



Final SIP

Ally Grishko, Lillian Kurz, Isaiah Sarria, Randy Koliha

```
similar_streets.append(other_value)
value == other_value:
similar_streets.append(other_value)
similar_streets
```

ll be a column whose rows (account) contains a list of strings that are similar to that account's street name, and it will be done for every row. The function takes an arbitrary value and a data series' column. It creates an empty list. A second value is created that becomes each entry in the column. If the value doesn't equal that entry and if an arbitrary value's string is in the row's entry or vice versa, then add the row's value from the column to the empty list.

```
similar_streets_list = []
df["Street"] = df["Street"]
similar_streets_list = check_similar_streets(value, df["Street"])
similar_streets_list.append(similar_streets)
df["Street"] = similar_streets_list
```

to create an empty list to store our results from the function we created later. Remember, our function takes an arbitrary value as one of the arguments. So, we create that value as it iterates through the 'Street' column to get compared to other_value, another iterated value created from the column. We store the results of the function in the empty list and then turn that list into a new column in the data frame.

```
def check_similar_streets_id(value, series, customer_series):
similar_streets = []
for other_value in enumerate(series):
if value != other_value and (value in other_value or other_value in value):
similar_streets.append(customer_series[i])
value == other_value:
similar_streets.append(customer_series[i])
similar_streets
```

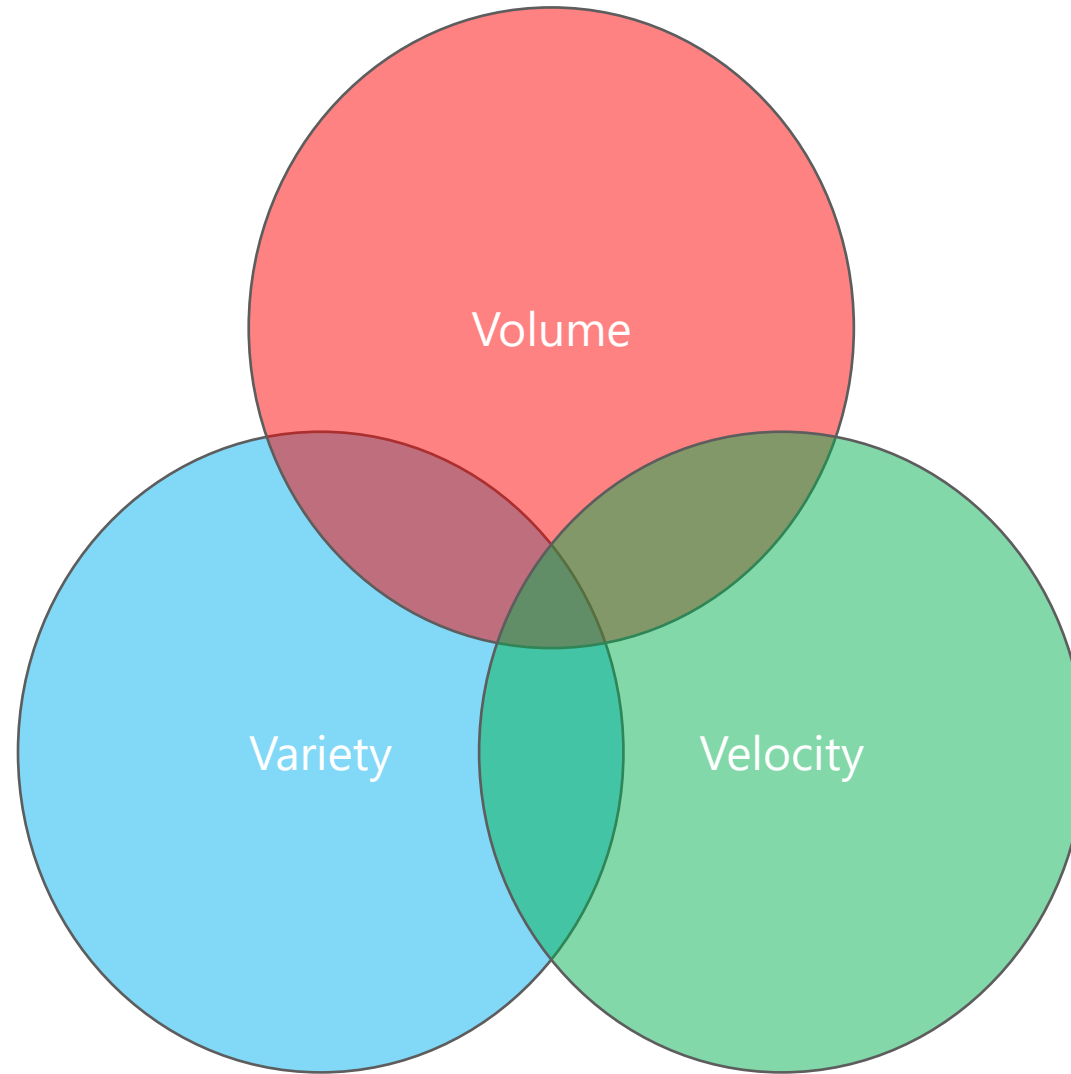
to match the similar names we found with their Customer IDs. This function takes an arbitrary value, and two data series. The first series will be what we want to match the Customer IDs to, and the second series will be the Customer IDs. The function will create an empty list to store the results. It takes an arbitrary value in an enumerated series. An enumerated series tells us the index and value of each row/account in a column. Then we check if the value is similar to other values like we did in the previous function created. For only those similar values, when then only add to the empty list. Finally, we return the list of indexes in the Customer column so that they could be matched to the ID.

