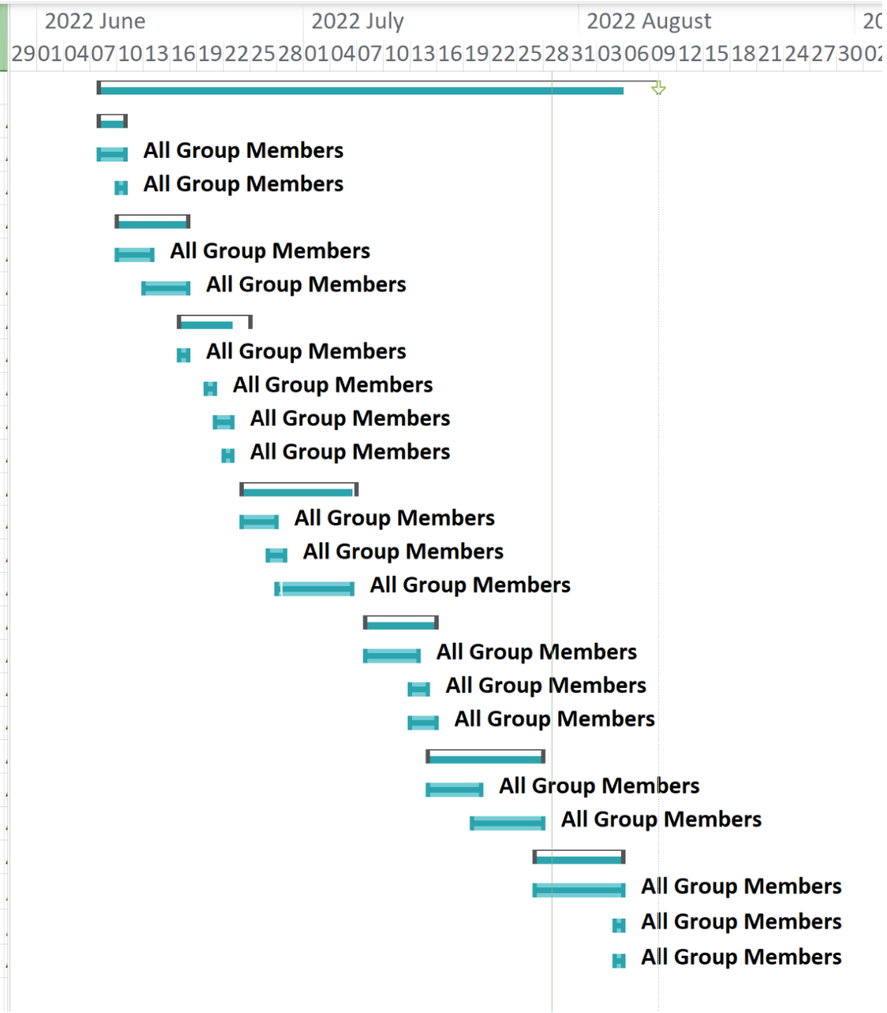
Capstone Project - Canada Post

Prepared by: Ming Liu , Xiaomei Lin, Mi Gan

Project Plan (Gantt Chart)

GANTT CHART

	Task Mode	WBS	Task Name	Duration	Start	Finish
1		CP	**Canada Post Capstone**	**45 days**	**Wed 22-06-08**	**Tue 22-08-09**
2		1	**Phase 1: Business Understanding**	**3 days**	**Wed 22-06-08**	**Fri 22-06-10**
3		1.1	Understand the Capstone requirement	3 days	Wed 22-06-08	Fri 22-06-10
4		1.2	Set up project meet schedule	1 day	Fri 22-06-10	Fri 22-06-10
5		2	**Phase 2: Data Understanding**	**6 days**	**Fri 22-06-10**	**Fri 22-06-17**
6		2.1	Creat Project Plan	2 days	Fri 22-06-10	Mon 22-06-13
7		2.2	Collect Data from Client	5 days	Mon 22-06-13	Fri 22-06-17
8		3	**Phase 3: Data Preparation**	**6 days**	**Fri 22-06-17**	**Fri 22-06-24**
9		3.1	Clean Data	1 day	Fri 22-06-17	Fri 22-06-17
10		3.2	Construt New Data	1 day	Mon 22-06-20	Mon 22-06-20
11		3.3	Integrat Data	2 days	Tue 22-06-21	Wed 22-06-22
12		3.4	Formatt Data	1 day	Wed 22-06-22	Wed 22-06-22
13		4	**Phase 4: Model**	**9 days**	**Fri 22-06-24**	**Wed 22-07-06**
14		4.1	Select Model Techniques	2 days	Fri 22-06-24	Mon 22-06-27
15		4.2	Generat a Test Design	2 days	Mon 22-06-27	Tue 22-06-28
16		4.3	Build Models	6 days	Tue 22-06-28	Wed 22-07-06
17		5	**Phase 5: Evaluation**	**6 days**	**Fri 22-07-08**	**Fri 22-07-15**
18		5.1	Evaluat the Results	4 days	Fri 22-07-08	Wed 22-07-13
19		5.2	Review Process	2 days	Wed 22-07-13	Thu 22-07-14
20		5.3	Refine models	3 days	Wed 22-07-13	Fri 22-07-15
21		6	**Phase 6: Deploymnet and Visualization**	**9 days**	**Fri 22-07-15**	**Wed 22-07-27**
22		6.1	Deploy models	4 days	Fri 22-07-15	Wed 22-07-20
23		6.2	Visualizate Final Report	6 days	Wed 22-07-20	Wed 22-07-27
24		7	**Phase 7: Presentation and Delivery**	**8 days**	**Wed 22-07-27**	**Fri 22-08-05**
25		7.1	Prepare final Reports and Models	8 days	Wed 22-07-27	Fri 22-08-05
26		7.2	Present final Report	1 day	Fri 22-08-05	Fri 22-08-05
27		7.3	Delivery Models and Report	1 day	Fri 22-08-05	Fri 22-08-05



Milestones:

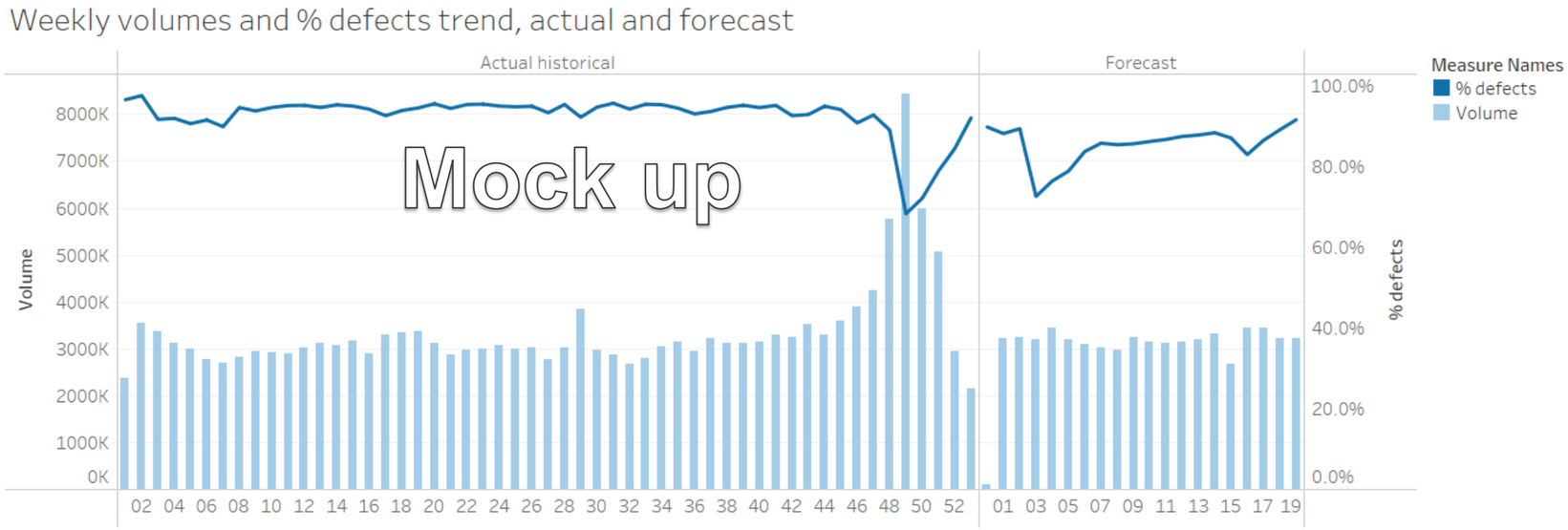
- Phase 1: Business Understanding
- Phase 2: Data Understanding
- Phase 3: Data Preparation
- Phase 4: Model
- Phase 5: Evaluation
- Phase 6: Deployment and Visualization
- Phase 7: Presentation and Delivery （ July 29 Friday)

Phase 1: Business Understanding

- **Project deliverable:**

  Predict % of defects (daily / weekly) based on volume forecast and historical results

  - Predict % of defects by site by day
  - Aggregate daily volume and % defects forecasts by site into weekly
  - Aggregate weekly volume and % defects forecasts by site into National weekly
  - Aggregate National weekly into overall forecast for the full period
  - Output in Tableau or Power BI file



Phase 1: Business Understanding

- **Our project plan:**
  - Understand and merge data from different files provided by clients
  - Add new features to the dataset from other data sources
  - Build machine learning models with different algorithms
  - Tune the models and compare the performances
  - Use the best model as the final model to predict the target
  - Visualize the final result with Tableau

Phase 2: Data Understanding

- **Main data file**
  - Excel sheet 1: 7 columns, over 20,000 records
  - Excel sheet 2: 2 columns, over 1000 records

