



**INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER**  
**SURVEY RESEARCH OPERATIONS**  
UNIVERSITY OF MICHIGAN

# **The Use of Python to Add Power and Flexibility to Data Management**

Emily Blasczyk and Minako Edgar (Presenter: Emily Blasczyk)

TSG



# Overview

- What is Python, why use it?
- What we do with Python at UM-SRO:
  - Reporting
  - Data Processing
  - Tool building

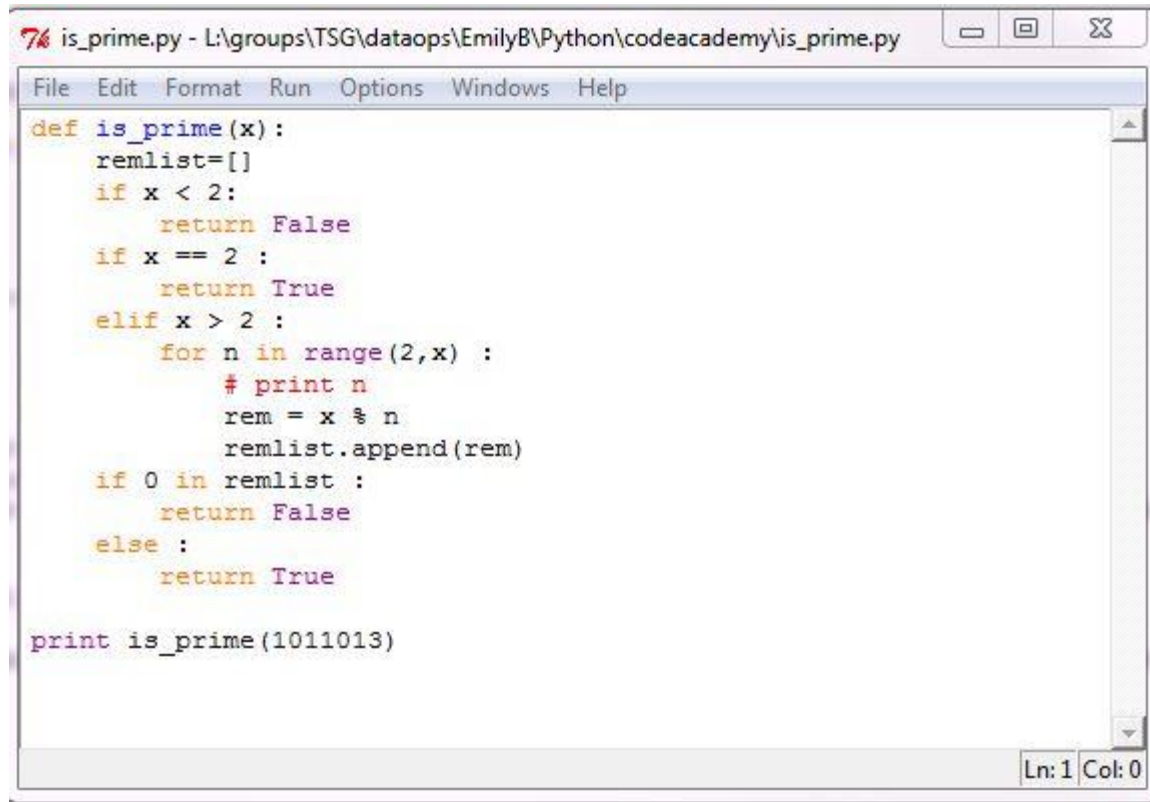


# What is Python, Why Use it?

- Python is an open source programming language. It is free to use!
- Borrows syntax from C so experienced programmers can transfer knowledge.
- Yet it is easy and intuitive for new programmers.
- Flexible, multi-purpose: Can do everything from statistical analysis to application programming.

# What is Python, Why Use it?

## Example of whitespace



```
7% is_prime.py - L:\groups\TSG\dataops\EmilyB\Python\codeacademy\is_prime.py
File Edit Format Run Options Windows Help
def is_prime(x):
    remlist=[]
    if x < 2:
        return False
    if x == 2 :
        return True
    elif x > 2 :
        for n in range(2,x) :
            # print n
            rem = x % n
            remlist.append(rem)
    if 0 in remlist :
        return False
    else :
        return True

print is_prime(1011013)
Ln: 1 Col: 0
```



# What is Python, Why Use it?

- Modules contain executable statements or functions that perform specific tasks.
- Python has standard modules, but many, many more optional modules a user can install and utilize.