Information Governance Project: Customer Duplicates

Context



The Customer Data



Data Variety



CUSTOMER ID



NAMES



LOCATION



BUSINESS
ACCOUNT DETAILS



SUPPRESSION
REASON

Example Problems

Inconsistent Naming Practices

Customer		Name	Street
0	1	Mickey Mouse's BBQ	863 Disney Street
1	2	Mickey's BBQ	353 Disney St
2	3	Mario Market	863 Luigi Lane
3	4	Mario Bros Market	863 Luigi Ln

Redundancies

Customer		Name	Street	Suppression Reason
0	1	Bugs Bunny Brewery	1563 Looney Way	
1	2	Bugs Bunny Brewery	1563 Looney Way	
2	3	Bugs Bunny Brewery	1563 Looney Way	S

Incorrect Suppression

Customer	Suppression Reason	Equipment?	Secondary Volume
500314783		N	0
600396599	S	Y	10

Customer Duplicate Removal

Problem Statement

Duplicate customer accounts are causing operational **inefficiencies**.

Commercial Capabilities

2.4.0 Incentives

5.1.0 Equipment Management

Advantages

Trusted data for associates

Better equipment management

Decrease costs of data maintenance



Methods

The Python Algorithm



Pulling and Importing



RAW CUSTOMER
DATA



EQUIPMENT



VOLUME



COORDINATES



LAST DELIVERY DATE

611	500623215	M	018	A3939	A3939	20	1013	1013	INCL	SECONI	N	-80.67	28.37	0	10	TRUE	TRUE	[1520 N [50062 [EMER [50062 1520 N	693	3	693_M	3		
4340	600337030	C	014	A0302	A3939	20	1001	1013	INCL	SUPP R	N	-80.6	28.41	0	9	FALSE	FALSE	[1 CHRS [60033 [CPARN [60033 1 CHRS	9	3	3_CAR	1		
4384	600339204	RI	000	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.77	28.41	0	0	FALSE	TRUE	[101GE [60033 [TRICKS [60033 101GEO	30	1	30_RIC	1		
3630	600339016	TE	013	A3939	A3939	20	1001	1013	INCL	SUPP R	N	-80.62	28.05	0	9	FALSE	FALSE	[1520 S [60033 [TROPR [60033 1520 S	9	300	300_TE	1		
1483	501087057	TE	020	A3939	A3939	20	1013	1013	INCL	SECONI	N	-80.63	28.04	0	0	FALSE	FALSE	[1521 BI [501081 [TED R [501081 1521BR	301	1	301_TEI	1		
4368	600237865	KE	014	A3939	A3939	20	1001	1013	INCL	SECONI	Y	-80.65	28.09	0	4	FALSE	FALSE	[1523 W [60033 [KERES [60033 1523 W	302	1	302_KE	1		
1465	501050515	AL	020	A3939	A3939	20	1013	1013	INCL	SECONI	N	-80.63	28.04	0	25	FALSE	FALSE	[1528 P [501081 [ALPIS [501081 1528 PA	303	1	303_AL	1		
2574	6003391472	SL	007	A0023	N1252	20	1001	1013	SREP	CLUB	C	-80.57	28.11	570	0	6003391472	6/8/2023	FALSE	TRUE	[1530 N [500515 [SUN F [60033 1530 N	304	1	304_SL	1
