Applying Data Science – Business Cases Case Studies using SAS

By Gerhard Svolba

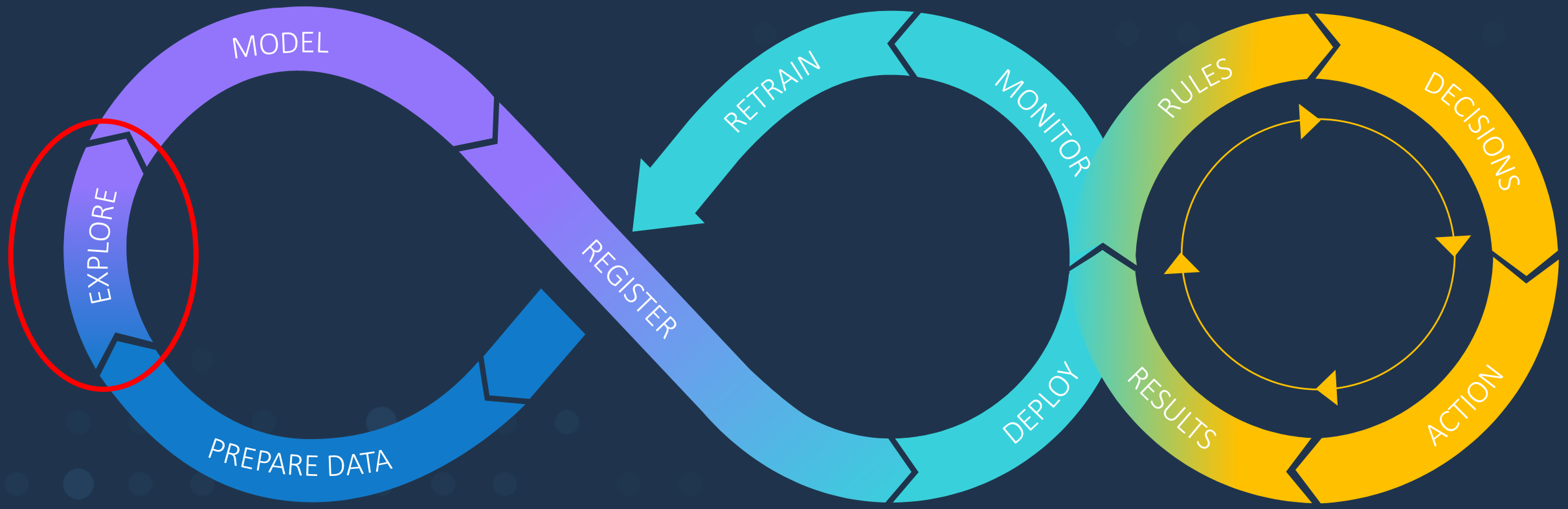
Presented by Jacob Mardfelt

Agenda – 2 Use Cases

Data exploration by using Association Analysis

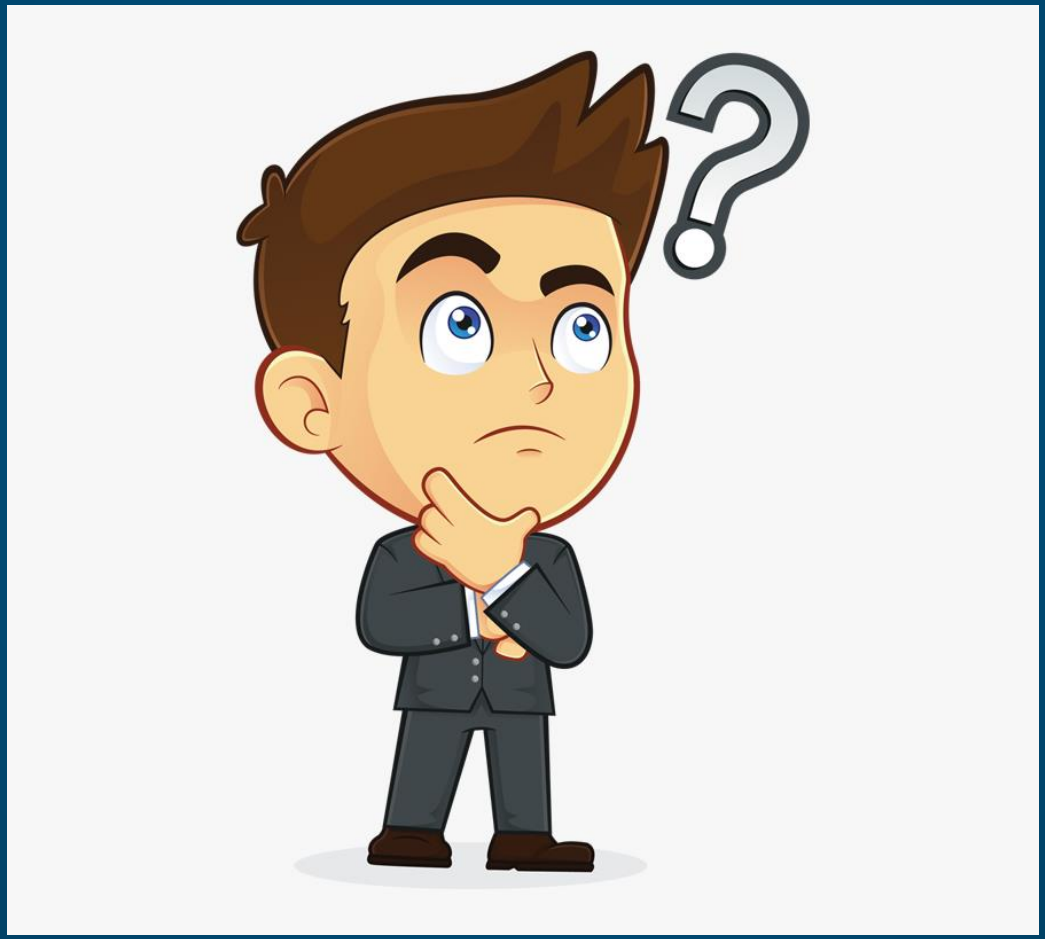
A quick look at finding Fraud with the Benford Distribution

THE DECISIONING PROCESS



ANALYTICS — IT — BUSINESS

Can our data tell us something about or subject with us not asking any explicit questions?



Use Case – Car Insurance Claims

Variable	Feature
Age	Young, Midwife, Old
Gender	Male, Female
Density	Highly Urban, Urban, Highly Rural, Rural
Car Type	Van, Sports Car, SUV, Sedan, Pick Up
Car Usage	Private, Commercial
Claim Flag	Claim, No Claim

Data exploration outcome from association analysis

Unsupervised learning to find relations between the variables

Data Quality from Business Point of view – Does it make sense?

This is explanatory analysis – No hypothese testing

Will uncover obvious rules – Great!

This will give you information before creating a Predictive Model