# How UM-SRO Technical Services Uses Python

- Reports
  - Field Progress
  - Payment
- Data Processing
  - Swapping IDs
- Tools
  - Interviewer-PSU Geo-coding tool
  - Password generator



# Reports: Field Progress Reports

- Summarizes current field production status outstanding sample, finalized cases, refusals and response rates.
- Key Python: `pyodbc`, `openpyxl`



# Reports: Field Progress Reports

## Summary report

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Study:04/16/2015	TOT SAMP	NO ATMT	NO CONT	CONT NOREIS	CONT RESIS	APPT	TRACK	HOLD	FI	NIREF	NINO CONT	NIOTH	NS	CONT	RESIS	CONTR	IW RATE	RESP RATE	COMP RATE	COOP RATE	ATMTS	ATMTS HOURS	AVG ATMTS IW	AVG IWLEN	PROD HOURS IW	PROD HOURS
2	Total	2301	231	189	14	33	17	300	9	870	100	136	26	376	1338	210	1144	0.75	0.45	0.51	0.87	8458	2.51	9.72	-	3.88	3372
3	R-ALL	1200	0	86	12	18	10	250	7	538	67	133	21	58	862	144	716	0.69	0.47	0.65	0.86	6184	-	11.49	76.23	-	-
4	R-W1	246	0	25	6	6	1	76	1	42	25	46	2	16	129	46	86	0.35	0.18	0.49	0.61	566	-	13.48	75.04	-	-
5	R-W2	954	0	61	6	12	9	174	6	496	42	87	19	42	733	98	630	0.76	0.54	0.69	0.89	5618	-	11.33	76.33	-	-
6	P	1101	231	103	2	15	7	50	2	332	33	3	5	318	476	66	428	0.86	0.42	0.35	0.9	2274	-	6.85	43.55	-	-
7	R-comp	416	0	25	0	4	2	34	1	315	15	8	8	4	386	39	362	0.9	0.76	0.84	0.93	2749	-	8.73	77.99	-	-
8	R-match	541	0	46	8	11	5	142	4	178	43	66	10	28	372	85	286	0.58	0.35	0.57	0.77	2793	-	15.69	75.17	-	-
9	R-ongoing	243	0	15	4	3	3	74	2	45	9	59	3	26	104	20	68	0.38	0.21	0.49	0.79	642	-	14.27	68.03	-	-
10	Lab-Tracking-All	958	0	78	12	12	12	300	7	286	44	135	9	63	579	95	437	0.59	0.32	0.51	0.84	4161	-	14.55	63.41	-	-
11	Lab-Tracking-R	750	0	60	11	8	9	250	6	189	33	133	8	43	430	79	307	0.51	0.27	0.49	0.82	3419	-	18.09	73.07	-	-
12	Lab-Tracking-P	208	0	18	1	4	3	50	1	97	11	2	1	20	149	16	130	0.84	0.52	0.55	0.89	742	-	7.65	44.59	-	-





# Reports: Field Progress Reports

By interviewer report

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Study:04/16/2015	person type	TOT SAMP	NO ATMT	NO CONT	CONT NOREIS	CONT RESIS	APPT	TRACK	HOLD	FI	NIREF	NINO CONT	NIOTH	NS	CONT	RESIS	CONTR	IW RATE	RESP RATE	COMP RATE	COOP RATE	ATMTS	ATMTS HOURS	AVG IW	AVG IWLEN	PROD HOURS IW	PROD HOURS
2	Total		2301	231	189	14	33	17	300	9	870	100	136	26	376	1338	210	1144	0.75	0.45	0.51	0.87	8458	2.51	9.72	-	3.88	3372
3	Leigh Baines	Total	65	0	1	0	0	0	1	0	57	2	0	0	4	259	20	158	0.97	0.93	0.91	0.97	657	3.3	11.53	-	3.49	199
4		R	41	0	0	0	0	0	1	0	35	1	0	0	4	106	14	106	0.97	0.95	0.88	0.97	476	-	13.6	70.26	-	-
5		P	24	0	1	0	0	0	0	0	22	1	0	0	0	52	6	52	0.96	0.92	0.96	0.96	181	-	8.23	39.69	-	-
6	Amanda Brighton	Total	102	0	18	0	1	0	21	0	59	2	0	0	1	217	17	155	0.95	0.58	0.61	0.97	433	1.68	7.34	-	4.36	257
7		R	75	0	6	0	1	0	21	0	45	1	0	0	1	121	11	121	0.96	0.61	0.63	0.98	351	-	7.8	72.96	-	-
8		P	27	0	12	0	0	0	0	0	14	1	0	0	0	34	6	34	0.93	0.52	0.56	0.93	82	-	5.86	38.71	-	-
9	Leigh deRamos	Total	18	0	0	0	1	0	0	0	17	0	0	0	0	56	3	37	0.94	0.94	1.0	1.0	109	1.56	6.41	-	4.12	70
10		R	14	0	0	0	1	0	0	0	13	0	0	0	0	32	3	32	0.93	0.93	1.0	1.0	93	-	7.15	89.75	-	-
11		P	4	0	0	0	0	0	0	0	4	0	0	0	0	5	0	5	1.0	1.0	1.0	1.0	16	-	4.0	42.08	-	-
12	Nancy Dryden	Total	225	0	21	2	4	2	41	0	124	24	1	5	1	456	45	297	0.78	0.55	0.7	0.81	772	1.53	6.23	-	4.06	504
13		R	144	0	13	1	3	2	35	0	69	17	0	4	0	192	32	192	0.74	0.48	0.65	0.77	516	-	7.48	78.82	-	-
14		P	81	0	8	1	1	0	6	0	55	7	1	1	1	105	13	105	0.85	0.69	0.8	0.87	256	-	4.65	46.36	-	-
15	Thomas Frank	Total	133	0	11	0	10	0	10	1	73	26	0	0	2	495	69	300	0.67	0.56	0.82	0.74	999	3.78	13.68	-	3.62	264
16		R	80	0	7	0	6	0	10	0	42	13	0	0	2	190	43	190	0.69	0.54	0.76	0.76	725	-	17.26	72.57	-	-
17		P	53	0	4	0	4	0	0	1	31	13	0	0	0	110	26	110	0.65	0.58	0.91	0.7	274	-	8.84	41.0	-	-
18	Kelly Groves	Total	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	0.0	0.0	-	0.0	0
19		R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	-	0.0	0.0	-	-
20		P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	-	0.0	0.0	-	-
21	Mia Hamm	Total	48	0	14	0	1	0	19	3	4	4	0	0	3	25	6	15	0.44	0.09	0.19	0.5	96	8.73	24.0	-	2.75	11
22		R	17	0	1	0	0	0	5	3	4	3	0	0	1	13	3	13	0.57	0.25	0.41	0.57	86	-	21.5	78.33	-	-
23		P	31	0	13	0	1	0	14	0	0	1	0	0	2	2	3	2	0.0	0.0	0.06	0.0	10	-	0.0	0.0	-	-