	A	B	C	D	E	F	G
1	Year	Date	Weekday	Site	Volume (historical)	% of defects	Volume (forecasted)
2	2019	2019-04-29	Monday	Site 1	85552	11.5%	
3	2019	2019-04-29	Monday	Site 2	67843	9.5%	
4	2019	2019-04-29	Monday	Site 3	29843	8.5%	
5	2019	2019-04-29	Monday	Site 4	31268	12.1%	
6	2019	2019-04-29	Monday	Site 5	19470	11.2%	
7	2019	2019-04-29	Monday	Site 6	202101	11.1%	
8	2019	2019-04-29	Monday	Site 7	22208	12.0%	
9	2019	2019-04-29	Monday	Site 8	23163	12.9%	
10	2019	2019-04-29	Monday	Site 9	52541	15.0%	
11	2019	2019-04-29	Monday	Site 10	37777	17.4%	
12	2019	2019-04-29	Monday	Site 11	12571	8.2%	
13	2019	2019-04-29	Monday	Site 12	4208	12.7%	
14	2019	2019-04-29	Monday	Site 13	19133	7.1%	
15	2019	2019-04-29	Monday	Site 14	12277	10.6%	
16	2019	2019-04-29	Monday	Site 15	11999	17.8%	
17	2019	2019-04-29	Monday	Site 16	2399	16.3%	
18	2019	2019-04-29	Monday	Site 17	325425	14.8%	
19	2019	2019-04-29	Monday	Site 18	191729	11.2%	
20	2019	2019-04-29	Monday	Site 19	18152	13.2%	

	A	B
1	Month, Day, Year of Origin Day	Week #
2	2 January, 2019	1
3	3 January, 2019	1
4	4 January, 2019	1
5	5 January, 2019	1
6	6 January, 2019	2
7	7 January, 2019	2
8	8 January, 2019	2
9	9 January, 2019	2
10	10 January, 2019	2
11	11 January, 2019	2
12	12 January, 2019	2
13	13 January, 2019	3
14	14 January, 2019	3
15	15 January, 2019	3
16	16 January, 2019	3
17	17 January, 2019	3
18	18 January, 2019	3
19	19 January, 2019	3
20	20 January, 2019	4

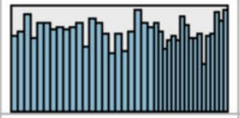
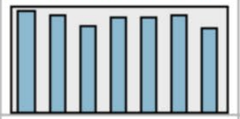
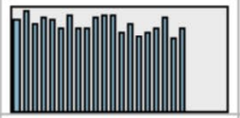
Phase 2: Data Understanding

- **Supporting data file**
  - Flag data, 79 columns, over 3000 records
  - With redundant information, select important features for next step

	A	B	C	D	E	F	G	H	I	J	K	L	M	I
1	CALENDAR_DATE	NATIONAL_HOLIDAY_FLG	PROVINCIAL_HOLIDAY_FLG	PROVINCIAL_HOLIDAY_LIST	IMPACT_DAY_FLG	IMPACT_DATE_DESC_EN	IMPACT_DATE_DESC_FR	IMPACT_DATE_SHORT_DESC_EN	IMPACT_DATE_SHORT_DESC_FR	ONTIME_DP1_FLG	DAYS_TO_CURRENT	CALENDAR_DAY	CAL_DAY_LONG_DESC_EN	CAL_DAY_LO
2	2021-08-28	F	F		F					T	-292	7	Saturday	Samedi
3	2013-01-16	F	F		F					F	-3438	4	Wednesday	Mercredi
4	2013-01-14	F	F		F					F	-3440	2	Monday	Lundi
5	2016-03-13	F	F		F					F	-2286	1	Sunday	Dimanche
6	2015-02-18	F	F		F					F	-2675	4	Wednesday	Mercredi
7	2016-05-15	F	F		F					F	-2223	1	Sunday	Dimanche
8	2019-04-28	F	F		F					F	-1145	1	Sunday	Dimanche
9	2019-12-02	F	F		T	Cyber Monday				T	-927	2	Monday	Lundi
10	2020-06-22	F	F		F					F	-724	2	Monday	Lundi
11	2015-05-19	F	F		F					F	-2585	3	Tuesday	Mardi
12	2017-07-24	F	F		F					F	-1788	2	Monday	Lundi
13	2018-10-21	F	F		F					F	-1334	1	Sunday	Dimanche
14	2018-09-16	F	F		F					F	-1369	1	Sunday	Dimanche
15	2013-08-13	F	F		F					F	-3229	3	Tuesday	Mardi
16	2016-01-16	F	F		F					F	-2343	7	Saturday	Samedi
17	2016-09-30	F	F		F					F	-2085	6	Friday	Vendredi
18	2019-10-03	F	F		F					F	-987	5	Thursday	Jeudi
19	2019-04-08	F	F		F					F	-1165	2	Monday	Lundi
20	2016-08-15	F	T	YT	F					F	-2131	2	Monday	Lundi
21	2016-04-07	F	F		F					F	-2261	5	Thursday	Jeudi
22	2014-09-07	F	F		F					F	-2839	1	Sunday	Dimanche
23	2015-11-22	F	F		F					F	-2398	1	Sunday	Dimanche
24	2022-05-11	F	F		F					F	-36	4	Wednesday	Mercredi
25	2017-06-10	F	F		F					F	-1832	7	Saturday	Samedi
26	2020-02-18	F	F		F					F	-849	3	Tuesday	Mardi
27	2018-12-12	F	F		F					T	-1282	4	Wednesday	Mercredi
28	2013-07-20	F	F		F					F	-3253	7	Saturday	Samedi
29	2018-06-19	F	F		F					F	-1458	3	Tuesday	Mardi
30	2014-03-04	F	F		F					F	-3026	3	Tuesday	Mardi
31	2021-04-10	F	F		F					T	-432	7	Saturday	Samedi
32	2015-09-06	F	F		F					F	-2475	1	Sunday	Dimanche
33	2013-09-01	F	F		F					F	-3210	1	Sunday	Dimanche
34	2018-04-09	F	F		F					F	-1529	2	Monday	Lundi
35	2014-12-15	F	F		F					F	-2740	2	Monday	Lundi
36	2016-05-05	F	F		F					F	-2233	5	Thursday	Jeudi
37	2018-11-20	F	F		F					T	-1304	3	Tuesday	Mardi
38	2022-06-19	F	F		T	Father's Day				F	3	1	Sunday	Dimanche
39	2020-09-19	F	F		F					F	-635	7	Saturday	Samedi
40	2018-05-27	F	F		F					F	-1481	1	Sunday	Dimanche
41	2013-12-19	F	F		F					F	-3101	5	Thursday	Jeudi
42	2014-05-17	F	F		F					F	-2952	7	Saturday	Samedi
43	2021-03-16	F	F		F					T	-457	3	Tuesday	Mardi
44	2020-06-08	F	F		F					F	-738	2	Monday	Lundi
45	2019-12-26	T	F	AB,BC,MB,NB,NL,NS,NT,NU,ON,PE,QC,SK,YT	F					T	-903	5	Thursday	Jeudi

Phase 3: Data Preparation

- **Used IBM SPSS Modeler to observe data**

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Year		Continuous	2019.000	2023.000	2020.950	1.283	0.023	--	31080
Date		Continuous	2019-04-28	2023-07-29	--	--	--	--	31080
Weekday		Categorical	--	--	--	--	--	7	31080
Site		Categorical	--	--	--	--	--	20	31080
Volume (historical)		Continuous	246.000	691572.000	59358.382	82389.930	2.314	--	21840
% of defects		Categorical	--	--	--	--	--	--	21838
Volume (forecasted)		Continuous	183.000	494722.000	49112.315	67274.374	2.302	--	9240

