LAB 360

BRINGING INSIGHTS TO LIFE

LEARNING WITH SAS

SUNZ 2015

Learning with SAS – Three ways

User Story: Programming Whiz

Extraction and analysis

Solves the pipeline problem, easy to work with large data sets

Move between point and click and coding

Macro language can be hard to debug

Careful with the SAS/Oracle interface e.g. padding strings/truncation/need for explicit casting

SAS throws an error but still executes, “SAS goes on, no matter what”

Log to datasets – they contain more information and are easier to navigate than the log

SAS has lots of tools for quickly evaluating whether the dataset you generated is what you expected, “PROC FREQ AND PROC MEANS are awesome”

“OPTIONS ERRORABEND is your friend”. It immediately terminates your program upon error, which will keep the erroneous dataset in its broken state, making it easier to debug.

User Story:
Experienced Scientist

MySQL

Easy to access and import data from different systems and file types e.g. Oracle, Excel, Access, Text files

DATA STEP iterative processing allows you to do some really neat things

Macro variables make dynamic coding easy

PROC SQL is a non-standard SQL implementation so be careful to check it's doing what you think

Diagnosing reasons for poor code performance can be difficult (e.g. used to working with SQL explain plan)

Treatment of missing values (e.g. missing < -10 is TRUE)

Use the SAS online courses to get started

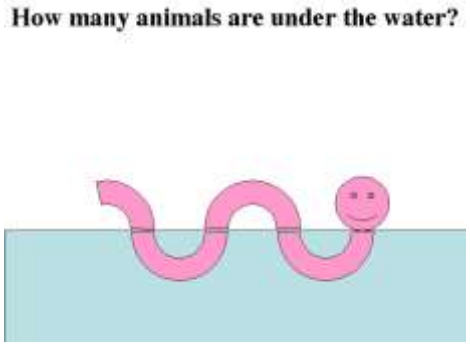
Find a good 'cheat sheet' and reference to help you get on top of syntax and basic procedure usage

support.sas.com documentation pages are really useful

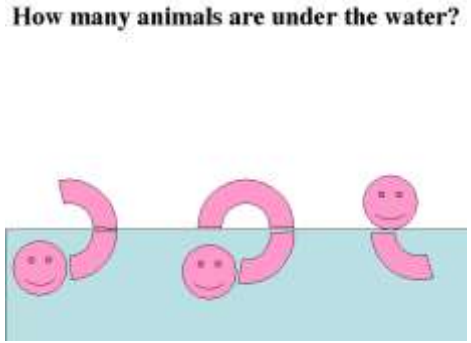
Use the GUI to prototype and generate SAS code

What lies beneath? Factor Analysis

- We often have a need to work with survey data
- Reduce a large number of variables down to a smaller number of factors that capture the variation
- Searches for joint variation in observed response due to unobserved (latent) variables
- Technique used in the behavioural and social sciences, marketing, product management, operations research



OR!
?



Getting the data into shape

1

Favourite TV channel				
Q2_TV_BEST	Frequency	Percent	Cumulative Frequency	Cumulative Percent
TV1				
TV2				
TV3				
C4				
Prime				
Other FTA				
Pay Sport				
Pay Movies				
Pay (other)				
Don't watch				

