Designing Shorting Strategies with Benford's Law

Sedrick Scott Keh

Supervised by Dr. David Rossiter

Contents

Abstract

Benford's Law is an observation about the frequency distribution of leading digits of many real-life sets of numerical data. It states that in many datasets, most elements (around 30%) will have 1 as the first digit, around 17% will have 2 as the first digit, and so on, with a decreasing trend until 9, of which only 4.5% of elements in the dataset will have it as its first digit. In this report, we investigate the applications of Benford's Law on financial data. More specifically, we verify that the closing prices of S&P500 stocks indeed closely follow the Benford distribution. We provide an analysis by sector and explore the Enron scandal as a case study of a dataset that deviates from the Benford distribution. This Benford model is then used as the motivation to design short sell recommendation strategies. Finally, we apply these strategies to potential stocks (Joyy Inc., eHealth Inc.) suggested by Muddy Waters Research.

1 Introduction

In today's robust financial industry, there are countless methods used to analyze stock prices. From charting technical indicators to scrutinizing disclosed financial statements, each of these factors influence the way that these stocks are valued. Recently, a class of heuristics-based algorithms has surfaced for analyzing financial data. Although relatively less verified than traditional methods, they arguably provide valuable insight that can be used to complement findings from various existing methods, as well as to give possible leads on areas to investigate.

One notable such heuristic-based algorithm is **Benford's Law**.

1.1 Benford's Law

Benford's Law is not strictly a law; rather, it is a formalized statement of an observed phenomenon in the real world. It states that in a collection of numbers (listings, tables, etc.) the distribution of the leading digits is not a uniform distribution where each number from 1 to 9 appears roughly 11.11% of the time. Contrary to intuition, the distribution is skewed to the left, with the digit 1 appearing as a leading digit roughly 30% of the time, the digit 2 appearing 17% of the time, and so on, decreasing until the digit 9, which appears 4.5% of the time.

More formally, the estimated frequency of digit d in a dataset is expressed as

$$\log_{10}\left(1 + \frac{1}{d}\right), \quad 1 \leq d \leq 9$$

Following this formula, a Benford distribution is shown in Figure 1.

1.2 Objectives

The main objectives of this report are two-fold:

1. We analyze the leading digits of various stock prices over a period of time and determine which stocks most closely follow the ideal Benford distribution. We verify whether or not the Benford model is accurate for stock price data. This will also include an analysis on how various factors such as the sector/industry affect the "Benford-ness" of a stock.
2. We then apply the findings to real-life stock data. More specifically, we analyze historical data of stocks that do not fit the ideal Benford distribution, as well as past incidents of fraud. This idea can be exploited to craft a shorting strategy.

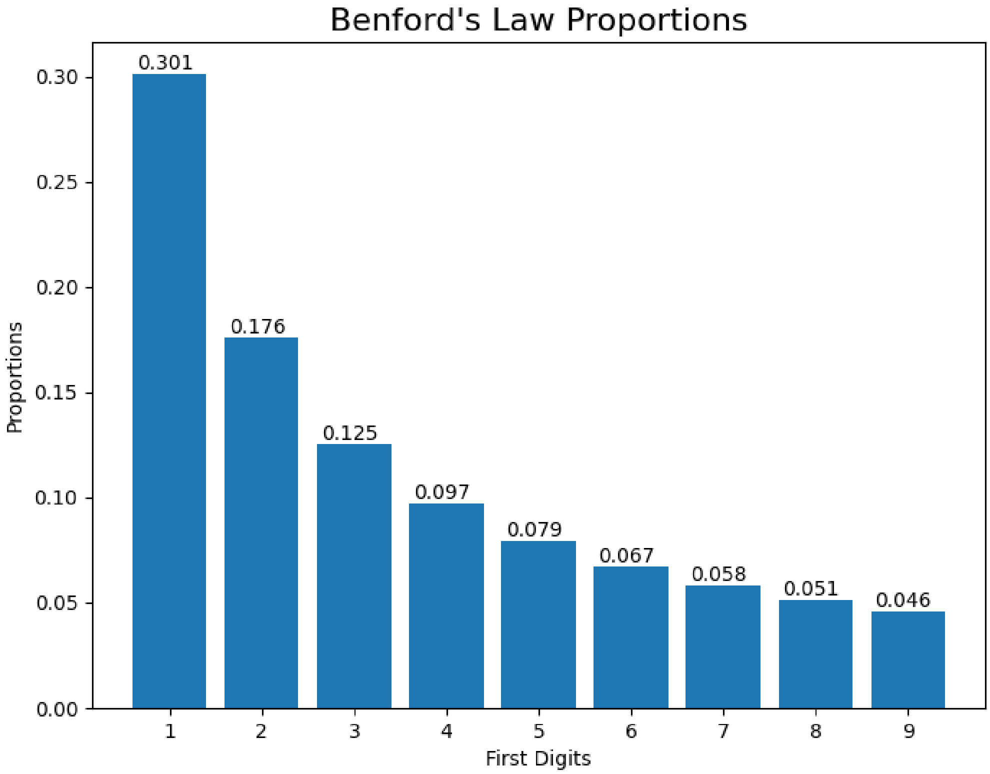

FIGURE 1: Benford's Law Distributions

2 Understanding Benford's Law

2.1 Background

The "first digit phenomenon" was first discovered by mathematician and astronomer Simon Newcomb in 1881, when he discovered in a book of logarithm tables that the first few pages were significantly more worn out than the rest of the pages. With this observation, Newcomb hypothesized that people performed more calculations with the lower digits than the higher ones.