Phase 3: Data Preparation

- Used Excel to merge data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Date	Weekday	Sit	Plant	Provir	ProvinceCo	Metri	Defect	Defect100	Volume(h	SiteClosu	NationalHolidayFl	ImpactDayFl	OntimeDP1Fl	WeekendFl	PeakWeekFl	CalenderYe	CalenderQuar	CalenderMor	CalenderWeek	CalenderDayOfMor	CalenderWeekd	Seas	
2	2019-04-28	Sunday	1	CALGARY MPP	AB	1	0.935	0.0650	650	80352	0	0	0	0	1	0	2019	2	4	17	28	1	1	
3	2019-04-28	Sunday	2	EDMONTON MPP	AB	1	0.96	0.0400	400	63051	0	0	0	0	1	0	2019	2	4	17	28	1	1	
4	2019-04-28	Sunday	3	HALIFAX MPP	NS	6	0.984	0.0160	160	27999	0	0	0	0	1	0	2019	2	4	17	28	1	1	
5	2019-04-28	Sunday	4	HAMILTON MPP	ON	7	0.97	0.0300	300	31810	0	0	0	0	1	0	2019	2	4	17	28	1	1	
6	2019-04-28	Sunday	5	KITCHENER MPP	ON	7	0.983	0.0170	170	24127	0	0	0	0	1	0	2019	2	4	17	28	1	1	
7	2019-04-28	Sunday	6	LEO BLANCHETTE MPP	QC	8	0.964	0.0360	360	143181	0	0	0	0	1	0	2019	2	4	17	28	1	1	
8	2019-04-28	Sunday	7	LONDON MPP	ON	7	0.984	0.0160	160	21757	0	0	0	0	1	0	2019	2	4	17	28	1	1	
9	2019-04-28	Sunday	8	MONCTON MPP	NB	4	0.932	0.0680	680	17826	0	0	0	0	1	0	2019	2	4	17	28	1	1	
10	2019-04-28	Sunday	9	OTTAWA MPP	ON	7	0.961	0.0390	390	30625	0	0	0	0	1	0	2019	2	4	17	28	1	1	
11	2019-04-28	Sunday	10	QUEBEC MPP	QC	8	0.943	0.0570	570	20028	0	0	0	0	1	0	2019	2	4	17	28	1	1	
12	2019-04-28	Sunday	11	REGINA MPP	SK	9	0.952	0.0480	480	9647	0	0	0	0	1	0	2019	2	4	17	28	1	1	
13	2019-04-28	Sunday	12	SAINT JOHN MPP	NB	4	0.963	0.0370	370	4049	0	0	0	0	1	0	2019	2	4	17	28	1	1	
14	2019-04-28	Sunday	13	SASKATOON MPP	SK	9	0.965	0.0350	350	11757	0	0	0	0	1	0	2019	2	4	17	28	1	1	
15	2019-04-28	Sunday	14	ST. JOHNS MPP	NL	5	0.962	0.0380	380	5941	0	0	0	0	1	0	2019	2	4	17	28	1	1	
16	2019-04-28	Sunday	15	SUDBURY MPP	ON	7	0.786	0.2140	2140	9437	0	0	0	0	1	0	2019	2	4	17	28	1	1	
17	2019-04-28	Sunday	16	THUNDER BAY MPP	ON	7	0.809	0.1910	1910	2382	0	0	0	0	1	0	2019	2	4	17	28	1	1	
18	2019-04-28	Sunday	17	TORONTO GATEWAY	ON	7	0.972	0.0280	280	201191	0	0	0	0	1	0	2019	2	4	17	28	1	1	
19	2019-04-28	Sunday	18	VANCOUVER MPP	BC	2	0.973	0.0270	270	135912	0	0	0	0	1	0	2019	2	4	17	28	1	1	
20	2019-04-28	Sunday	19	VICTORIA MPP	BC	2	0.981	0.0190	190	9443	0	0	0	0	1	0	2019	2	4	17	28	1	1	
21	2019-04-28	Sunday	20	WINNIPEG MPP	MB	3	0.975	0.0250	250	51891	0	0	0	0	1	0	2019	2	4	17	28	1	1	
22	2019-04-29	Monday	1	CALGARY MPP	AB	1	0.885	0.1150	1150	85552	0	0	0	0	0	0	2019	2	4	17	29	2	1	
23	2019-04-29	Monday	2	EDMONTON MPP	AB	1	0.905	0.0950	950	67843	0	0	0	0	0	0	2019	2	4	17	29	2	1	
24	2019-04-29	Monday	3	HALIFAX MPP	NS	6	0.915	0.0850	850	29843	0	0	0	0	0	0	2019	2	4	17	29	2	1	
25	2019-04-29	Monday	4	HAMILTON MPP	ON	7	0.879	0.1210	1210	31268	0	0	0	0	0	0	2019	2	4	17	29	2	1	
26	2019-04-29	Monday	5	KITCHENER MPP	ON	7	0.888	0.1120	1120	19470	0	0	0	0	0	0	2019	2	4	17	29	2	1	
27	2019-04-29	Monday	6	LEO BLANCHETTE MPP	QC	8	0.889	0.1110	1110	202101	0	0	0	0	0	0	2019	2	4	17	29	2	1	
28	2019-04-29	Monday	7	LONDON MPP	ON	7	0.88	0.1200	1200	22208	0	0	0	0	0	0	2019	2	4	17	29	2	1	
29	2019-04-29	Monday	8	MONCTON MPP	NB	4	0.871	0.1290	1290	23163	0	0	0	0	0	0	2019	2	4	17	29	2	1	
30	2019-04-29	Monday	9	OTTAWA MPP	ON	7	0.85	0.1500	1500	52541	0	0	0	0	0	0	2019	2	4	17	29	2	1	
31	2019-04-29	Monday	10	QUEBEC MPP	QC	8	0.826	0.1740	1740	37777	0	0	0	0	0	0	2019	2	4	17	29	2	1	
32	2019-04-29	Monday	11	REGINA MPP	SK	9	0.918	0.0820	820	12571	0	0	0	0	0	0	2019	2	4	17	29	2	1	
33	2019-04-29	Monday	12	SAINT JOHN MPP	NB	4	0.873	0.1270	1270	4208	0	0	0	0	0	0	2019	2	4	17	29	2	1	
34	2019-04-29	Monday	13	SASKATOON MPP	SK	9	0.929	0.0710	710	19133	0	0	0	0	0	0	2019	2	4	17	29	2	1	
35	2019-04-29	Monday	14	ST. JOHNS MPP	NL	5	0.894	0.1060	1060	12277	0	0	0	0	0	0	2019	2	4	17	29	2	1	
36	2019-04-29	Monday	15	SUDBURY MPP	ON	7	0.822	0.1780	1780	11999	0	0	0	0	0	0	2019	2	4	17	29	2	1	
37	2019-04-29	Monday	16	THUNDER BAY MPP	ON	7	0.837	0.1630	1630	2399	0	0	0	0	0	0	2019	2	4	17	29	2	1	
38	2019-04-29	Monday	17	TORONTO GATEWAY	ON	7	0.852	0.1480	1480	325425	0	0	0	0	0	0	2019	2	4	17	29	2	1	
39	2019-04-29	Monday	18	VANCOUVER MPP	BC	2	0.888	0.1120	1120	191729	0	0	0	0	0	0	2019	2	4	17	29	2	1	
40	2019-04-29	Monday	19	VICTORIA MPP	BC	2	0.868	0.1320	1320	18152	0	0	0	0	0	0	2019	2	4	17	29	2	1	
41	2019-04-29	Monday	20	WINNIPEG MPP	MB	3	0.916	0.0840	840	63552	0	0	0	0	0	0	2019	2	4	17	29	2	1	
42	2019-04-30	Tuesday	1	CALGARY MPP	AB	1	0.96	0.0400	400	86589	0	0	0	0	0	0	2019	2	4	17	30	3	1	
43	2019-04-30	Tuesday	2	EDMONTON MPP	AB	1	0.95	0.0500	500	78865	0	0	0	0	0	0	2019	2	4	17	30	3	1	
44	2019-04-30	Tuesday	3	HALIFAX MPP	NS	6	0.932	0.0680	680	29808	0	0	0	0	0	0	2019	2	4	17	30	3	1	
45	2019-04-30	Tuesday	4	HAMILTON MPP	ON	7	0.918	0.0820	820	34968	0	0	0	0	0	0	2019	2	4	17	30	3	1	

Table (23 fields, 31,080 record

Phase 3: Data Preparation

- **Used Excel to check missing values and strange values**

A	B	C	D	E	F	G	H	I	J
Ye	Date	Weekday	Site	Volume (historica	% of defects	Volume (forecastec	Week #	National_Holiday_I	Impact_Day_I
2021	2021-11-29	Monday	19	19731	#N/A		48	0	1
2021	2021-11-30	Tuesday	19	19425	#N/A		48	0	1
2021	2021-12-01	Wednesday	18	257904	#N/A		48	0	1
2021	2021-12-01	Wednesday	19	20384	#N/A		48	0	1
2021	2021-12-02	Thursday	19	19481	#N/A		48	0	1

A	B	C	D	E	F	G	H
Ye	Date	Weekday	Site	Volume (historica	% of defects	Volume (forecastec	Week #
2019	2019-08-05	Monday	15	8192	0.0%		32

Ye	Date	Weekday	Site	Volume (historica	% of defects	Volume (forecastec	Week #
2020	2020-03-24	Tuesday	14	1090	99.9%		13
2020	2020-03-25	Wednesday	14	837	100.0%		13
2020	2020-03-26	Thursday	14	727	100.0%		13
2021	2021-11-30	Tuesday	18	276478	100.0%		48
2021	2021-12-02	Thursday	18	242067	100.0%		48
2021	2021-12-03	Friday	19	18402	100.0%		48

Phase 3: Data Preparation

- **Add new features to dataset, e.g. Province and ProvinceCode**

Derive field:

Province

Derive as: Formula

Field type: <Default>

Formula:

```
if Site=1 or Site=2 then "AB"
elseif Site=18 or Site=19 then "BC"
elseif Site=20 then "MB"
elseif Site=8 or Site=12 then "NB"
elseif Site=14 then "NL"
elseif Site=3 then "NS"
elseif Site=6 or Site=10 then "QC"
elseif Site=11 or Site=13 then "SK"
else "ON"
endif
```

Derive field:

ProvinceCode

Derive as: Formula

Field type: <Default>

Formula:

```
if Site=1 or Site=2 then 1
elseif Site=18 or Site=19 then 2
elseif Site=20 then 3
elseif Site=8 or Site=12 then 4
elseif Site=14 then 5
elseif Site=3 then 6
elseif Site=6 or Site=10 then 8
elseif Site=11 or Site=13 then 9
else 7
endif
```

Phase 3: Data Preparation

- **Add new features to dataset, e.g. Lockdown_Flag**

AB

==========

2020: Lockdown and relaunch

March 5: Alberta identifies its first presumptive COVID-19 case, a Calgary woman returning from a California cruise.
March 11: The World Health Organization declares COVID-19 a pandemic, while Alberta begins to recommend against out-of-country travel. The following day, Alberta bans gatherings of more than 250 people.
March 15: Calgary is among Alberta cities to declare a state of local emergency, closing most non-essential businesses and services. Alberta cancels all school classes and declares a provincial public health state of emergency the following day.
March 19: Alberta records its first COVID-19 death, an Edmonton man in his 60s. The province's case count rises to 146, the majority of which originated through out-of-country travel.
April 6: Alberta chief medical officer of health Dr. Deena Hinshaw recommends the use of non-medical face masks when physical distancing is difficult. Modelling suggests a mid-May peak for COVID-19 in Alberta, with around 800,000 total cases in the province. Twenty-four Albertans have died of the virus, with almost 1,350 infections.
April 20: The Cargill meat-packing plant in High River temporarily closes down amid a COVID-19 outbreak. The outbreak would become the largest in Alberta, linked to more than 1,500 case and three deaths. Meanwhile, the JBS meat-processing plant in Brooks would record more than 650 cases.
May 3: Alberta records fewer than 100 cases of COVID-19 for the first time since mid-April, signalling the end of the province's first wave of the virus.
May 13: Alberta enters Stage 1 of its COVID-19 relaunch, letting businesses like restaurants and retailers reopen, but Calgary and Brooks are singled out of the relaunch due to ongoing meat-plant outbreaks. Both municipalities are allowed to join in the relaunch May 25.
June 5: Alberta records only seven new COVID-19 cases while also conducting its highest number of tests to date. Hinshaw lauds the efforts of Albertans to flatten the curve. The province has almost 7,100 virus cases and 151 deaths.
June 12: Stage 2 of Alberta's relaunch begins a week ahead of schedule, with businesses like massage clinics, theatres and libraries allowed to reopen. Calgary's local state of emergency expires after three months, with Alberta following on June 16.
Dec. 8: Lockdown-style restrictions come to Alberta, with many businesses forced to close and all indoor and outdoor social gatherings banned

2021

Jan. 14: Alberta eases restrictions for outdoor gatherings and allows personal services businesses to reopen as case counts slowly decline.

June 18: With over 70 per cent of eligible Albertans receiving at least one dose of vaccine, Kenney announced all restrictions would be lifted on July 1.

July 1: Rules around indoor and outdoor social gatherings, capacity limits in businesses and other venues, recreation, large events like concerts or sports and other settings were lifted.

