Programming for Data Analytics Project

FRAUD DETECTION WITH BENFORD'S LAW

M.sc in Data Analytics

2023/2024

Prepared For:

Dr Lauren Connell

Prepared By:

Peter Chukwuka Ibeabuchi

20007349

20007349@mydbs.ie

October 25, 2023

ABSTRACT

The authenticity of data is essential for effective decision-making. Fraudulent datasets can lead to incorrect conclusions and undermine the purpose of any analysis. This report focuses on the application of Benford's law to detect anomalies in a population dataset. By comparing the distribution of the data variables to the expected distribution predicted by Benford's law, we identify certain deviations that may indicate fraud.

INTRODUCTION

Fraud is one of the most critical challenges faced by data analysts today. Datasets can be intentionally altered or falsified to present values that do not represent reality. Ensuring that data is accurate is important for any analyst, as fraudulent data affect decisions, policies, and even resource allocation. Therefore, it is important to address the issue of fraud in a given data before any further critical analysis is performed.

Data analysts use a variety of fraud detection techniques, including statistical clustering and machine learning models. Benford's law is a mathematical principle that can help analysts identify anomalies in datasets. As defined by Tirthajyoti Sarkar,

"Benford's Law states that the first digits found in a data set are expected to be arranged in a way that the lowest digit, one, appears the most frequently, followed by two, three, etc. This law can be utilized to detect patterns, or lack thereof, in naturally occurring data sets, which can be used to help catch anomalies or fraud in data."

Sarkar's definition explains that Benford's law establishes a pattern in the first digits found in a dataset. This pattern can be used to detect anomalies or fraud in data, as human-forged data is likely to deviate from this pattern. Discovered by Simon Newcomb, this law states that leading ones 1s appear more times as often as leading 9s! Benford's law is also known as the First Digit Law. The graph below shows the distribution of leading digits according to Benford's law.

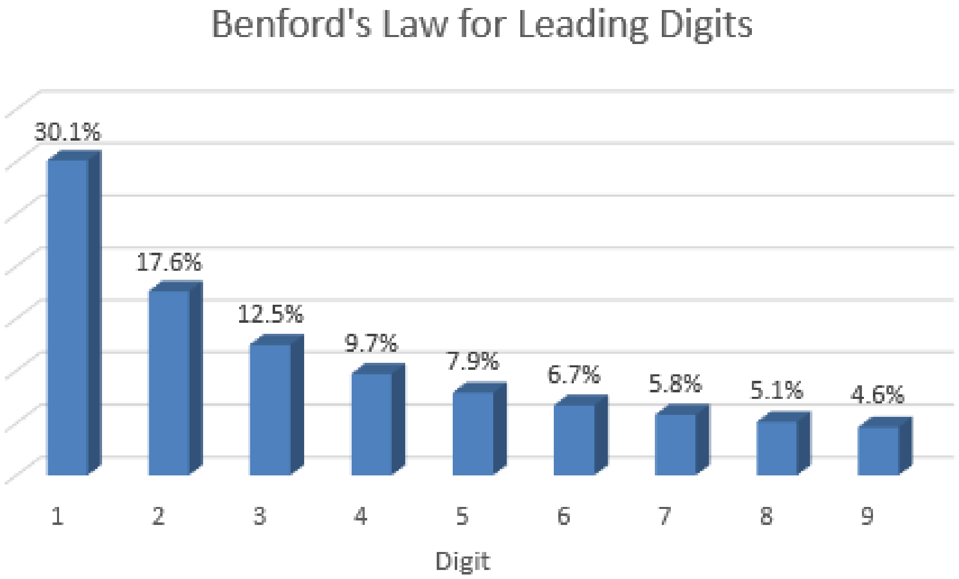


Figure 1: Jim Frost. Benford's Law for leading digits, 2022. https://statisticsbyjim.com/probability/benfords-law/

As seen in the chart above, the leading digit 1 appears the most frequently, followed by 2, 3, and so on. The probability of each leading digit decreases as the digit increases. The Benford's frequency chart can be used to compare the leading digits of a dataset to the expected Benford's distribution. If the dataset deviates significantly from the expected distribution, it may be a sign of fraud or other data manipulation.

The purpose of this report is to analyze how the first digits in five different columns of a population dataset conform to Benford's Law. We will examine the extent to which the data in these columns aligns with the expected Benford's Law percentages.

METHODOLOGY

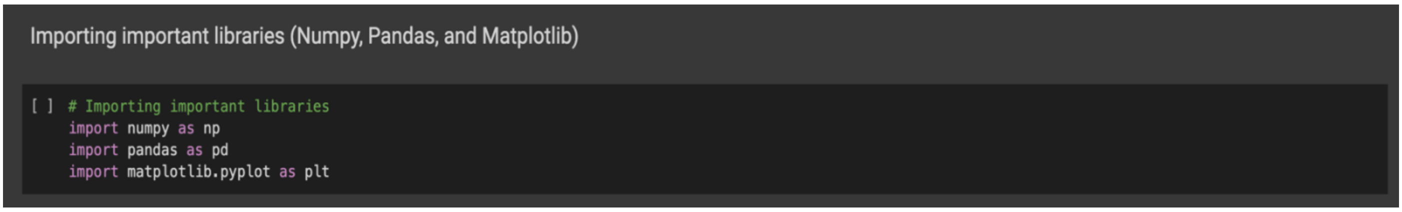
The primary language used in the analysis is the Python programming language. Our analysis includes the following process:

- **Importing Important Libraries:** Three key Python libraries are imported for the analysis: Pandas, NumPy and Matplotlib. These libraries are essential for data manipulation, numerical analysis, and visualization.
- **Loading the Dataset**: The dataset is loaded into the Pandas DataFrame.
- **First Digit Retrieval:** The Once the data is loaded into a Pandas DataFrame, the analysis focuses on retrieving the first digits of the data. This step is important for calculating leading digit percentages.
- **Calculating Leading Digit Percentages:** With the leading digits retrieved, the next step involves calculating the percentage of occurrence of each leading digit (1 through 9) within the dataset. This calculation is done for each column separately.
- **Comparison with Benford's Law:** After calculating the actual percentages of leading digits, the data is compared with the Benford's Law, which predicts the expected distribution of leading digits.
- **Identifying Deviations:** . Following the principles of Benford's law, a significant deviation from the expected distribution suggest that the column does not conform to the expected pattern, this guides the decision on the column, whether there is the presence of fraud or not. For better clarity, the Matplotlib library is used to create a chart to display the actual distribution of the leading digits of each column and how they compare to Benford's Law.

It is important to note, that the leading digits percentages as stipulated by Benford's law was hard coded into the notebook, as there are not Python libraries that makes provision for accessing them.

Importing Libraries:

As mentioned earlier, three libraries were used for the analysis, Pandas for data manipulation and analysis, NumPy for numerical analysis, and Matplotlib for visualizations.



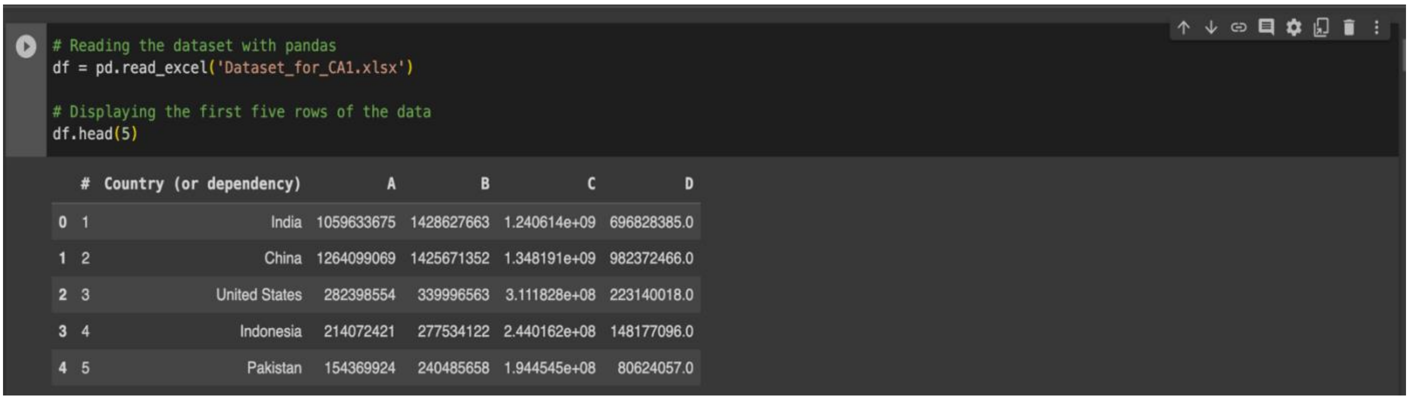
Importing important libraries (Numpy, Pandas, and Matplotlib)

```
[ ] # Importing important libraries
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
```

Figure 2: Importing Python Libraries

The Dataset

The dataset used in this analysis is provided by the course lecturer Dr Lauren Connell. The data shows the population of several countries at different intervals represented in columns A, B, C, and D. For this analysis, we do not carry out a descriptive analysis of the dataset, however, the dataset contains 5 rows, and 234 columns. There are no null values in the dataset, and it is in a clean format. To load the DataFrame, we used the `pd.read_excel()` Python function. Here is what the first five row of the looks like:



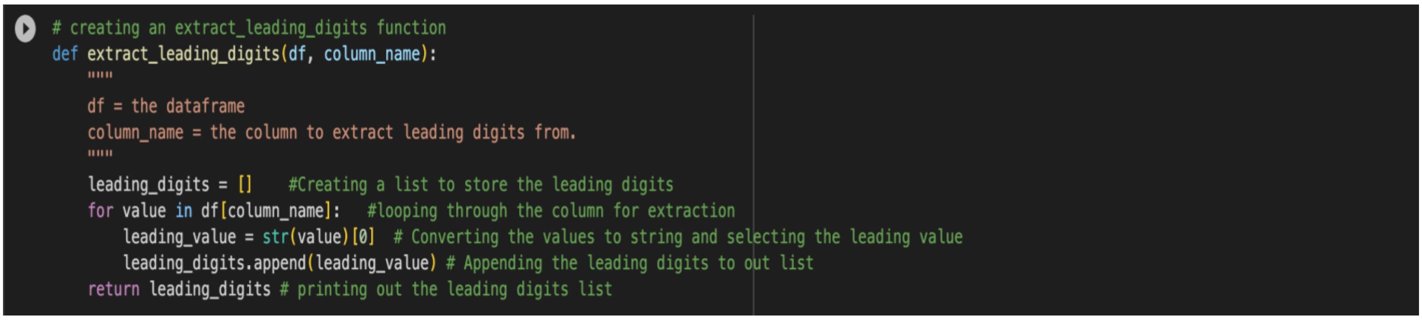
```
# Reading the dataset with pandas
df = pd.read_excel('Dataset_for_CA1.xlsx')

# Displaying the first five rows of the data
df.head(5)
```

