numbers are strings that can take up much of the data size. Then the failure data and the corresponding SMART data of interest will be imported.

One of the validations that we will need to do is to ensure that there is no duplicated serial numbers in each row of the observation. This simple check can be done by counting the distinct number of serial numbers of each csv, and see if it is equal to the number of rows.

Data Analysis

As of the writing of this report, here are the details of all the data available:

Directory	Data Available	Total Data Size	Each File Size	Number of Drives on Shelf (from first day to last day in the directory)
data_2013	266 days (starting 2013-04-10)	738 MB	3 - 4 MB	21,195 - 27,223
data_2014	365 days	2.81 GB	7 - 10 MB	27,223 - 41,213
data_2015	365 days	4.19 GB	10 - 14 MB	41,213 - 57,544
data_Q1_2016	91 days	1.32 GB	14 - 15 MB	57545 - 62,952
data_Q2_2016	91 days	1.44 GB	15 - 18 MB	62,952 - 70,403

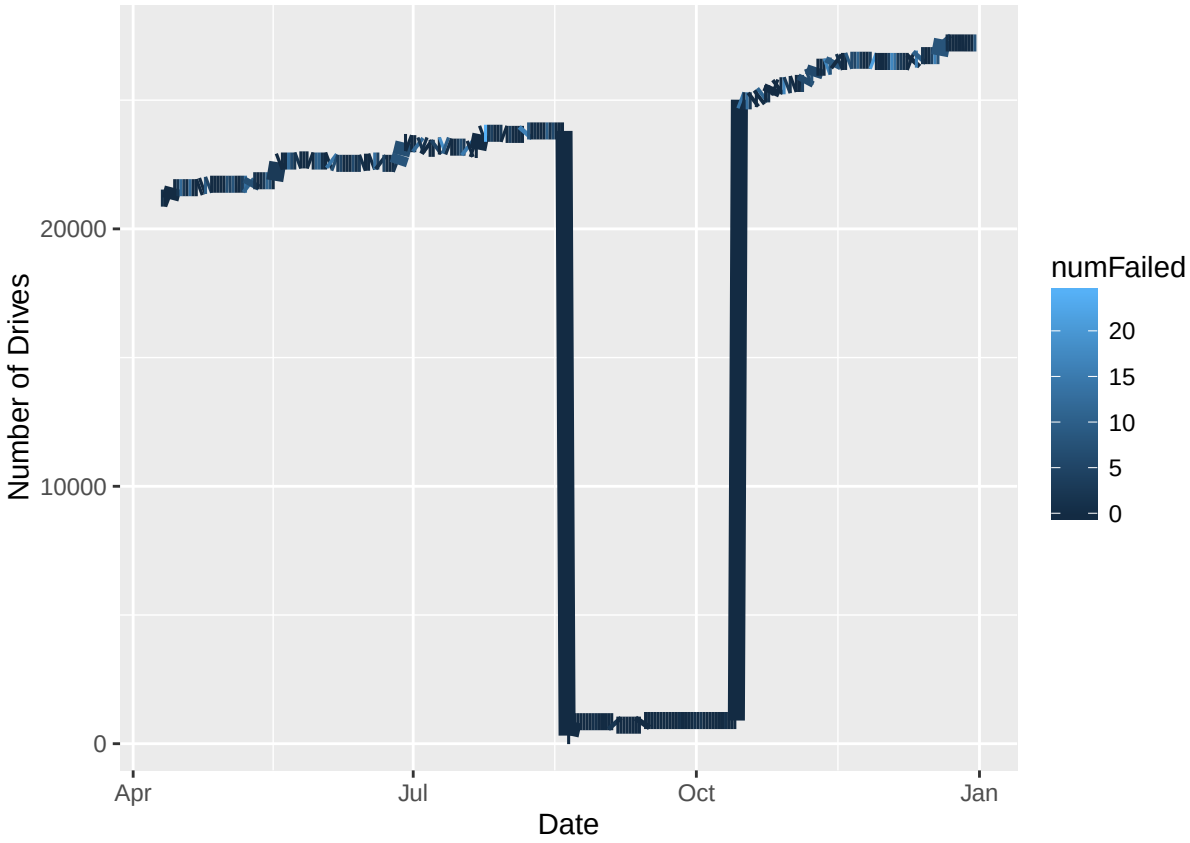
As can be seen, the data involved is huge. We first perform some data exploration by looking at the number of hard drives on the shelf over time. Subsequently, we also look at the number of hard drives of each brand and size on the shelf over time.

Number of Hard Drives Over Time

As mentioned previously, only the required data is calculated to create the RData files. In this case, the number of drives of each day and the number of failures are extracted and created a new data frame of only date, number of drives (number of rows) and number of failures (sum of number of 1's in failure columns). Here are the results:

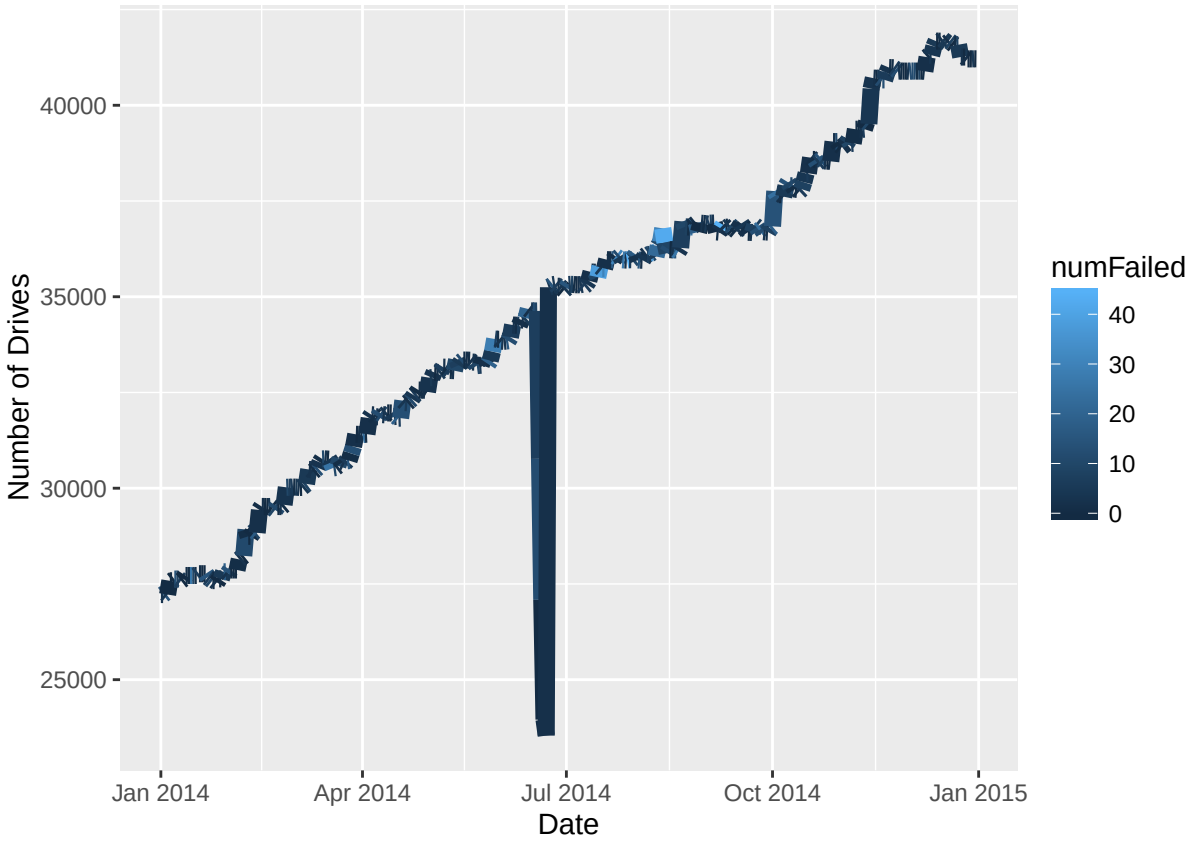
Year 2013 (from 2013-04-10 to 2013-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2013/2013/hd2013_totalDriveCnt_Failu
ggplot(stat_totalDriveCnt_FailureCnt_df, aes(x = date_col, y = numDrive, colour = numFailed)) +
  geom_line(size = 3) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



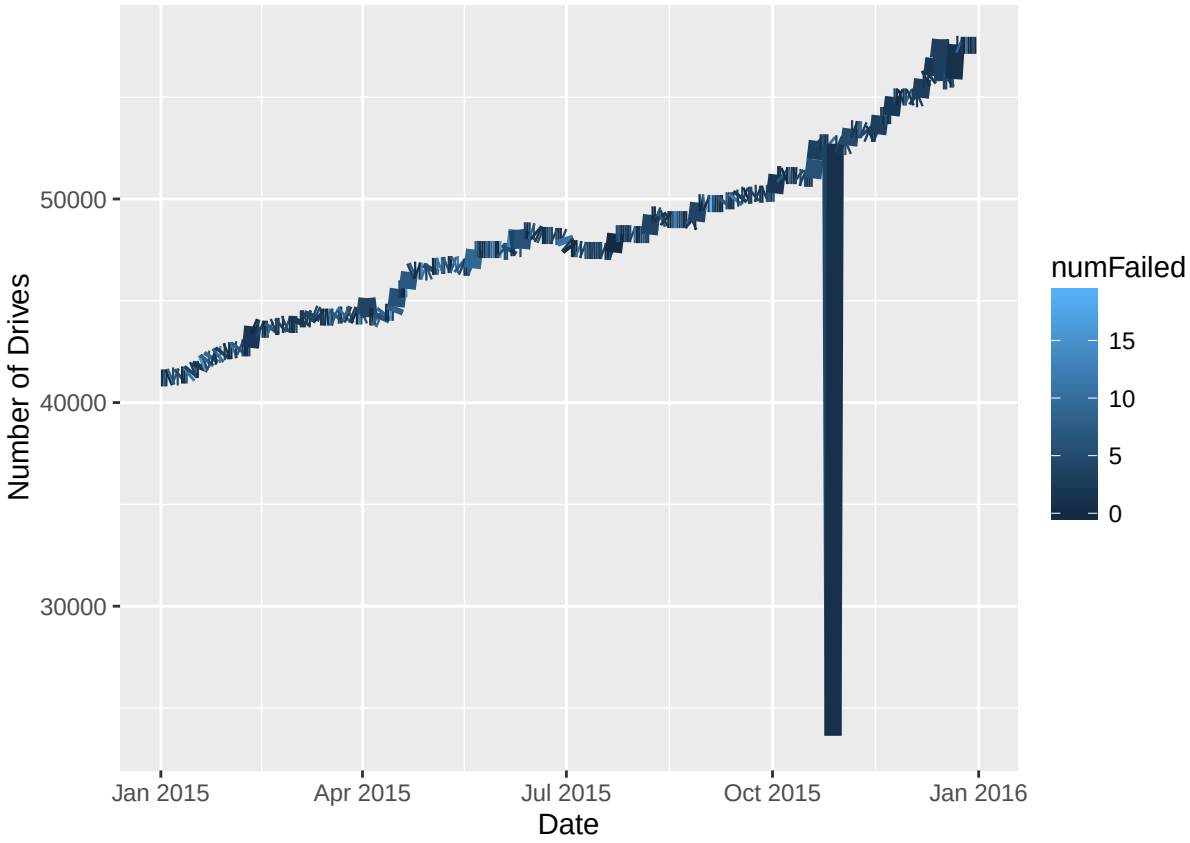
Year 2014 (from 2014-01-01 to 2014-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2014/2014/hd2014_totalDriveCnt_Failu
ggplot(stat_totalDriveCnt_FailureCnt_df, aes(x = date_col, y = numDrive, colour = numFailed)) +
  geom_line(size = 3) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