Sept. 15: Kenney declares a state of public health emergency in Alberta, and introduces a number of new restrictions. He also announced the implementation of a vaccine passport program – also known as the Restriction Exemption Program.

Oct. 14: The United States announces it will lift restrictions at its land borders with Canada and Mexico for fully vaccinated foreign nationals in early November. The land-border closures had been in place since March 2020.

Nov. 8: The U.S. land border fully reopens for those who are fully vaccinated. It was the first time in 597 days Canadians could drive into the United States for non-essential travel.

BC	
2020-02-20	2020-05-19
2021-03-17	2021-06-30
2021-11-30	2022-01-18
AB	
2020-03-05	2020-05-13
2021-01-14	2021-06-18
2021-09-15	2022-03-01
SK	
2020-03-18	2020-05-19
2020-11-16	2021-05-04
MB	
2020-03-12	2020-05-04
2020-11-12	2021-02-01
2021-06-01	2022-03-01
QC	
2020-03-14	2020-05-21
2020-12-03	2021-02-01
2021-12-20	2022-05-14
NS	
2020-03-22	2020-07-03
2020-10-29	2020-11-14
2021-11-09	2022-03-01
NB	
2020-03-22	2020-07-03
PE	
2020-03-16	2020-07-03
NL	
2020-03-14	2020-07-03
2021-02-01	2021-03-27
2021-12-23	2022-03-14

```
df['LockdownFlag'] =  0
```

```
conditions = [
    (df['Province'] == 'ON') & (df['Date'] > pd.to_datetime(date(2020,3,17))) & (df['Date'] < pd.to_datetime(date(2020,7,17))),
#     (df['Province'] == 'ON') & (df['Date'] > pd.to_datetime(date(2020,12,26))) & (df['Date'] < pd.to_datetime(date(2021,2,10))),
    (df['Province'] == 'QC') & (df['Date'] >= pd.to_datetime(date(2020,3,14))) & (df['Date'] <= pd.to_datetime(date(2020,5,21))),
#     (df['Province'] == 'QC') & (df['Date'] >= pd.to_datetime(date(2020,12,3))) & (df['Date'] <= pd.to_datetime(date(2021,2,1))),
    (df['Province'] == 'BC') & (df['Date'] >= pd.to_datetime(date(2020,2,20))) & (df['Date'] <= pd.to_datetime(date(2020,5,19))),
    (df['Province'] == 'AB') & (df['Date'] >= pd.to_datetime(date(2020,3,5))) & (df['Date'] <= pd.to_datetime(date(2020,5,13))),
    (df['Province'] == 'SK') & (df['Date'] >= pd.to_datetime(date(2020,3,18))) & (df['Date'] <= pd.to_datetime(date(2020,5,19))),
    (df['Province'] == 'MB') & (df['Date'] >= pd.to_datetime(date(2020,3,12))) & (df['Date'] <= pd.to_datetime(date(2020,5,4))),
    (df['Province'] == 'NS') & (df['Date'] >= pd.to_datetime(date(2020,3,22))) & (df['Date'] <= pd.to_datetime(date(2020,7,3))),
    (df['Province'] == 'NB') & (df['Date'] >= pd.to_datetime(date(2020,3,22))) & (df['Date'] <= pd.to_datetime(date(2020,7,3))),
    (df['Province'] == 'NL') & (df['Date'] >= pd.to_datetime(date(2020,3,14))) & (df['Date'] <= pd.to_datetime(date(2020,7,3)))
]

values = [
    1,
    1,
    1,
    1,
    1,
    1,
    1,
    1,
    1
]
```

```
df['LockdownFlag'] = np.select(conditions, values)
```

Phase 4: Model

- **Algorithms**
  - Regression:
    - Polynomial Regression
    - Random Forest / Decision Tree
    - SVM
    - Neural Network
    - AutoML
  - Time Series:
    - LSTM
    - Auto Forecast with Power BI
  - Other:
    - KNN
- **Tools:**
  - Python
  - IBM SPSS Modeler
  - Power BI

Phase 4: Model

- **LSTM with Python**

```
import pandas as pd

pd.set_option('display.max_columns', 1000)
pd.set_option('display.width', 1000)
pd.set_option('display.max_colwidth', 1000)

import matplotlib.pyplot as plt
from pandas import read_excel
import numpy as np
import pandas as pd
from pandas import DataFrame
from pandas import concat
from sklearn.preprocessing import MinMaxScaler
from keras.models import Sequential
from keras.layers import LSTM,Dense,Dropout
from numpy import concatenate
from sklearn.metrics import mean_squared_error,mean_absolute_error,r2_score
from math import sqrt

import keras.backend as K
from keras.callbacks import LearningRateScheduler
```

```
def scheduler(epoch):

    if epoch % 5 == 0 and epoch != 0:
        lr = K.get_value(model.optimizer.lr)
        K.set_value(model.optimizer.lr, lr * 0.1)
        print("lr changed to {}".format(lr * 0.1))
    return K.get_value(model.optimizer.lr)
```

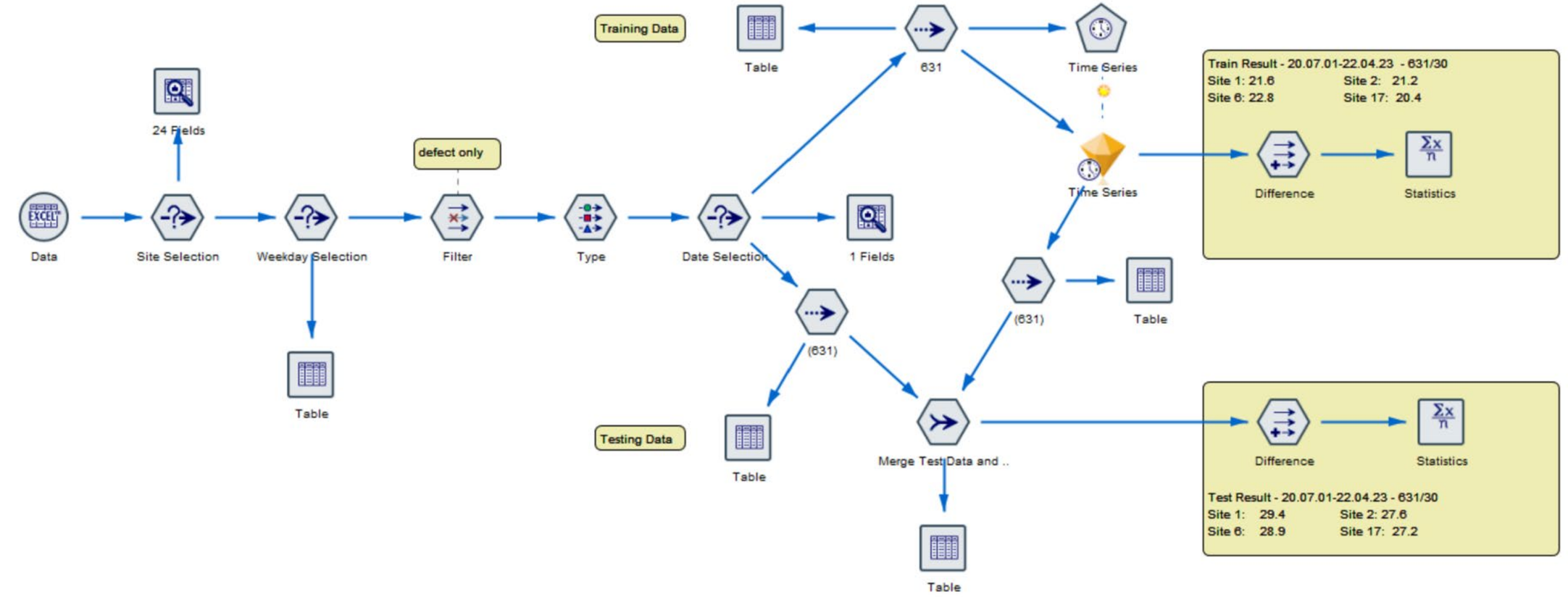
```
res_df = pd.DataFrame({'site':sites, 'mae':maes, 'mse':mses, 'rmse':rmses, 'r_square':r_s})
res_df.to_csv('result_new.csv')
res_df.head()
```

A1

	A	B	C	D	E	F	G	H
1		site	mae	mse	rmse	r_square	mape	
2	0	Site 1	0.058213	0.005949	0.077127	0.278658	46.41583	
3	1	Site 2	0.063922	0.008482	0.092098	0.027533	40.94877	
4	2	Site 3	0.09046	0.01161	0.107751	0.551142	52.39172	
5	3	Site 4	0.095217	0.01308	0.114366	0.268825	74.64878	
6	4	Site 5	0.089573	0.011685	0.108099	0.319062	65.05853	
7	5	Site 6	0.043224	0.003578	0.059815	0.453756	46.74289	
8	6	Site 7	0.107583	0.016853	0.129818	-0.02696	74.64954	
9	7	Site 8	0.086872	0.011172	0.105698	0.03154	129.1438	
10	8	Site 9	0.099733	0.013319	0.115409	-0.54754	88.9155	
11	9	Site 10	0.06997	0.007354	0.085756	-0.15309	86.66208	
12	10	Site 11	0.065278	0.009114	0.095469	0.527076	inf	
13	11	Site 12	0.084153	0.010858	0.104202	0.368944	inf	
14	12	Site 13	0.062322	0.007736	0.087953	0.718228	29.7697	
15	13	Site 14	0.11016	0.017868	0.133671	0.2999	inf	
16	14	Site 15	0.10344	0.015956	0.126315	-0.07131	inf	
17	15	Site 16	0.119858	0.02208	0.148593	0.109895	86.74351	
18	16	Site 17	0.097421	0.012528	0.111929	-0.51958	91.73904	
19	17	Site 18	0.079058	0.016327	0.127779	0.424802	inf	
20	18	Site 19	0.123173	0.030942	0.175904	0.358594	160.206	
21	19	Site 20	0.076099	0.010566	0.102793	0.187005	55.46776	
22								