This idea was further investigated by physicist Frank Benford in 1938. The pattern was shown to appear in numerous real world datasets such as baseball statistics, city weather statistics, and census population data. Roger Pinkham later expounded on this by relating digit frequencies with the idea of scale invariance, hence deriving the formula for the law.

2.2 When is Benford's Law Applicable?

Before proceeding, it is important to note that Benford's Law, while often applicable in the real world, is usually inaccurate in theoretical cases, especially in the case of true randomness. Consider, for example, the example of randomly selecting n integers with replacement between 1 and 99, and plotting the distributions of first digits. An elementary counting argument reveals that each digit from 1 to 9 appears as the leading digit with equal frequency, meaning that it would be a truly uniform distribution, and not a Benford one.

On the other hand, Benford's Law also fails with real-world cases that are overly constrained. Consider, for example, the ages of presidents around the world. It is very unlikely that many of these will begin with a 1, contrary to the statement of Benford's Law. This is because age is a very constrained measurement with a very narrow range.

In other words, Benford's law requires two main conditions in order to hold:

1. The data is not totally random.
2. There are minimal constraints on the range and spread of the data.

Under these reasonable assumptions, Benford's Law will apply. Fortunately, stock prices satisfy both of these conditions. Most prices can vary greatly through time in an unconstrained manner, and the distribution can be considered to be non-random because it is affected by various real-life factors. It is hence sensible to apply Benford's Law to analyze stock prices.

3 Methodology

3.1 Data Collection and Analysis

In this study, the stocks analyzed are from the S&P 500 index. This index was selected because most of the listed corporations here have a relatively long history, hence providing more data to work with. The S&P 500 also has a substantial number of companies across different industries, which will allow for more substantial comparison and analysis.

3.1.1 Data Collection

The data was collected from the official Bloomberg data using a Bloomberg Terminal machine. Data collection was done on September 29, 2020. During the process, the closing price of each stock was extracted, together with its corresponding sector. The dataset has a total of 505 listed companies.

3.1.2 Data Preprocessing

To ensure consistency and relevancy among the companies, starting dates were capped at January 1, 1980. This means that the date range covered for each stock is January 1, 1980 to September 29, 2020. Some companies started listing after January 1, 1980. In these cases, we simply consider their time frame to start from their first listed date.

In Table 1, we present a table containing some statistics discovered while performing exploratory data analysis, including a breakdown by sector.

	Total	Average (Price)
Number of sectors	12	
Number of companies	**505**	47.43
Industrial	73	40.91
Information Technology	71	41.25
Financials	66	51.69
Health Care	62	50.51
Consumer Discretionary	61	72.86
Consumer Staples	33	34.62
Real Estate	31	49.90
Materials	28	42.77
Utilities	28	33.38
Energy	26	33.29
Communication Services	26	64.68
Length of Series (i.e. Trading days from 01/01/1980 to 09/29/2020)	10631	
Number of entries	3423481	6779

TABLE 1: Preliminary Data Statistics of S&P500 Stocks

The exploratory data analysis reveals that the prices are indeed quite varied, with a mean price of **47.43** and a standard deviation of **103.19**. The 25, 50, and 75 quartiles are 10.68, 26.64, and 52.85 respectively, so the interquartile range is 41.17, which is very significant, considering that the mean is only 47.43.

Because the prices are very varied, this further confirms the suitability of applying Benford's Law to S&P500 closing price data, as it satisfies both the nonrandom and minimal constraints conditions.

3.2 Benford's Law Algorithm

After processing the data, the next step is to plot out the distributions of the first digits and compare it with the Benford distribution. The algorithm to compute how close a list of numbers is to the ideal Benford distribution is relatively straighforward and is detailed below.

Algorithm 1: Calculate how close list *arr* is to Benford distribution.

```
Result: Distance between the leading digit distribution of arr and the ideal
        Benford distribution
Input: arr;
Let $B$ be the ideal Benford distribution;
Let $d(\cdot, \cdot)$ denote a given distance function between distributions;
Create digit_counts[9], initialized to all 0;
for num in arr do
    // Find first non-zero digit.
    first_digit = None;
    for digit in num do
        if $digit \neq .$ and $digit \neq 0$ then
            first_digit = digit;
            break;
        end
    end
    // Update digit_counts.
    digit_counts[first_digit-1] += 1;
end
// Normalize digit_counts.
digit_counts_normalized = digit_counts / sum(digit_counts);
Return d(digit_counts_normalized, B);
```

This algorithm returns the distance between the ideal Benford distribution and the leading digit distribution of any array.

3.2.1 Choice of Distance Metric

One point of note in the algorithm is the choice of our distance function $d(\cdot, \cdot)$. The idea is that this distance function should measure the closeness between the two given distributions. We consider the following distance functions:

- **Mean Absolute Error:** This is the sum of the absolute differences between corresponding frequencies in the discrete distribution. Mathematically, this is expressed as $d_{MAE}(a, b) = \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|$. It is the simplest metric we use.
- **Mean Squared Error:** MAE has the disadvantage that large errors are not penalized that much. To resolve this, MSE instead squares the differences between corresponding frequencies. Mathematically, this is expressed as $d_{MSE}(a, b) = \frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2$.