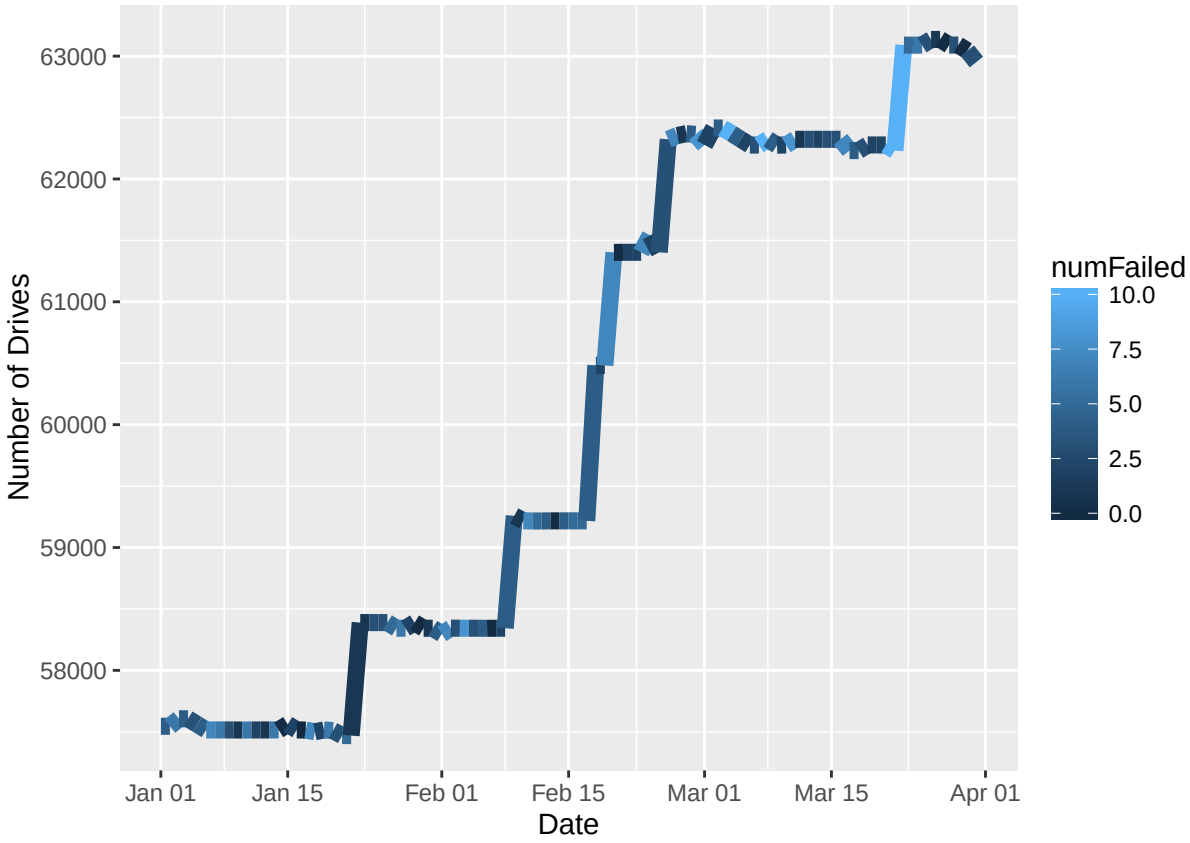
Year 2015 (from 2015-01-01 to 2015-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2015/2015/hd2015_totalDriveCnt_Failu
ggplot(stat_totalDriveCnt_FailureCnt_df, aes(x = date_col, y = numDrive, colour = numFailed)) +
  geom_line(size = 3) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



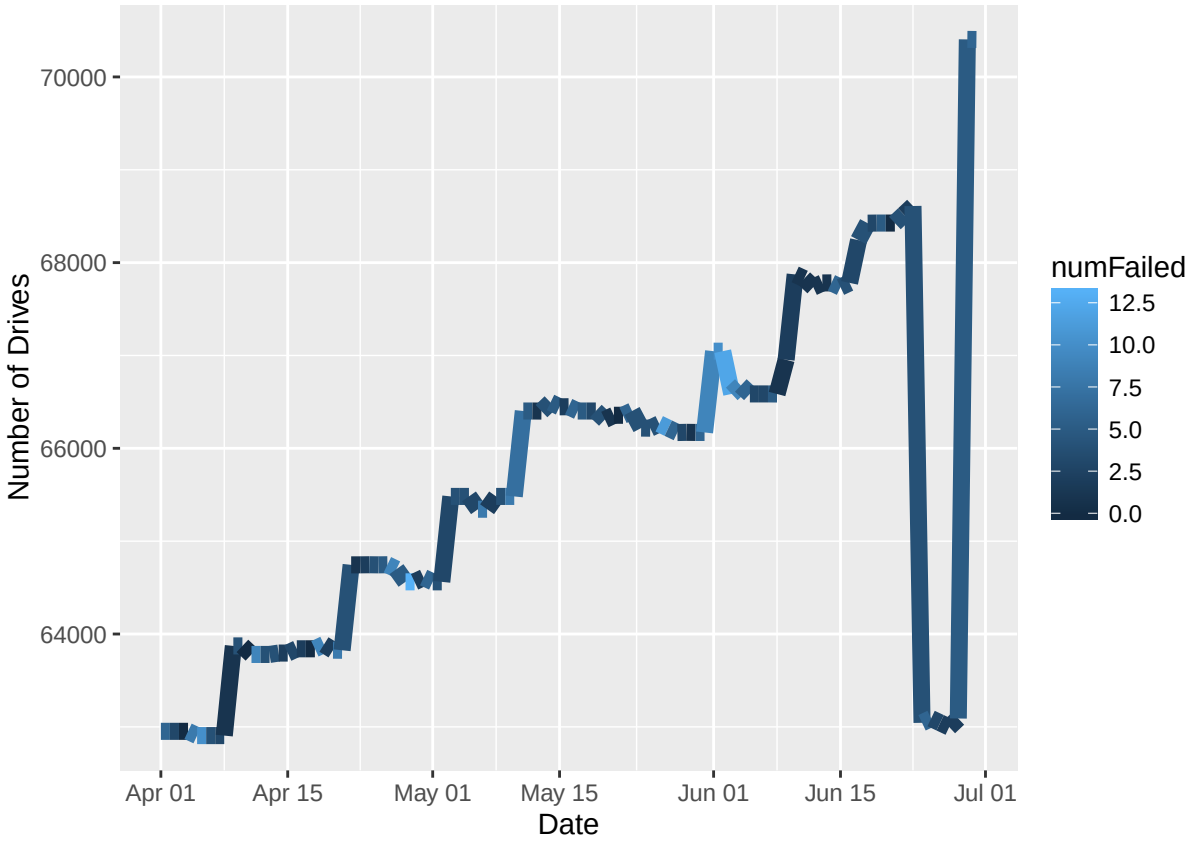
Q1 of Year 2016 (from 2016-01-01 to 2016-03-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_Q1_2016/data_Q1_2016/hdQ1_2016_tota
ggplot(stat_totalDriveCnt_FailureCnt_df, aes(x = date_col, y = numDrive, colour = numFailed)) +
  geom_line(size = 3) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



Q2 of Year 2016 (from 2016-04-01 to 2016-06-30):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_Q2_2016/data_Q2_2016/hdQ2_2016_tota
ggplot(stat_totalDriveCnt_FailureCnt_df, aes(x = date_col, y = numDrive, colour = numFailed)) +
  geom_line(size = 3) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



As can be noted from the above graphs, the data logged is not without issue. There are a few files logged 0 kb, which means that there was no data on those particular days. There are files also have sudden drop of the number of hard drives (some hard drives details are not logged), which may be due to the logging issue on those days. Fortunately this is not imposing a huge issue since the sampling is assumed to be in a day, and we may just lose the data of some of the subjects (data is "censored" from the survival analysis point of view).

Number of Hard Drives of Different Brand/Manufacturer Over Time

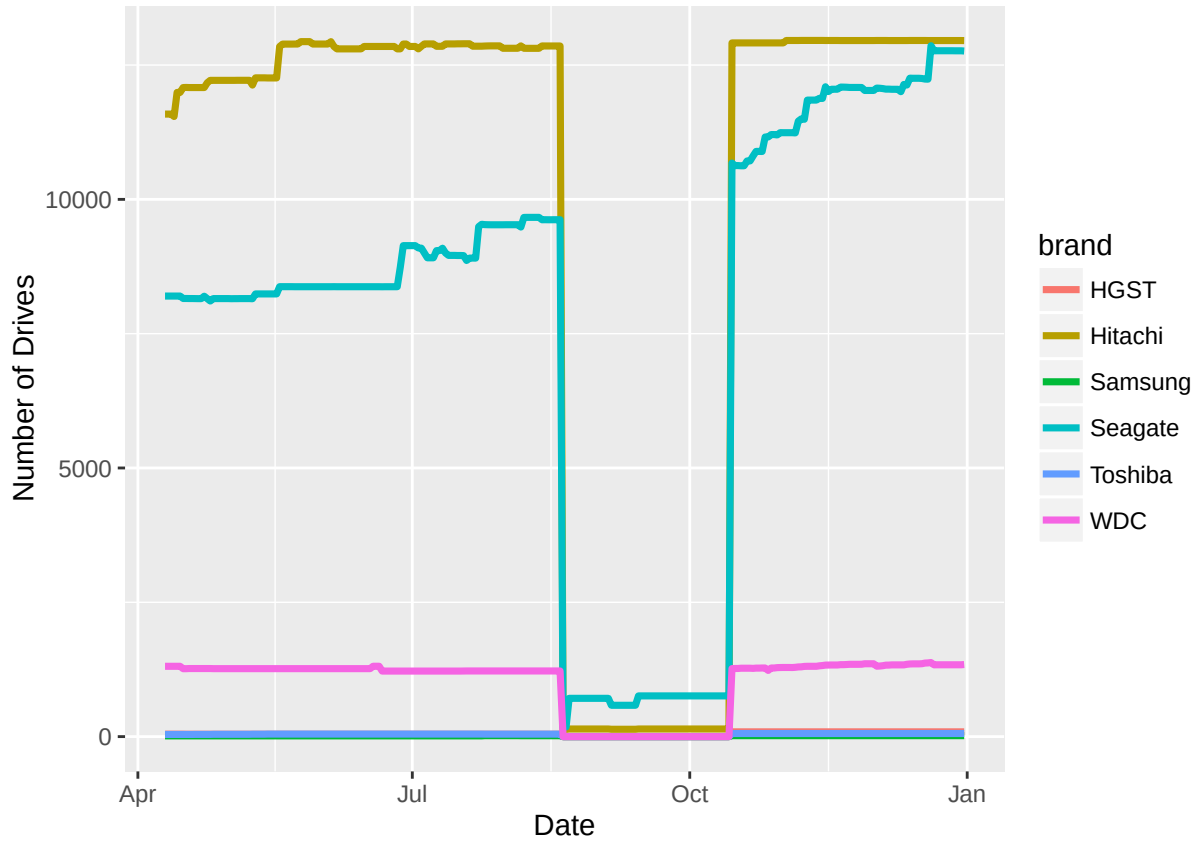
It is also interesting to see the number of different brand/manufacturer of the drives being consumed. We use grep function to identify the following expressions in model to assign to a new column brand:

- "HGST"" -> "HGST"
- "Hitachi" -> "Hitachi"
- "SAMSUNG" -> "Samsung"
- "TOSHIBA" -> "Toshiba"
- "WDC" -> "Western Digital"
- "^ST" -> "Seagate"

The only tricky part is all Seagate hard drive has prefix "ST", while "ST" is part of "HGST" that requires special distinction. Fortunately in "HGST" hard drives, "ST" won't be prefixes.