	#	Country (or dependency)	A	B	C	D
0	1	India	1059633675	1428627663	1.240614e+09	696828385.0
1	2	China	1264099069	1425671352	1.348191e+09	982372466.0
2	3	United States	282398554	339996563	3.111828e+08	223140018.0
3	4	Indonesia	214072421	277534122	2.440162e+08	148177096.0
4	5	Pakistan	154369924	240485658	1.944545e+08	80624057.0

Figure 3 First five rows of the dataset

First Digit Retrieval:

To retrieve the leading digit of each column, we write a function that creates a list to store the leading digits. Then, it loops through the column, converting each value to a string and selecting the leading digit. Finally, the function appends the leading digit to the list.



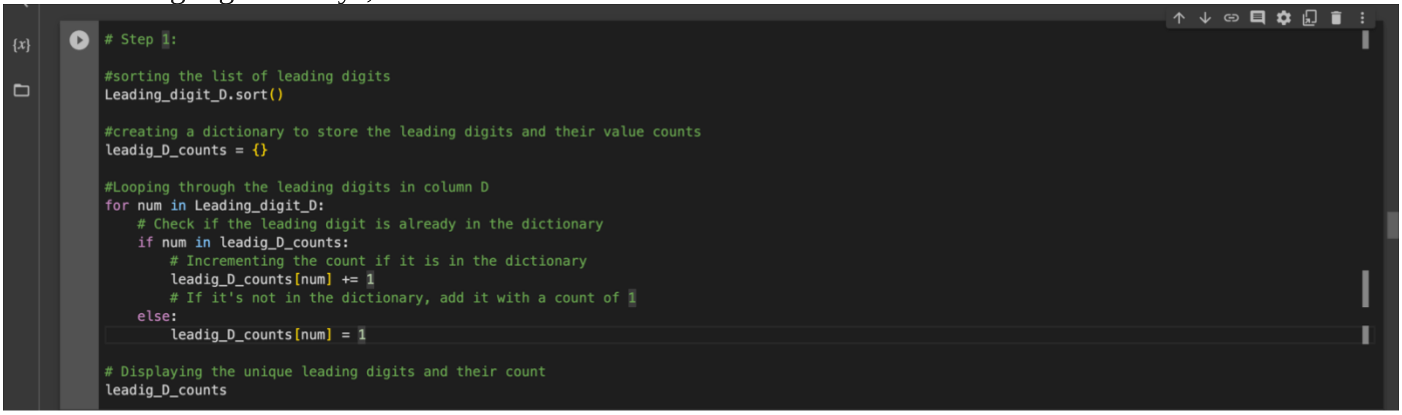
```
# creating an extract_leading_digits function
def extract_leading_digits(df, column_name):
    """
    df = the dataframe
    column_name = the column to extract leading digits from.
    """
    leading_digits = []    #Creating a list to store the leading digits
    for value in df[column_name]:   #looping through the column for extraction
        leading_value = str(value)[0]  # Converting the values to string and selecting the leading value
        leading_digits.append(leading_value) # Appending the leading digits to out list
    return leading_digits # printing out the leading digits list
```

Figure 4: Leading digits function

Leading Digit Percentage:

After retrieving the leading digit, the values are converted from the string data type to the integer data type, to allow for easy analysis. The process of retrieving the leading digit percentage follows two steps:

1. **Create a dictionary to store the leading digits and their value counts**: First we sort the list of leading digits. This makes it easier to identify the unique leading digits and their counts. Once the list of leading digits is sorted, a dictionary is created to store the leading digits and their value counts. The dictionary has a key for each unique leading digit and a value for the count of that leading digit. We loop through the list of leading digits and increment the count of each digit in the dictionary when they appear on the list. At the end of this process, we have a dictionary containing the leading digits as keys, and their count as values.



