Homework 8

Part 1: Description of Benford's Law

What is Benford's Law?

Benford's law is an observation about the frequency distribution of the first digits in many real-life sets of numerical data. The law states that in many naturally occurring collections of numbers, the leading significant digit is likely to be small. For example, in sets that obey the law, the number 1 appears as the leading significant digit about 30% of the time, while 9 appears as the leading significant digit less than 5% of the time.

The distribution curve of the first digits, $f(x)$, is defined as:

$$f(x) = \frac{1}{x \ln 10}$$

The probability that a first digit is Y, $P(Y)$, is defined as:

$$P(Y) = \log_{10}(Y+1) - \log_{10} Y$$

Following the $P(Y)$ function, the distribution of the first non-zero digits in real-life numerical data sets can be represented in the table below:

Distribution of First Digits Under Benford's Law

First digit	Probability
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

Why is it a potential fraud indicator?

In some real-life non-fraud numerical datasets, the distribution of the first digits obeys the nonintuitive Benford's law. If a potential fraudster is making up a large amount of numbers intuitively (e.g. assumes that the first digits are uniformly distributed), this action will disturb the natural layout of the first digits and gives an anomalous distribution that is different from what Benford's Law would expect. Thus, the more different the first digit distribution is from the Benford's Law distribution, the more anomalous is the group of records.

What did I do?

- **Data Cleaning**: removed all *Fedex* transactions. Because the number of *Fedex* transaction is large and most of the amount of these transaction starts with the same digits (i.e. violates the Benford's Law), these records would be identified as anomaly even though they are not. Thus, the existence of these records would disturb the anomaly detection process under Benford's Law and should be removed.
- **Calculate the N(low)/N(high) ratio under Benford's Law**:
  To quantify how different the first digit distribution from the Benford's Law distribution, the first digits are binned into groups. As binning these digits into 9 groups will results in low or zero number of records for the entities (Merch number/Card number), the first digits are instead binned into 2 groups. The low bin includes records with first digit being 1 or 2, and the high bin includes records with first digit being 3,4,5,6,7,8 and 9.
  Under Benfold's Law, the N(low)/N(high) ratio can be calculated as 52.3%/47.7% = 1.096.
- **Extract the first non-zero digit of amount**. If one of the amounts is under 1, eg. 0.02, then the first non-zero digit would be 2 instead of 0.
- **Group transactions separately by the two entities (Cardnum & Merchnum)**
- **For each resulting group**, look at the distribution of the first digit of the purchase amount and **calculate U***, the measure for unusualness. The calculation process of U* is explained below
  - For each group, count the number of first digits benning with either 1 or 2, i.e. n_{low}. Then $n_{\text{high}} = n - n_{\text{low}}$. if either of n_{low} or n_{high} is zero, set it to 1 to avoid dividing by zero.
  - $R = \frac{1.096 \times n_{\text{low}}}{n_{\text{high}}}$ should be about 1, so a measure of unusualness could be U = max(R, 1/R).
  - But sometimes the number of samples in each group is too small that the measure is not trustworthy. Thus, a better measure is a smoothed U*. The formula of U* is defined below (with c=3 and n_{mid} =15)

    $$U^* = 1 + \left(\frac{U-1}{1+\exp^{-t}}\right) \quad t = (n - n_{mid})/c$$

- **Sort the two table by U* and submit the top 40**.

Part 2: Result Table

Top 40 Cardnum (potential fraud based on Benford's law)