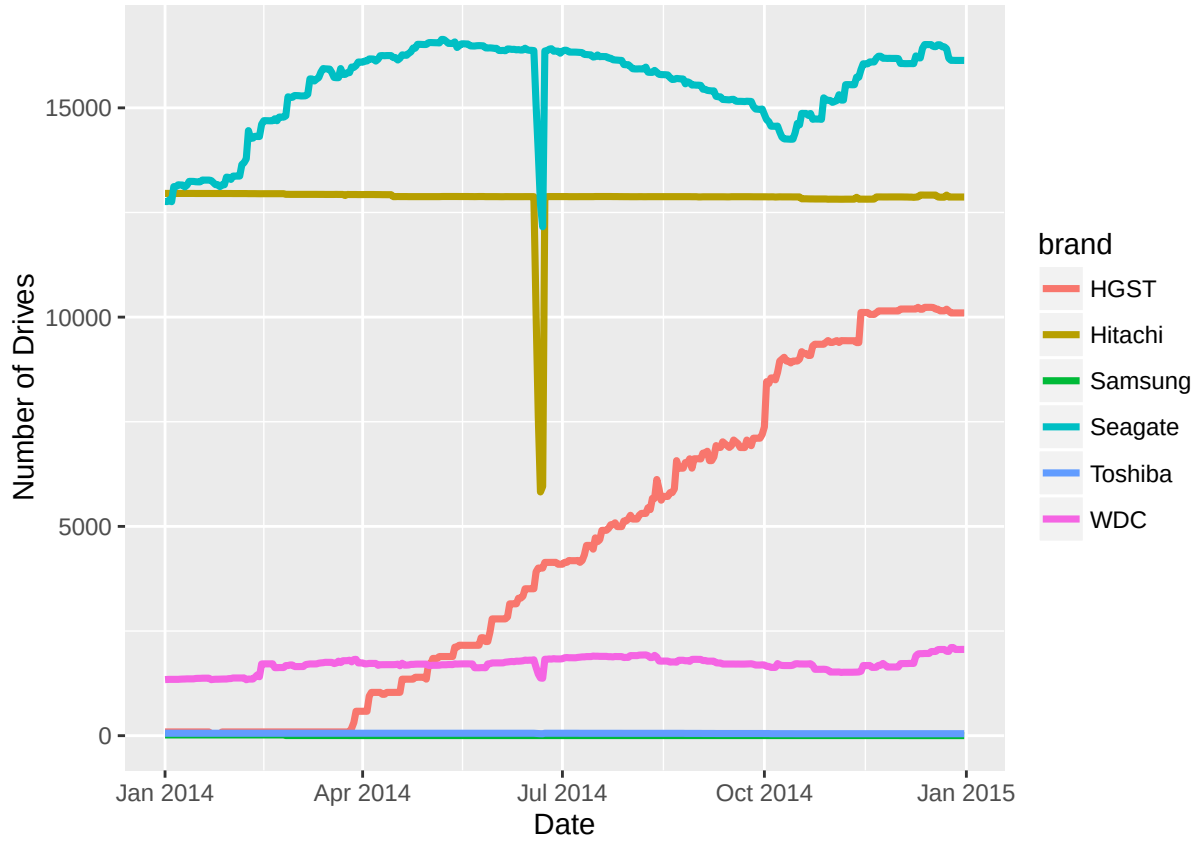
Year 2013 (from 2013-04-10 to 2013-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2013/2013/hd2013_brandDriveCnt_Failu
ggplot(stat_brandDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = brand), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



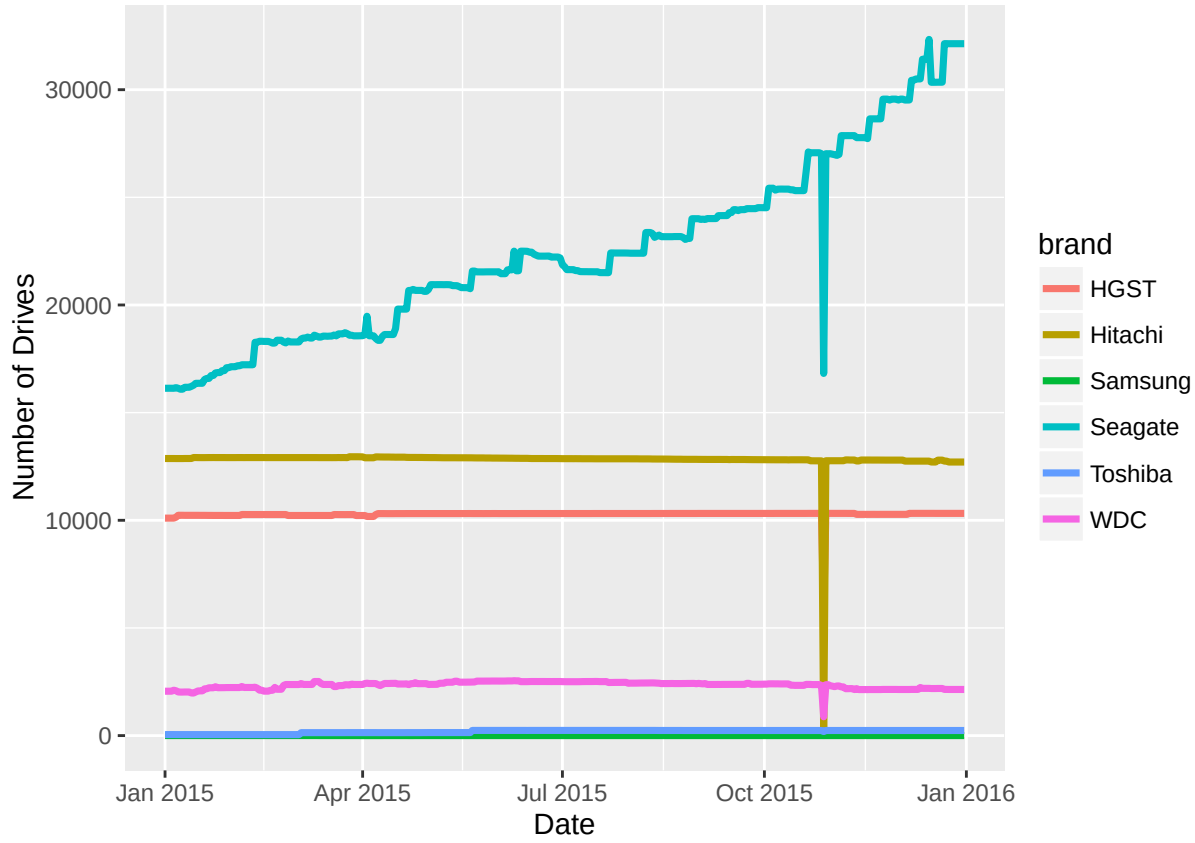
Year 2014 (from 2014-01-01 to 2014-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2014/2014/hd2014_brandDriveCnt_Failu
ggplot(stat_brandDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = brand), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



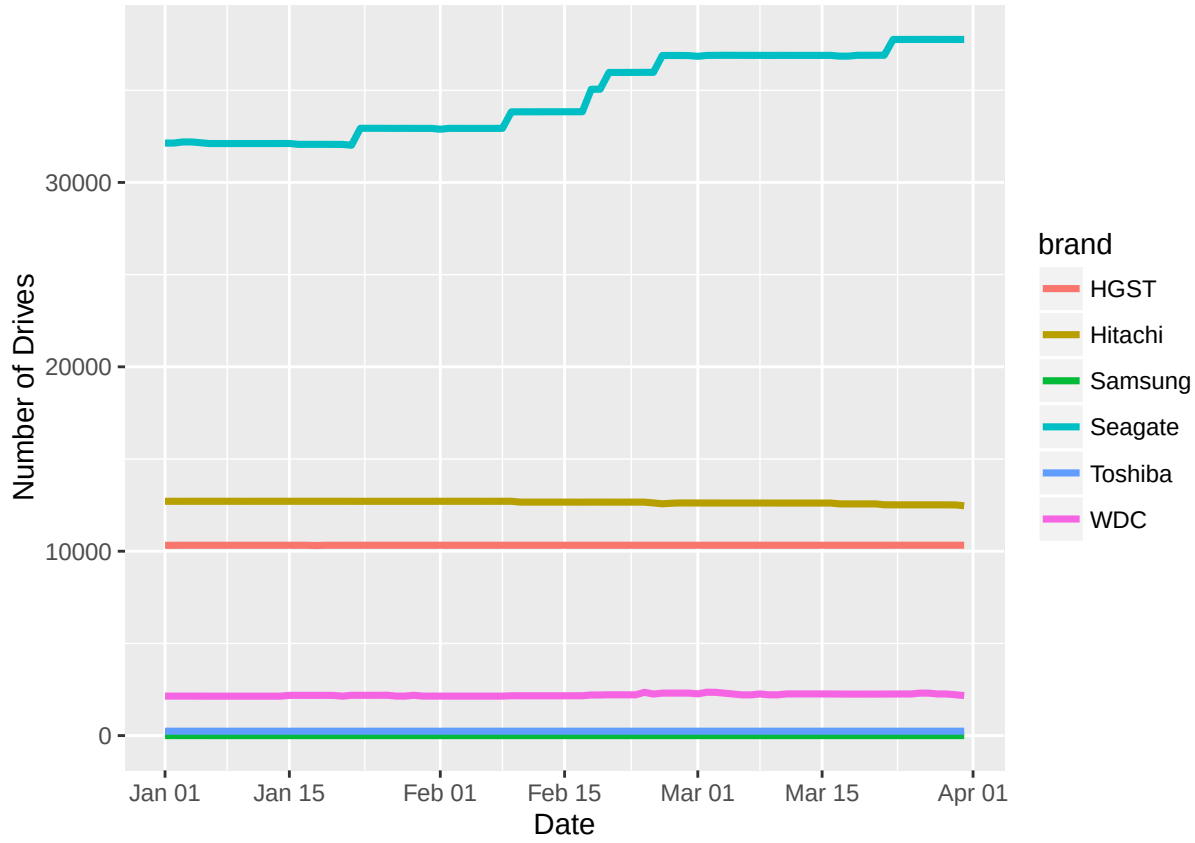
Year 2015 (from 2015-01-01 to 2015-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2015/2015/hd2015_brandDriveCnt_Failu
ggplot(stat_brandDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = brand), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



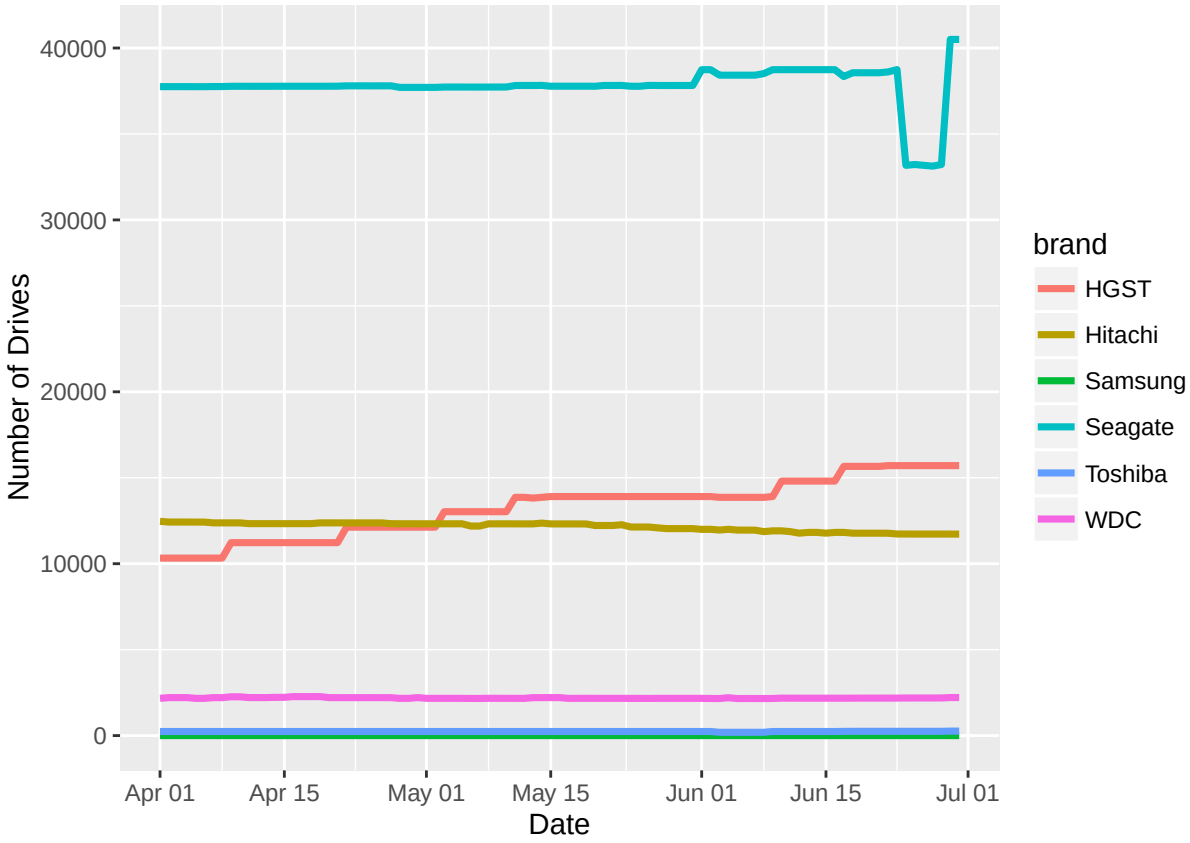
Q1 of Year 2016 (from 2016-01-01 to 2016-03-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_Q1_2016/data_Q1_2016/hdQ1_2016_brand
ggplot(stat_brandDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = brand), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