- **Kullback-Leibler (KL) Divergence:** These next two distance metrics (KL and JS) are centered on information theory and hence are usually considered to be more robust and accurate than MAE or MSE. The KL divergence measures the relative entropy in the difference of the two distributions. Mathematically, this is expressed as $d_{KL}(a||b) = \sum_{i=1}^{n} \left(a_i \cdot \log_{10} \left(\frac{a_i}{b_i} \right) \right)$.
- **Jensen-Shannon (JS) Divergence:** One disadvantage of using the KL divergence is that it is not symmetric, i.e. $d_{KL}(a||b) \neq d_{KL}(b||a)$. The JS divergence is basically a symmetric and smoothed version of the KL divergence, denoted by $d_{JS}(a, b) = \frac{1}{2} \left(d_{KL}(a||M) + d_{KL}(b||M) \right)$ where $M = \frac{a+b}{2}$.

3.2.2 Results

To verify the applicability of Benford's Law on stock prices, we carry out Algorithm 1 on each of the 505 companies in our dataset. As an example, we consider McDonald's Corporation (MCD), with 10631 entries. The comparison between its first digit distributions and the ideal Benford distribution is shown below:

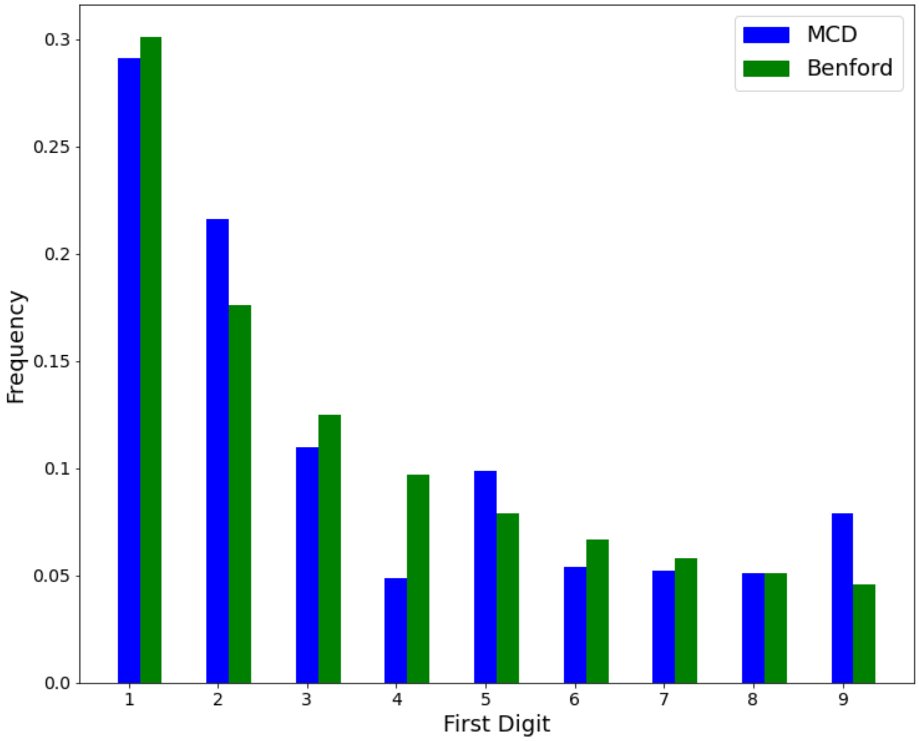

FIGURE 2: First Digit Distributions: MCD vs Ideal Benford

The figure above shows a very close match between the two distributions. This confirms that the stock price of MCD can indeed be modelled accurately using a Benford distribution.

To further quantify this, evaluation is performed separately using each of the 4 metrics mentioned previously (MAE, MSE, KD, JS). For MCD, these distances are 0.021, 0.00066, 0.049, and 0.092 for MAE, MSE, KD, and JS respectively. Note that these metrics represent distance, so a lower score is better.

We repeat this operation on each of the stocks. The results are then tallied and a histogram is plotted:

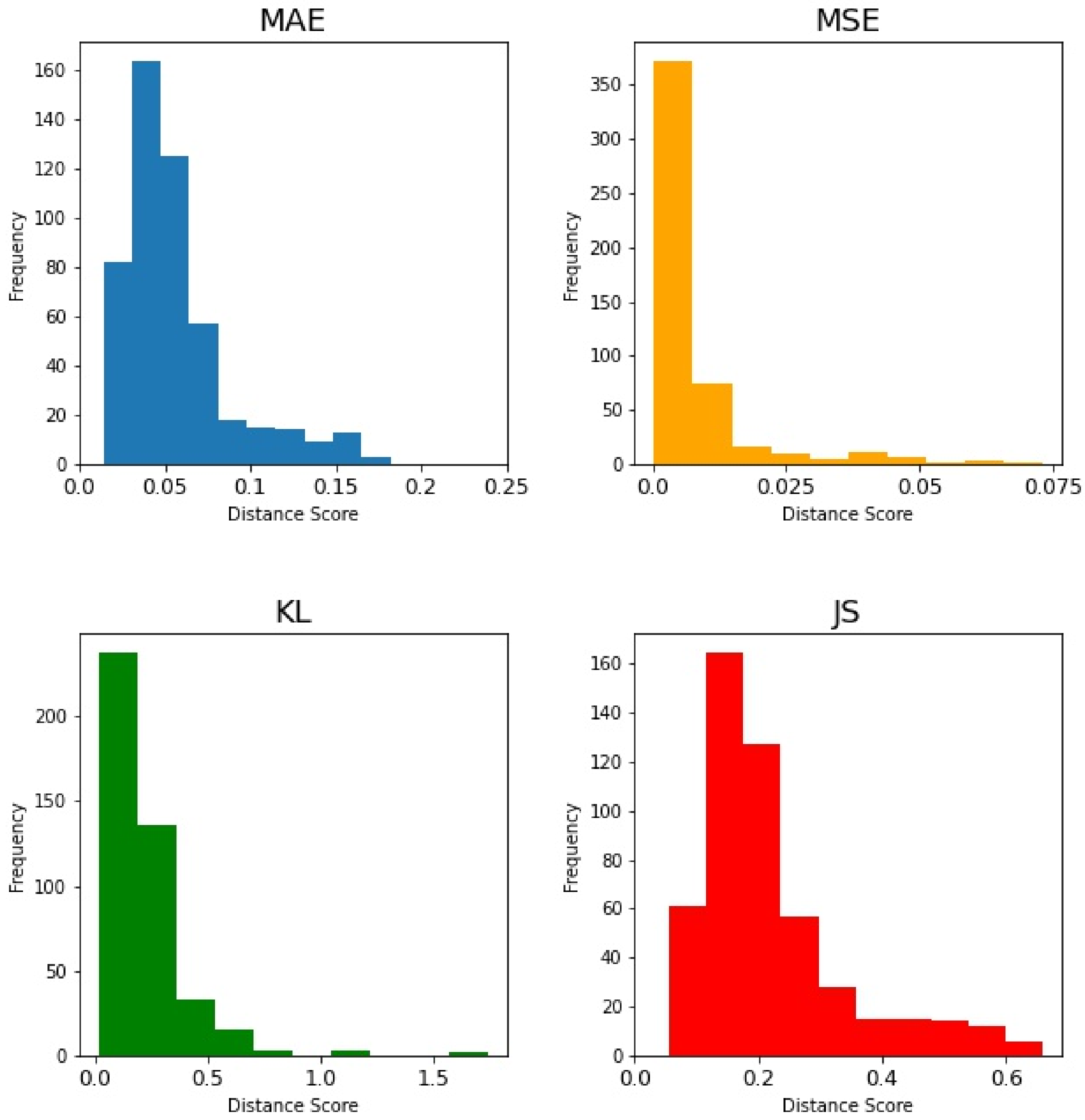

FIGURE 3: Histogram Tallies of Distance Evaluation Results

The diagrams in Figure 4 are all very skewed to the left and close to 0, revealing that most of the distances are quite small, i.e. the first digit distributions are very close to the Benford distributions. **This confirms our hypothesis that we can indeed use a Benford distribution to model the closing price data of these stocks.**

3.2.3 Industry Analysis

In this part, we consider the distances to the ideal Benford distribution, similar to the previous part, but divided by sector.

Sector	**MAE**	**MSE**	**KL**	**JS**
Communication Services	0.0714	0.0141	0.162	0.285
Consumer Discretionary	0.0522	0.0064	0.190	0.201
Consumer Staples	0.0614	0.0092	0.163	0.244
Energy	0.0640	0.0098	0.290	0.245
Financials	0.0548	0.0070	0.196	0.217
Health Care	0.0415	0.0035	0.154	0.166
Industrials	0.0583	0.0074	0.147	0.225
Information Technology	0.0541	0.0066	0.190	0.209
Materials	0.0615	0.0095	0.182	0.239
Real Estate	0.0559	0.0076	0.244	0.225
Utilities	0.0679	0.0113	0.183	0.264

TABLE 2: Distances to Ideal Benford Distribution (by Sector)

From Table 2, we see that sector closest to the ideal Benford distribution is Health Care, followed by Consumer Discretionary and Information Technology. Meanwhile, sectors like Communication Services, Energy, and Utilities have a larger distance as compared to the ideal Benford distribution. This is consistent with our intuition, as products like oil are generally much more volatile due to international sourcing and trading demands.

Comparing Metrics: When comparing metrics, we also see that most of the time, the different metrics perform similarly on a relative basis even though the magnitudes themselves may be very different. There are, however, some exceptions. For instance, in the Communication Services sector, the JS divergence is higher than the KL divergence, while in the Energy sector, the KL divergence is higher. These occasional exceptions make sense because the metrics after all are measuring different things (e.g. MSE penalizes big penalties more than MAE).

3.2.4 Volatility Analysis

Expounding on some ideas discussed earlier, it would seem like the closeness of a stock's price to an ideal Benford distribution is indicative of its relative stability.

To test this hypothesis, we explore the relationship between a stock's volatility and its distance to the Benford distribution. We use the **Cboe Volatility (VIX) Index** to quantify volatility. This measures the market's expectation of 30-day forward-looking volatility.

As we have previously established that the various evaluation metrics work similarly, we simply consider the JS divergence here for convenience. Below is a correlation plot of stocks' VIX Index against their JS-distance from the Benford distribution.

