Project 2

Luka Tumbas

We will first load and look at the data needed.

```
#company_financials <- read.csv("C:/Users/lukat/Downloads/company_financials.csv")
company_financials <- read.csv("company_financials.csv")

head(company_financials,10)
```

```
   CompanyID  Revenue Expenditure
1       A001 29674.06     5972.18
2       A002 20272.92     4962.98
3       A003 32487.58     1000.00
4       A004 54930.16     3025.12
5       A005 19139.47     3819.06
6       A006 19139.66     4260.30
7       A007 56813.41      884.40
8       A008 34907.63     5011.31
9       A009 16619.22     6416.31
10      A010 30501.65     4299.36
```

Now look at the structure of the object.

```
str(company_financials)
```

```
'data.frame':   200 obs. of  3 variables:
 $ CompanyID  : chr  "A001" "A002" "A003" "A004" ...
 $ Revenue    : num  29674 20273 32488 54930 19139 ...
 $ Expenditure: num  5972 4963 1000 3025 3819 ...
```

Let’s now summarize some of this data to get a feel for the revenue and expenditure numbers in the dataset.

```r
Revenue_summary <- summary(company_financials$Revenue)
Expenditure_summary <- summary(company_financials$Expenditure)
print("Revenue Summary")
```

```
[1] "Revenue Summary"
```

```r
head(Revenue_summary)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
  4574.04  14428.03  21971.48  25199.16  29748.11 112656.82 
```

```r
print("Expenditure Summary")
```

```
[1] "Expenditure Summary"
```

```r
head(Expenditure_summary)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
  426.340  1999.457  3145.320  3783.820  4756.870 30080.700 
```

We are now ready to analyze if the dataset conforms to Benfords law. Let us do that for Revenue first. We will extract the revenue data and filter out any NA values.

```r
Revenue <- company_financials$Revenue #Extract the revenue

Valid_revenue <- Revenue[!is.na(Revenue)] #Filter out 0 and NA revenue values
```

Now let us find the leading digits for the revenue.

```r
revenue_leading_digits <- as.numeric(substr(as.character(Valid_revenue), 1, 1))
```

We will do the same thing for the expenditures now.

```r
Expenditure <- company_financials$Expenditure #Extract the expenditure

Valid_expenditure <- Expenditure[!is.na(Expenditure)] #Filter out 0 and NA expenditure values

expenditure_leading_digits <- as.numeric(substr(as.character(Valid_expenditure), 1, 1))
```

Now, let us compute the frequencies with which each digit occurs within revenue and expenditures, and crate a data frame to see how those frequencies compare with what is expected of them according to Benford's law.

```
# First let us tabulate empirical frequencies
emp_freq_revenue <- table(revenue_leading_digits) / length(revenue_leading_digits)
emp_freq_expenditure <- table(expenditure_leading_digits) / length(expenditure_leading_digit

# Create a data frame with empirical and Benford probabilities
Benford <- data.frame(
  Digit = 1:9,
  Empirical_Freq_Revenue = as.numeric(emp_freq_revenue[1:9]),
  Empirical_Freq_Expenditure = as.numeric(emp_freq_expenditure[1:9]),
  Benford_Prob = log10(1 + 1 / (1:9)))
```

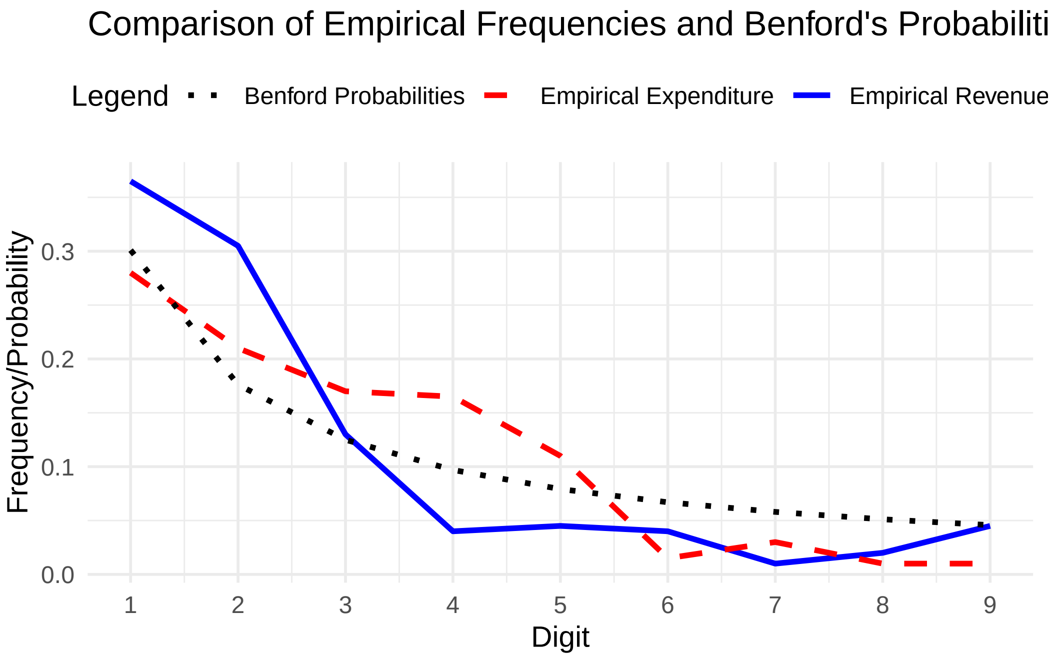
Finally, to see the Benford data frame we just created, we will import the knitr package and call the table

```
library(knitr)
knitr::kable(Benford)
```