Q2 of Year 2016 (from 2016-04-01 to 2016-06-30):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_Q2_2016/data_Q2_2016/hdQ2_2016_brand
ggplot(stat_brandDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = brand), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```

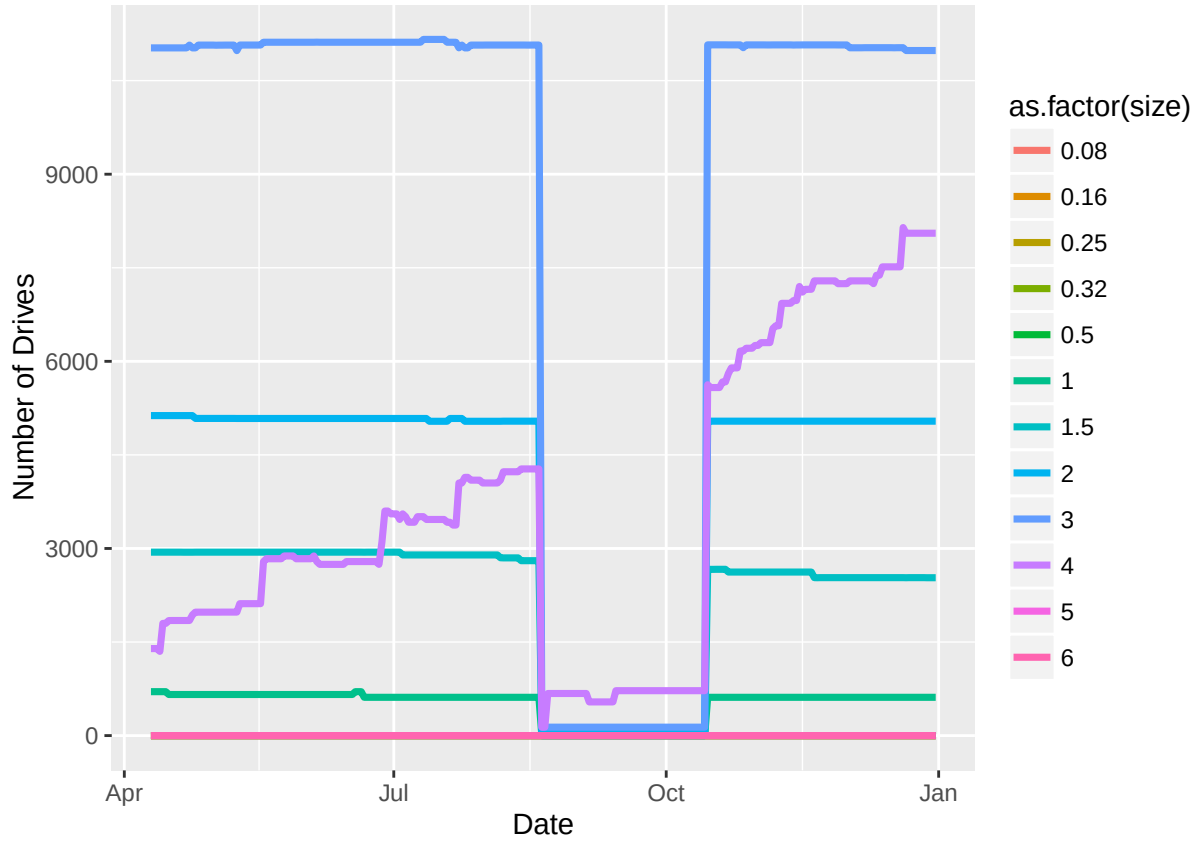


One can see that Backblaze does favor some brands. In 2013, Hitachi and Seagate were mainly used. Then we saw increased installation of Seagate over time, and became the majority of them on the shelf. We also saw some increased installation of HGST hard drives in 2014. The rest of the brands were not changed much.

Number of Hard Drives of Different Capacity Over Time

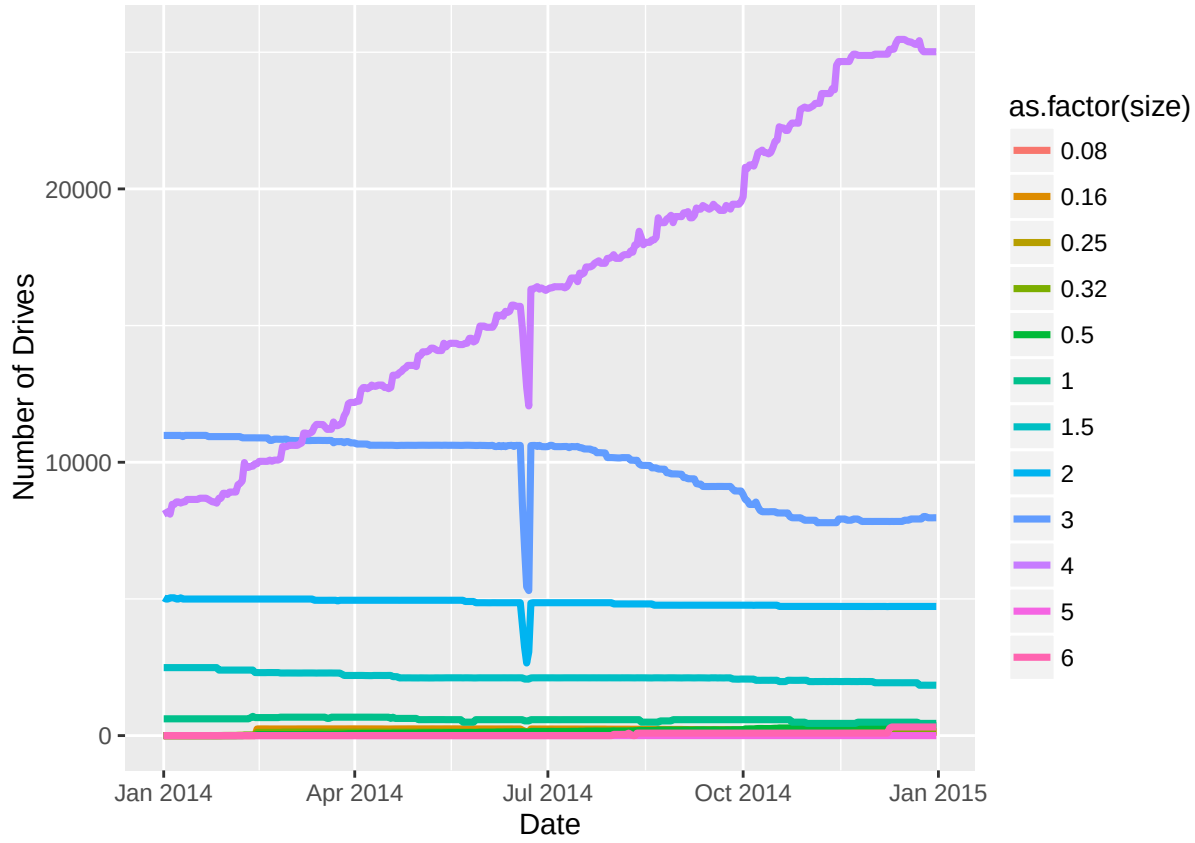
Year 2013 (from 2013-04-10 to 2013-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2013/2013/hd2013_sizeDriveCnt_Failu
ggplot(stat_sizeDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = as.factor(size)), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



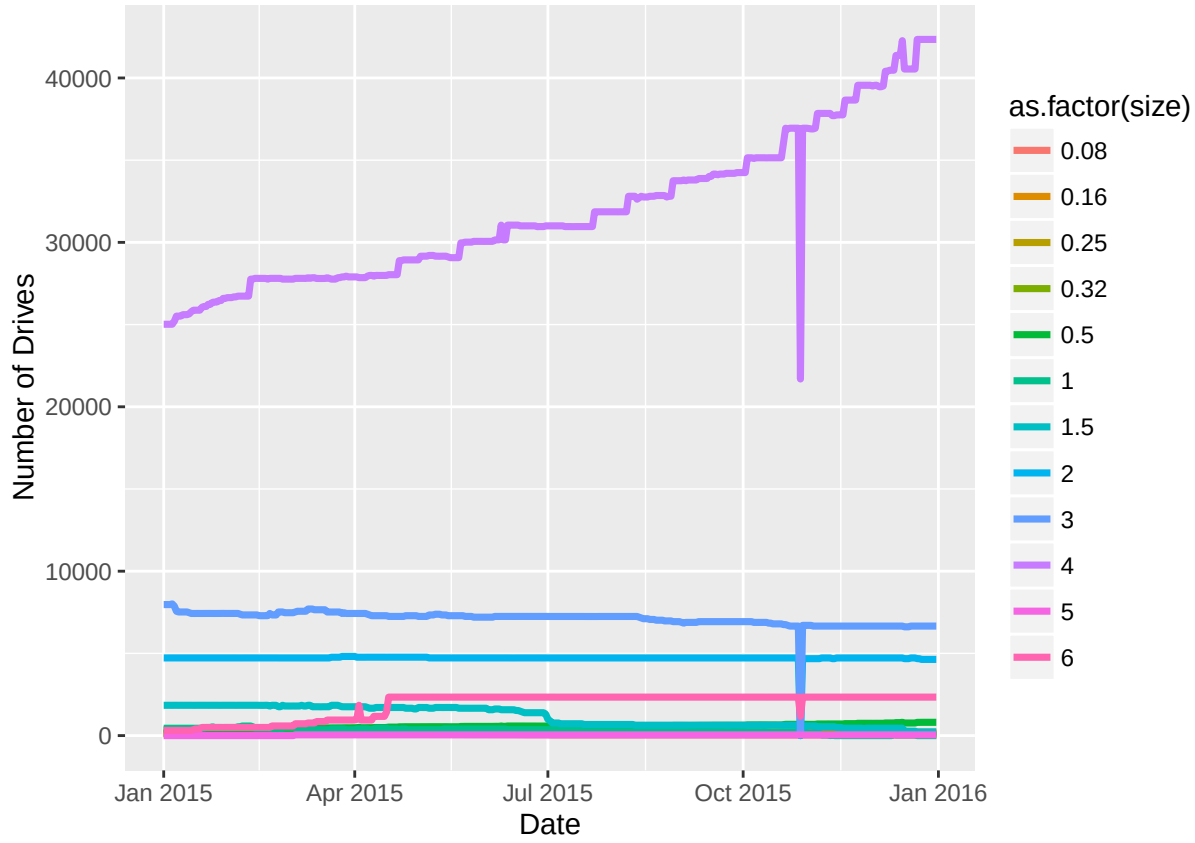
Year 2014 (from 2014-01-01 to 2014-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2014/2014/hd2014_sizeDriveCnt_Failu
ggplot(stat_sizeDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = as.factor(size)), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