# Reports: Field Progress Reports

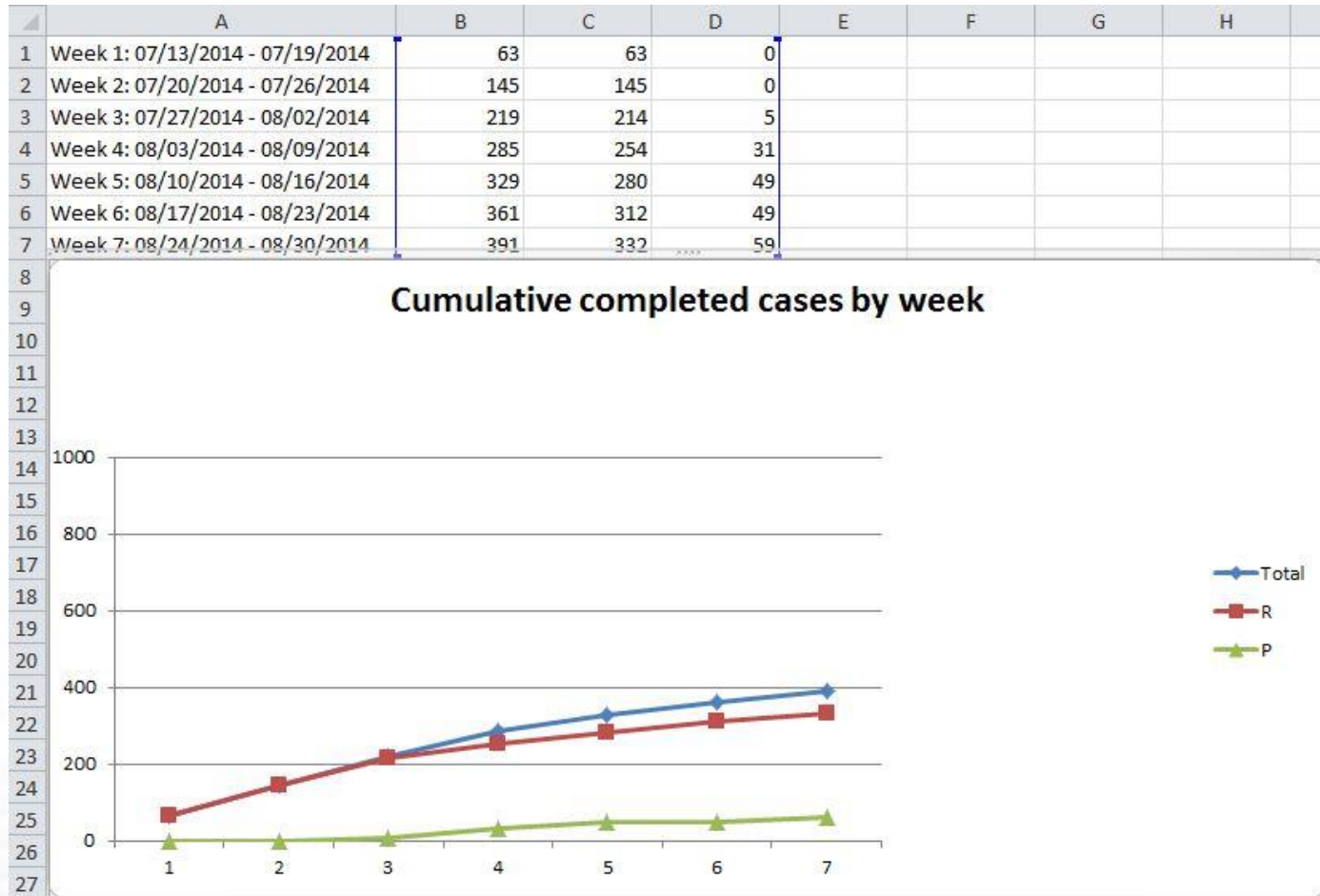
## By week report

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Study:04/16/2015	person type	TOT SAMP	NO ATMT	NO CONT	CONT NOREIS	CONT RESIS	APPT	TRACK	HOLD	FI	NIREF	NINO CONT	NIOTH	NS	CONT	RESIS	CONTR	IW RATE	RESP RATE	COMP RATE	COOP RATE	ATMTS	ATMTS HOURS	AVG ATMTS IW	AVG IWLEN	PROD HOURS IW	PROD HOURS
2	Total		2301	231	189	14	33	17	300	9	870	100	136	26	376	1338	210	1144	0.75	0.45	0.51	0.87	8458	2.51	9.72	-	3.88	3372
3	Week 1: 07/13/2014 - 07/19/2014	Total	78	6	0	0	0	0	0	0	63	2	0	3	4	235	11	185	0.93	0.85	0.87	0.93	499	3.28	7.92	-	2.41	152
4		R	72	0	0	0	0	0	0	0	63	2	0	3	4	185	11	185	0.93	0.93	0.94	0.93	499	-	7.92	81.0	-	-
5		P	6	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	-	0.0	0.0	-	-
6	Week 2: 07/20/2014 - 07/26/2014	Total	95	7	0	0	0	0	0	0	82	0	0	3	3	297	10	228	0.96	0.89	0.89	0.96	659	2.93	8.04	-	2.74	225
7		R	88	0	0	0	0	0	0	0	82	0	0	3	3	228	10	228	0.96	0.96	0.97	0.96	659	-	8.04	79.12	-	-
8		P	7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	-	0.0	0.0	-	-
9	Week 3: 07/27/2014 - 08/02/2014	Total	112	12	0	0	0	0	0	0	74	0	0	1	25	292	16	221	0.99	0.85	0.67	0.99	666	3.03	9.0	-	2.97	220
10		R	75	0	0	0	0	0	0	0	69	0	0	1	5	203	15	203	0.99	0.99	0.93	0.99	630	-	9.13	76.19	-	-