Association analysis / Market Basket Analysis – How does it work?

<u>Rule</u>: Saint Agur Cheese –> Crackers (LHS -> RHS)

<u>Confindence</u>: Customers who buys Cheese (LHS) buys in 23 % of the cases also Crackers

<u>Support</u>: relative frequency of ,, Cheese + Crackers Combinations is all baskets , eg 2,67 %

<u>Lift</u>: 3.5 factor that the rule ”Cheese + Crackers” appear more frequently

Two things to look for: Lift >> 1 and Lift << 1

Association analysis – Transactional structure Needed

Policynumber	Claim flag	Car Use	Car Type	Age	Gender	Density
160	No Claim	Private	Sedan	57	Male	Urban
334	Claim	Private	Van	22	Female	Highly Rural
13431	Claim	Commercial	SUV	45	Male	Rural
1212	No Claim	Private	Sedan	55	Female	Urban
122	No Claim	Private	SUV	34	Male	Rural
3535	No Claim	Private	Sedan	43	Male	Urban
22	Claim	Private	Sedan	57	Male	Urban

One-row-per-subject (ABT)

Policynumber	Feature
160	Private
160	Sedan
160	Old
160	Male
160	Urban
160	No Claim
334	Claim
334	Van
334	young
334	Female
334	Highly Rural
334	Claim

Multiple-rows-per-subject (Transacational data)

Let’s head into SAS Studio!

