# CS 3983 - Assignment 2

Nathan McGugan - 3714794

December 11, 2023

## 1 Introduction and Background

The goal of this project is to transform and analyze data collected from Backblaze about the hard drives in their data center in order to test a hypothesis. Backblaze is a cloud storage provider that has been publishing hard drive statistics since 2013, offering data on the performance and reliability of hard drives in a large-scale environment. This project will use the years 2016 to 2023. Their dataset includes information from hundreds of thousands of hard drives and solid-state drives spanning various models and manufacturers. The data is provided as a daily CSV file, and contains columns such as the model, serial number, state (failed or not), data, and capacity. Additionally, the CSV contains the SMART (Self-Monitoring, Analysis, and Reporting Technology) data for each drive on that day. However, this project will not be utilizing the SMART columns.

## 2 Hypothesis

The hypothesis for this project was that hard drives are more likely to fail as time goes on than they were at the beginning. The reason that this hypothesis is logical is the idea of wear and tear; as hard drives are used more, they should deteriorate in quality, leading to a hard drive failure.

With the many years of data provided by Backblaze and the hundreds of thousands of hard drives they have used, a trend that either proves or disproves this hypothesis should appear. In order to test this hypothesis, a correlation heat map can be created, and logistic regression can be performed. These two methods can show how related operational days and failure rates are, and create a prediction model for those variables.

## 3 Data Analysis

The first step in analyzing the data is first to transform it into a usable format. Initially, each day has a separate CSV for each day. By concatenating each column into one data frame and grouping by serial number, the data shrinks

substantially. However, in order to preserve the most data possible, three additional columns are derived: operational days, first day seen, and last day seen. The 'failure' column is then set to one if the drive failed, and zero if it never did. This data frame is then saved in the feather file format, as it is more efficient for data analysis than the CSV file type.[1]

Once the data was fully loaded, a heat map could be generated. This involved using the Python packages 'seaborn' for computing the correlation and 'matplotlib' for creating the graph. The correlation was calculated on the two columns 'failure' and 'operational_days', and created this graph:

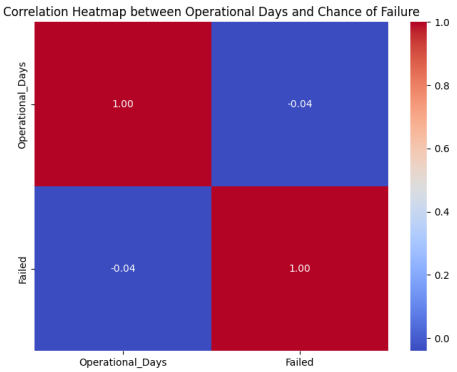


Figure 1: Correlation Heat Map

The graph showed that the failure of a hard drive is negatively uncorrelated with how long the drive has been in use.

The next step in analyzing the data was to create a predictive model for failure and operational days. Since failure is categorical in nature (either failed or did not), logistic regression is the best choice for the model. In order to create the model, the Python package 'statsmodels' can be used, and 'matplotlib' can once again be used to graphically represent the model. This is the model that was created from the response variable 'failure' and the predictor variable 'operational_days':

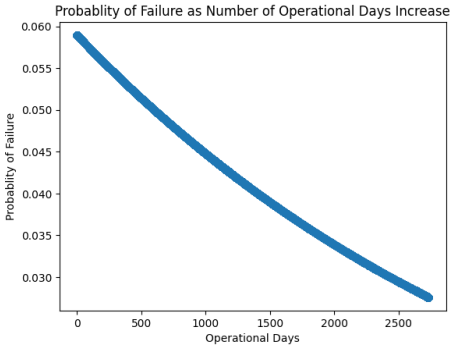


Figure 2: Logistic Regression Model

This model shows that as a hard drive gets older, it actually becomes less likely to fail. This could be due to a couple of factors, but it is most likely due to defects in the manufacturing of hard drives that cause them to fail early. This would cause there to be more failures at low operational days and fewer failures later on when most of the bad hard drives have already failed.

As an additional form of analysis, a bar chart can be created that represents the failure percentage for various hard drive brands. By deriving the brand of a drive through its model, the following graph can be produced:

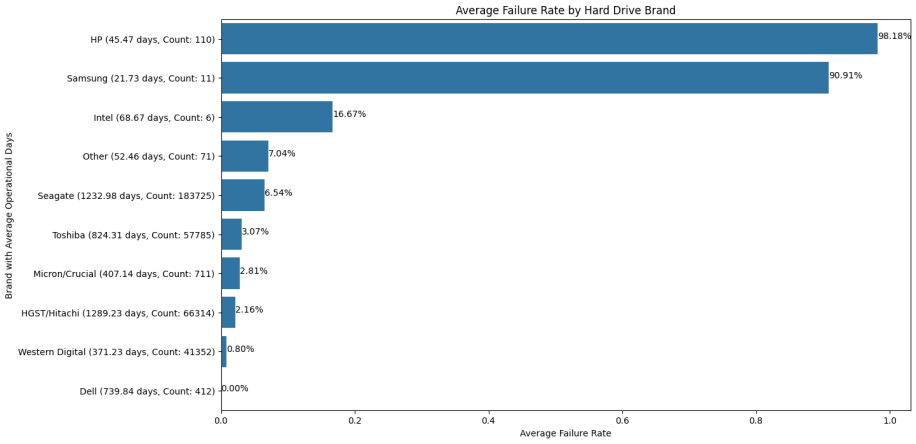


Figure 3: Brand Failures

This shows

# 4 Conclusion

testsetst

# References

[1] Ilia Zaitsev. The best format to save pandas data. *Towards Data Science*, Mar 2019.