11		P	37	12	0	0	0	0	0	0	5	0	0	0	20	18	1	18	1.0	0.29	0.14	1.0	36	-	7.2	38.66	-	-
12	Week 4: 08/03/2014 - 08/09/2014	Total	77	3	0	0	0	0	0	0	66	0	0	3	5	282	16	191	0.96	0.92	0.9	0.96	688	3.51	10.42	-	2.97	196
13		R	44	0	0	0	0	0	0	0	40	0	0	1	3	123	11	123	0.98	0.98	0.93	0.98	515	-	12.88	78.59	-	-
14		P	33	3	0	0	0	0	0	0	26	0	0	2	2	68	5	68	0.93	0.84	0.85	0.93	173	-	6.65	42.21	-	-
15	Week 5: 08/10/2014 - 08/16/2014	Total	59	2	0	0	0	0	0	0	44	0	0	2	11	241	10	152	0.96	0.92	0.78	0.96	631	3.67	14.34	-	3.91	172
16		R	31	0	0	0	0	0	0	0	26	0	0	2	3	111	9	111	0.93	0.93	0.9	0.93	528	-	20.31	87.11	-	-
17		P	28	2	0	0	0	0	0	0	18	0	0	0	8	41	1	41	1.0	0.9	0.64	1.0	103	-	5.72	41.13	-	-
18	Week 6: 08/17/2014 - 08/23/2014	Total	62	13	0	0	1	0	0	0	32	1	0	1	14	167	9	102	0.91	0.67	0.56	0.94	432	3.32	13.5	-	4.06	130
19		R	35	0	0	0	1	0	0	0	32	1	0	1	0	99	9	99	0.91	0.91	1.0	0.94	429	-	13.41	78.23	-	-
20		P	27	13	0	0	0	0	0	0	0	0	0	14	3	0	3	0	0.0	0.0	0.0	0.0	3	-	0.0	0.0	-	-
21	Week 7: 08/24/2014 - 08/30/2014	Total	42	4	0	0	0	0	0	1	30	1	0	1	5	118	9	80	0.94	0.81	0.76	0.94	277	2.25	9.23	-	4.1	123
22		R	26	0	0	0	0	0	0	1	20	1	0	1	3	50	6	50	0.91	0.87	0.85	0.91	201	-	10.05	82.39	-	-
23		P	16	4	0	0	0	0	0	0	10	0	0	0	2	30	3	30	1.0	0.71	0.63	1.0	76	-	7.6	42.84	-	-