2

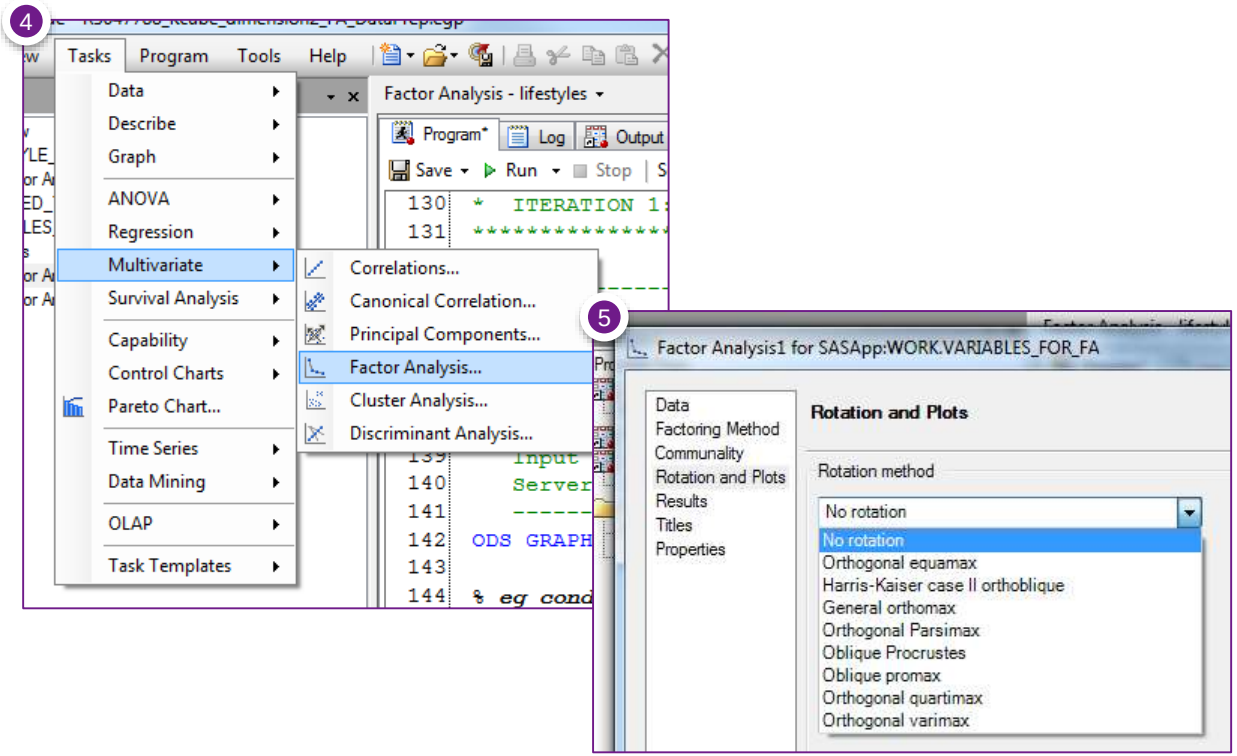
alyze Export Send To

IA_NONE	Q2_TV_BEST	Q3_MAGS_READ_ART	Q3_MAGS_READ_BOAT
0	9	0	0
0	1	0	0
0	1	0	0
0	5	0	0
0	1	0	0
0	2	0	0
0	9	0	0
0	1		0
0	1		
0	1		
0	9		
0	7		
0	10		
0	10		
0	1		
0	6		
0	1		

3

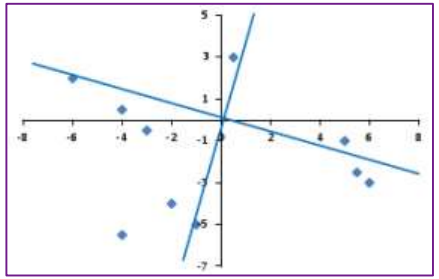
Q2_TV_BEST_TV1	Q2_TV_BEST_TV2	Q2_TV_BEST_TV3	Q2_TV_BEST_C4	Q2_TV_BEST_Prime
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	1	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	1	0	0
0	0	0	0	0

Running the analysis



Exploring and interpreting

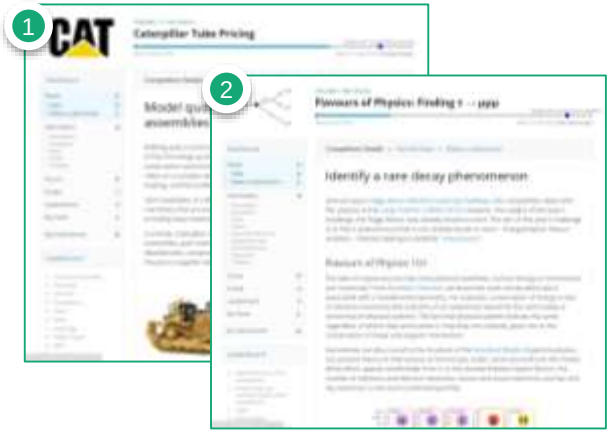
- Interpreting and characterizing the factors
- Trying different rotations
- Apply clustering after the Factor Analysis
- Profile by original variables
- What do you **learn**?