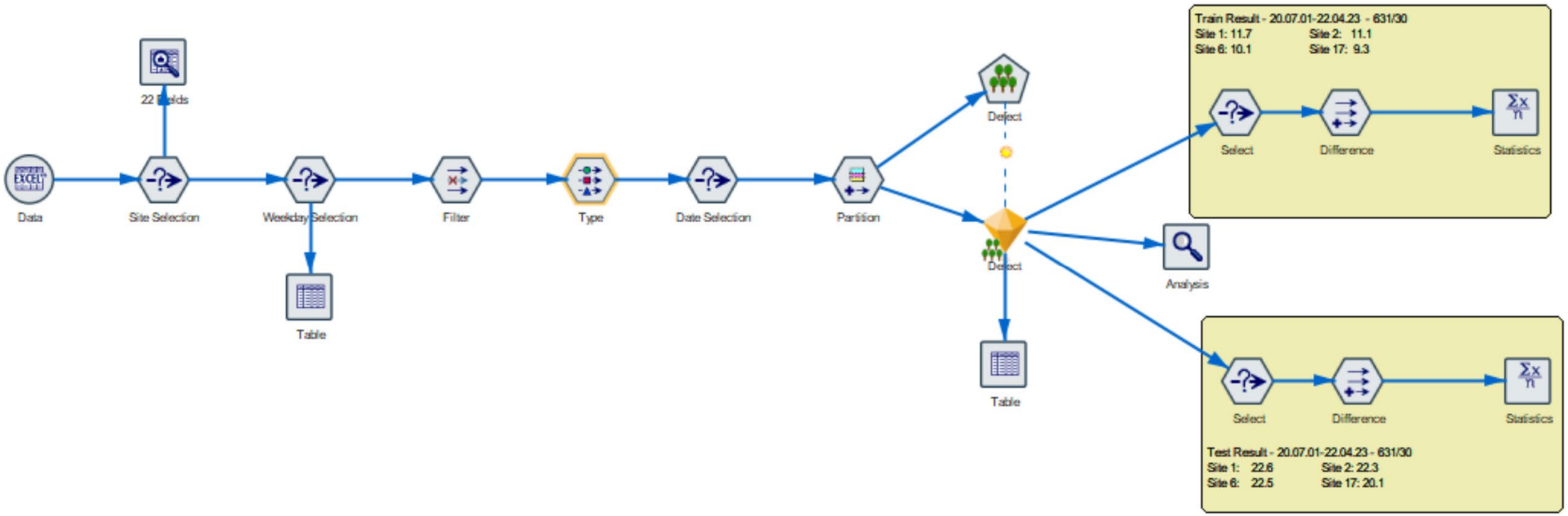
Phase 4: Model

- **Time Series with SPSS**



Phase 4: Model

- **XGBoost Tree with SPSS**



Phase 4: Model

- **Neural Network with Python**

```
from sklearn.model_selection import GridSearchCV
from sklearn.neural_network import MLPRegressor
MLPRegressor().get_params()
```

```
{'activation': 'relu',
 'alpha': 0.0001,
 'batch_size': 'auto',
 'beta_1': 0.9,
 'beta_2': 0.999,
 'early_stopping': False,
 'epsilon': 1e-08,
 'hidden_layer_sizes': (100,),
 'learning_rate': 'constant',
 'learning_rate_init': 0.001,
 'max_fun': 15000,
 'max_iter': 200,
 'momentum': 0.9,
 'n_iter_no_change': 10,
 'nesterovs_momentum': True,
 'power_t': 0.5,
 'random_state': None,
 'shuffle': True,
 'solver': 'adam',
 'tol': 0.0001,
 'validation_fraction': 0.1,
 'verbose': False,
 'warm_start': False}
```

```
params = {'hidden_layer_sizes':[(7,7,7), (5,5,5), (7,7), (7,5), (5,5), (5,7)],
          'activation':['identity', 'logistic', 'tanh', 'relu'],
          'solver':['sgd', 'adam'],
          'max_iter':[200, 400, 600],
          'alpha': [0.0001, 0.001],
          'random_state':[123]}
```

```
grid = GridSearchCV(MLPRegressor(), params, cv=5, scoring='neg_mean_absolute_error')
grid.fit(Xtrain, ytrain)
```

```
grid.best_params_
```

```
{'activation': 'identity',
 'alpha': 0.0001,
 'hidden_layer_sizes': (5, 5, 5),
 'max_iter': 200,
 'random_state': 123,
 'solver': 'adam'}
```

```
mape_train_nn = MAPE(ytrain, ypred_train)
mape_test_nn = MAPE(ytest, ypred_test)
mape_train_nn, mape_test_nn
```

```
(24.75558938807629, 24.958772639455944)
```

Phase 4: Model

- **AutoML with Python**

```
from flaml import AutoML
```

```
automl = AutoML(time_budget = 180, metric= 'mape', task ='regression')
```

```
automl.fit(X_train = xtrain, y_train = ytrain, verbose=1)
```

```
print(automl.model)
```

```
<flaml.model.XGBoostLimitDepthEstimator object at 0x0000026E50EC38E0>
```

Phase 4: Model

- **Model Tuning**
  - Feature Selection:
    - Correlation Matrix and Heat Map
    - AutoML: *Feature_Importances_*
  - Hyperparameter:
    - GridSearchCV
    - Auto tuning with AutoML
  - Divide dataset into sub-datasets:
    - Working day/Weekend/Holiday
    - Sites with good/bad performance
    - Sites with big/small volume

Phase 5: Evaluation

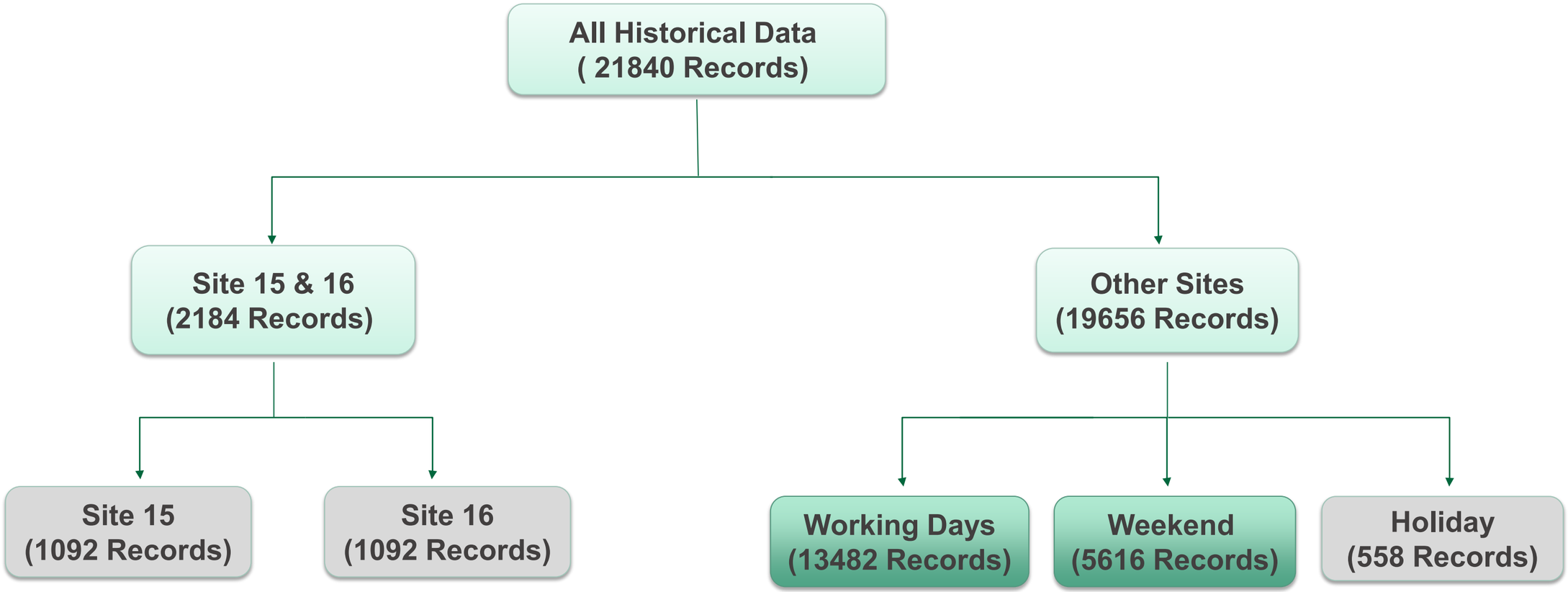
- **Comparison of MAPE**
  - Performance ranking:
    - AutoML
    - Random Forest
    - Decision Tree
    - Other

Rank	Model	Training MAPE(%)	Test MAPE(%)
1	AutoML	5.1	15.1
2	Random Forest	6.8	18.8
3	Decision Tree	13.7	22.9
4	Polynomial Regression	23.3	24.6
5	Neural Network	24.7	24.9
6	SVM	26.3	27.3
7	LSTM	29.8	30.3
8	Auto Time Series	31.3	50.1

- **Final Model:**
  - **AutoML**

Phase 5: Evaluation

- **Further accuracy improvement**
  - Grouping for sub-datasets

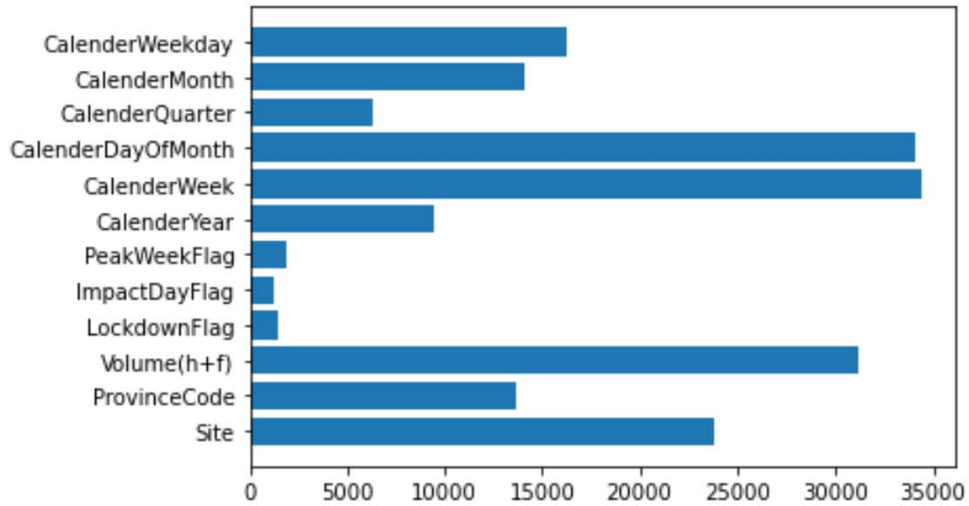


Phase 5: Evaluation

- **Further accuracy improvement**
  - Input feature selection:

```
import matplotlib.pyplot as plt
plt.barh(xtrain_WorkingDay_GoodSites.columns, automl_WorkingDay_GoodSites.model.estimator.feature_importances_)
```

```
<BarContainer object of 12 artists>
```