# Reports: Field Progress Reports

Cumulative completed cases graph





# Reports: Payment Reports

- Study has “journals” that respondent completes, payment amount varies based on number of journals completed.
- Key Python: pyodbc, openpyxl



# Reports: Payment Reports

	A	B	C	D	E	F	G	H	I
1	Study: 04/22/2015	personType	COMP_JOURNAL	EARN_AMT	PAID_AMT	PAID_DATE	PEND_AMT	REF_AMT	REFUSED
2	Total		2007	34640	31940		2700	45	
3	2021040120	R	7	75	70	2015-03-12	5		
4	1078064120	R	4	50	50	2015-03-12	0		
5	1042127120	R	0	25	25	2014-07-21	0		
6	1078049110	R	0	25	25	2014-07-27	0		
7	1085102120	R	5	55	50	2015-03-12	5		
8	1096019120	R	0	25	25	2014-07-27	0		
9	1038216120	R	1	30	30	2015-04-12	0		
10	1077051120	R	0	25	25	2014-09-21	0		
11	1034093120	R	2	40	40	2015-03-15	0		
12	1071006120	R	0	25	25	2014-11-23	0		
13	1005026120	R	0	25	25	2014-12-21	0		
14	1066214120	R	5	50	50	2015-04-12	0		
15	1058082120	R	0	25	25	2014-08-18	0		
16	1023116120	R	0	40	40	2015-03-22	0		
17	2097050120	R	2	40	40	2015-03-15	0		
18	1095002120	R	6	60	55	2015-03-12	5		
19	1058067120	R	0	25	25	2014-08-24	0		
20	1011058110	R	1	30	30	2015-04-12	0		
21	2069092110	R	2	40	40	2015-03-15	0		
22	1047155120	R	0	25	25	2014-08-04	0		
23	1035051110	R	1	35	35	2015-04-12	0		
24	1056023140	R	7	75	70	2015-03-12	5		
25	1088114120	R	0	25	25	2014-09-07	0		
26	1010166120	R	1	30	30	2015-04-19	0		
27	1020166120	R	0	25	25	2014-08-10	0		
28	1013090120	R	0	25	25	2014-08-04	0		
29	1024039120	R	7	75	70	2015-03-12	5		
30	1031044120	R	2	40	40	2015-03-15	0		
31	1035087120	R	0	40	40	2015-04-05	0		
32	1006015120	R	2	35	35	2015-04-19	0		





# Data Processing: Swapping IDs

- Study has an internal ID and a “research” id.
- Data comes in with internal ID, needs “research” id.
- Files are tab delimited from bio specimen data and are very large, around 20GB. Programs like SAS cannot handle easily.

# Data Processing: Swapping IDs

- 3 Different Python scripts depending on data structure:
  - The ID to be swapped is in the column headers and does **not** need to be parsed
  - The ID to be swapped is in the column headers and **needs** to be parsed. (i.e. file header contains [OLDID].[TEXT] )
  - The ID to be swapped is a column of data with an id for each row
- Key Python: All three scripts rely on a crosswalk .txt file with OldID-NewID pairs. Python's dictionary function is used.



# Tools: Interviewer Location Geo-Coding

- Used to map the closest Primary Sample Units (PSU) for interviewing staff applicants.
- Inputs:
  - Excel file with interviewer location information: Address, City, State, Zip code, Latitude & Longitude.
  - Excel file of PSU location information: Zip code, Latitude & Longitude.
- Outputs:
  - Excel file with 5 closest PSU to the interviewer's location based on latitude and longitude.
  - XML file of 4 closest PSU for uploading to website
  - Google map of interviewer and PSU location.
- Key Python: geopy, wxpython





# Tools: Interviewer Location Geo-Coding

## Applicant Address File

	A	B	C	D	E	F	G	H	I	J
1	rec_recruit_ID	Project	Appl_ID	UM_Employed	Resume	LastName	FirstName	MiddleName	Address1	Address2
2		1 FAKE	A0017998			Panda		Blue	300 Packard Street	
3		2 FAKE	A0017999			Lion		Red	3158 Williamsburg Rd	
4		3 FAKE	A0018000			Monkey		Green	12048 W Ellsworth Rd	

K	L	M	N	O	P	Q	R	S	T	U
City	State	Zip	PSU	PSUName	PSU_2	PSUName_2	PSU_3	PSUName_3	PSU_4	PSUName_4
Ann Arbor	MI	48105	'000	TEMP HOLDER						
Ann Arbor	MI	48108	'000	TEMP HOLDER						
Manchester	MI	48158	'000	TEMP HOLDER						



# Tools: Interviewer Location Geo-Coding

Study PSU file

	A	B	C	D	E
1	<b>PSUno</b>	<b>PSUName</b>	<b>ZIP</b>	<b>Latitude</b>	<b>Longitude</b>
2	101	PSU-101	48105	42.27758	-83.7337
3	102	PSU-102	48103	42.23323	-83.874
4	103	PSU-103	48198	42.20844	-83.639
5	104	PSU-104	48239	42.3794	-83.2911
6	105	PSU-105	48210	42.331	-83.1279



# Tools: Interviewer Location Geo-Coding

DCO geocoding tool

----- STEP1 Interviewer location -----

Input Files: L:\groups\TSG\dataops\EmilyB\IFDTC\addr\_test.xlsx

Select Sheet: 000s

Iwer ID: Appl\_ID

Address1: Address1

City: City

State: State

ZipCode: Zip

Status: RecruitStatusText



# Tools: Interviewer Location Geo-Coding

DCO geocoding tool

----- STEP2 PSU location -----

PSU Files: L:\groups\TSG\dataops\EmilyB\IFDTC\psu\_test.xlsx Browse...