Q4_ACTIVITY_GENERAL_ART	Arts and crafts
Q5_ACTIVITY_SPORT_ART	Arts and crafts
Q7_LIFESTYLE_DESCR_CREATE	Creative
Q5_ACTIVITY_KNIT	Knitting/cross-stitch
Q5_ACTIVITY_PHOTO	Photography
Q4_ACTIVITY_GENERAL_EXERCISE	Exercise or gym
Q5_ACTIVITY_SPORT_GYM	Gym work-outs
Q5_ACTIVITY_SPORT_RUN	Running
Q7_LIFESTYLE_DESCR_ACTIVE	Active
Q5_ACTIVITY_SPORT_MARTIAL	Martial arts
Q3_MAGS_READ_WHEELS	Wheels
Q5_ACTIVITY_SPORT_MOTOR	Motor sports
Q5_ACTIVITY_SPORT_CARS	Working on cars

Kaggling with SAS

- Predictive modelling platform with more than 300,000 members
- Some competitions allow you to enter with minimal data work
- A fun way to play with automated modelling
  - RPM Basic and Advanced
- We entered two competitions:
  1. Caterpillar Tube Pricing
  2. Flavours of Physics



1. Caterpillar Tube Pricing

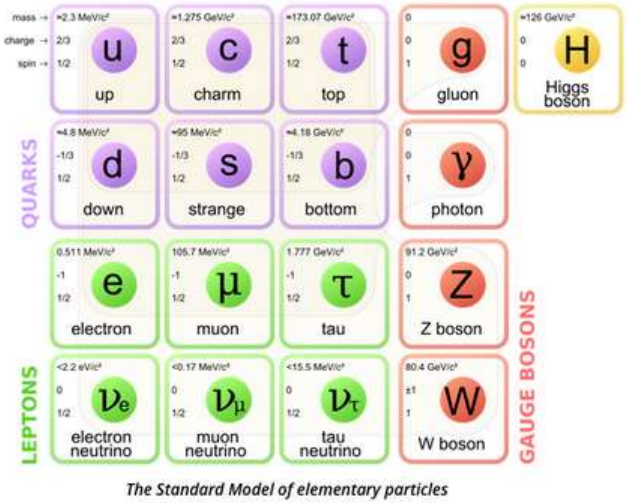
- Enormous variety of construction and mining equipment
  - Each machine relies on a complex set of tubes
  - These tubes can vary across a number of dimensions including base materials, number of bends, bend radius, bolt patters and end types

- Challenge: Predict the price a supplier will quote for a particular tube assembly

- Data extract from relational database

2. Flavours of Physics

- Aim of the challenge is to find a phenomenon not already known to exist – charged lepton flavour violation
- Real data from the Large Hadron Collider, combined with simulated data of the decay
- 50 variables
- Binary response (signal)



Putting it into battle

- We wanted to test a very naïve approach
- No feature engineering and minimal data cleaning, simply throwing a flat file at SAS Rapid Predictive Modeller (RPM)
- SAS RPM – decided which transformations to use, which variables to select or drop, which algorithms
- Quick to run
- **So, how did we go?**

Putting it into battle

- **Caterpillar (numeric response)**
  - Decision model
  - 950th place (out of 1206)

- **Flavour of physics (binary response)**
  - Ensemble Champion model
  - Lift of 1.6 on the top 10%
  - 293rd place (out of 337 entries)
  - Just below a gradient boosting benchmark that Kaggle provides

- **Conclusions**
  - Better than the bottom 15% in each competition
  - Quick and easy to run Advanced RPM
  - Helpful to have the Enterprise Miner Report
  - Still a big human component! Understanding the domain and feature engineering

THANK YOU

@pietabrown