Walkthrough Data Cleansing

Pangkalan Data Perlindungan Sosial

Perisian untuk kegunaan Hands-on

http://colab.research.google.com/

Perisian untuk kegunaan Hands-on

http://colab.research.google.com/

https://bit.ly/11klasPython

https://bit.ly/11klasData

https://www.sololearn.com/Play/Python/hoc

Login > “tocolab”

1. Login to GMAIL
2. > https://bit.ly/11klasPython
3. https://github.com/booluckgmie/training/blob/main/GColab_and_Intro_to_Python.ipynb
4. https://githubtocolab.com/booluckgmie/training/blob/main/GColab_and_Intro_to_Python.ipynb

Instructor Introduction

- Name: Ahmad Najmi Ariffin
- Email: najmi.ariffin@dosm.gov.my

- Main research focus:
  - Analyzing Data by using Machine Learning algorithms

Course Logistics

Day	Time	Activities
Day 1/2	2:30pm – 3:45pm (1hr 15min)	Afternoon Session 1
	3:45pm – 4:00pm	Break
	4:00pm – 5:30pm (1hr 30min)	Afternoon Session 2
Day 2/2	9:30am – 11:00am (1hr 30min)	Morning Session 1
	11:00am – 11:15am	Morning break
	11:15am -12:45pm (1hr 30min)	Morning Session 2
	12:45pm – 2:30pm	Lunch
	2:30pm – 3:45pm (1hr 15min)	Afternoon Session 1
	3:45pm – 4:00pm	Break
	4:00pm – 5:30pm (1hr 30min)	Afternoon Session 2

Course Outcomes

- After completing this course, you will be able to
  - understand the features of Python Programming
  - understand the concept of variables
  - write simple python programs using flow control
  - understand the concept of collections
  - **use some python libraries**
  - **understand program structure**

Course Content

- Introductions to the Features of Python Programming
- Working Variables in Python
- Flow Control in Python
- Using Python Collection
- **Working in Libraries in Python**
- **Program Structure**

Working with Libraries in Python

Python Libraries

- Pandas
  - Pandas
  - Modin.pandas
  - Pandas Profiling
- Numpy
- Sidetable
- OneTable
- PyGeocoder

Google Maps

We submit a business name as input and the program gives the complete address as the output. The module uses data from google maps in the background to retrieve the result.

```
pip install pygeocoder
```

```
from pygeocoder import Geocoder


business_name = "Oracle Malaysia"
print ("Searching %s" %business_name)
results = Geocoder.geocode(business_name)
for result in results:
        print (result)
```