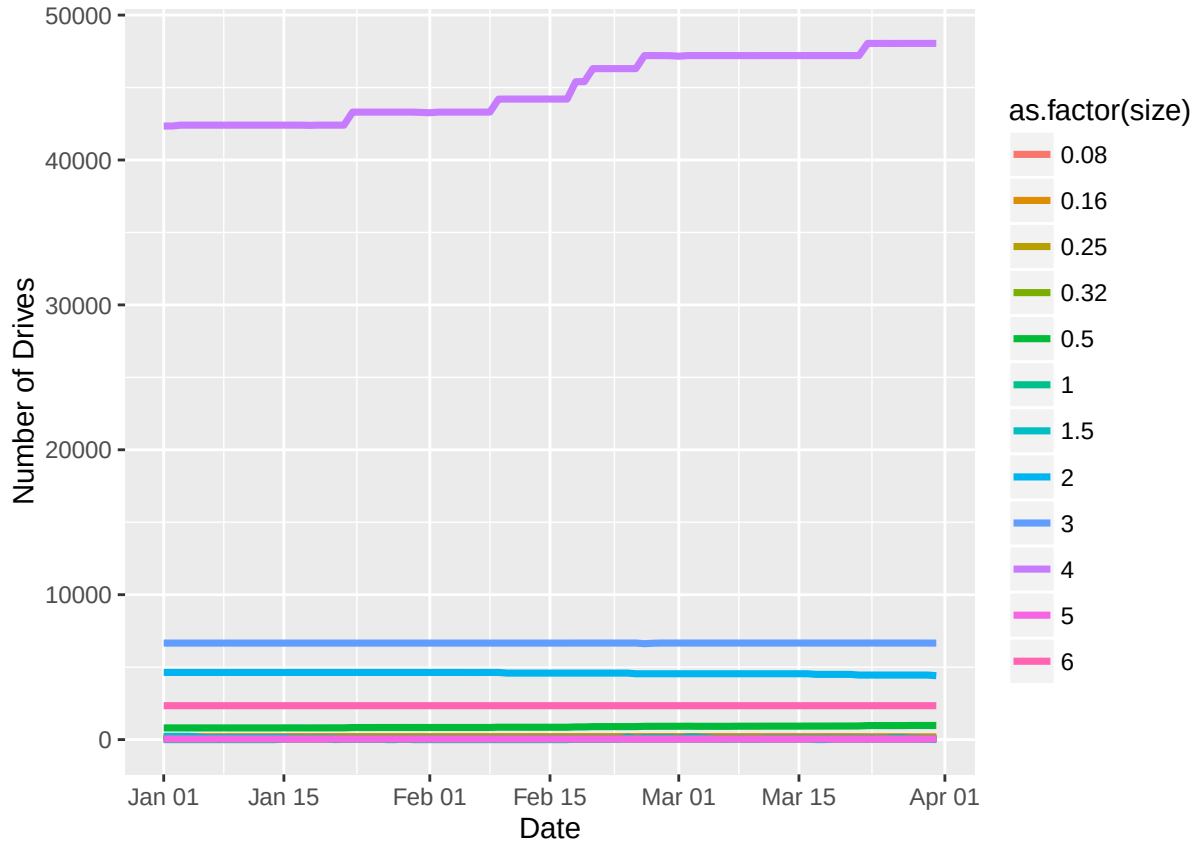
Year 2015 (from 2015-01-01 to 2015-12-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_2015/2015/hd2015_sizeDriveCnt_Failu
ggplot(stat_sizeDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = as.factor(size)), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



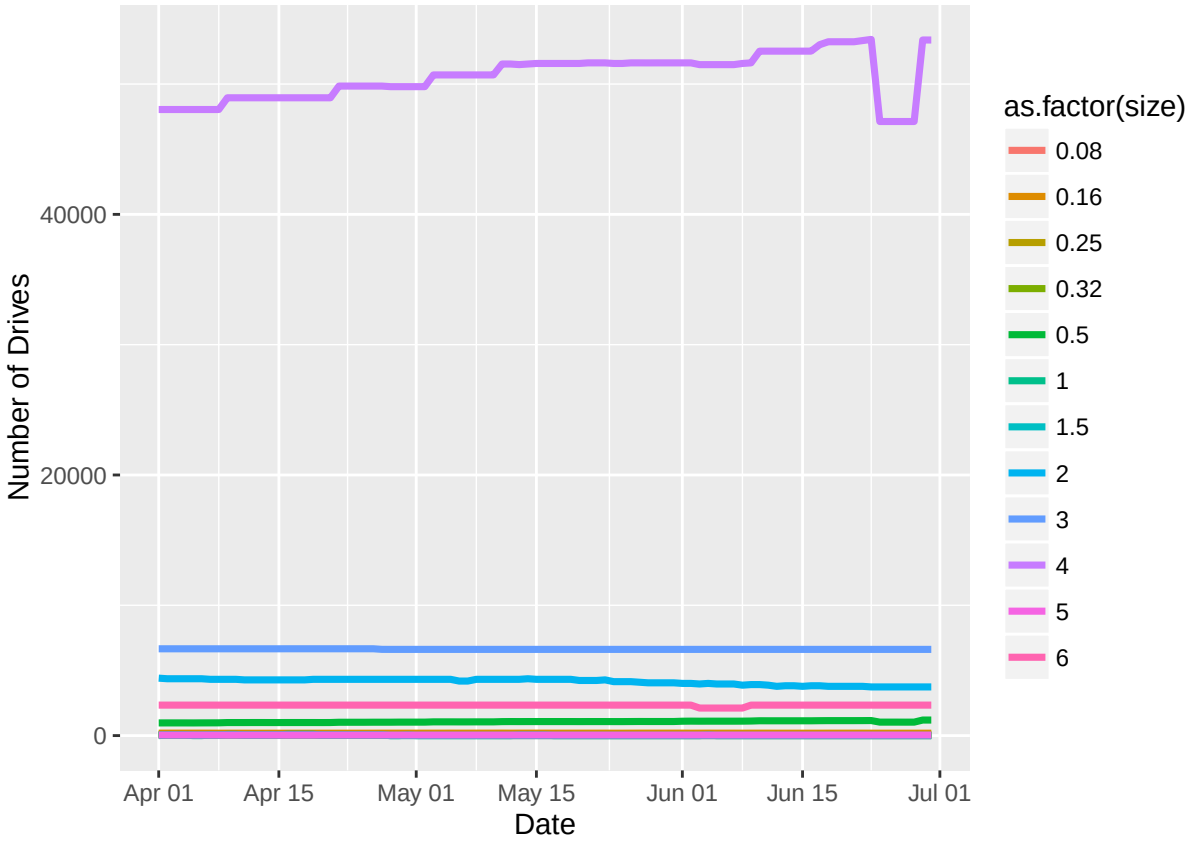
Q1 of Year 2016 (from 2016-01-01 to 2016-03-31):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_Q1_2016/data_Q1_2016/hdQ1_2016_sizeI
ggplot(stat_sizeDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = as.factor(size)), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



Q2 of Year 2016 (from 2016-04-01 to 2016-06-30):

```
library(ggplot2)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/data_Q2_2016/data_Q2_2016/hdQ2_2016_sizeI
ggplot(stat_sizeDriveCnt_FailureCnt_df, aes(x = date, y = total)) +
  geom_line(aes (color = as.factor(size)), size = 1.25) +
  scale_x_date("Date") +
  scale_y_continuous("Number of Drives")
```



Similarly, we can see that the 3 GB drives are mainly used in 2013, then the steady increased installation of 4 GB drives over the years.

Hard Drive Reliability Performance

Survival Analysis

The suitable analysis to be studied on the data is the survival analysis. Survival analysis is to analyze the expected duration of time until an event occurs. R has a package "survival" to perform the analysis. There are some good resources on survival analysis:

- **Econometrics videos and notes by Ani Katchova**: Very accessible videos to walk through the fundamentals and examples using R. https://sites.google.com/site/econometricsacademy/econometrics-models/survival-analysis
- **David Madigan Notes from Columbia University**: More in-depth explanation with R examples. http://www.stat.columbia.edu/~madigan/W2025/notes/survival.pdf
- **"Survival Analysis - A Self-Learning Text, 3rd Edition" Book**: Springer textbook by David G. Kleinbaum and Mitchel Klein. http://www.springer.com/us/book/9781441966452

The key of the survival analysis is the following concept. Suppose the probability that the duration time of less than t is $F(t) = Pr(T \leq t) = \int_0^t f(s)ds$, then the *survival function* is the probability that the duration of at least t as $S(t) = 1 - F(t) = Pr(T \geq t)$. *Hazard rate* is defined as the probability that the duration will end after time t given that it has lasted until time t as $\lambda(t) = \frac{f(t)}{S(t)}$. The hazard rate is the probability that a subject experiences the event at time t while that subject is at risk for experiencing that event.

In the case of hard drive, our interest is to understand the expected duration of the hard drive can "survive" on the shelf before it fails. This is important since the datacenter like BackBlaze can: (a) favorbly procure the hard drives that have high "survival" period, (b) optimize the time for procuring the replacement, i.e. not procuring too early to store too much inventory, or too late that affect the uptime.

Survival and Hazard Models

This section provides very high level description of survival model analysis. Please consult the references in the previous subsection for more information.

Generally, the survival function is decreasing over time. There are three types of models to model such survival functions and the corresponding hazard rate: (a) non-parametric models, (b) parametric models, and (c) semi-parametric.

The **non-parametric models** involves arranging the durations of the observations from smallest to largest, then calculate the corresponding *survival function* and *hazard rate* for each duration. The Kaplan-Meier calculation can be plotted on a graph to see the decrease of survival function over the duration. Then the Nelson-Aalen calculation can be used to calculate the corresponding hazard rate. See the above references for details.

The **parametric models** are mainly just fit them in different models. The popular ones are *exponential*, *Weibull*, *Gompertz* and *log-logistics*. In advantages of these are the performance can be compared using parametric values. In this project, we will not focus on this method as the interpretation of the parameters can be highly dependent on the model used. We think that the nonparametric models are more visual.

The **semiparametric** Cox proportional hazard model is also another more intuitive (and less ambiguous) method of evaluating the hazard rate. While details can also be found in the above references, mainly the determined coefficients can be interpreted as follows:

- Positive coefficient means hazard rate of > 1, which means that it has lower duration and higher hazard rate (more likely for the event to happen).
- Negative coefficient means hazard rate of between 0 and 1, which means that it has higher duration and lower hazard rate (less likely for the event to happen).

We will adopt the following to study the hard drive reliability performance:

- **Kaplan-Meier** survival function: survfit(Surv(spell, event) ~ group)
- **Nelson-Aalen** hazard function: survfit(coxph(Surv(spell, event) ~ group), type = "aalen")
- **Cox proportional** hazard function: coxph(Surv(spell, event) ~ groups, method = "breslow")

For our analysis, a few covariates/groups (basically properties that affect the survival) can be looked into. First is more on the hard drive physical properties itself like manufacturer (brand) and capacity. The reason why this is physical is that it is something that doesn't change over time.

For future work, we can explore the properties changed over time (time series data) such as using the SMART data. This is more difficult since typically covariates are discrete variables, while SMART data are time series integers.