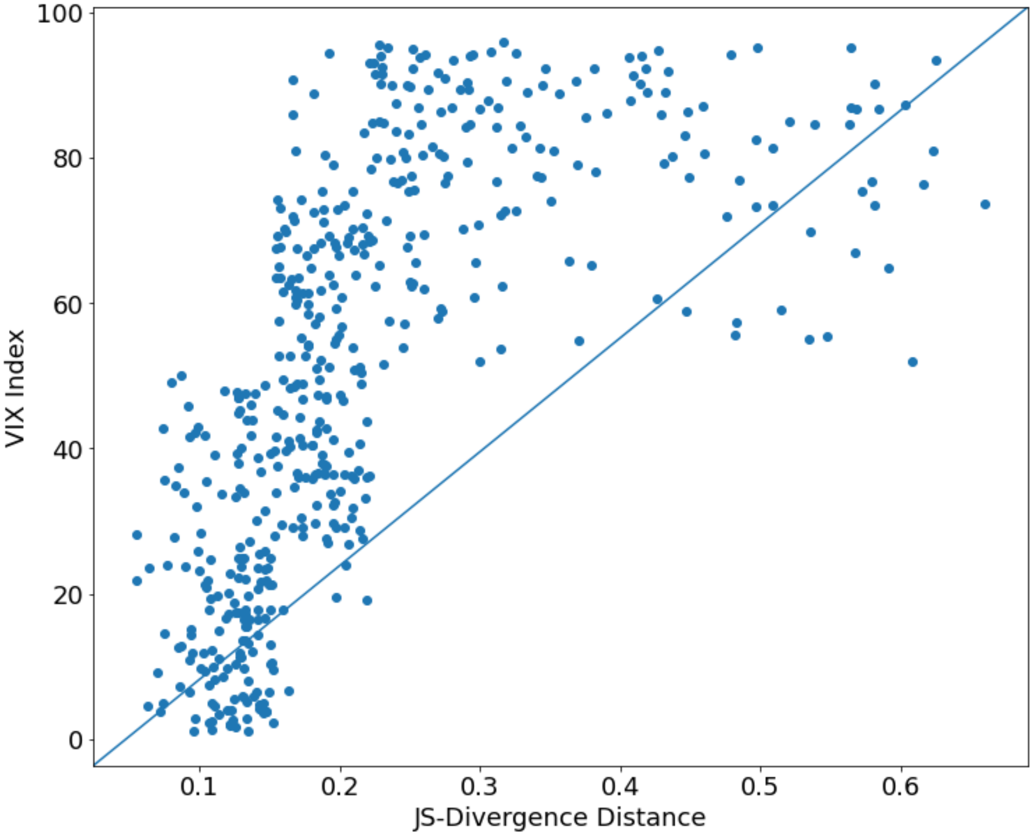

FIGURE 4: Correlation Plot (VIX Index vs JS-Divergence Distance)

The correlation of this plot is **0.68**, which is a moderately strong correlation. This confirms our previous hypothesis that indeed, the distance of a stock price's leading digit distribution with the ideal Benford distribution is a good indicator of its stability. We can then apply this idea to identify unstable or fraudulent stocks, which forms the core of our shorting strategy.

4 Shorting Strategy: Identifying Benford Outliers

Now that we have confirmed the relationship between a stock's "Benford-ness" and its volatility, we can apply Benford's Law to identify outlier stocks that deviate greatly from the ideal distribution. These large deviations may give us possible leads for further investigation, which can result in identifying suitable stocks to short sell.

4.1 Case Study: Enron Fraud

Enron was an American energy company that went bankrupt in 2001, after it was revealed that its strong financial status was the result of corporate corruption and accounting fraud. Enron operated in the utility sector, which we previously showed has a relatively large deviation from the Benford distribution.

In this analysis, we gather Enron's closing prices starting from January 1, 1998 until December 31, 2001, which is the primary period of Enron's fraudulent activities. We then extract the leading digits from each of the closing prices in these dates. These are plotted side-by-side with the Benford distribution in Figure 5.