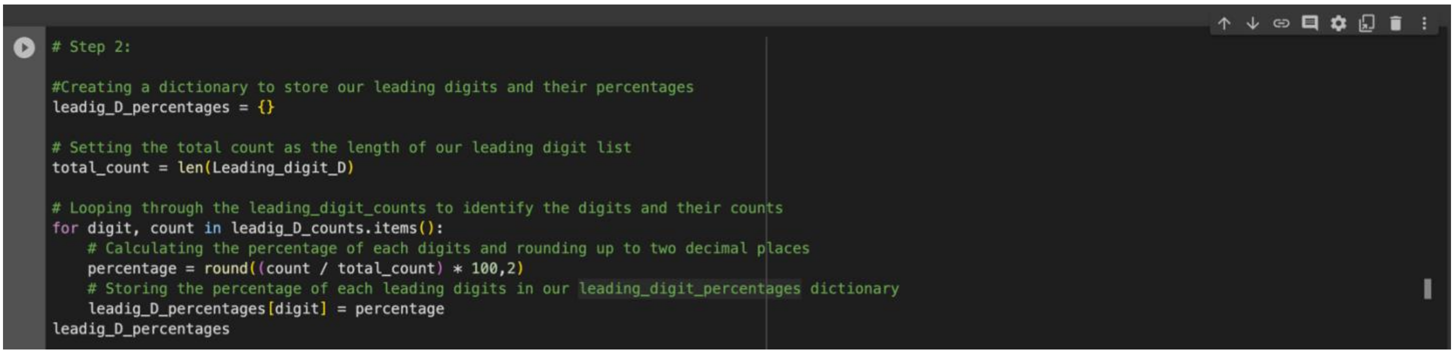
```
# Step 1:

#sorting the list of leading digits
Leading_digit_D.sort()

#creating a dictionary to store the leading digits and their value counts
leadig_D_counts = {}

#Looping through the leading digits in column D
for num in Leading_digit_D:
    # Check if the leading digit is already in the dictionary
    if num in leadig_D_counts:
        # Incrementing the count if it is in the dictionary
        leadig_D_counts[num] += 1
        # If it's not in the dictionary, add it with a count of 1
    else:
        leadig_D_counts[num] = 1

# Displaying the unique leading digits and their count
leadig_D_counts
```

Figure 5: Leading digit counts

2. **Create a dictionary to store the leading digits and their percentages:** In this step, we create a dictionary to store the leading digits and their percentages. We first find the length of the column, which is also the total number of rows in the dataset. We then iterate through the leading digit count dictionary (from step one) and calculate their percentages using the following formula:

   Percentage = (count/sum of counts) * 100

   Once the percentages of each digit are calculated, we assign them as values to the leading digits in our percentage dictionary.

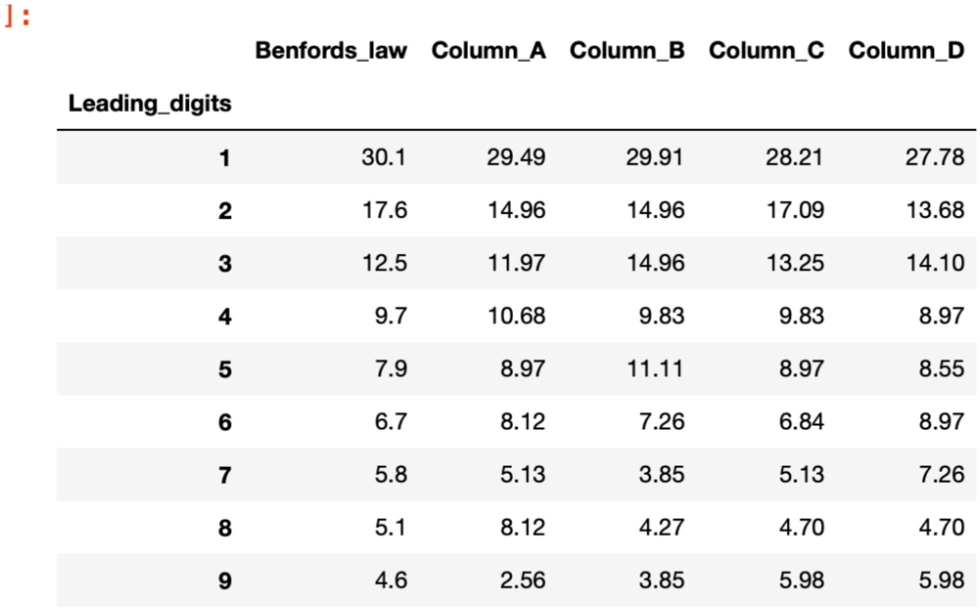


```
# Step 2:

#Creating a dictionary to store our leading digits and their percentages
leadig_D_percentages = {}

# Setting the total count as the length of our leading digit list
total_count = len(Leading_digit_D)

# Looping through the leading_digit_counts to identify the digits and their counts
for digit, count in leadig_D_counts.items():
    # Calculating the percentage of each digits and rounding up to two decimal places
    percentage = round((count / total_count) * 100,2)
    # Storing the percentage of each leading digits in our leading_digit_percentages dictionary
    leadig_D_percentages[digit] = percentage
leadig_D_percentages
```

Figure 6: Leading digits percentage

Comparison with Benford's Law:

Having identified the leading digit percentage in each of the columns, we compare them to that of the expected percentages as stipulated by Benford’s law. First, we confirm that the percentage distribution or each column follows the Benfords pattern, then we identify how each leading digit aligns or deviate from the expected percentages.