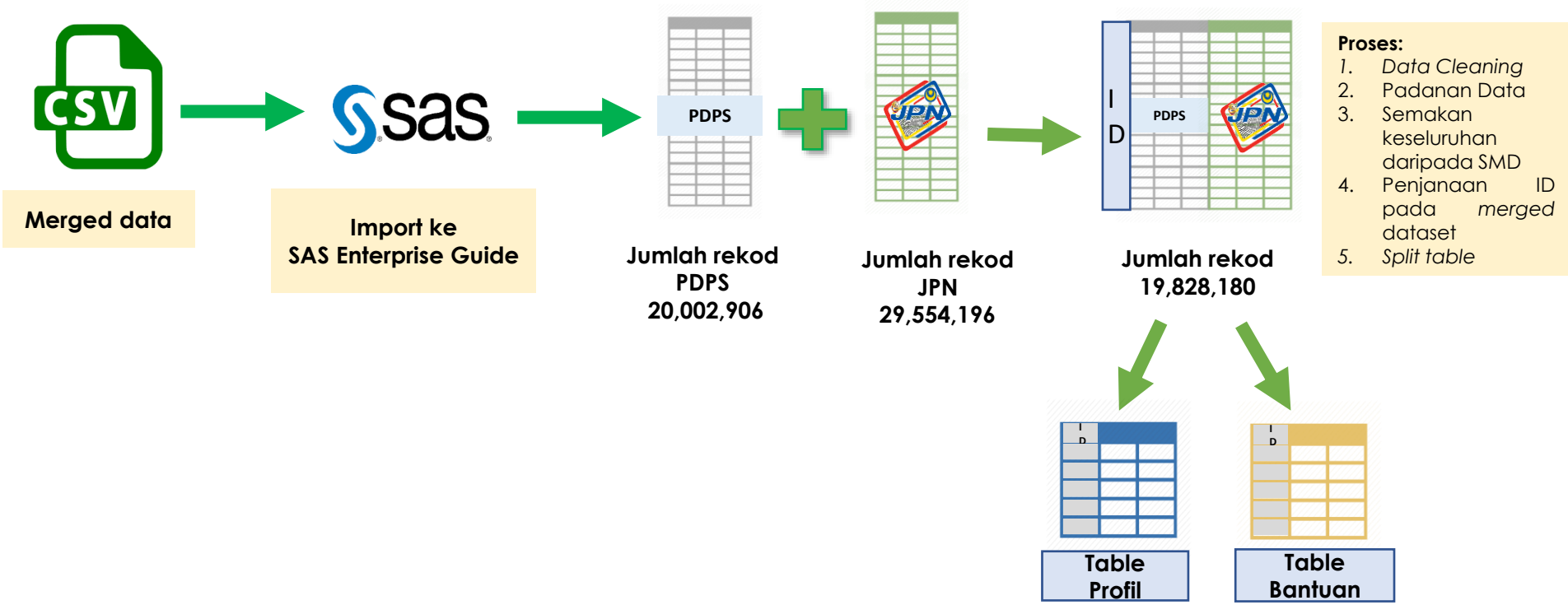
Open/Read Data

Recode

Unique Profil

Process Flow Data Integration

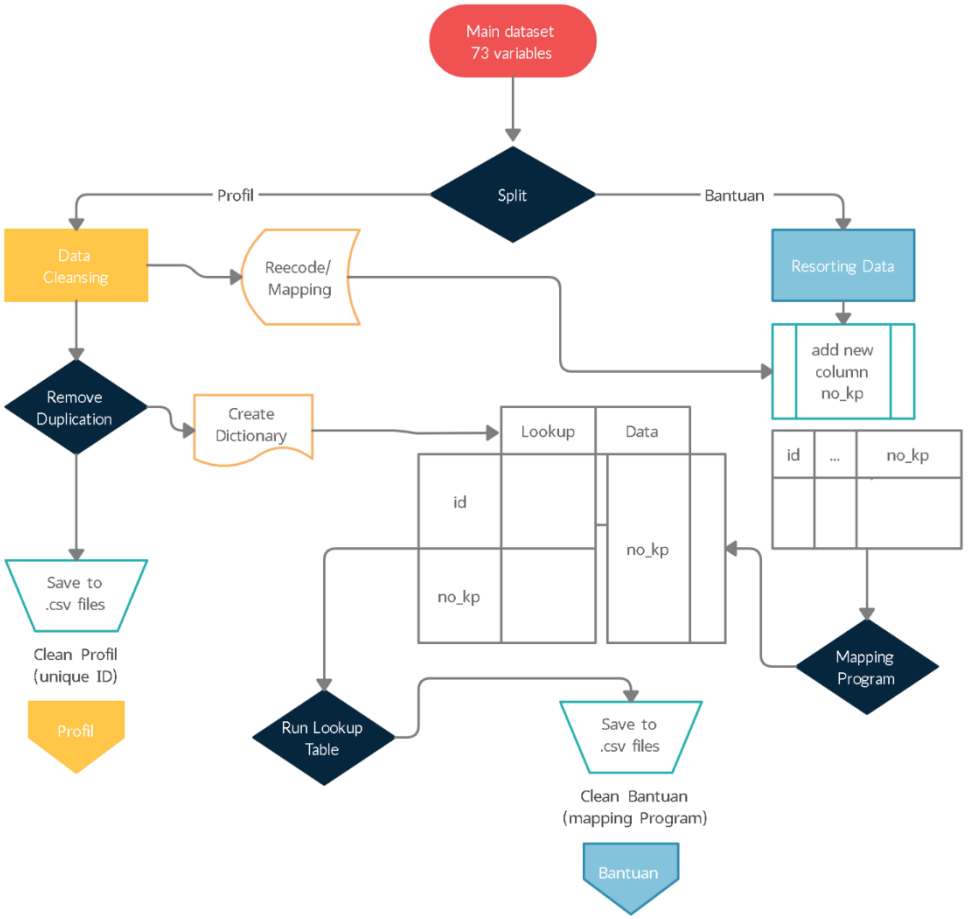
PROCESS FLOW



Legend:

SAS EG

PROCESS FLOW (SAMB.)