268	60015157	SL	018	A3939	A3939	20	1001	1013	INCL	SUPP R	N	-80.57	28.1	0	0	FALSE	TRUE	[1530 N [500515 [SUN F [60033 1530 N	305	1	305_SL	1		
2188	501322909	M	022	A3939	A3939	20	1013	1013	INCL	SECONI	N	-80.72	28.33	0	0	FALSE	FALSE	[1530 U [50133 [NANC [50133 1530 US	306	1	306_M	1		
2336	600334250	V	391	A0149	A3939	20	1001	1013	SREP	VINND	Y	-80.84	28.64	14,768	0	FALSE	TRUE	[1535 N [60033 [VINND [60033 1535 N	307	1	307_V	1		
5070	600339465	V	002	A0556	A3939	20	1001	1013	SREP	SUPP R	N	-80.85	28.64	542	0	600339465	6/3/2023	FALSE	FALSE	[1537 M [60033 [YAMM [60033 1537 M	309	1	309_V	1
5036	600339350	ET	336	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.61	28.41	0	0	FALSE	TRUE	[101GE [60033 [STERL [60033 101GEO	31	1	31_ETE	1		
2575	6003391477	JC	007	A0415	A3939	20	1001	1013	EDI	SUPP R	N	-80.65	28.09	0	0	FALSE	FALSE	[1537 W [60033 [JOAHM [60033 1537 W	1	310	310_JO	1		
3645	600339566	OE	310	A0432	A3939	20	1001	1013	INCL	SECONI	N	-80.65	28.09	0	0	FALSE	FALSE	[1538 M [60033 [OFFIC [60033 1538 HA	311	1	311_OFF	1		
2279	501433320	AL	022	A0224	A3939	20	1001	1013	SREP	ALDI	L	-80.65	28.09	1,357	0	FALSE	FALSE	[1539 W [50143 [ALDI B [50143 1539 W	312	1	312_ALI	1		
3436	600334381	DA	014	A0376	A3939	20	1001	1013	ECOM	DOHBY	Y	-80.65	28.64	616	0	600334381	6/14/2023	FALSE	FALSE	[1540 N [60033 [DOHBY [60033 1540 N	313	1	313_DO	1
5385	600400635	TE	333	A3939	A3939	20	1001	1013	SREP	SUPP R	Y	-80.63	28.04	0	0	FALSE	FALSE	[1540 P [60040 [TEKAX [60033 1540 PA	314	1	314_TE	1		
4401	600339266	M	014	A3939	A3939	20	1001	1013	INCL	SUPP R	N	-80.67	28.36	0	0	FALSE	FALSE	[1545 E [60033 [TWKE [50065 1545 E	315	1	315_M	1		
4618	600337371	SL	015	A3939	A3939	20	1001	1013	INCL	LIMIT	Y	-80.63	28.13	0	0	FALSE	FALSE	[1547 M [60033 [TUSHI [50033 1547 M	316	1	316_SL	1		
3263	600333396	AL	331	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.66	28.25	0	0	FALSE	FALSE	[1551 U [60033 [ALLEG [60033 1551 U	317	1	317_AL	1		
2007	501027850	M	021	B2392	N1233	20	1000	1013	ECOM	SUPP R	N	-80.59	28.17	0	0	FALSE	TRUE	[1550 FI [501301 [TMR PL [501301 1550 FL	318	1	318_M	1		
2436	501050409	SE	023	B2392	A3939	20	1001	1013	ECOM	BROWP	N	-80.06	28.7	0	20	FALSE	TRUE	[1550 FI [501301 [TEASH [501305 1550 FL	319	1	319_SE	1		
2684	501435141	DI	022	A0115	A3939	20	1001	1013	SREP	DOLLAR	Y	-80.78	28.43	1,069	0	501435141	6/14/2023	FALSE	FALSE	[101 CO [50143 [DOLLAR [50143 101 COC	32	1	32_DI	1