Association analysis – Discussing the results

WORK.CLAIMS_EXPLORATION ▾ Columns: 6 of 6 | Total rows: 278 | Rows 1 to 200

Enter expression

	COUNT	SUPPORT	CONF	conf_exp	LIFT	RULE
1	30	0.29	2.54	46.18	0.06	Sports Car ==> Male
2	99	0.96	3.43	46.18	0.07	SUV ==> Male
3	99	0.96	2.08	27.98	0.07	Male ==> SUV
4	40	0.39	4.69	53.82	0.09	Panel Truck ==> Female
5	117	1.14	12.7	53.82	0.24	Van ==> Female
6	117	1.14	2.11	8.94	0.24	Female ==> Van
7	32	0.31	1.17	4.93	0.24	Claim ==> Highly Rural
8	32	0.31	6.3	26.66	0.24	Highly Rural ==> Claim
9	10	0.1	8.33	34.93	0.24	Young ==> Highly Urban
10	102	0.99	3.71	15.18	0.24	Claim ==> Rural
11	102	0.99	6.52	26.66	0.24	Rural ==> Claim
12	200	1.94	5.28	11.44	0.46	Commercial ==> Sports Car
13	200	1.94	16.96	36.78	0.46	Sports Car ==> Commercial

Customer base – Mostly Woman that drives Sports Car?

Data Quality Issue?

The sports car is the 2nd or 3rd vehicle in the family, registrated to the female (German example ☺)

Competitors has a better priced policy for men

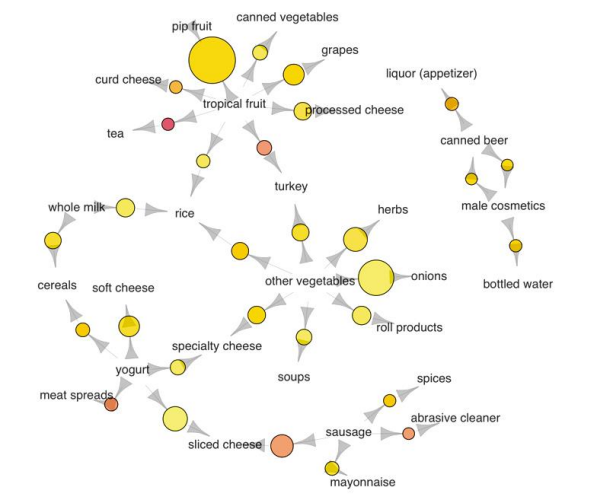
Association analysis / Market Basket Analysis wrap up

Get a more objective picture of the data

Get explicit result, no needle in a haystack

Recieve findings automatically instead of manually

Do you regulary use AA / MBA for data exploration analysis?



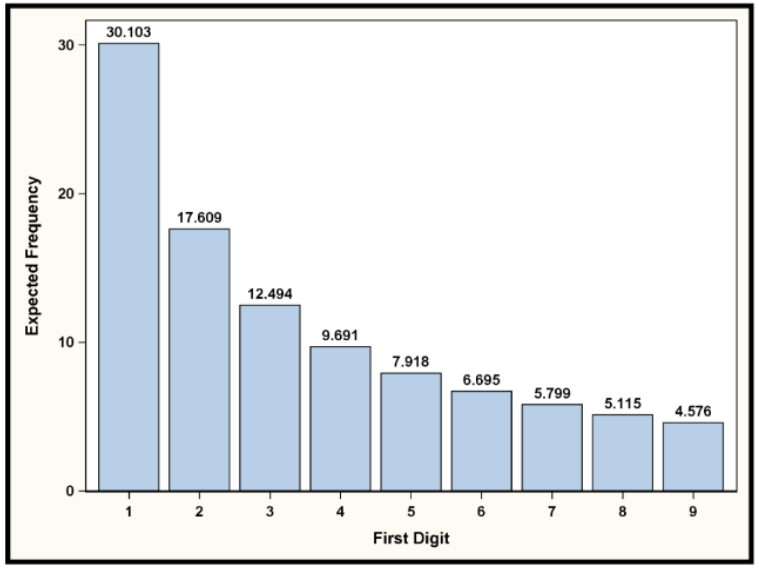
Agenda – 2 cases

Data exploration by using Association Analysis

A quick look at finding Fraud with the Benford Distribution

Benford's Law

Benford's Law – Distribution of the Digits 1-9

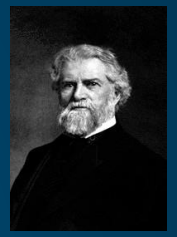



```
1,323.23
   43.00
  622.12
    1.10
   89.09
2,592.22
    7.40
   82.10
  620.19
   30.00
```

$$P(d) = \log10(1 + 1/d)$$

Only in sets of natural occurring numbers!