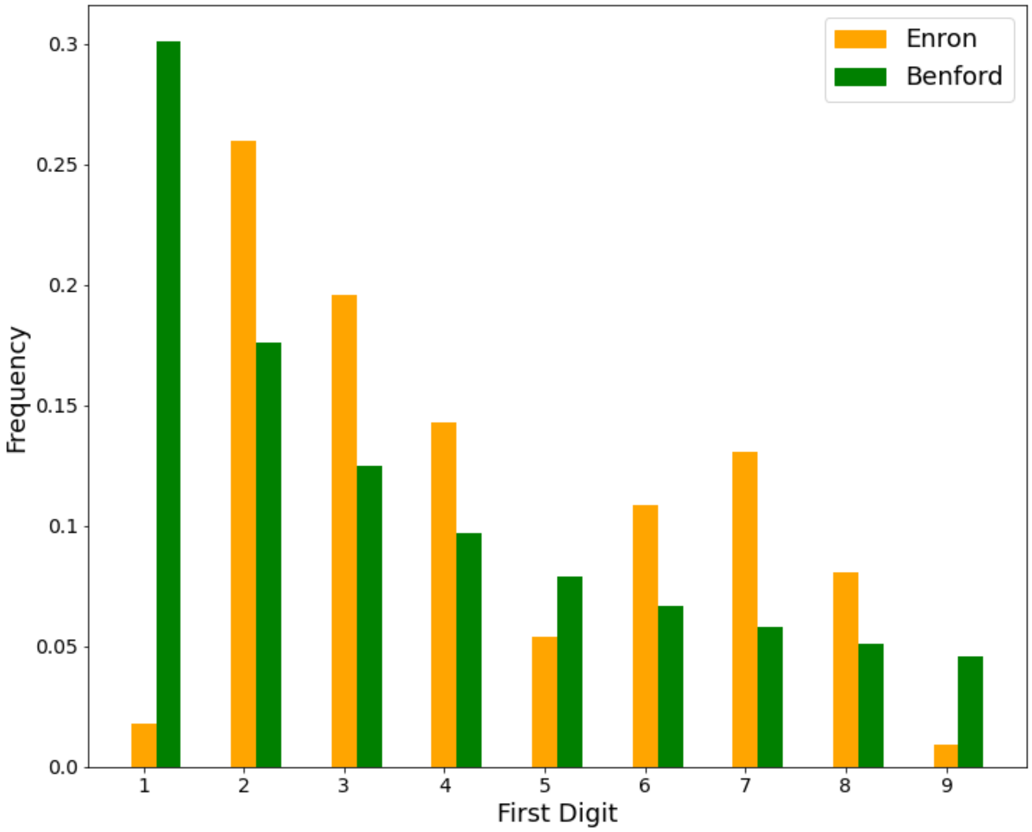

FIGURE 5: Enron Leading Digit Distribution from Jan. 1998 to Dec. 2001

Figure 5 indeed reveals a significant deviation from the ideal Benford distribution. In fact, instead of 1 being the most frequent leading digit, it was actually the second least frequent, appearing as the first digit roughly 1.8% of the time. Except for 1 and 9, most of the other digits appear as a leading digit much more often than expected.

A significant deviation from Benford numbers may possibly suggest very volatile or unstable stocks. In the case of Enron, this happens to be true. The volatility manifested itself when Enron's stock prices crashed drastically in 2001, and it eventually filed for bankruptcy. This highlights the effectiveness of our proposed shorting strategy.

4.2 Shorting Strategy

Note that as a heuristic strategy, this shorting strategy is not something that should be strictly followed. Rather, it should only serve as a starting point to give possible leads on which stocks to short, as well as the most suitable times to do so. Needless to say, further investigation needs to be carried out for every decision to complement this Benford heuristic.

With that, we propose the following shorting strategy:

- Denote d_{stock} as the measured distance between the closing price leading digit and the ideal Benford distribution, as given in the result of Algorithm 1. We use JS-Divergence for this algorithm.
- Refer to Table 2 for the sector Benford statistics. Denote μ_{sector} and σ_{sector} as the mean and standard deviation of the sector's distances to the ideal Benford distribution.
- **Short Sell Condition:** If $|d_{stock} - \mu_{sector}| > \sigma_{sector}$, consider shorting the stock, as the significant deviation from the mean may be an indicator of volatility.
- **Buy Condition:** If $|d_{stock} - \mu_{sector}| < \sigma_{sector}$ and you have previously shorted the stock, consider cashing out, as the prices may begin to stabilize and rise soon.

5 Potential Applications: Muddy Waters Short Sell Recommendations

Muddy Waters Research is a short-selling research firm that provides information on certain companies that have questionable financial statuses.

In this section, we explore short sell recommendations from the comprehensive catalog of Muddy Waters. We identify certain companies whose historical stock prices deviate significantly from the Benford distribution, and we illustrate how our proposed shorting strategy can apply these stocks.

5.1 Application: Joyy Inc. (YY US)

Joyy is a global video-based social media platform. This short sell report from Muddy Waters comes at the heels of Baidu's decision to acquire YY Live from Joyy. Muddy Waters called Joyy's strong financial standing "a mirage" and claimed that it is "90% fraudulent", citing the expansive network of bots creating an illusion of traffic in Joyy's services.

In line with this, we investigate Joyy's closing stock prices below. The dates begin on November 21, 2012 until December 18, 2020.