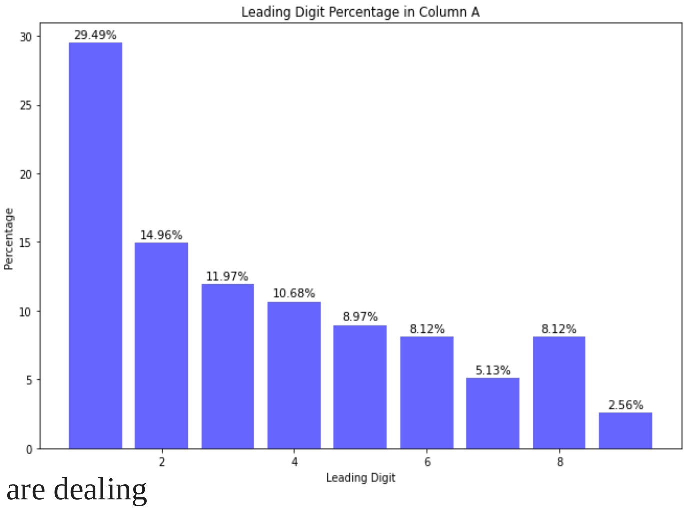
]:

	Benfords_law	Column_A	Column_B	Column_C	Column_D
Leading_digits					
1	30.1	29.49	29.91	28.21	27.78
2	17.6	14.96	14.96	17.09	13.68
3	12.5	11.97	14.96	13.25	14.10
4	9.7	10.68	9.83	9.83	8.97
5	7.9	8.97	11.11	8.97	8.55
6	6.7	8.12	7.26	6.84	8.97
7	5.8	5.13	3.85	5.13	7.26
8	5.1	8.12	4.27	4.70	4.70
9	4.6	2.56	3.85	5.98	5.98

Figure 7: Leading Digits Percentages of all Columns

Comparing Leading Digit Distribution in Column A

The leading digits percentage in column A shows pattern quite similar to that of Benford's law, with the highest highest percentage of occurrences being digit 1 (29.49%), followed by a continuous decline to 9 (2.56%). However, there is an anomaly in digit 8, which has a higher percentage (8.12%) than digit 7 (5.13%).



This is quite a deviation from the Benford's pattern, and it could imply several things, one being that the population data given in column A could have been manipulated in some way, and thus fraudulent. There is also the fact that we are dealing with a population dataset, and it is possible that there are simply more countries with population sizes that start with the digit 8 than there are countries with population sizes that start with the digit 7. In any case, this calls for further investigation of the dataset and more analysis.

While we looked at the pattern of the percentage distribution of column A, we are also concerned with how much each digit in the column deviates or is in line with the actual percentages by the Benfords law.

Leading Digits	Benford's %	Column A %	Deviation (%)
1	30.1%	29.49%	0.61%
2	17.6%	14.96%	2.64%
3	12.5%	11.97%	0.53%
4	9.7%	10.68%	-0.98%
5	7.9%	8.97%	-1.07%
6	6.7%	8.12%	-1.42%
7	5.8%	5.13%	-0.67%
8	5.1%	8.12%	-3.02%
9	4.6%	2.56%	2.04%