Data Wrangling/Manipulation

In order to perform survival analysis, the data needs to be coverted such that they can be suitable for the analysis. The serial number is an important element of the hard drive, which serves as the identifier of the "subjects" to be studied. Since each subject's situation is spreaded across the csv files, all these rows of data with appropriately picked/computed/grepped columns will need to be binded to create a large data frame.

Then, the serial numbers can be grouped (using “group_by” in “dplyr”), then calculate the corresponding “spell”, “event” and “group” or “groups” (using “summarise” in “dplyr”). Finally, these “summarised” data will need to be joined column-wise (using “left_join” in “dplyr”) to create the format to be used with the survival function.

As discussed, we will look into:

1. Overall survival of the hard drives in Backblaze server: take only serial number and failure binary columns, and bind all the data in rows for each day over the whole duration. For example:

	serial_number	spell	event
1	13H2B97AS	789	0
2	13H3012AS	329	0
3	13H32WEAS	1103	0
4	13H6A0DGS	444	0
5	13H6A21GS	1077	0
6	13H7JUNAS	352	0
7	13H7X2HAS	508	0
8	13H80PNGS	664	0
9	13H85BMGS	933	0
10	13H87YWAS	1081	0
11	13H883WAS	756	1
12	13H89KSGS	1082	0
13	13H89U2GS	779	0
14	13H8AB0GS	881	0
15	13H8B2RGS	789	0
16	13H8B38GS	663	0
17	13H8B3SGS	1086	0
18	13H8B43GS	801	0
19	13H8B48GS	446	0
20	13H8D70AS	722	0
21	2511K0Q3FMYB	478	0

2. The survival/hazard based on different brands/manufacturer: take model info and use grep to create brand columns of TRUE and FALSE, delete the model info to save file size, then combine to serial number and failure binary columns. The resulting columns will be “serial number (character)” - “failure (binary)” - “HGST (binary)” - “Hitachi (binary)” - “Samsung (binary)” - “Toshiba (binary)” - “WDC (binary)” - “Seagate (binary)”. Again, bind all the data in rows for each day over the whole duration. For example:

	serial_number	spell	event	HGST	Hitachi	Samsung	Toshiba	WDC	Seagate	brand
1	13H2B97AS	789	0	0	0	0	1	0	0	Toshiba
2	13H3012AS	329	0	0	0	0	1	0	0	Toshiba
3	13H32WEAS	1103	0	0	0	0	1	0	0	Toshiba
4	13H6A0DGS	444	0	0	0	0	1	0	0	Toshiba
5	13H6A21GS	1077	0	0	0	0	1	0	0	Toshiba
6	13H7JUNAS	352	0	0	0	0	1	0	0	Toshiba
7	13H7X2HAS	508	0	0	0	0	1	0	0	Toshiba
8	13H80PNGS	664	0	0	0	0	1	0	0	Toshiba
9	13H85BMGS	933	0	0	0	0	1	0	0	Toshiba
10	13H87YWAS	1081	0	0	0	0	1	0	0	Toshiba
11	13H883WAS	756	1	0	0	0	1	0	0	Toshiba
12	13H89KSGS	1082	0	0	0	0	1	0	0	Toshiba
13	13H89U2GS	779	0	0	0	0	1	0	0	Toshiba
14	13H8AB0GS	881	0	0	0	0	1	0	0	Toshiba
15	13H8B2RGS	789	0	0	0	0	1	0	0	Toshiba
16	13H8B38GS	663	0	0	0	0	1	0	0	Toshiba
17	13H8B3SGS	1086	0	0	0	0	1	0	0	Toshiba
18	13H8B43GS	801	0	0	0	0	1	0	0	Toshiba
19	13H8B48GS	446	0	0	0	0	1	0	0	Toshiba
20	13H8D70AS	722	0	0	0	0	1	0	0	Toshiba

3. The survival/hazard based on different sizes/capacity: take the capacity info and create the binary details of each of them (like the brand details above), then combine them with serial number and failure binary columns. For example:

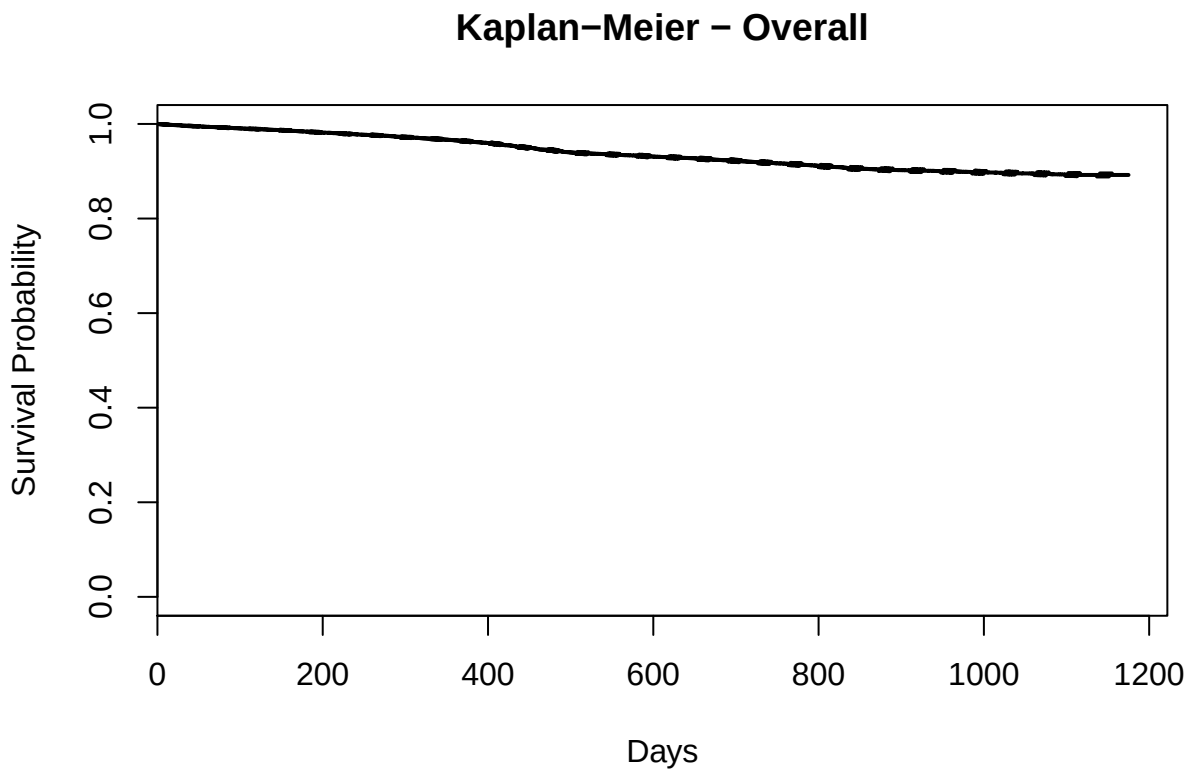
	serial_number	spell	event	size0p08TB	size0p16TB	size0p25TB	size0p32TB	size0p50TB	size1p00TB	size1p50TB	size2p00TB	size3p00TB	size4p00TB	size5p00TB	size6p00TB	TB
1	13H2B97AS	789	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
2	13H3012AS	329	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
3	13H32WEAS	1103	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
4	13H6A0DGS	444	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
5	13H6A21GS	1077	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
6	13H7JUNAS	352	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
7	13H7X2HAS	508	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
8	13H80PNGS	664	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
9	13H85BMGS	933	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
10	13H87YWAS	1081	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
11	13H883WAS	756	1	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
12	13H89KSCS	1082	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
13	13H89U2CS	779	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
14	13H8AB0CS	881	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
15	13H8B2RCS	789	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
16	13H8B38CS	663	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
17	13H8B3SCS	1086	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
18	13H8B43CS	801	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
19	13H8B48CS	446	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
20	13H8D70AS	722	0	0	0	0	0	0	0	0	0	1	0	0	0	3.00 TB
21	2511K0Q3FMYB	478	0	0	0	0	0	0	0	0	0	0	1	0	0	4.00 TB
22	2511K0Q5FMYB	478	0	0	0	0	0	0	0	0	0	0	1	0	0	4.00 TB
23	2511K0Q6FMYB	478	0	0	0	0	0	0	0	0	0	0	1	0	0	4.00 TB

Hard Drive Reliability Performance Results

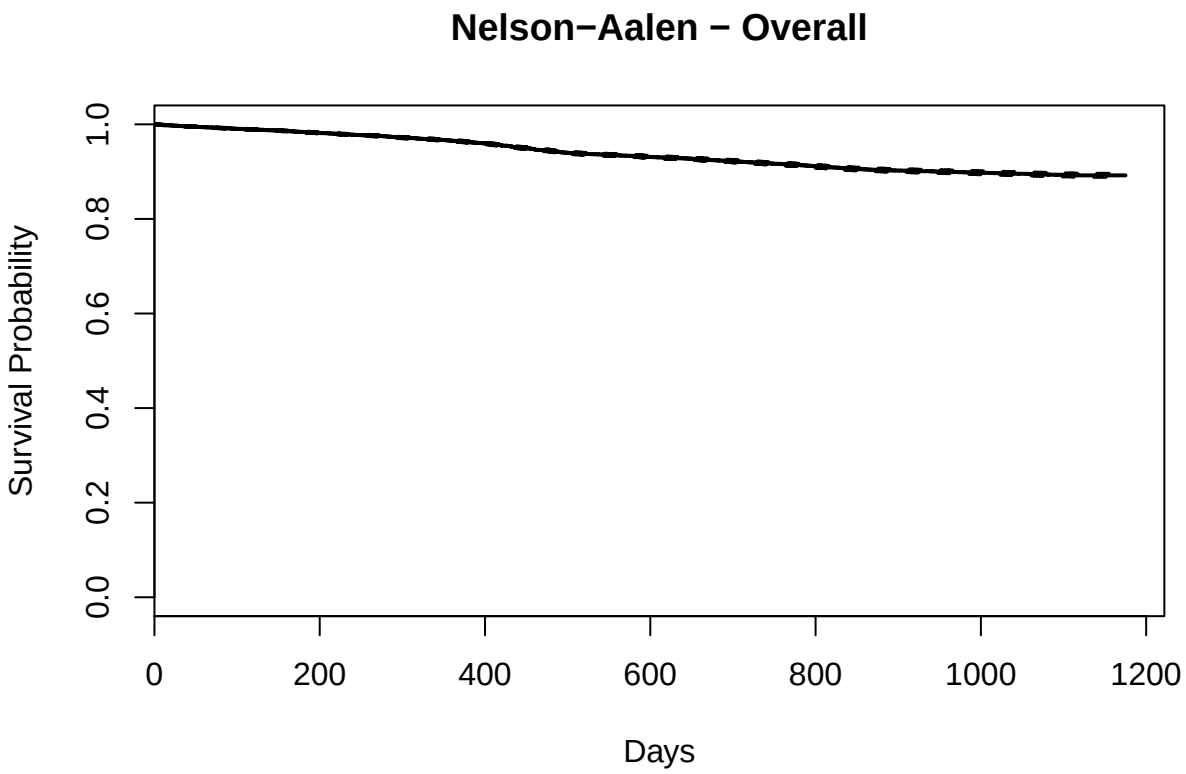
Overall Survival of Hard Drives

```
library(survival)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/SurvivalAnalysisData/allSurvival.RData")

# Kaplan-Meier non-parametric analysis
kmsurvival <- survfit(Surv(allSurvival$spell, allSurvival$event) ~ 1)
#summary(kmsurvival)
plot(kmsurvival, lwd = 2.0, xlab = "Days", ylab = "Survival Probability")
title("Kaplan-Meier - Overall")
```




```
# Nelson-Aalen non-parametric analysis
nasurvival <- survfit(coxph(Surv(allSurvival$spell, allSurvival$event) ~ 1), type = "aalen")
#summary(nasurvival)
plot(nasurvival, lwd = 2.0, xlab = "Days", ylab = "Survival Probability")
title("Nelson-Aalen - Overall")
```




```
# Cox proportional hazard model - coefficients and hazard rates
coxph <- coxph(Surv(allSurvival$spell, allSurvival$event) ~ 1, method = "breslow")
summary(coxph)
```

```
## Call:  coxph(formula = Surv(allSurvival$spell, allSurvival$event) ~
##     1, method = "breslow")
##
## Null model
##   log likelihood= -54830.07
##   n= 84058
```