Select Sheet: FAMID

PSU ID: PSUno

PSU Name: PSUName

PSU Latitude: Latitude

PSU Longitude: Longitude

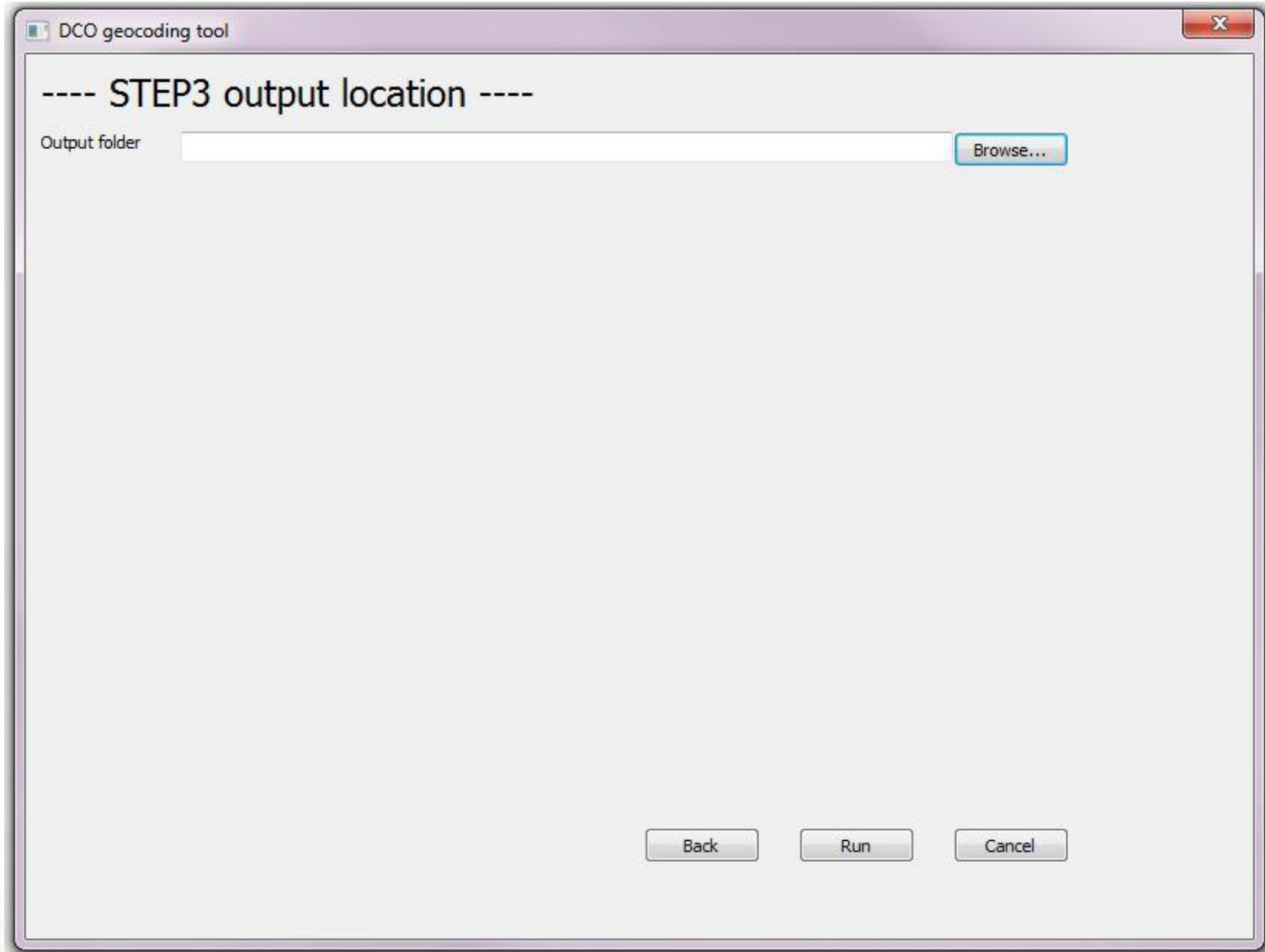
Calculate travel distance

PSU Radius: 50 miles

Back Next Cancel



# Tools: Interviewer Location Geo-Coding





# Tools: Interviewer Location Geo-Coding

Applicant file with 5 closest PSU

	A	B	C	D	E	F	G	H
1	Appl_ID	Address1	City	State	Zip	Status	Latitude	Longitude
2	A0017998	300 Packard Street	Ann Arbor	MI	48105	Screening Done	42.27553	-83.7463162
3	A0017999	3158 Williamsburg Rd	Ann Arbor	MI	48108	Screening Done	42.23427	-83.6962266
4	A0018000	12048 W Ellsworth Rd	Manchester	MI	48158	Screening Done	42.22467	-83.9643605