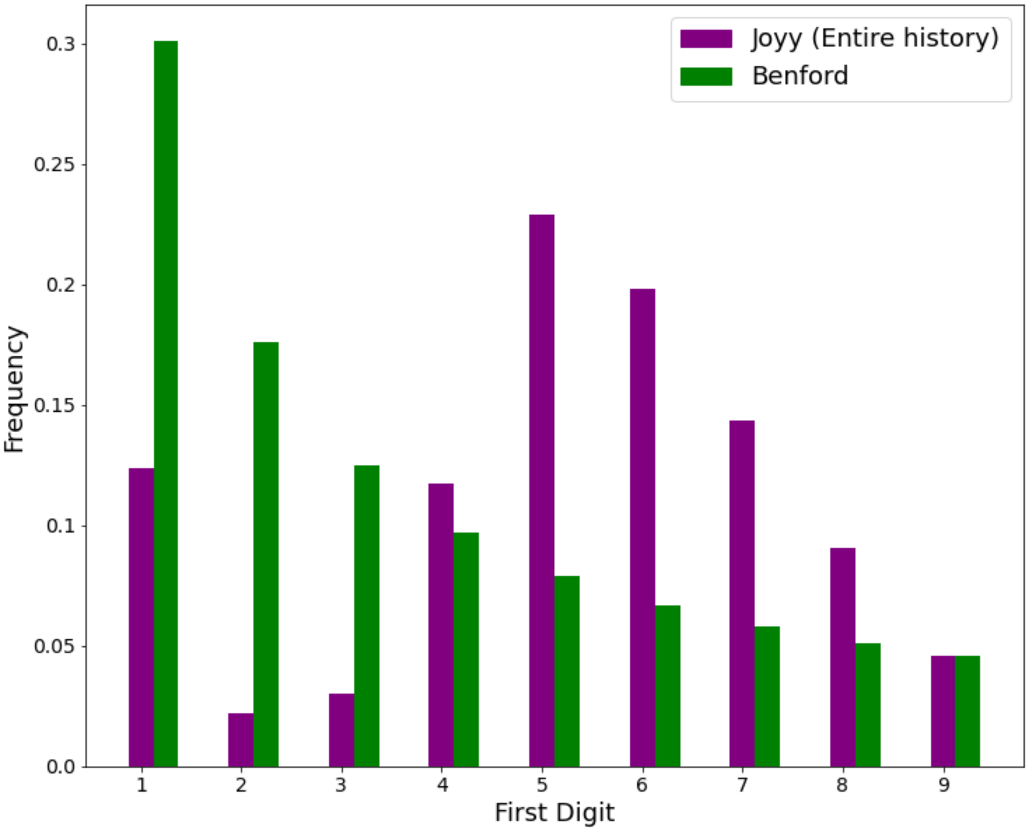

FIGURE 6: Joyy Leading Digit Distribution from **Nov. 2012 to Dec. 2020**

As suspected, Figure 6 shows a very significant deviation from the ideal Benford distribution. The leading digits 1, 2, and 3 are quite low. In comparison, the digits 5, 6, 7, and 8 appear as leading digits much more often than expected. This supports the hypothesis of Muddy Waters that there may be something unsable with Joyy.

To apply our shorting strategy, we need to further explore the history of Joyy. Consider, for example, the first digit distributions of Joyy closing prices from 2012 to 2015, as shown in Figure 7.

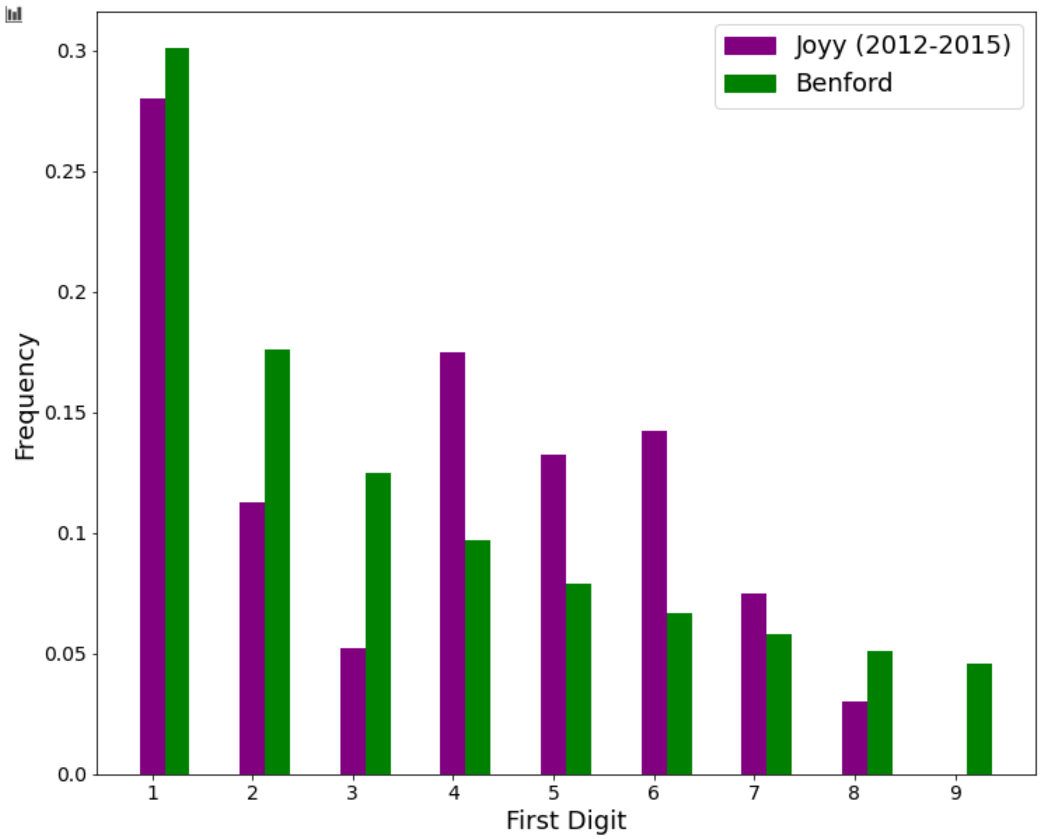

FIGURE 7: Joyy Leading Digit Distribution from **Nov. 2012 to Dec. 2015**

Surprisingly, the distribution seems much closer to the Benford distribution when considering only the years 2012 to 2015. This suggests that from 2012 to 2015, Joyy may have been a stable and strong company, but somewhere between 2016 and 2020, it may have become more unstable and/or fraudulent.

Since Joyy operates in the Information Technology sector, then in our short selling algorithm, $\mu_{sector} = 0.209$ and $\sigma_{sector} = 0.084$. In our data, Joyy's prices first exceed $\mu_{sector} + \sigma_{sector}$ on February 16, 2017, so this date would have been a suitable point to start taking a serious look at Joyy and evaluate whether it is worth shorting.

5.2 Application: eHealth Inc. (EHTH US)

eHealth Inc. (EHTH) is a health insurance company. Muddy Waters claimed that although the company tries to project a strong financial standing, the underlying business model and practices are very unprofitable. A key argument of their claim is that EHTH was burning through cash very quickly to generate unsustainable growth, as its revenues are still expected to be collected in 2029.