Phase 5: Evaluation

- **Further accuracy improvement**
  - Hyperparameter tuning for AutoML
    - Time budget / Estimator_list / Metric

```
automl_WorkingDay_GoodSites = AutoML(time_budget = 5 * time_budget,
                                     metric= 'r2',
                                     task ='regression',
                                     estimator_list=['xgboost', 'lgbm'])
```

```
automl_Weekend_GoodSites = AutoML(time_budget = time_budget, metric= 'r2', task ='regression')
```

Phase 5: Evaluation

- **MAPE for models with good sites and working days**

```
print('Working Day & Good Sites', automl_WorkingDay_GoodSites.model)
mape_train_WorkingDay_GoodSites = MAPE(ytrain_WorkingDay_GoodSites,
                                       automl_WorkingDay_GoodSites.predict(xtrain_WorkingDay_GoodSites))
mape_test_WorkingDay_GoodSites = MAPE(ytest_WorkingDay_GoodSites,
                                      automl_WorkingDay_GoodSites.predict(xtest_WorkingDay_GoodSites))
print('train_error_percentage', mape_train_WorkingDay_GoodSites,
      'test_error_percentage', mape_test_WorkingDay_GoodSites)
```

```
Working Day & Good Sites <flaml.model.XGBoostSklearnEstimator object at 0x00000241D86F65B0>
train_error_percentage 4.454336333263354 test_error_percentage 14.590903268324453
```

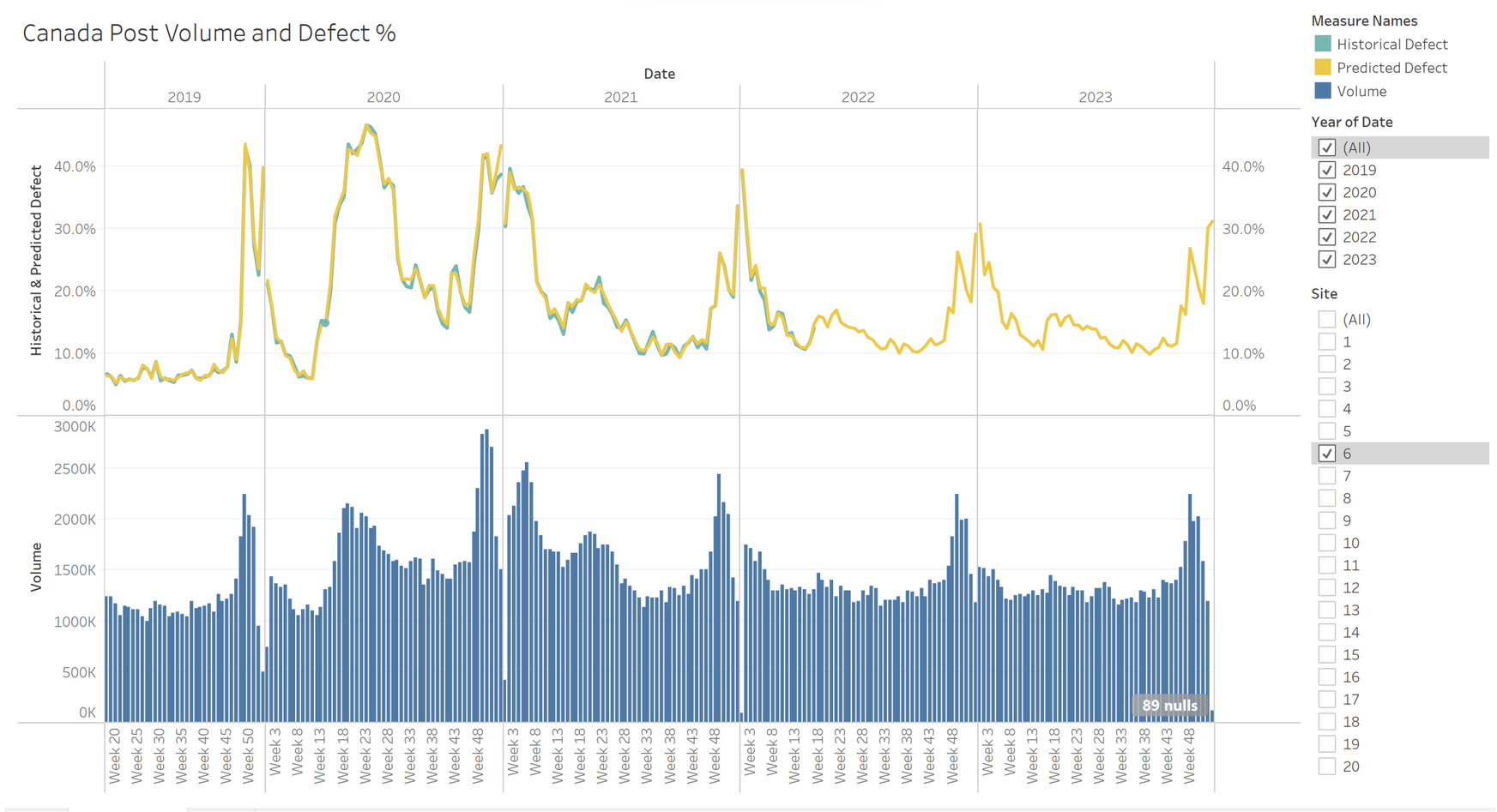
Phase 6: Deployment and Visualization

- **Deploy model to predict historical events and make future forecasts**

	A	B	C	D	E	F	G
1	Date	Site	Plant	Province	olume(h+	Defect	Defect(Predict)
2	2019-04-28	1	CALGARY MPP	AB	80352	0.065	0.056814
3	2019-04-28	2	EDMONTON MPP	AB	63051	0.04	0.055035576
4	2019-04-28	3	HALIFAX MPP	NS	27999	0.016	0.091480657
5	2019-04-28	4	HAMILTON MPP	ON	31810	0.03	0.012811425
6	2019-04-28	5	KITCHENER MPP	ON	24127	0.017	0.004733069
7	2019-04-28	6	LEO BLANCHETTE MPP	QC	143181	0.036	0.032635596
8	2019-04-28	7	LONDON MPP	ON	21757	0.016	0.005857045
9	2019-04-28	8	MONCTON MPP	NB	17826	0.068	0.000816036
10	2019-04-28	9	OTTAWA MPP	ON	30625	0.039	0.000374047
11	2019-04-28	10	QUEBEC MPP	QC	20028	0.057	0.085428804
12	2019-04-28	11	REGINA MPP	SK	9647	0.048	0.002724159
13	2019-04-28	12	SAINT JOHN MPP	NB	4049	0.037	0.001219241
14	2019-04-28	13	SASKATOON MPP	SK	11757	0.035	0.002724159
15	2019-04-28	14	ST. JOHNS MPP	NL	5941	0.038	0.003907009
16	2019-04-28	15	SUDBURY MPP	ON	9437	0.214	0.219972521
17	2019-04-28	16	THUNDER BAY MPP	ON	2382	0.191	0.178399071
18	2019-04-28	17	TORONTO GATEWAY	ON	201191	0.028	0.000374047
19	2019-04-28	18	VANCOUVER MPP	BC	135912	0.027	0.016801957
20	2019-04-28	19	VICTORIA MPP	BC	9443	0.019	0.006873338
21	2019-04-28	20	WINNIPEG MPP	MB	51891	0.025	0.004602515
22	2019-04-29	1	CALGARY MPP	AB	85552	0.115	0.102324776
23	2019-04-29	2	EDMONTON MPP	AB	67843	0.095	0.092421025
24	2019-04-29	3	HALIFAX MPP	NS	29843	0.085	0.096696787
25	2019-04-29	4	HAMILTON MPP	ON	31268	0.121	0.120359249
26	2019-04-29	5	KITCHENER MPP	ON	19470	0.112	0.119145073
27	2019-04-29	6	LEO BLANCHETTE MPP	QC	202101	0.111	0.110036351
28	2019-04-29	7	LONDON MPP	ON	22208	0.12	0.113144286
29	2019-04-29	8	MONCTON MPP	NB	23163	0.129	0.128313616
30	2019-04-29	9	OTTAWA MPP	ON	52541	0.15	0.156724155
31	2019-04-29	10	QUEBEC MPP	QC	37777	0.174	0.164915055
32	2019-04-29	11	REGINA MPP	SK	12571	0.082	0.087303653