Cardnum	n_low	n_high	R	1/R	U	n	t	U*
5142253356	61	5	13.37	0.07	13.37	66	17.00	13.37
5142299705	25	3	9.13	0.11	9.13	28	4.33	9.03
5142197563	15	134	0.12	8.15	8.15	149	44.67	8.15
5142194617	5	33	0.17	6.02	6.02	38	7.67	6.02
5142288241	1	13	0.08	11.86	11.86	14	-0.33	5.53
5142239140	16	3	5.85	0.17	5.85	19	1.33	4.83
5142144931	6	29	0.23	4.41	4.41	35	6.67	4.41
5142192606	13	2	7.12	0.14	7.12	15	0.00	4.06
5142204384	199	54	4.04	0.25	4.04	253	79.33	4.04
5142284940	21	6	3.84	0.26	3.84	27	4.00	3.78
5142189113	6	24	0.27	3.65	3.65	30	5.00	3.63
5142225308	4	17	0.26	3.88	3.88	21	2.00	3.53
5142116864	58	18	3.53	0.28	3.53	76	20.33	3.53
5142293257	2	13	0.17	5.93	5.93	15	0.00	3.47
5142173286	2	13	0.17	5.93	5.93	15	0.00	3.47
5142246929	79	25	3.46	0.29	3.46	104	29.67	3.46
5142224699	7	25	0.31	3.26	3.26	32	5.67	3.25
5142847398	10	35	0.31	3.19	3.19	45	10.00	3.19
5142273608	6	21	0.31	3.19	3.19	27	4.00	3.15
5142147267	22	76	0.32	3.15	3.15	98	27.67	3.15
5142224769	15	5	3.29	0.30	3.29	20	1.67	2.92
5142242241	16	51	0.34	2.91	2.91	67	17.33	2.91
5142260984	265	101	2.88	0.35	2.88	366	117.00	2.88
5142113192	2	12	0.18	5.47	5.47	14	-0.33	2.87
5142126842	38	15	2.78	0.36	2.78	53	12.67	2.78
5142191416	18	7	2.82	0.35	2.82	25	3.33	2.76
5142194228	11	2	6.03	0.17	6.03	13	-0.67	2.71
5142308889	11	2	6.03	0.17	6.03	13	-0.67	2.71
5142195887	12	3	4.38	0.23	4.38	15	0.00	2.69
5142212038	12	3	4.38	0.23	4.38	15	0.00	2.69
5142225184	27	11	2.69	0.37	2.69	38	7.67	2.69
5142257356	142	58	2.68	0.37	2.68	200	61.67	2.68
5142216493	14	5	3.07	0.33	3.07	19	1.33	2.64
5142239106	8	23	0.38	2.62	2.62	31	5.33	2.62
5142144593	4	14	0.31	3.19	3.19	18	1.00	2.60
5142117315	7	20	0.38	2.61	2.61	27	4.00	2.58
5142218798	21	9	2.56	0.39	2.56	30	5.00	2.55
5142180432	58	25	2.54	0.39	2.54	83	22.67	2.54
5142264155	27	12	2.47	0.41	2.47	39	8.00	2.47
5142294614	5	15	0.37	2.74	2.74	20	1.67	2.46

Top 40 Merchnum (potential fraud based on Benford's law)

Merchnum	n_low	n_high	R	1/R	U	n	t	U*
991808369338	1	181	0.01	165.15	165.15	181	55.33	165.15
8078200641472	59	1	64.66	0.02	64.66	60	15.00	64.66
308904389335	1	53	0.02	48.36	48.36	53	12.67	48.36
3523000628102	34	1	37.26	0.03	37.26	34	6.33	37.20
808998385332	1	36	0.03	32.85	32.85	37	7.33	32.83
55158027	27	1	29.59	0.03	29.59	28	4.33	29.22
8916500620062	1	31	0.04	28.28	28.28	31	5.33	28.15
3910694900001	25	1	27.40	0.04	27.40	26	3.67	26.74
8889817332	24	1	26.30	0.04	26.30	25	3.33	25.43
881145544	24	1	26.30	0.04	26.30	24	3.00	25.10
5600900060992	1	27	0.04	24.64	24.64	28	4.33	24.33
6844000608436	23	1	25.21	0.04	25.21	23	2.67	23.64
5803301245621	21	1	23.02	0.04	23.02	22	2.33	21.07
92891948003	1	24	0.05	21.90	21.90	24	3.00	20.91
3433000017263	52	3	19.00	0.05	19.00	55	13.33	19.00
467615916337	1	22	0.05	20.07	20.07	22	2.33	18.39
817004638227	19	1	20.82	0.05	20.82	20	1.67	17.67
2376700063599	30	2	16.44	0.06	16.44	32	5.67	16.39
993620816222	1	19	0.06	17.34	17.34	20	1.67	14.74
993620810220	5	76	0.07	13.87	13.87	81	22.00	13.87
8999000079657	1	18	0.06	16.42	16.42	19	1.33	13.21
5000006000095	253	23	12.06	0.08	12.06	276	87.00	12.06
9420966064460	1	17	0.06	15.51	15.51	18	1.00	11.61
5186264200136	1	17	0.06	15.51	15.51	18	1.00	11.61
600000201284	4	47	0.09	10.72	10.72	51	12.00	10.72
5600000060302	1	16	0.07	14.60	14.60	17	0.67	9.99
7080606900600	1	16	0.07	14.60	14.60	17	0.67	9.99
6070095870009	26	3	9.50	0.11	9.50	29	4.67	9.42
999960264339	3	28	0.12	8.52	8.52	31	5.33	8.48
555400670006	1	15	0.07	13.69	13.69	16	0.33	8.39
881894855	1	15	0.07	13.69	13.69	16	0.33	8.39
1960400470068	23	3	8.40	0.12	8.40	26	3.67	8.22
993620559229	5	43	0.13	7.85	7.85	48	11.00	7.85
6000330043193	13	1	14.25	0.07	14.25	14	-0.33	6.53
2644006060269	13	1	14.25	0.07	14.25	14	-0.33	6.53
8124906575841	29	5	6.36	0.16	6.36	34	6.33	6.35
6880098906148	23	157	0.16	6.23	6.23	180	55.00	6.23
6020094951312	3	21	0.16	6.39	6.39	24	3.00	6.13
2586000448258	1	14	0.08	12.77	12.77	14	-0.33	5.91
604901367333	1	14	0.08	12.77	12.77	14	-0.33	5.91