2835	600332388	XX	005	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.53	28.17	0	0	FALSE	FALSE	[1550 H [60033 [YCCAP [60033 1550 HC	320	1	320_XX	1		
3000	600335435	CE	011	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.61	28.07	0	0	FALSE	TRUE	[1550 M [60033 [CORAN [60033 1550 MC	321	1	321_CE	1		
5006	600339247	E	339	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.64	28.13	0	0	FALSE	TRUE	[1550 M [60033 [EALU G [60033 1550 MC	322	1	322_E	1		
2514	600339259	HE	006	A0400	A3939	20	1001	1013	ECOM	BROWP	N	-80.61	28.34	418	326	600339259	6/15/2023	FALSE	TRUE	[1550 N [60033 [EHLTO [60033 1550 N	323	1	323_HE	1
2738	600332081	HH	008	A0400	A3939	20	1001	1013	INCL	SUPP R	N	-80.61	28.34	0	0	FALSE	TRUE	[1550 N [60033 [EHLTO [60033 1550 N	324	1	324_HH	1		
4554	600237755	HH	015	A3939	A3939	20	1001	1013	SRVC	SERVIC	N	-80.61	28.34	0	0	FALSE	TRUE	[1550 N [60033 [EHLTO [60033 1550 N	325	1	325_HH	1		
2650	600237775	HE	008	A3939	A3939	20	1001	1013	SRVC	SERVIC	N	-80.61	28.34	0	0	FALSE	TRUE	[1550 N [60033 [EHLTO [60033 1550 N	326	1	326_HE	1		
3783	600335455	SE	011	A3939	A3939	20	1001	1013	SRVC	SERVIC	N	-80.7	28.33	0	0	FALSE	FALSE	[1550 N [60033 [ENBIA [60033 1550 N	327	1	327_SE	1		
170	600427177	CL	01T	A3939	A3939	20	1001	1013	INCL	LIMIT	Y	-80.14	28.38	0	0	FALSE	TRUE	[1551 N [50042 [TEL JET [50042 1551 N	328	1	328_CL	1		
4643	600339063	TE	015	A3939	A3939	20	1001	1013	CALL	SUPP R	N	-80.14	28.38	0	0	FALSE	TRUE	[1551 N [50042 [TEL JET [50042 1551 N	329	1	329_TE	1		
4187	600339465	BE	013	A3939	A3939	20	1001	1013	INCL	SUPP R	N	-80.62	28.05	0	0	FALSE	TRUE	[1551 N [50042 [TEL JET [50042 1551 N	330	1	330_BE	1		
4654	600338078	PI	015	A0628	A3939	20	1001	1013	SREP	FIVE DE	Y	-80.65	28.08	283	0	600338078	6/14/2023	FALSE	FALSE	[1551 W [60033 [FIVE E [60033 1551 W	331	1	331_PI	1
304	600339459	BE	013	A3939	A3939	20	1001	1013	INCL	LIMIT	Y	-80.16	28.36	0	350	FALSE	FALSE	[1551 W [60033 [FIVE E [60033 1551 W	332	1	332_BE	1		
4035	600336375	DI	013	A0115	A3939	20	1001	1013	SREP	DOLLAR	Y	-80.82	28.56	1,509	0	600336375	6/17/2023	FALSE	FALSE	[1555 C [60033 [TOWKE [60033 1555 CH	332	1	332_DI	1
3864	600339566	M	011	A3939	A3939	20	1001	1013	ECOM	SUPP R	N	-80.67	28.36	0	0	FALSE	FALSE	[1555 E [60033 [TWKE E [60033 1555 E	333	1	333_M	1		