Environment : Python
Libraries : Modin.pandas

Proses

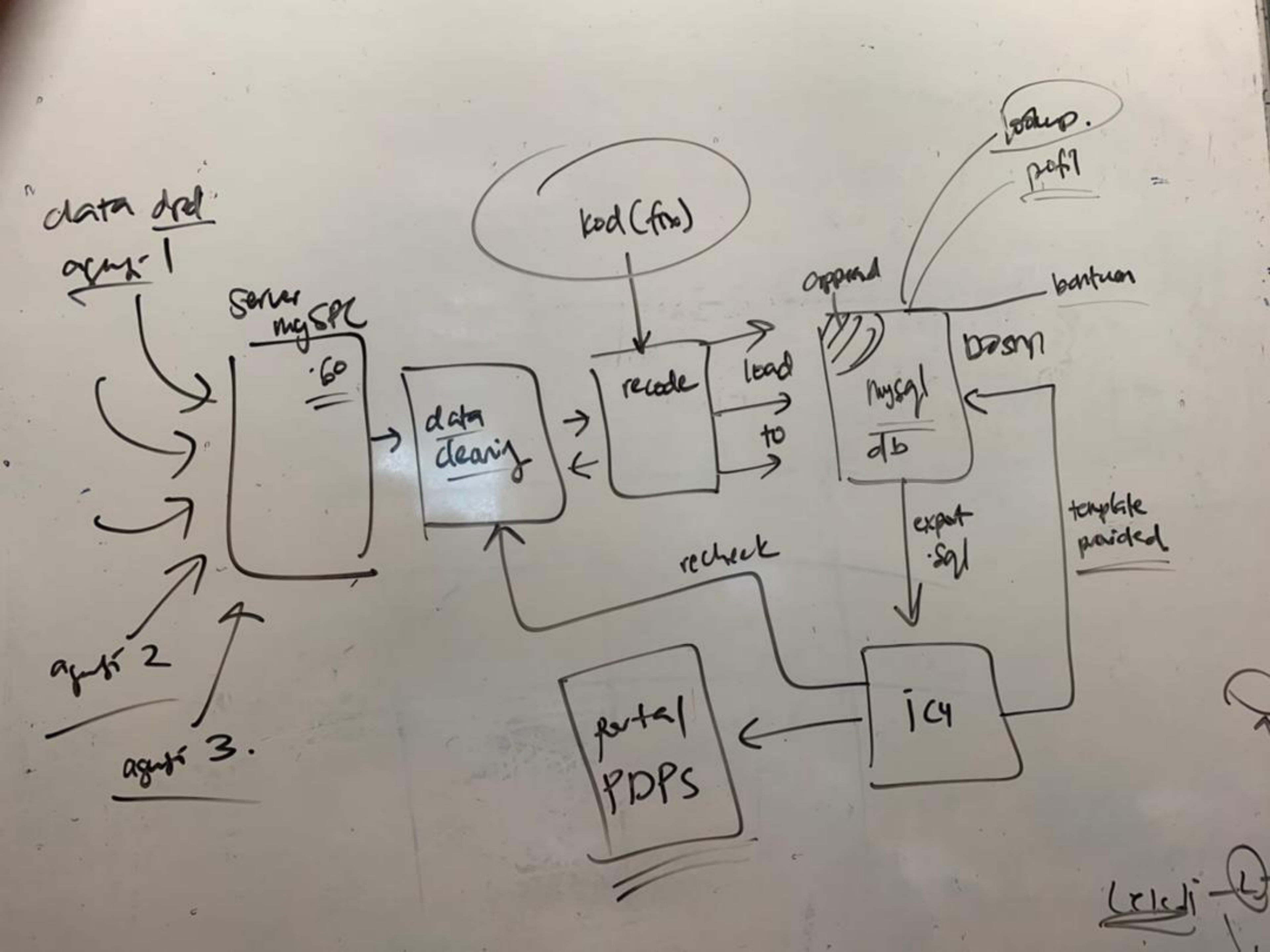
- *Profil*
  1. *Data cleansing (transform to UPPERCASE, leading zero, Unicode issues)*
  2. *Diwujudkan dictionary reference ID menggunakan variable no_kp*
  3. *Remove duplikasi*
  4. *Simpan data profil (|)*