In Figure 8 we see the leading digit distributions of eHealth from October 2006 to December 2020. This is actually a much longer time frame than the previous companies we have explored. This also means that the results will likely be more conclusive.

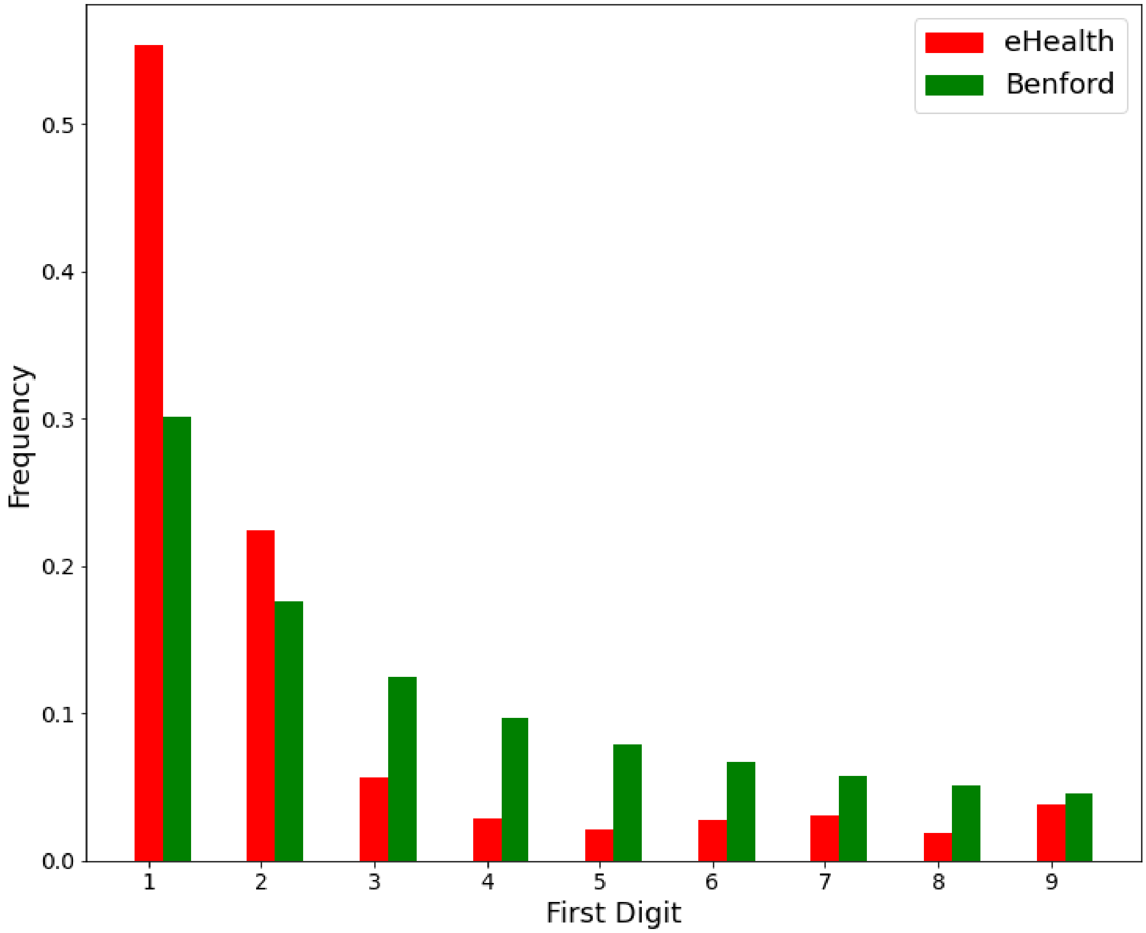

FIGURE 8: eHealth Leading Digit Distribution from **Oct. 2006 to Dec. 2020**

This is a different pattern than the ones explored before. Here, the appearance of 1 and 2 as the leading digits are actually much higher than the Benford distribution, while the rest of the digits are much lower. Surprisingly, 1 appears as the leading digit more than 50% of the time, which is very unusual and is a significant deviation from Benford's Law.

Since eHealth operates in the Health Care sector, then in our short selling algorithm, $\mu_{sector} = 0.166$ and $\sigma_{sector} = 0.077$. In our data, eHealth prices first exceed $\mu_{sector} + \sigma_{sector}$ on December 8, 2012. Note that in this case, it would not make sense to short the stock for a very long period of time, so more investigation will need to be conducted to determine the optimal times to short sell. This Benford model, however, provides a suitable lead that makes eHealth worth investigating.

6 Conclusion

In conclusion, we indeed confirm that the Benford distribution is applicable to financial data (more specifically, S&P500 stock closing prices). The short sell recommendation strategy motivated by the Benford model shows great promise in identifying potential stocks which have questionable financial statuses. Though this method is still lacking in terms of rigor and thoroughness, it nevertheless provides a valuable complementary piece to many buy/sell analyses, most especially in terms of shorting.

References

[1] Collins, J. (2017). Using Excel and Benford's Law to detect fraud. *Journal of Accountancy*.

[2] Linder, D. (2020). Enron stock price and data. *Famous Trials*.

[3] MW is short Joyy Inc. (2020). *Muddy Waters Research.*

[4] MW is short eHealth Inc. (2020). *Muddy Waters Research.*

[5] Walthoe, J. (1999). Look out for number one. *Plus Magazine*.