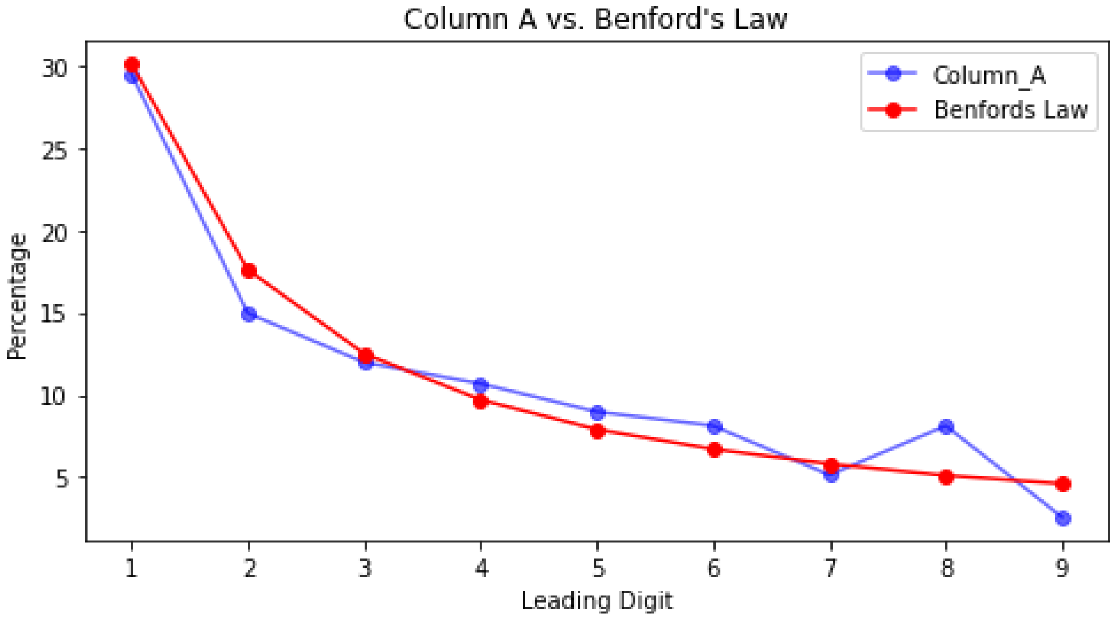


Figure 8 Benford's percentages vs Column B percentages

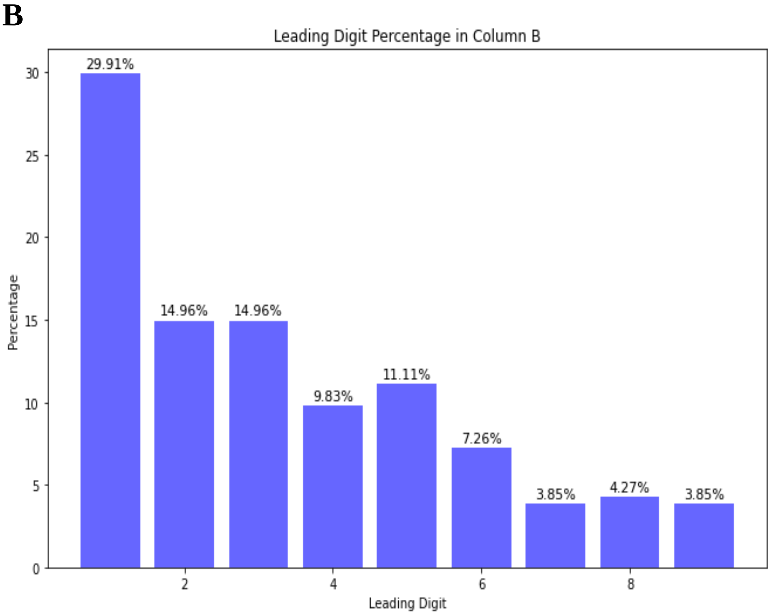
From the figure above we can see that there is a deviation in every leading digit in column A from that of Benford’s Law. Overall, our analysis reveals the following key findings for columns A:

- For digits 1, 3, 4, and 7, the differences between the expected Benford's percentages and the percentages in Column A are relatively small, ranging from -0.98% to 0.61%. These differences are generally within a reasonable margin of error and may not be indicative of fraud. Small variations are common in real-world data.
- For digits 2, 5,6, 8, and 9 the differences are more significant, ranging from 1.07% to -3.02%. In particular, digits 8 show notable deviations from Benford's Law.These deviations might warrant further investigation.

Comparing Leading Digit Distribution in Column B

Unlike column A, the leading digit percentages in column B show a complete deviation from Benford's law. Digits 2 and 3 appear with the same frequency (14.96%), while digit 5 appears more frequently (11.11%) than digit 4 (9.83%). Similarly, digit 8 appears more frequently (4.27%) than digit 7 (3.85%).

The deviation in the pattern from the Benford’s pattern suggest that Column B may not be a true reflection of the population dataset. While we recognize that we are dealing with a population dataset and some countries are more populous than others, this column exhibits a significant number of anomalies, suggesting fraud.



For better comparison, we also looked at the exact percent deviation of each digit in column B from the specified percentages by Benford.

Leading Digits	Benford’s %	Column B %	Deviation (%)
1	30.1%	29.91%	0.19%
2	17.6%	14.96%	2.64%
3	12.5%	14.96%	-2.46%
4	9.7%	9.83%	-0.13%
5	7.9%	11.11%	-3.21%
6	6.7%	7.26%	-0.56%
7	5.8%	3.85%	-1.95%
8	5.1%	4.27%	-0.83%
9	4.6%	3.85%	0.75%