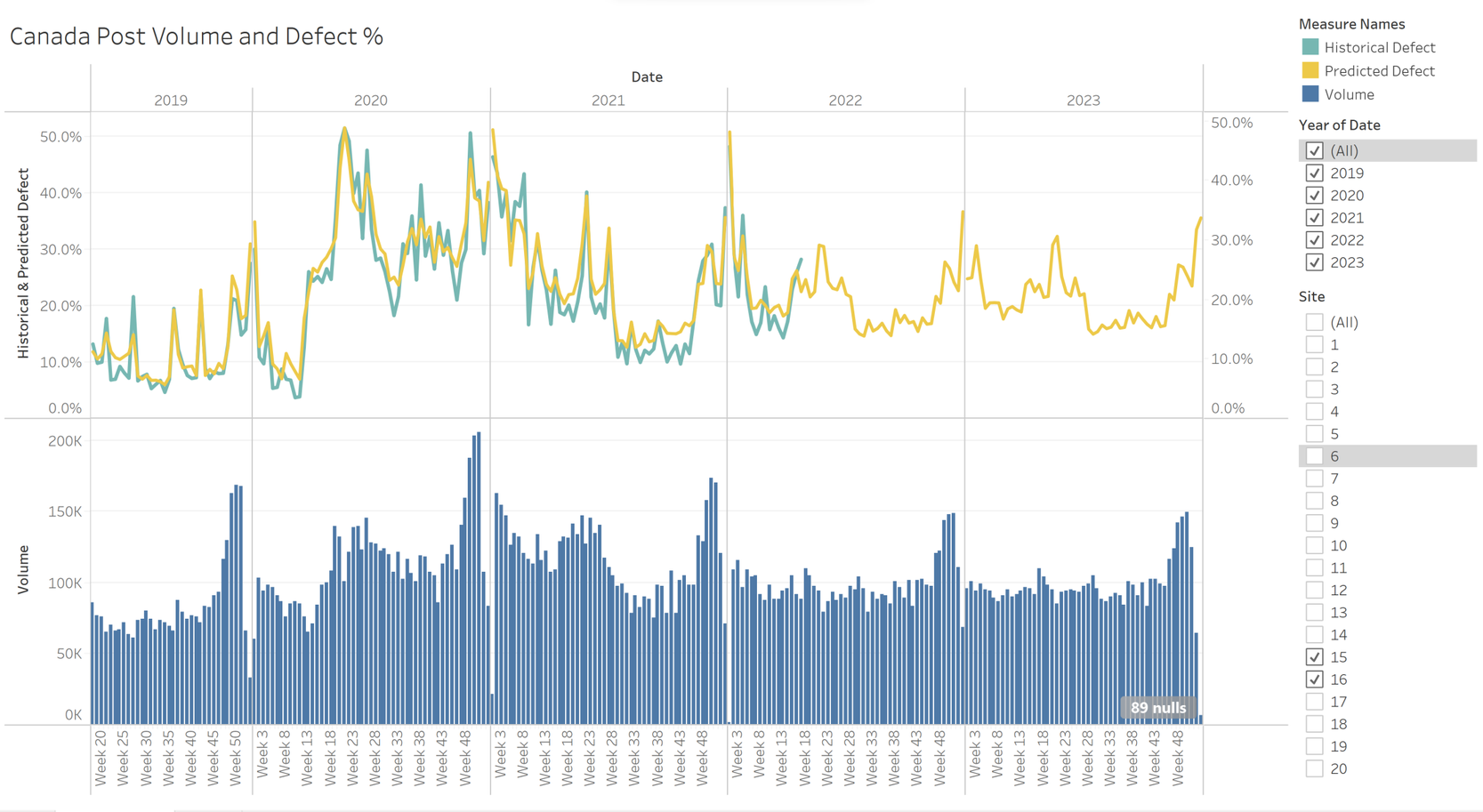
Phase 6: Deployment and Visualization

- **Historical + Forecast Weekly Volume & Defect % of Site 6**



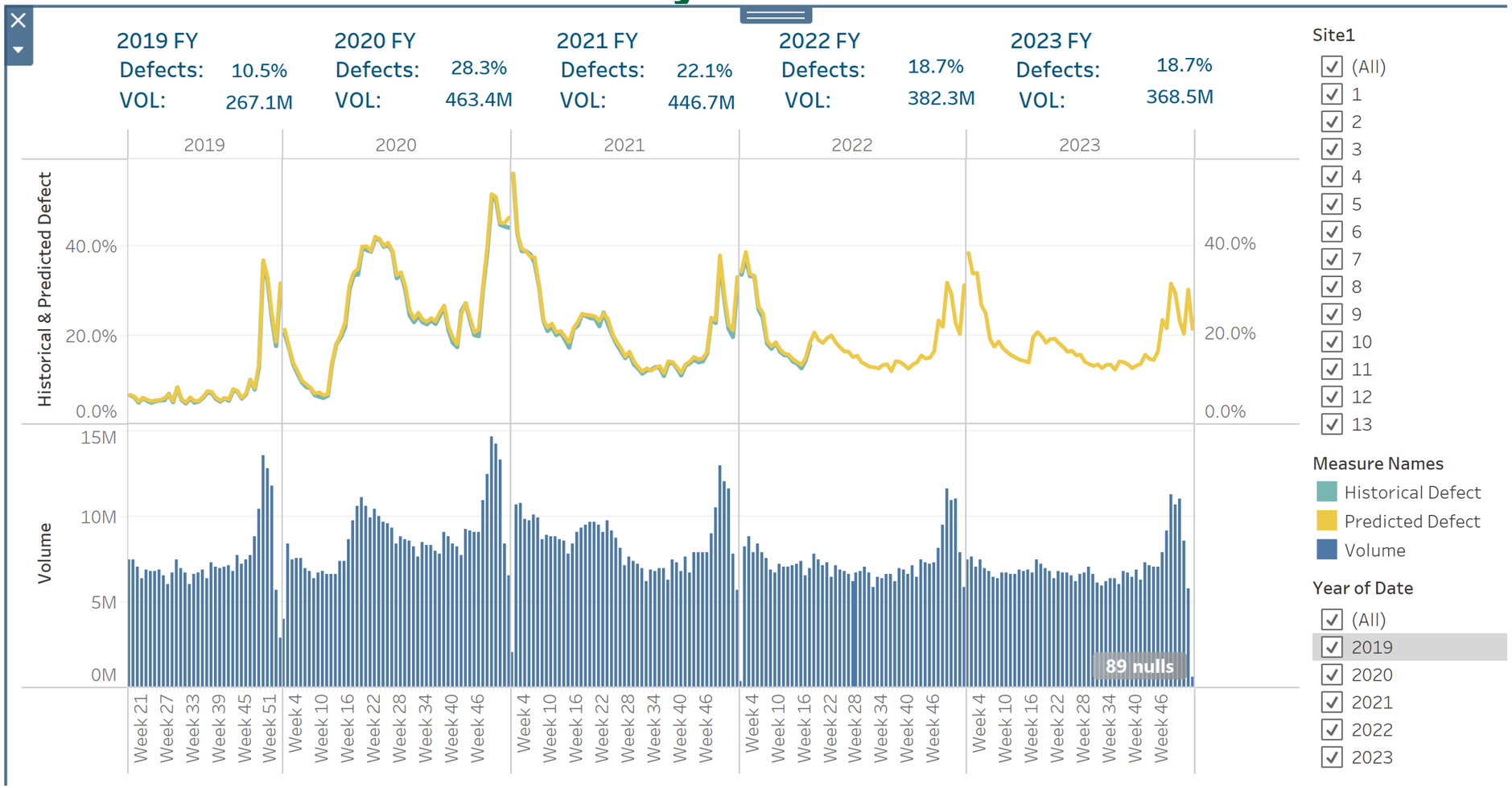
Phase 6: Deployment and Visualization

- **Historical + Forecast Weekly Volume & Defect % of Site 15 & 16**



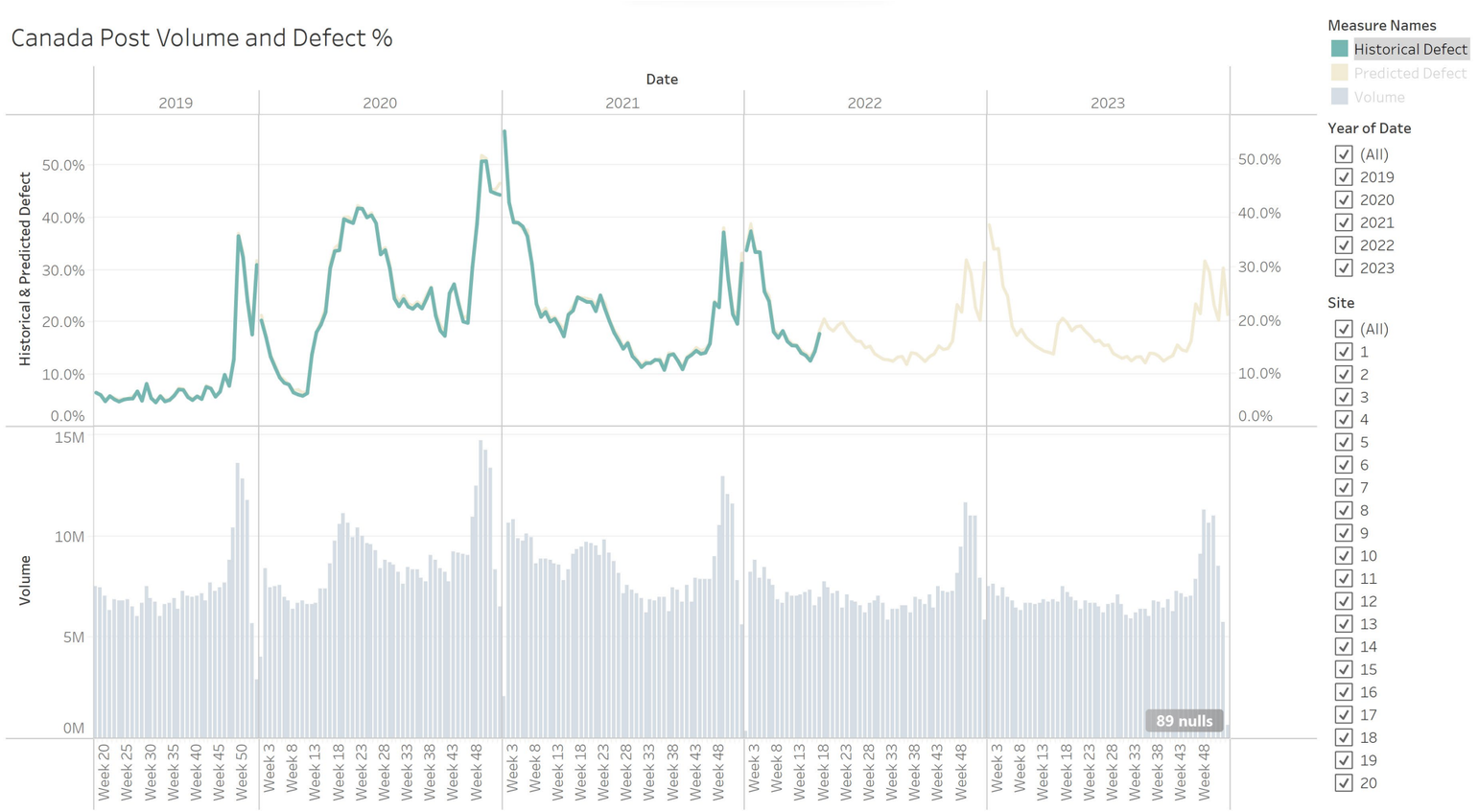
Phase 6: Deployment and Visualization

- **Historical + Forecast Weekly Volume & Defect % of All Nation**



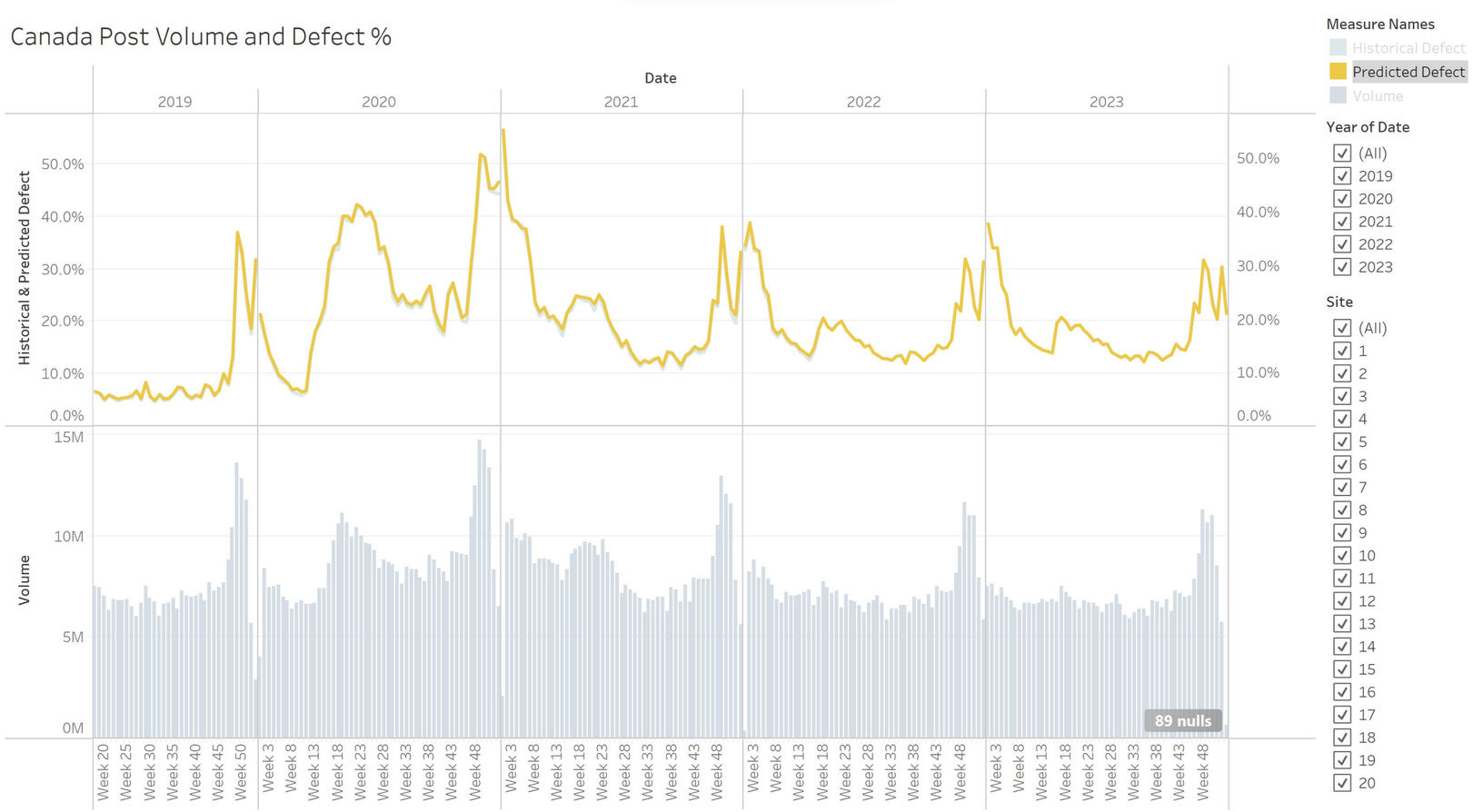
Phase 6: Deployment and Visualization

- **Historical + Forecast Weekly Volume & Defect % of All Nation**



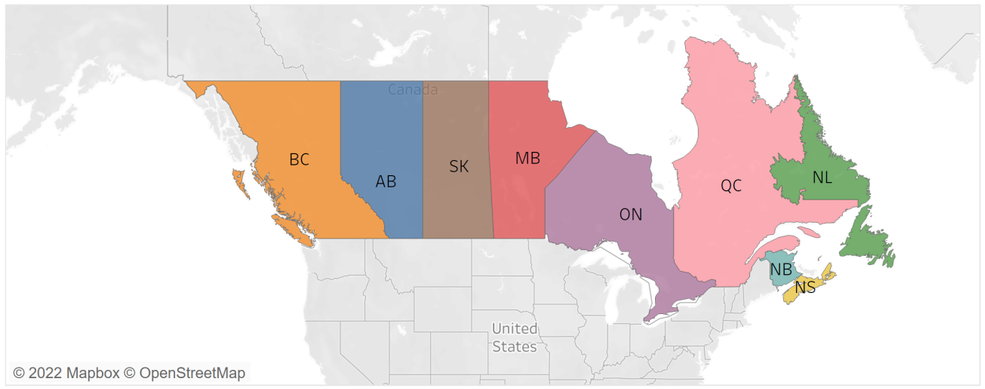
Phase 6: Deployment and Visualization