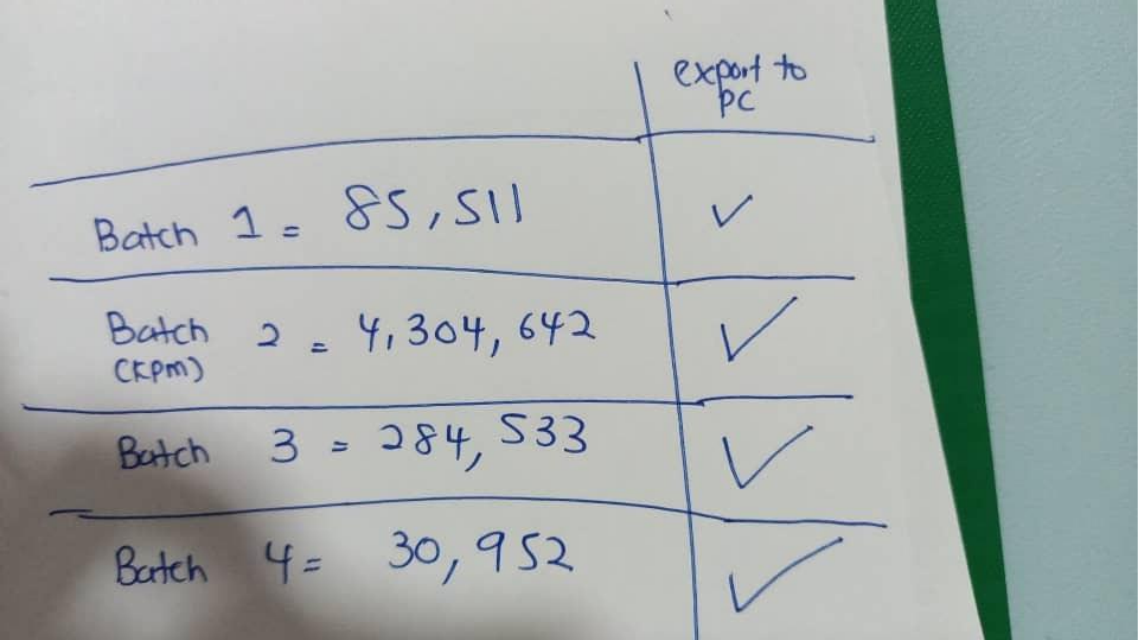
- *Bantuan*
  1. *Resort data (penambahan column no_kp : reference ID)*
  2. *Padanan rekod table program ke table bantuan*
  3. *Padanan dictionary reference ID table bantuan*
  4. *Simpan data bantuan (|)*

CodeShare Realtime

https://codeshare.io/BAOw8x

Split Table


server
mySQL
recode
load
to
append
bantuan
mysql
db
export
.sql
template
provided
recheck
portal
PDPS



Batch 1 = 85,511
Jakoa = 22,235
Mitra = 43,806
Lkim = 4,198
Kpt = 15,272

Batch 2 = 4,304,642
Kpm = 4,304,642

Batch 3 = 284,533
Kesedar = 1,342
Kpkt = 1,962
Lppkn = 14,871
Mara = 35,939
Pprt kplb = 15,466
Risda bmt = 214,953