4514	600337830	C	015	A3939	A3939	20	1001	1013	ECOM	SUPP R	N	-80.6	28.02	0	0	FALSE	FALSE	[1555 P [60033 [CARGO [60033 1555 P	334	1	334_C	1		
2757	600339250	BE	008	A0348	A3939	20	1000	1013	EDI	SUPP R	N	-80.65	28.08	0	0	FALSE	FALSE	[1555 W [60033 [EED B [60033 1555 W	335	1	335_BE	1		
3001	600400468	DA	001	A3939	A3939	20	1001	1013	CALL	SUPP R	N	-80.8	28.53	0	0	FALSE	FALSE	[1556 H [60040 [HMRG E [60040 1556 HA	336	1	336_DA	1		
2572	600339167	DA	007	A3939	A3939	20	1001	1013	CALL	SUPP R	N	-80.85	28.64	0	0	FALSE	FALSE	[1557 N [60033 [EDAIRY [60033 1557 N	337	1	337_DA	1		
3056	600339404	TE	391	A3939	A3939	20	1001	1013	CALL	SUPP R	Y	-80.7	28.36	0	0	FALSE	FALSE	[1558 N [60033 [EDAIRY [60033 1558 N	338	1	338_TE	1		
3968	600336000	BE	012	A3939	A3939	20	1001	1013	INCL	SUPP R	N	-80.85	28.64	0	12	FALSE	TRUE	[1561 N [50078 [BIGLE [60033 1561 N	339	1	339_BE	1		
163	500756424	RE	018	A3939	A3939	20	1013	1013	INCL	SECONI	N	-80.62	28.05	0	0	FALSE	TRUE	[1561 N [50078 [BIGLE [60033 1561 N	340	1	340_RE	1		
93	500756424	RE	018	B2392	N1233	20	1001	1013	ECOM	BROWP	N	-80.65	28.64	39	0	500756424	5/3/2023	FALSE	TRUE	[1561 N [50078 [BIGLE [60033 1561 N	341	1	341_RE	1
435	600336483	CH	013	A3939	A3939	20	1001	1013	SRVC	SERVIC	N	-80.65	28.08	0	103	TRUE	FALSE	[1562 W [60033 [CHPO [50030 1562 W	341	1	341_CH	1		
3336	600334832	V	331	A0149	A3939	20	1001	1013	SREP	VINND	Y	-80.82	28.53	6,733	0	600334832	6/13/2023	FALSE	FALSE	[1564 H [60033 [VINND [60033 1564 HA	342	1	342_V	1
103	500351542	LA	007	A3939	A3939	20	1001	1013	CALL	SUPP R	N	-80.64	28.12	0	0	FALSE	FALSE	[1565 S [50035 [LATIN [50035 1565 SA	343	1	343_LA	1		
3738	600339528	DA	011	A3939	A3939	20	1001	1013	INCL	SECONI	Y	-80.67	28.08	0	80	FALSE	FALSE	[1567 DA [60033 [EDAIRY [60033 1567 DA	344	1	344_M	1		
4157	600336430	AA	013	B2392	N1233	20	1001	1013	ECOM	BROWP	Y	-80.82	28.53	0	0	FALSE	FALSE	[1570 H [60033 [TAA A [60033 1570 HA	345	1	345_AA	1		
420	500387868	CF	007	A3939	A3939	20	1001	1013	SREP	SUPP R	N	-80.71	28.38	0	0	FALSE	TRUE	[1570 M [60033 [LOROS [60033 1570 M	346	1	346_CF	1		
2608	600339530	H	007	A3939	A3939	20	1001	1013	INCL	SUPP R	N	-80.71	28.38	0	0	FALSE	TRUE	[1570 N [60033 [TH AND [60033 1570 N	347	1	347_H	1		
1235	500973654	TH	019	B2392	N1237	20	1000	1013	ECOM	SUPP R	N	-80.71	28.39	0	0	FALSE	FALSE	[1570 N [50097 [TROPR [50097 1570 N						