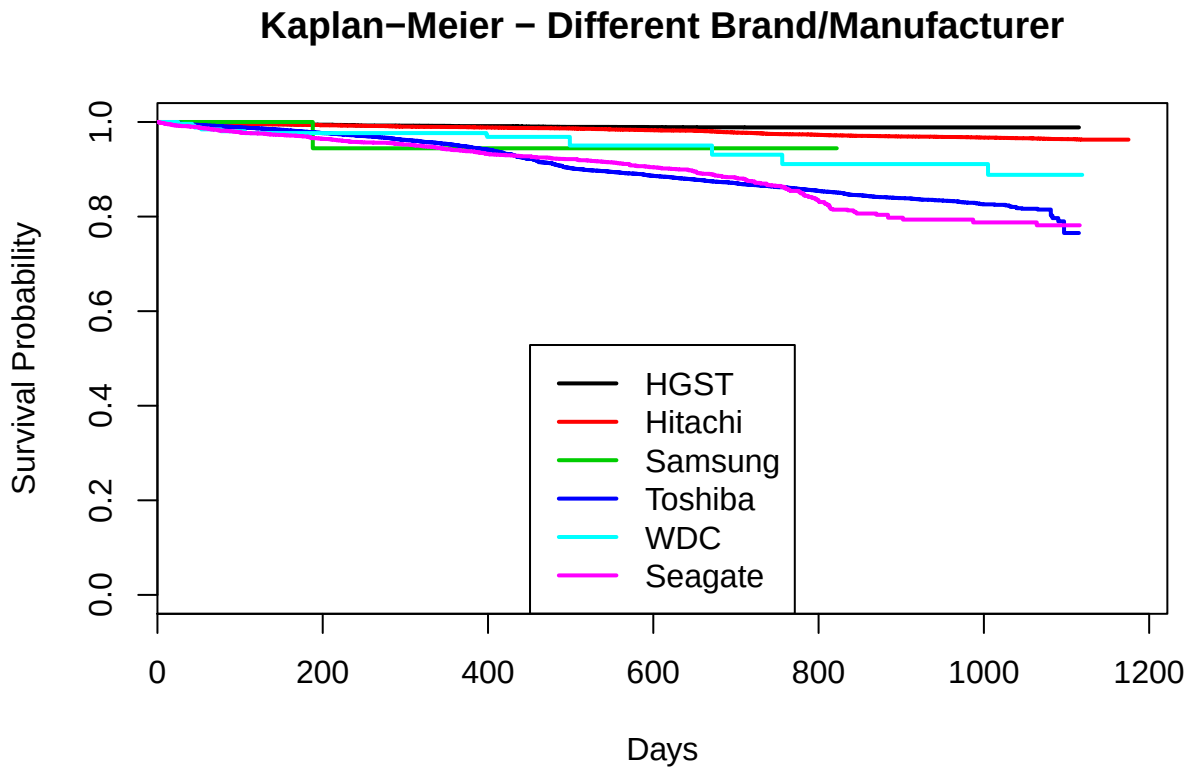
Overall, it can be seen that hard drives are pretty reliable products where the survival probability stays above 85% at about 1200 days (3+ years) of operation.

Survival based on Different Brand/Manufacturer

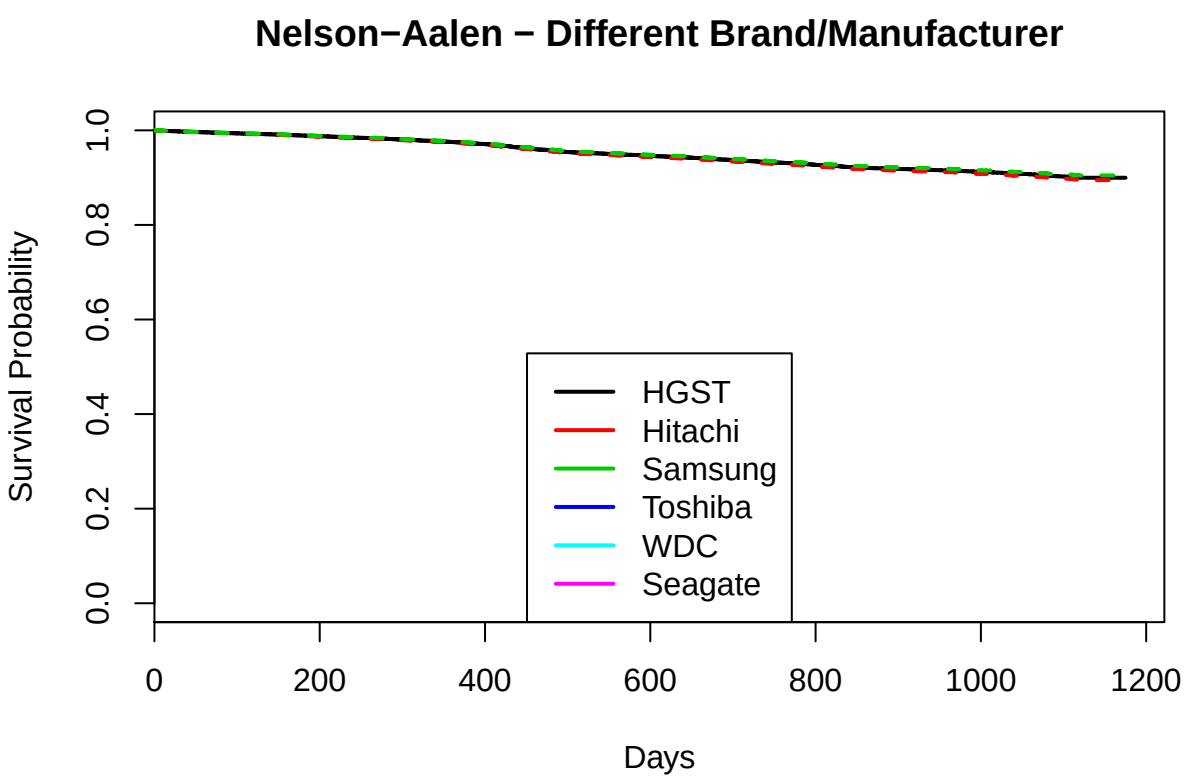
```
library(survival)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/SurvivalAnalysisData/allBrandSurvival.RDa

# Kaplan-Meier non-parametric analysis
kmsurvival <- survfit(Surv(allBrandSurvival$spell, allBrandSurvival$event) ~ allBrandSurvival$brand)
#summary(kmsurvival)
plot(kmsurvival, col = c(1:6), lwd = 2.0, xlab = "Days", ylab = "Survival Probability")
```

```
legend("bottom", c("HGST", "Hitachi", "Samsung", "Toshiba", "WDC", "Seagate"), col = c(1:6), lwd = 2.0)
title("Kaplan-Meier - Different Brand/Manufacturer")
```




```
# Nelson-Aalen non-parametric analysis
nasurvival <- survfit(coxph(Surv(allBrandSurvival$spell, allBrandSurvival$event) ~ allBrandSurvival$bra
#summary(nasurvival)
plot(nasurvival, col = c(1:6), lwd = 2.0, xlab = "Days", ylab = "Survival Probability")
legend("bottom", c("HGST", "Hitachi", "Samsung", "Toshiba", "WDC", "Seagate"), col = c(1:6), lwd = 2.0)
title("Nelson-Aalen - Different Brand/Manufacturer")
```




```
# Cox proportional hazard model - coefficients and hazard rates
coxph <- coxph(Surv(allBrandSurvival$spell, allBrandSurvival$event) ~ allBrandSurvival$brand, method =
summary(coxph)
```

```
## Call:
## coxph(formula = Surv(allBrandSurvival$spell, allBrandSurvival$event) ~
##     allBrandSurvival$brand, method = "breslow")
##
##   n= 84058, number of events= 5077
##
##                                   coef exp(coef) se(coef)      z Pr(>|z|)
## allBrandSurvival$brandHitachi  0.66091   1.93655  0.10076  6.559 5.41e-11
## allBrandSurvival$brandSamsung  2.57339  13.11015  1.00397  2.563   0.0104
## allBrandSurvival$brandSeagate  2.41341  11.17204  0.08962 26.931  < 2e-16
## allBrandSurvival$brandToshiba  1.66336   5.27700  0.30195  5.509 3.61e-08
## allBrandSurvival$brandWDC      2.46722  11.78959  0.10072 24.495  < 2e-16
##
## allBrandSurvival$brandHitachi ***
## allBrandSurvival$brandSamsung *
## allBrandSurvival$brandSeagate ***
## allBrandSurvival$brandToshiba ***
## allBrandSurvival$brandWDC     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                               exp(coef) exp(-coef) lower .95 upper .95
```

```
## allBrandSurvival$brandHitachi     1.937    0.51638     1.589     2.359
## allBrandSurvival$brandSamsung    13.110    0.07628     1.832    93.797
## allBrandSurvival$brandSeagate    11.172    0.08951     9.372    13.317
## allBrandSurvival$brandToshiba     5.277    0.18950     2.920     9.537
## allBrandSurvival$brandWDC        11.790    0.08482     9.678    14.363
##
## Concordance= 0.665  (se = 0.004 )
## Rsquare= 0.035   (max possible= 0.729 )
## Likelihood ratio test= 2952  on 5 df,   p=0
## Wald test            = 1850  on 5 df,   p=0
## Score (logrank) test = 2514  on 5 df,   p=0
```

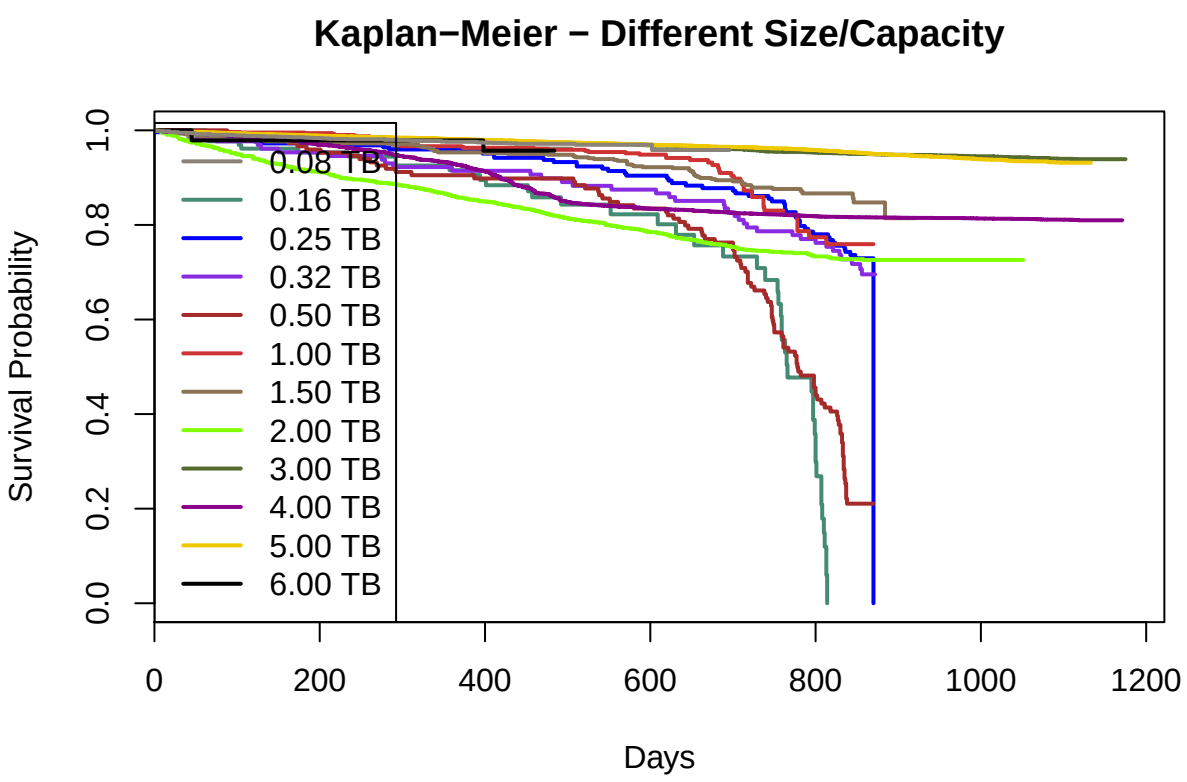
It can be seen that HGST and Hitachi hard drives have the highest survival probability, while Toshiba and Seagate have the lowest survival probability. However, when looking at the hazard rate, Samsung has more risk, but it has the least history amount the other drives. With comparable history WDC probably has the highest risk. Hence, Hitachi may be the highest reliability hard drive.