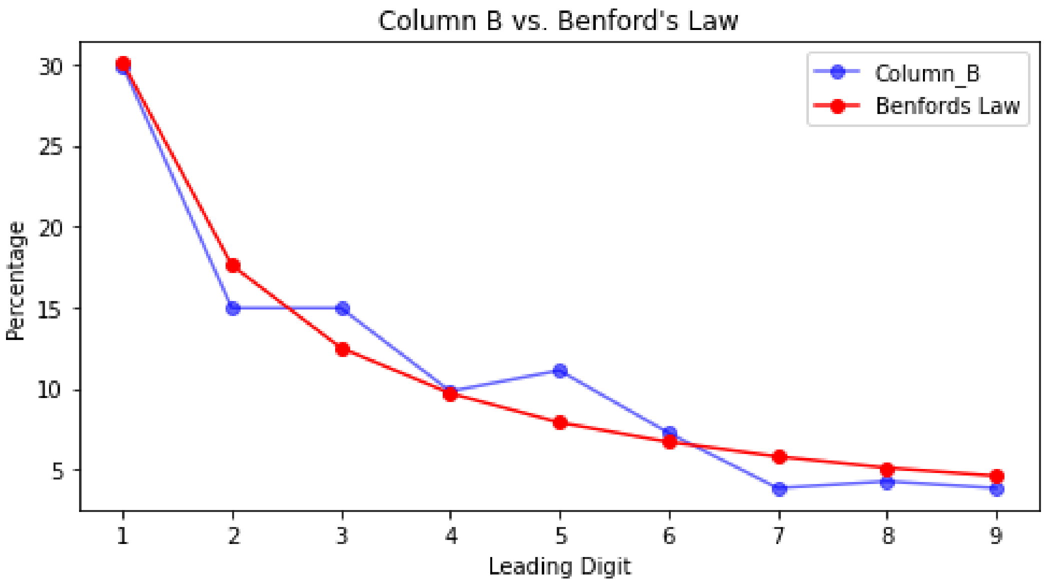


Figure 9: Benfords percentage vs Column B percentage

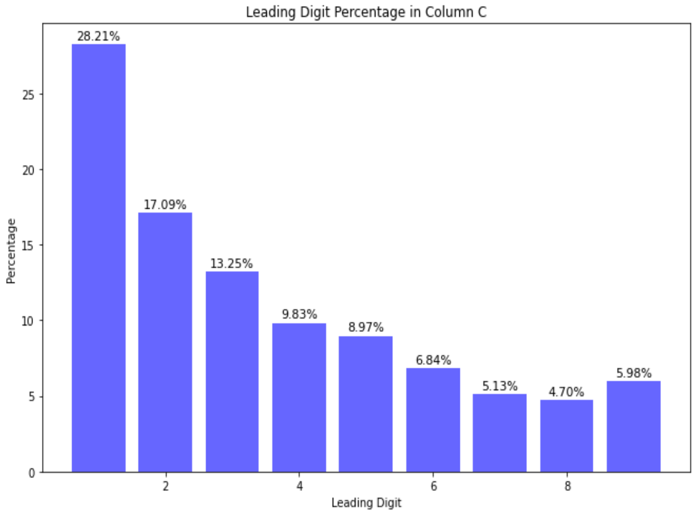
Here are our findings in columns B:

- For digits 1, 4, 6, 8, and 9, the differences between the expected Benford's percentages and the percentages in Column A are relatively small, ranging from -0.83% to 0.75%. These differences are within a reasonable margin of error and do not suggest significant deviations from Benford's Law.
- For digits 2,3, 5, and 7the differences are more significant, with variations of -2.46% and -3.21%, respectively. These deviations are relatively large and may warrant further investigation.

Comparing Leading Digit Distribution in Column C

Just as we see in column A, the leading digit percentage distribution follows closely with that of Benford’s pattern. The frequency has a continuous decline from digit 1(28.21%) to digit 9(5.98%). However, digit 9 deviate from this rule as it has a higher occurrence (5.98%) compared to digit 8(4.70%).

As we have seen in subsequent analysis of other columns, it is not enough that the columns follow the Benford’s pattern. It is important to know how much each digit actually deviate from the stipulated Benford’s percentages.



Leading Digits	Benford’s %	Column C %	Deviation (%)
1	30.1%	28.21%	1.89%
2	17.6%	17.09%	0.51%
3	12.5%	13.25%	-0.75%
4	9.7%	9.83%	-0.13%
5	7.9%	8.97%	-1.07%
6	6.7%	6.84%	-0.14%
7	5.8%	5.13%	-0.67%
8	5.1%	4.70%	-0.4%
9	4.6%	5.98%	-1.38%