	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
s	PSU_no1	PSU_name1	PSU_dist1	PSU_no2	PSU_name2	PSU_dist2	PSU_no3	PSU_name3	PSU_dist3	PSU_no4	PSU_name4	PSU_dist4	PSU_no5	PSU_name5	PSU_dist5
0	101	PSU-101	0.659846	102	PSU-102	7.167935631	103	PSU-103	7.193128151	104	PSU-104	24.39183994	105	PSU-105	31.91551842
0	103	PSU-103	3.43552	101	PSU-101	3.55413709	102	PSU-102	9.118112458	104	PSU-104	23.04646892	105	PSU-105	29.883451
0	102	PSU-102	4.673263	101	PSU-101	12.37674171	103	PSU-103	16.73131953	104	PSU-104	36.11110049	105	PSU-105	43.49750868





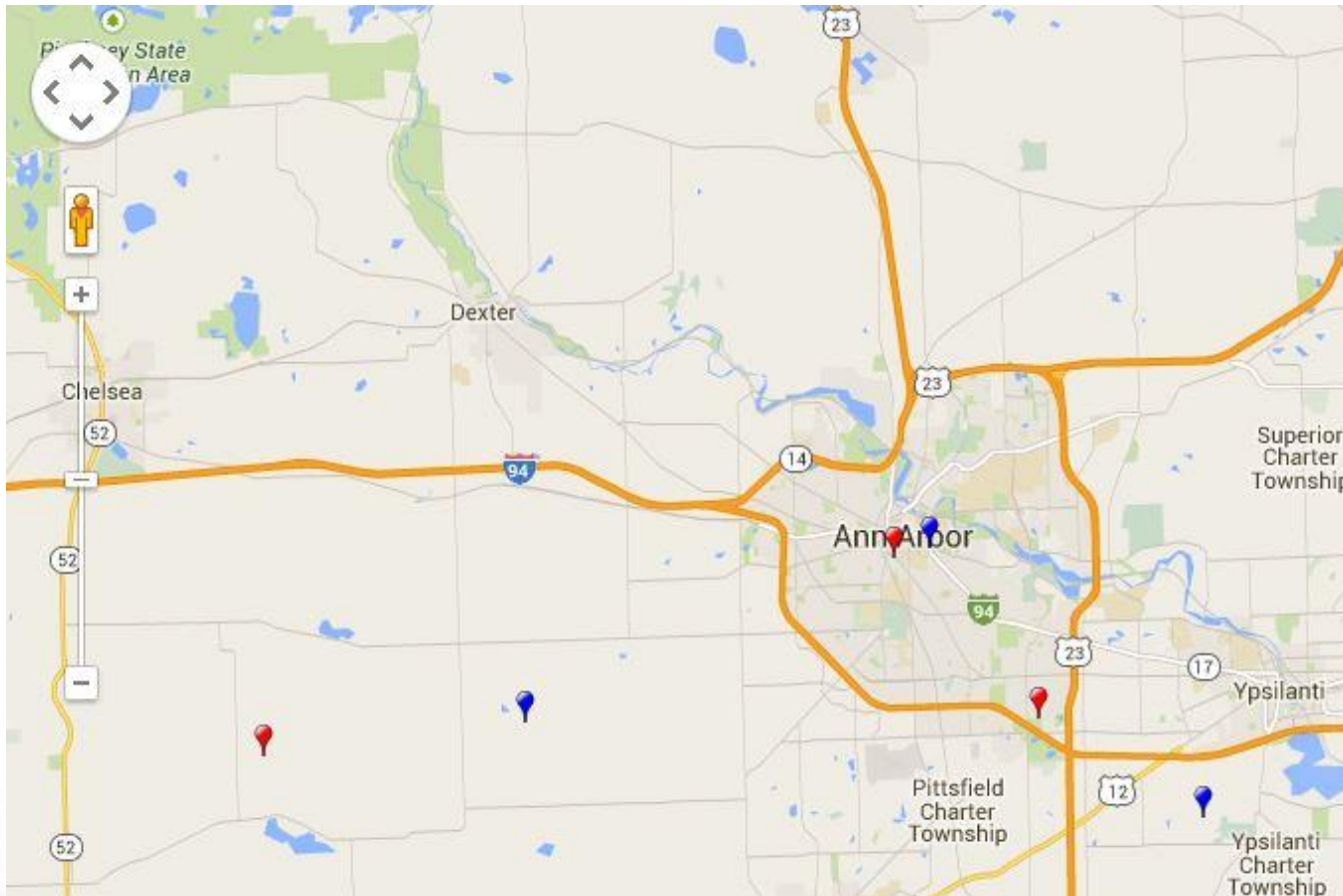
# Tools: Interviewer Location Geo-Coding