Survival based on Different Size/Capacity

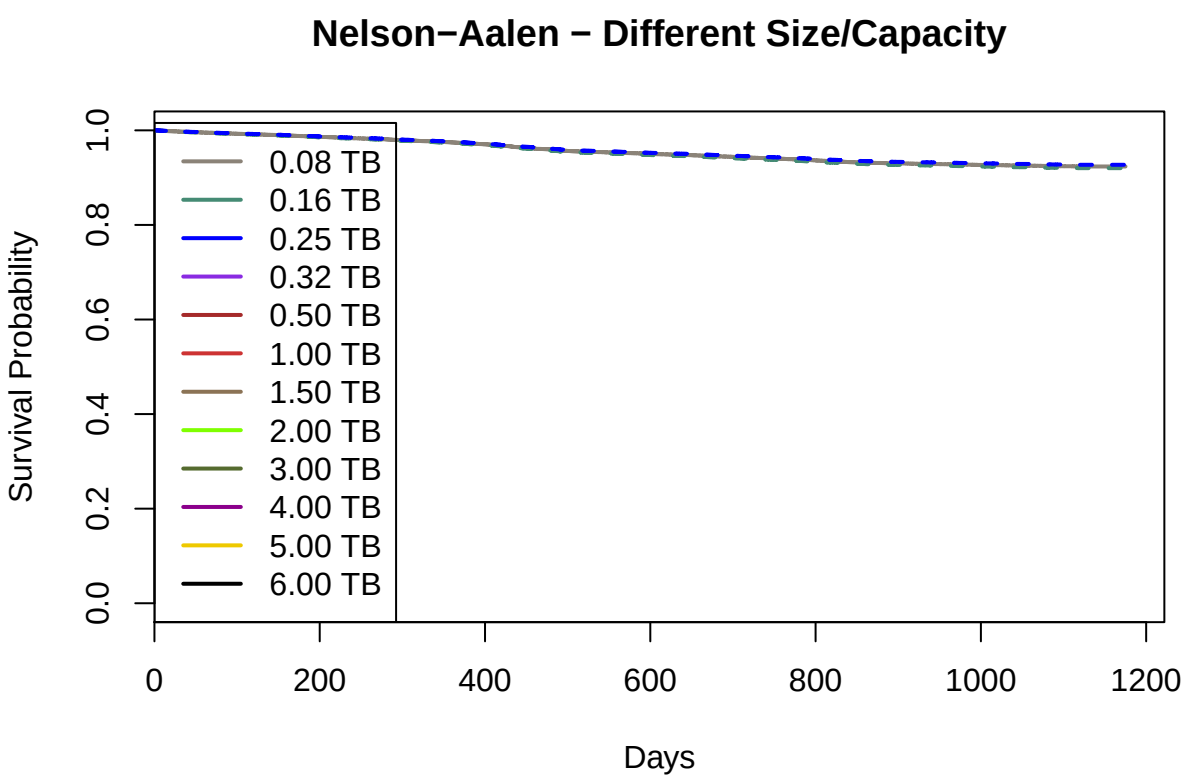
```
library(survival)
load(file = "C:/Users/Chinpei/Documents/R/HardDriveReliability/SurvivalAnalysisData/allSizeSurvival.RDa

color12 <- c("#8B8378", "#458B74", "#0000FF", "#8A2BE2", "#A52A2A", "#CD3333", "#8B7355", "#7FFF00", "#

# Kaplan-Meier non-parametric analysis
kmsurvival <- survfit(Surv(allSizeSurvival$spell, allSizeSurvival$event) ~ allSizeSurvival$TB)
#summary(kmsurvival)
plot(kmsurvival, col = color12, lwd = 2.0, xlab = "Days", ylab = "Survival Probability")
legend("bottomleft", c("0.08 TB", "0.16 TB", "0.25 TB", "0.32 TB", "0.50 TB", "1.00 TB", "1.50 TB", "2.
title("Kaplan-Meier - Different Size/Capacity")
```




```r
# Nelson-Aalen non-parametric analysis
nasurvival <- survfit(coxph(Surv(allSizeSurvival$spell, allSizeSurvival$event) ~ allSizeSurvival$TB), ty
#summary(nasurvival)
plot(nasurvival, col = color12, lwd = 2.0, xlab = "Days", ylab = "Survival Probability")
legend("bottomleft", c("0.08 TB", "0.16 TB", "0.25 TB", "0.32 TB", "0.50 TB", "1.00 TB", "1.50 TB", "2.0
title("Nelson-Aalen - Different Size/Capacity")
```




```r
# Cox proportional hazard model - coefficients and hazard rates
coxph <- coxph(Surv(allSizeSurvival$spell, allSizeSurvival$event) ~ allSizeSurvival$TB, method = "bresl
summary(coxph)
```

```
## Call:
## coxph(formula = Surv(allSizeSurvival$spell, allSizeSurvival$event) ~
##     allSizeSurvival$TB, method = "breslow")
##
##   n= 84058, number of events= 5077
##
##                               coef exp(coef) se(coef)      z Pr(>|z|)
## allSizeSurvival$TB0.08 TB  2.2964    9.9380   0.4720  4.865 1.14e-06 ***
## allSizeSurvival$TB0.16 TB  1.3742    3.9520   0.4696  2.927 0.003428 **
## allSizeSurvival$TB0.25 TB  1.5971    4.9389   0.4782  3.340 0.000838 ***
## allSizeSurvival$TB0.32 TB  2.5219   12.4518   0.4603  5.479 4.27e-08 ***
## allSizeSurvival$TB0.50 TB  0.5588    1.7486   0.4673  1.196 0.231770
## allSizeSurvival$TB1.00 TB  0.7552    2.1281   0.4636  1.629 0.103311
## allSizeSurvival$TB1.50 TB  1.6344    5.1263   0.4506  3.627 0.000286 ***
## allSizeSurvival$TB2.00 TB -0.1875    0.8290   0.4528 -0.414 0.678859
## allSizeSurvival$TB3.00 TB  1.1180    3.0588   0.4496  2.487 0.012888 *
## allSizeSurvival$TB4.00 TB -0.2766    0.7584   0.4496 -0.615 0.538440
## allSizeSurvival$TB5.00 TB  0.1094    1.1156   0.8376  0.131 0.896099
## allSizeSurvival$TB6.00 TB -0.2335    0.7917   0.4646 -0.503 0.615181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                             exp(coef) exp(-coef) lower .95 upper .95
## allSizeSurvival$TB0.08 TB    9.9380    0.10062    3.9402    25.065
## allSizeSurvival$TB0.16 TB    3.9520    0.25304    1.5744     9.920
## allSizeSurvival$TB0.25 TB    4.9389    0.20248    1.9345    12.609
## allSizeSurvival$TB0.32 TB   12.4518    0.08031    5.0520    30.690
## allSizeSurvival$TB0.50 TB    1.7486    0.57188    0.6997     4.370
## allSizeSurvival$TB1.00 TB    2.1281    0.46991    0.8578     5.280
## allSizeSurvival$TB1.50 TB    5.1263    0.19507    2.1198    12.397
## allSizeSurvival$TB2.00 TB    0.8290    1.20620    0.3413     2.014
## allSizeSurvival$TB3.00 TB    3.0588    0.32693    1.2673     7.383
## allSizeSurvival$TB4.00 TB    0.7584    1.31865    0.3141     1.831
## allSizeSurvival$TB5.00 TB    1.1156    0.89639    0.2160     5.761
## allSizeSurvival$TB6.00 TB    0.7917    1.26307    0.3185     1.968
##
## Concordance= 0.692  (se = 0.004 )
## Rsquare= 0.036   (max possible= 0.729 )
## Likelihood ratio test= 3068  on 12 df,   p=0
## Wald test            = 3233  on 12 df,   p=0
## Score (logrank) test = 4176  on 12 df,   p=0
```

It can be seen that, with comparable history, from highest to lowest survival probability, 3 TB has the highest survival probability, followed by 5 TB, then 4 TB. Large drive like 6 TB has only about 500 days of history, so it is a bit early to tell its survival performance. The smaller drives are less important since they are not very useful today. In terms of hazard rate, it can be seen that 2 TB, 4 TB and 6 TB have the best hazard performance. Hence, probably the 4 TB drives are recommended.

Conclusion

The study is very preliminary since only brand/manufacturer and size/capacity are evaluated. In this study, Hitachi has the best reliability, Toshiba and Seagate have the least reliability. 3 TB, 5 TB and 4 TB drives have high survival rates, but 2 TB, 4 TB and 6 TB drive have the lowest hazard rates. The recommendation is to use Hitachi and 4 TB drives with highest survival probabiliity and lowest hazard rate.

The analysis have some limitation. First, the reliability of the hard drive data collection may affect the results. As seen in the exploratory time-series plots, quite significant amount of data were dropped or missing. Then, the performance can also be heavily affected by some external factors, i.e. how the data center install the hard drives, and the quantity of each of the models. Overall, the project produces reasonable analytical results for decision making.

Future Work

There are a lot of future avenue for this work. Survival analysis of each model (about 100+ of them) for more serious consideration. The censoring of the data needs to be investigated more carefully. Finally, the time-series SMART data should be taken into account for much more advanced analysis.

Notes to the Codes

The codes can be found at: https://github.com/chinpeitang/Springboard_FDS/tree/master/CapstoneProject.

The hard drive data by Backblaze can be downloaded at: https://www.backblaze.com/b2/hard-drive-test-data.html.

In each of the codes, make sure you change the filePathRoot of where you save your data. Once you downloaded data_xxxx.zip file, extract it to the data_xxxx directory, and don't make any changes within the directory.

```
filePathRoot <- "C:/Users/Chinpei/Documents/R/HardDriveReliability/"
```

- harddrive_readdata.R: Create the RData files for time series plotting and survival analysis for each year.
- harddrive_quickloaddata.R: Quick plot of the imported data from harddrive_readdata.R.
- harddrive_allsurvivalanalysis_savedata.R, harddrive_brandSurvivalAnalysis.R, harddrive_sizesurvivalanalysis_savedata Create the RData file of full set of available data (2013, 2014, 2015 and 2016) for survival analysis. Due to time limitation, the authors did not make the convenience to select different function and filePathRoot setup.

Disclaimer: Use the codes and analysis results at your own risk. This project is mainly for exercise purpose. The codes and analysis results comes with no warranty. The author is not responsible for the consequence that may result in any form of lost should the reader use the result from this codes and analysis.

Acknowledgment

The author would like to acknowledge Backblaze huge effort in logging the hard drive data and making them open source. The work is impossible without their hard work. The author would also like to thank the mentor Jeff Lipkowitz for introducing survival analysis for this work. Finally, thanks to Springboard for providing the platform for the author to learn data science.