Digit	Empirical_Freq_Revenue	Empirical_Freq_Expenditure	Benford_Prob
1	0.365	0.280	0.3010300
2	0.305	0.210	0.1760913
3	0.130	0.170	0.1249387
4	0.040	0.165	0.0969100
5	0.045	0.110	0.0791812
6	0.040	0.015	0.0669468
7	0.010	0.030	0.0579919
8	0.020	0.010	0.0511525
9	0.045	0.010	0.0457575

We can also plot the data.

```
library(ggplot2)
ggplot(Benford, aes(x = Digit)) +
  geom_line(aes(y = Empirical_Freq_Revenue, color = "Empirical Revenue"), linewidth = 1) +
  geom_line(aes(y = Empirical_Freq_Expenditure, color = "Empirical Expenditure"), linewidth =
  geom_line(aes(y = Benford_Prob, color = "Benford Probabilities"), linewidth = 1, linetype =
  labs(
    title = "Comparison of Empirical Frequencies and Benford's Probabilities",
    x = "Digit",
    y = "Frequency/Probability",
    color = "Legend"
  ) +
  scale_color_manual(values = c("Empirical Revenue" = "blue", "Empirical Expenditure" = "red"
  scale_x_continuous(breaks = 1:9) +
  theme_minimal() +
  theme(legend.position = "top")
```



Comments and conclusions:

As we can see, both from the table and the plot, both the revenue and expenditures do not exactly follow Benford's law. This could be due to some of the following reasons:

- The dataset has been tampered with and numbers have been altered to, perhaps, show, more favorable business results (slightly increased revenues, decreased expenditures). Since we do not know anything about the industry in which these companies do business, we may only hint at such a possibility. We can also see (through summary statistics , but also by glancing at individual companies) that the profit margins are quite high. This is also dependent on the industry type, but profit margins of >80% are very rare to be seen across an industry. Thus, we may conclude that the dataset has been tampered with.

- This dataset may represent only the most successful companies (in terms of profit margins) across several industries, so it is not a complete dataset. Thus, it will not follow Benford's law, as there are obviously several cutoff points if that is the case.
- The dataset might simply be too small to follow Benford's law. When we compared the previous dataset (aapl_prices), we had about 9000 observations. In our case, we only have 200, which can be considered too small for it to be used to compute relative frequeincies to measure probability.

  In conclusion, we can conclude that the dataset does not follow Benford's law, but we are unsure if the data has been tampered with or if the dataset is simply too small to compute accurate realtive frequencies. Had we had a larger dataset both in amount and quality of data (industry, dates, company names, etc.), we might have been able to give a more accurate interpretation on if the dataset behaves in accordance with Benford's law.

Comments MS:

Your submission is excellent, and these suggestions aim to refine and expand your already strong analysis.

- Your visual analysis is clear and well-executed. To really quantify conformity of distributions would need more formal statistical testing though, something we did not discuss in this course. In case you are interested what could be done further here you might consider applying a **Chi-Square Goodness-of-Fit Test**. This test would quantify how well the observed distributions conform to Benford's Law and provide statistical evidence for or against conformity. You could calculate this for both revenue and expenditure to compare them systematically.
- You did a good job preparing the data, but it would be helpful to explicitly document your steps when filtering invalid entries (e.g., removing `NA`s or non-positive values). This not only improves reproducibility but also clarifies how the data cleaning process might influence your findings.
- While visual comparisons are a great starting point, including a table of absolute or relative differences between observed and theoretical frequencies for each leading digit would make your analysis more concrete. For example:

  ```
  Digit | Observed Frequency | Expected Frequency | Absolute Difference
  --------------------------------------------------------------
    1   |         0.31          |          0.30           |        0.01
    2   |         0.18          |          0.18           |        0.00
  ```

This helps pinpoint specific digits that deviate most and could indicate anomalies.

- Your use of `ggplot2` for visualization is excellent. The clean syntax and flexibility of `ggplot` make it an ideal choice for projects like this. Consider adding annotations or labels to your plots to highlight key insights, such as the leading digit with the largest deviation.
- You noted potential issues in the expenditure data, which is a strong observation. It might be useful to investigate specific digits (e.g., those with the largest deviations) or subsets of the data. For example:
- Are there particular ranges of expenditure that deviate more?
- Could rounding practices or reporting thresholds explain some of the anomalies?
- Your analysis would benefit from a brief discussion of what deviations might imply in real-world terms. For instance:
- Are the deviations in expenditure suggestive of anomalies, such as fraud or manipulation?
- Could legitimate practices (e.g., rounding, aggregation) explain the observed patterns?
- Connecting your findings to the purpose of financial forensics would give your conclusions greater depth and relevance.
- It's great that you analyzed revenue and expenditure separately. Adding a direct comparison—both visually and numerically—could provide further insights. For example, which data set conforms more closely to Benford's Law, and why might this be the case?
- Try to avoid strong wording and premature conclusions such as "data have been tampered with". While the analysis gives you a weak signal of something going on it is not yet clear evidence of tampering of any sorts.
- Consider adding a brief explanation for each step in your code. For instance, when creating plots or filtering data, a short comment or markdown note could clarify your intent.
- If possible, identify specific entries or companies that contribute most to the deviations in expenditure. Highlighting these can provide a starting point for deeper investigation.
- A plot showing the absolute differences between observed and theoretical frequencies for each leading digit could complement your existing bar plots.