Batch 4 = 30,952
Kkm = 20,145
Ketsa = 334
Jpa = 4,750

Batch 5 = 1,048,123
JKM = 1,048,123

Batch 6
BSH = 4,298,000

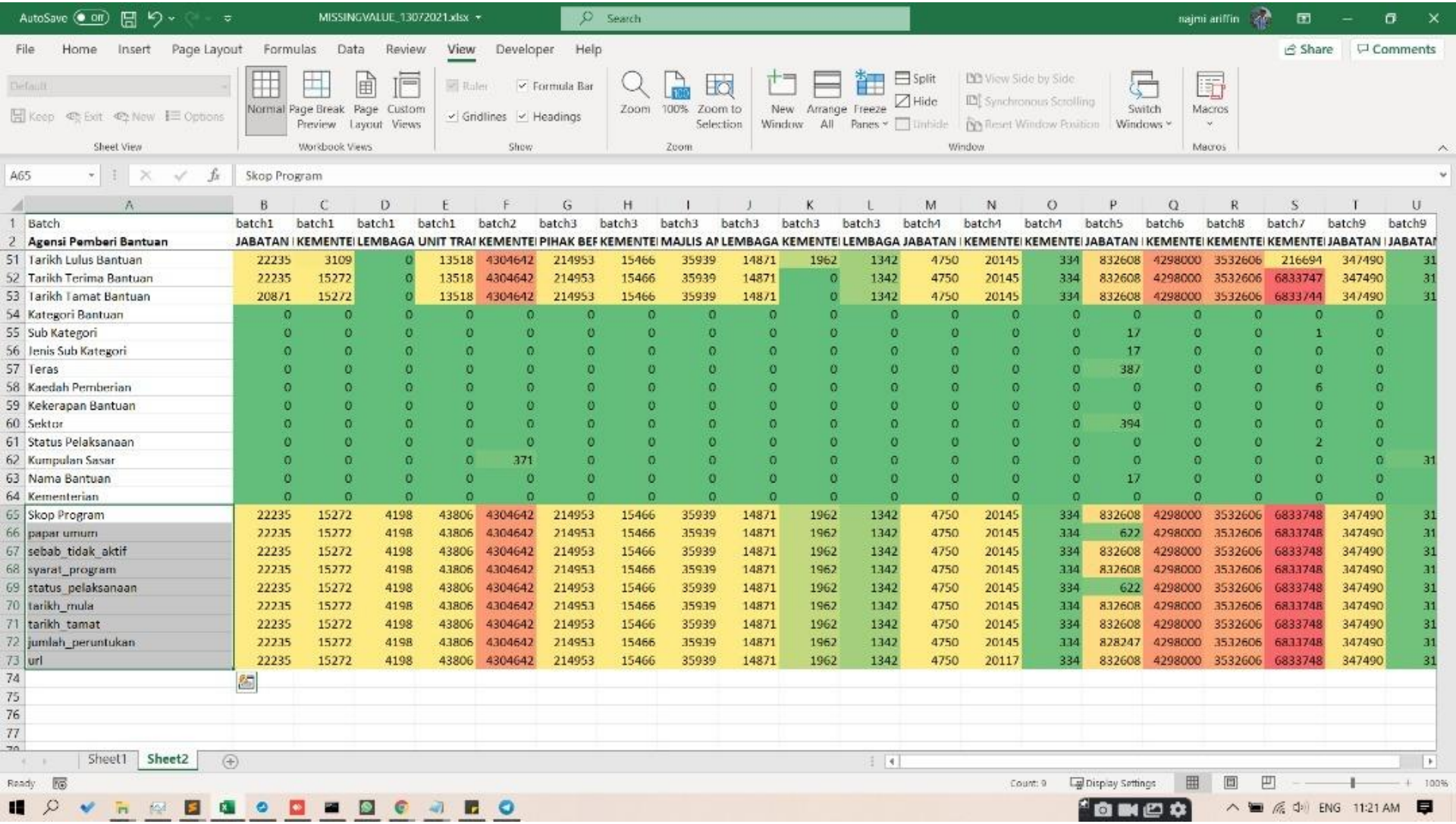
Batch 7