Similarities and Exact Matches

Isaiah's Ice
Cream Igloo
≠
Isaiah's Ice
Igloo

Customer	Name	Similar Names	Similar Names ID
1128	Isaiah's Igloo	<ul style="list-style-type: none">- Isaiah's Igloo- Isaiah's Igloo- Isaiah's Ice Igloo	<ul style="list-style-type: none">- 1128- 1999- 1691

0927 Randy
Rd
≠
0927
Randy Road

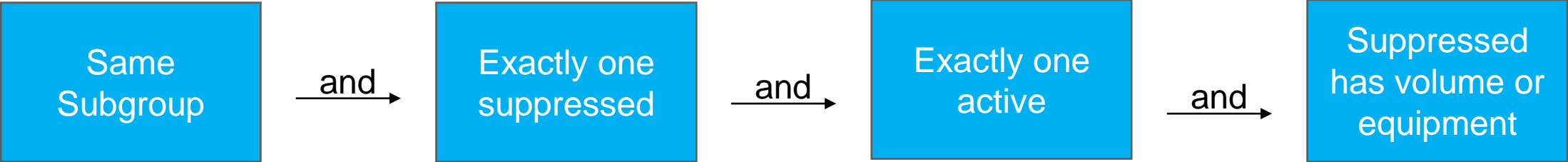
Grouping

Customer	Name	Street
1128	Isaiah's Igloo	0927 Randy E Rd
1999	Isaiah's Igloo	0927 Randy Rd

Customer	Name	Street	Combined	Group	Subgroup
1128	Isaiah's Igloo	0927 Randy E Rd	Isaiah's Igloo 0927 Randy Rd	305	305_ISA
1999	Isaiah's Igloo	0927 Randy Rd	Isaiah's Igloo 0927 Randy Rd	305	305_ISA

Nuances? Then why do it?

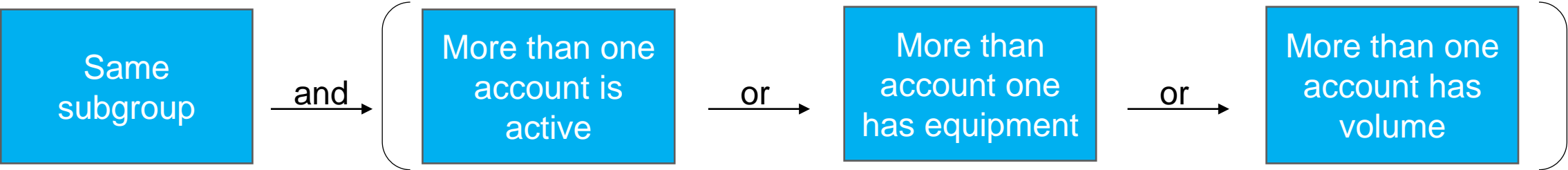
Conditioning



	Suppression Reason	Equipment?	Subgroup	# of Subgroup Members	Secondary Volume	action
15		N	1081_SUN	2	0	yes
1222	S	Y	1081_SUN	2	10	yes



Conditioning



	Name	Street	Subgroup	Suppression Reason	Equipment?	Secondary Volume	action
43	CENTER	WAY	221		N	0	duped
44	CENTER	WAY	221		N	0	duped
45	CENTER	WAY E	221	S	N	0	duped



What do we do with the work?

The Automation Algorithm

Feed in all
sales office
customer data

Import the
created
algorithm

Schedule for
automation

Act on
identified
anomalies



Thank you.
Questions?