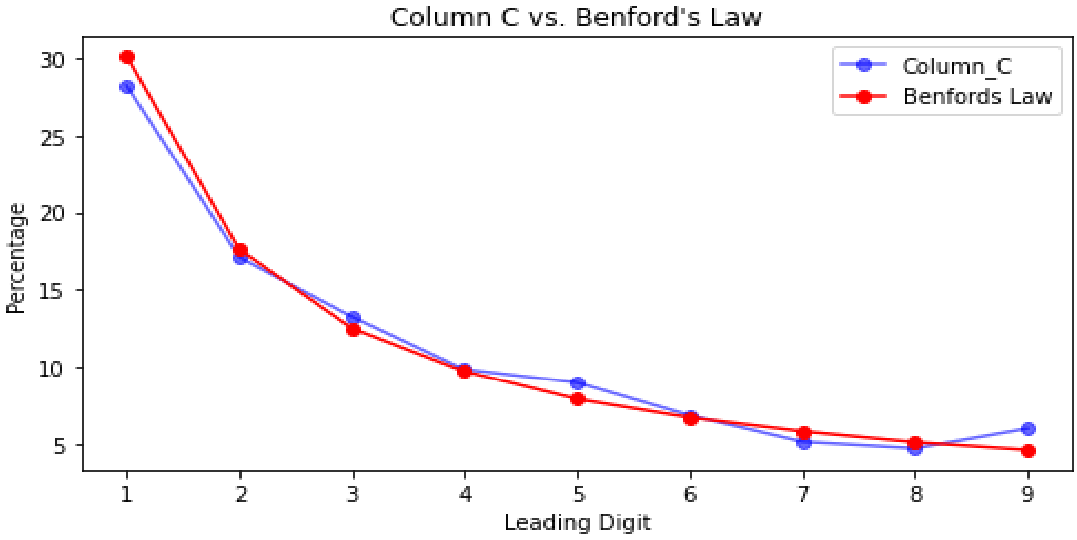


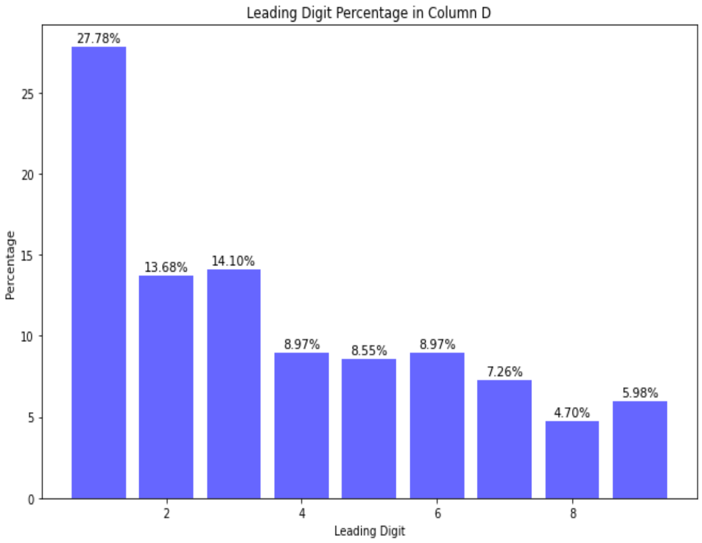
Figure 9: Benford's percentages vs Column C percentages

From the table above, we can see that so far, column C has conformed better to the Benford’s percentage than the previously analyzed columns. All but three digits (1, 5, and 9) deviate less than 1% from the expected Benford's percentages.

Comparing Leading Digit Distribution in Column D

The leading digit percentages in column D follow a similar pattern to those in column B, showing a complete deviation from Benford's law. The frequencies do not decrease throughout the chart, with digit 3 (14.10%) occurring more frequently than digit 2 (13.68%), and digit 6 (8.97%) occurring more frequently than digit 5 (8.55%).

Following Benford’s law, these deviations does suggest that the values in column D are not a true reflection of the countries in the data set.



While we have observed the pattern of the percentage distribution in the column, we also look at the actual deviation for each digit from the expected percentage by Benfords law.

Leading Digits	Benford’s %	Column D %	Deviation (%)
1	30.1%	27.78%	2.32%
2	17.6%	13.68%	3.92%
3	12.5%	14.10%	-1.6%
4	9.7%	8.97%	-0.73%
5	7.9%	8.55%	-0.65%
6	6.7%	8.97%	-2.27%
7	5.8%	7.26%	-1.46%
8	5.1%	4.70%	-0.4%
9	4.6%	5.98%	-1.38%

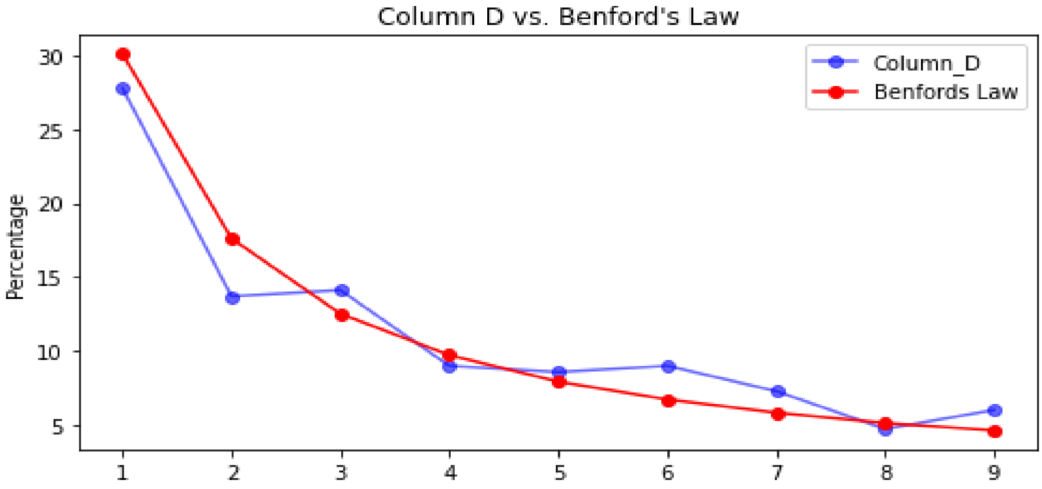


Figure 10: Benfords percentages vs ColumnD percentages

From the figure above, we see that all but three digits (4,5, and 8) in column D deviate more than 1% from the expected Benfords percentage, with digit 2 showing the highest level of deviation 3.92%. Following the principles of Benfords law, this column shows no randomness and may possibly have been manipulated.

Conclusion

Haven compared all columns with the percentages and principles stipulated by the Benfords law, we see that all columns in the dataset has shown certain levels of deviation, with column B and D having the most deviations. This suggest that the values in the dataset may not be true representation of the populations of the countries and are possibly manipulated. Notably, column C performed better than the rest of the columns and conform most to the expected distribution.

It is important to emphasize that deviations from Benford’s Law do not by themselves provide conclusive evidence of fraud or irregularities. However, for the purpose of this assessment, and following strictly the rules laid down by the Benfords law, **I conclude that all the columns are fraudulent.**

References

1. Benford's law. (2023)Wikipedia.Available at: https://en.wikipedia.org/wiki/Benford%27s_law.(Accessed: 22 October 2023).
2. Jim Frost (2022) Benford’s Law Explained with Examples. Available at: https://statisticsbyjim.com/probability/benfords-law/ (Accessed: 22 October 2023).
3. Tirthojyoti Sarker(2023) What is Benford’s Law and Why is it Important. Available at: https://builtin.com/data-science/benfords-law .(Accessed: 22 October 2023).