## XML file

```
datamapping.xml x
1  <?xml version="1.0" encoding="us-ascii" standalone="yes"?>
2  <data>
3  <o>
4      <v>A0017998</v>
5      <v>101</v>
6      <v>102</v>
7      <v>103</v>
8      <v>104</v>
9  </o>
10 <o>
11     <v>A0017999</v>
12     <v>103</v>
13     <v>101</v>
14     <v>102</v>
15     <v>104</v>
16 </o>
17 <o>
18     <v>A0018000</v>
19     <v>102</v>
20     <v>101</v>
21     <v>103</v>
22     <v>104</v>
23 </o>
24 </data>
```

# Tools: Interviewer Location Geo-Coding

Map



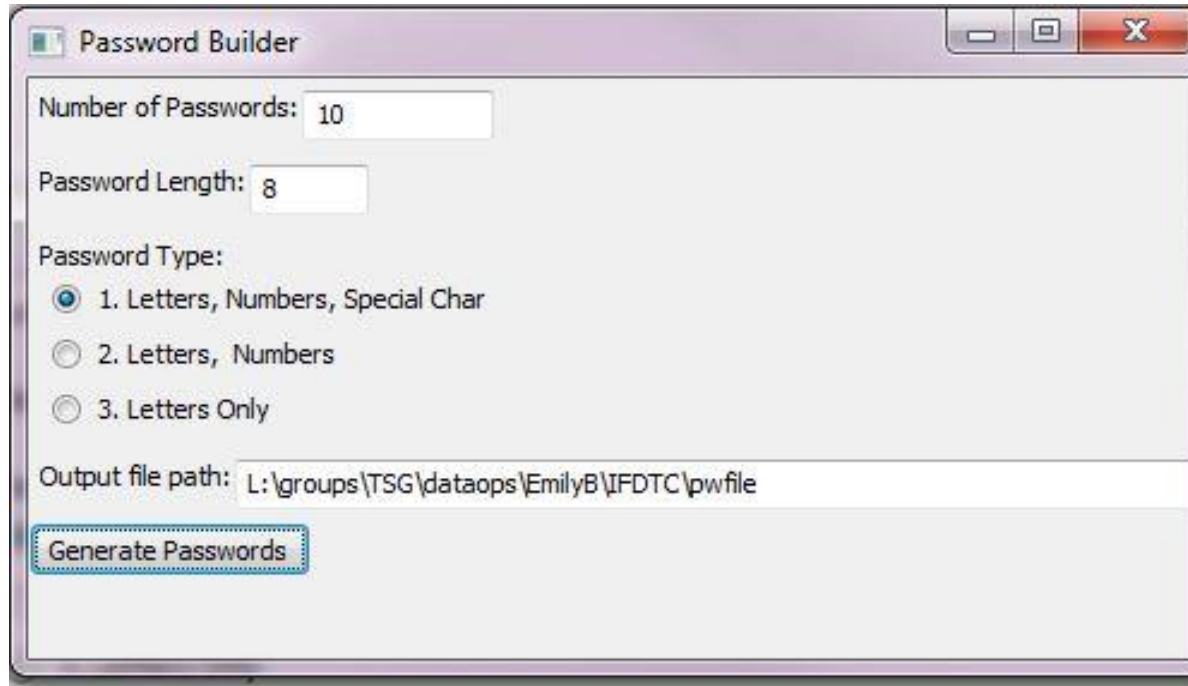


# Tools: Password Builder

- Used to create login credentials for web surveys.
- User defines:
  - Length (number of characters)
  - Characters allowed (Numbers, letters, special characters)
  - Number of passwords needed
  - Output file location
- Will not make duplicate passwords.
- Will use each character type specified.
- Key Python: random, wxpython



# Tools: Password Builder





# Tools: Password Builder

Outputs passwords

```
1 9ud7#rsa|
2 !8uktzuq
3 pjy%x6a4
4 2jmswb#k
5 pv#kerme
6 @3stn59a
7 hgk4*j5v
8 #hr6newy
9 u%88y6er
10 yk5mhw@8
```



# Some downsides...

- Setting up Python tools requires the correct modules are installed on each computer and the same version is used.
- Runs slower than other languages, such as C.



# Conclusion and Summary

- Python is flexible and multipurpose.
- Open source: free to use; builds a welcoming, supportive, and helpful online community of users.
- Many more options available, such as statistical analysis.



# Resources

- Install Python: <https://www.python.org/>
- Free Python Course:  
<http://www.codecademy.com/en/tracks/python>
- StackOverflow: Question and answer site for programmers <http://stackoverflow.com/>



**INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER**  
**SURVEY RESEARCH OPERATIONS**  
UNIVERSITY OF MICHIGAN

# Thank You!

Email: [emblasz@umich.edu](mailto:emblasz@umich.edu)