BPN = 6,833,744

Batch 8
BPN_B40IR = 3,532,606

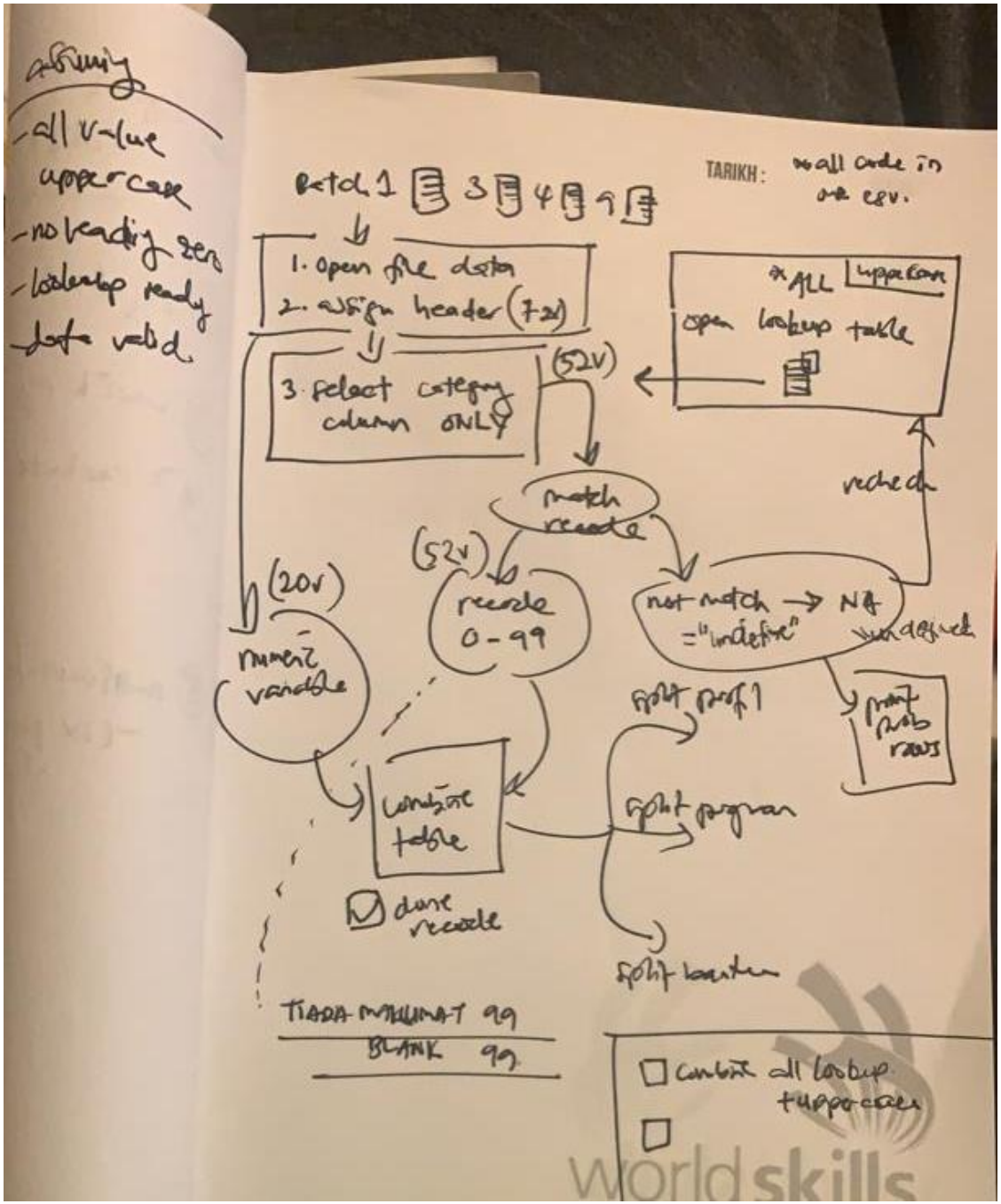
Batch 9 = 347,802
Kemas_GPK = 173,745
Kemas_PBMT = 173,745
JKM_BB = 312