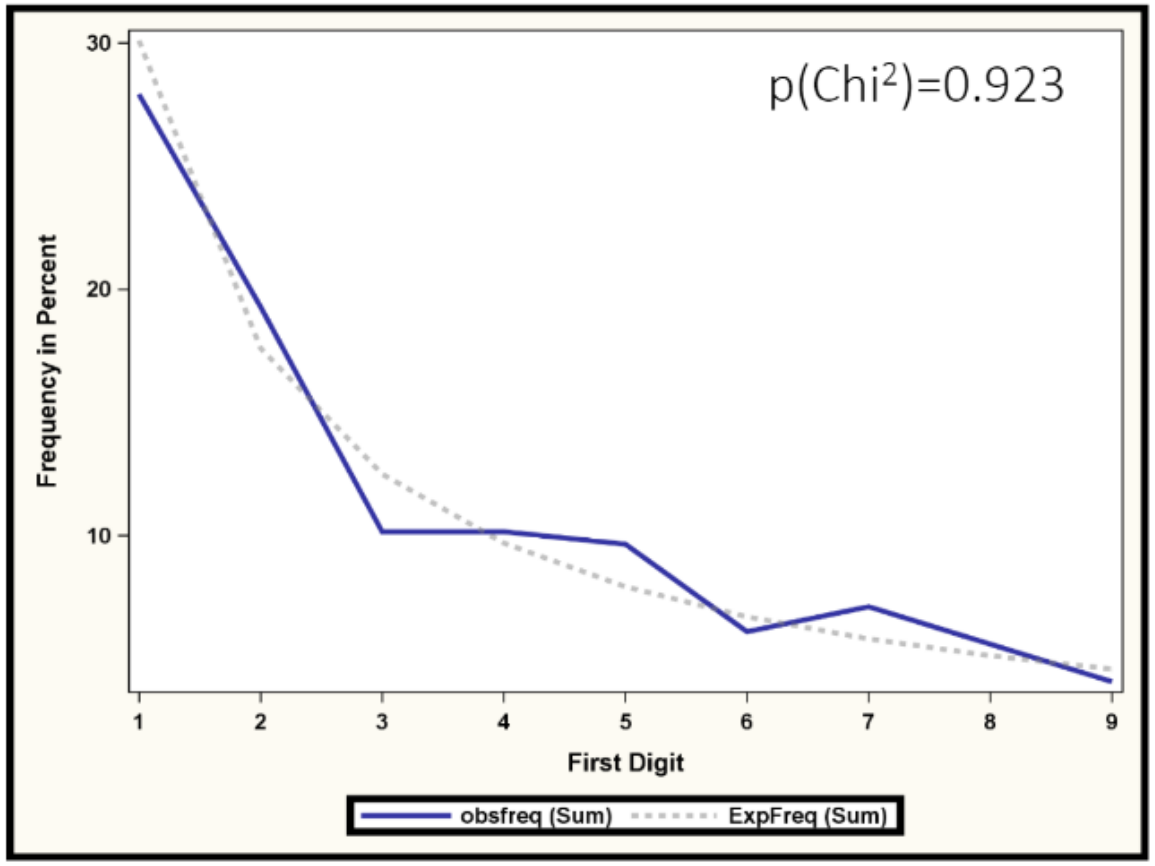
1881: Simon Newcomb



1938: Frank Benford

1972: Hal Varian

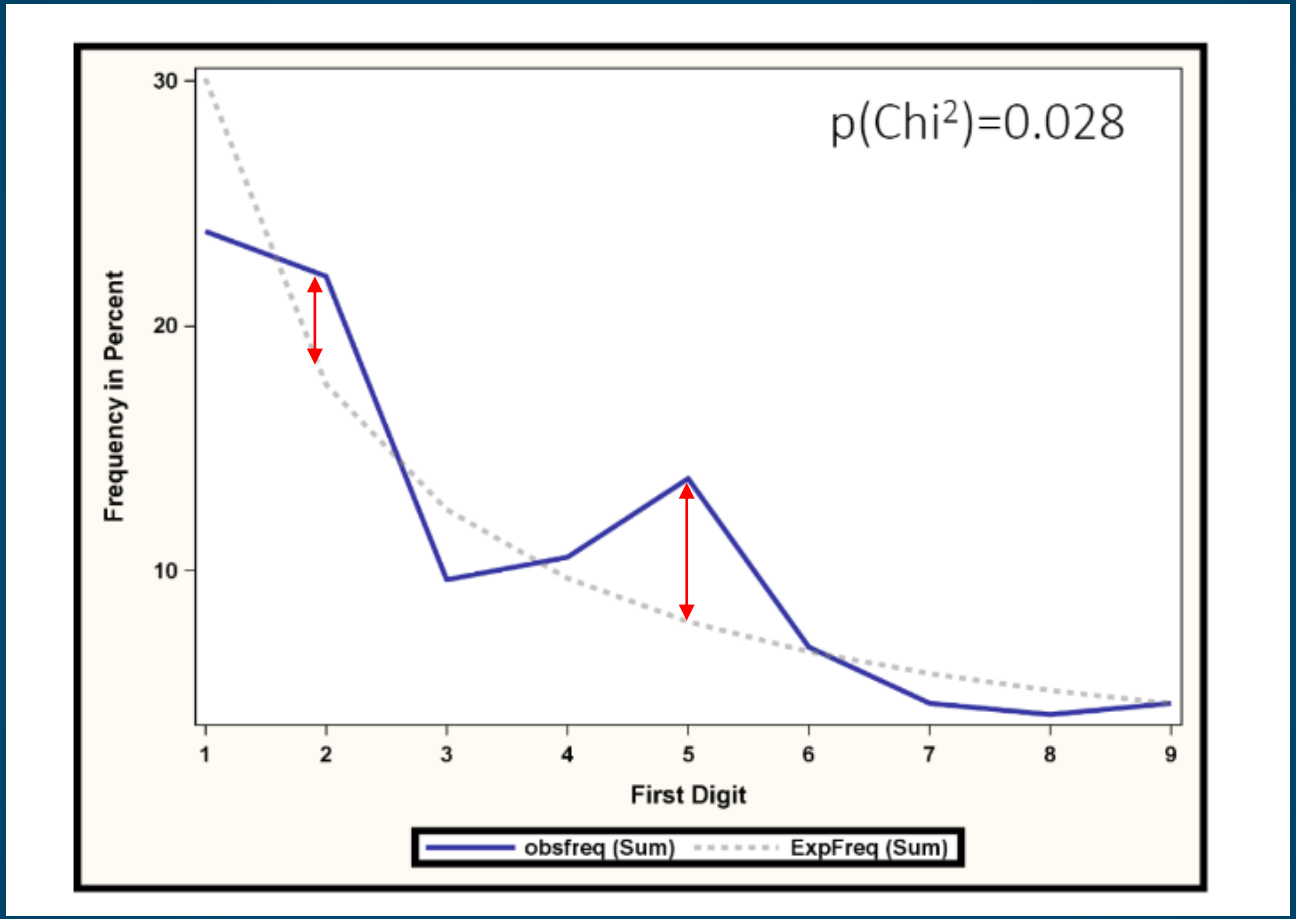
Transactions from an account A vs Benford Distribution



Chi² Independence test

It is used to determine whether there is a significant association between the two variables.

Transactions from an account B vs Benford Distribution



Rank customers by deviation from the expected distribution

Rank	CustomerID	Chi2_Value	P_Value
1	5000	42.3	0.000%
2	2000	33.4	0.005%
3	8000	28.3	0.042%
4	4000	28.0	0.048%
5	3000	27.1	0.068%
6	1000	26.4	0.090%
7	10000	25.2	0.145%
8	6000	23.0	0.341%
9	11000	17.9	2.207%
10	7000	15.0	5.898%
11	9000	10.4	23.95%

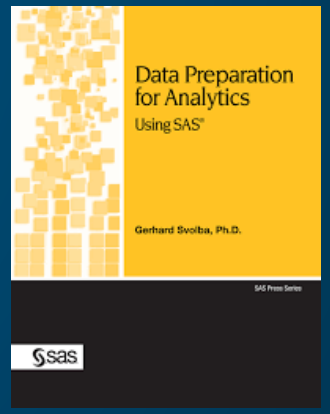
Start by investigation these customers!

Gerhald Svolba

His code is on Github

https://github.com/gerhard1050/Applying-Data-Science-Using-SAS

His book are available online



And case studies are now on youtube

https://www.youtube.com/playlist?list=PLdMxv2SumIKs0A2cQLeXg1xb9OVE8e2Yq