SE	LOAD P
kluster_etnik	1
penyakit	1
status_pekerjaan	1
teras	1
sijil	1
jenis_kenderaan	1
status_pelaksanaan	1
jenis_oku	1
strata	1
kumpulan_sasar	1
jenis_pekerjaan	1
sub_kategori	1
rawatan	1
sumber_pendapatan	1
agama	1
bank	1
jantina	1
jenis_akaun_bank	1
jenis_kemahiran	1
jenis_pemilikan_kenderaan	1
sektor	1
status_bantuan	1
status_kahwin	1
taraf_pendidikan	1
kategori_bantuan	1
kekerapan	1

kementerian	1
kaedah_pemberian	1
warganegara	1
negeri	1
jenis_sub_kategori	2
daerah	2
mukim	2
parlimen	2
kumpulan_etnik	2
agensi	2
profil	3
program	3
dun	3
oku	4
kemahiran	4
kenderaan	4
pendapatan	4
profil_penyakit_rawatan	4
program_jenis_sub_kategori	4
program_kumpulan_sasar	4
bantuan	5



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Batch	batch1	batch1	batch1	batch1	batch2	batch3	batch3	batch3	batch3	batch3	batch3	batch4	batch4	batch4	batch5	batch6	batch8	batch7	batch9	batch9
2	**Agensi Pemberi Bantuan**	**JABATAN**	**KEMENTE**	**LEMBAGA**	**UNIT TRA**	**KEMENTE**	**PIHAK BEF**	**KEMENTE**	**MAJLIS AP**	**LEMBAGA**	**KEMENTE**	**LEMBAGA**	**JABATAN**	**KEMENTE**	**KEMENTE**	**JABATAN**	**KEMENTE**	**KEMENTE**	**KEMENTE**	**JABATAN**	**JABATAN**
51	Tarikh Lulus Bantuan	22235	3109	0	13518	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	832608	4298000	3532606	216694	347490	31
52	Tarikh Terima Bantuan	22235	15272	0	13518	4304642	214953	15466	35939	14871	0	1342	4750	20145	334	832608	4298000	3532606	6833747	347490	31
53	Tarikh Tamat Bantuan	20871	15272	0	13518	4304642	214953	15466	35939	14871	0	1342	4750	20145	334	832608	4298000	3532606	6833744	347490	31
54	Kategori Bantuan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
55	Sub Kategori	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	1	0	
56	Jenis Sub Kategori	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	
57	Teras	0	0	0	0	0	0	0	0	0	0	0	0	0	0	387	0	0	0	0	
58	Kaedah Pemberian	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	
59	Kekerapan Bantuan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
60	Sektor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	394	0	0	0	0	
61	Status Pelaksanaan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	
62	Kumpulan Sasar	0	0	0	0	371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31
63	Nama Bantuan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	
64	Kementerian	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
65	Skop Program	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	832608	4298000	3532606	6833748	347490	31
66	papar umum	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	622	4298000	3532606	6833748	347490	31
67	sebab_tidak_aktif	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	832608	4298000	3532606	6833748	347490	31
68	syarat_program	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	832608	4298000	3532606	6833748	347490	31
69	status_pelaksanaan	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	622	4298000	3532606	6833748	347490	31
70	tarikh_mula	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	832608	4298000	3532606	6833748	347490	31
71	tarikh_tamat	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	832608	4298000	3532606	6833748	347490	31
72	jumlah_peruntukan	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20145	334	828247	4298000	3532606	6833748	347490	31
73	url	22235	15272	4198	43806	4304642	214953	15466	35939	14871	1962	1342	4750	20117	334	832608	4298000	3532606	6833748	347490	31
74																					
75																					
76																					
77																					


- all value uppercase
- no leading zero
- lookup ready
- data valid
TARIKH :
* all code in one csv.
Retak 1 3 4 7
1. open file data
2. assign header (7x)
3. select category column ONLY
(52V)
* ALL uppercase
open lookup table
match records
recheck
(52V)
(20V)
records 0-99
not match → NA = "undefine"
numeric variable
split prog 1
print prob rows
combine table
split program
done records
split location
TIADA MAKLUMAT 99
BLANK 99
combine all lookup uppercase
worldskills

Q & A

THANK YOU