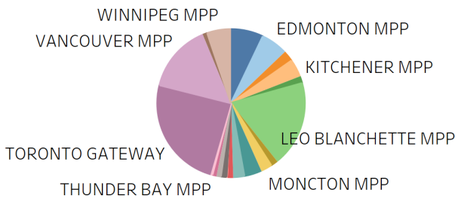
- **Historical + Forecast Weekly Volume & Defect % of All Nation**



Phase 6: Deployment and Visualization

- **Historical + Forecast Weekly Volume & Defect % of All Nation**

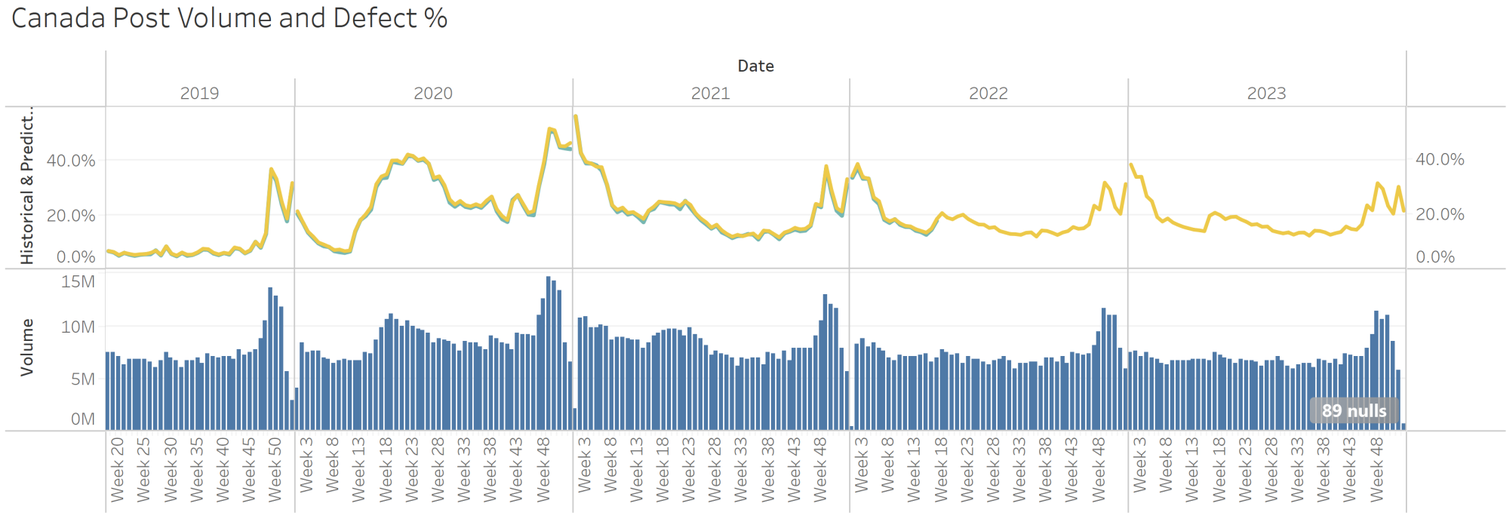




Year of Date
- [x] (All)
- [x] 2019
- [x] 2020
- [x] 2021
- [x] 2022
- [x] 2023

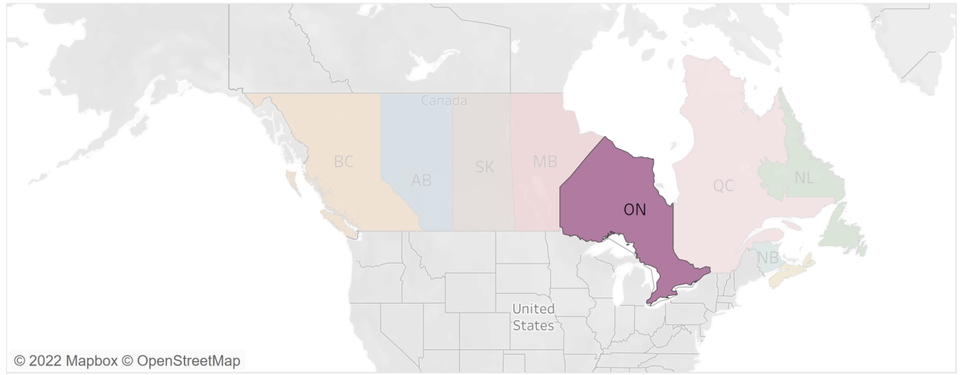
Measure Names
- Historical Defect
- Predicted Defect
- Volume

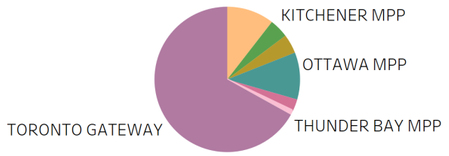
Site1
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20



Phase 6: Deployment and Visualization

- **Historical + Forecast Weekly Volume & Defect % of Ontario**

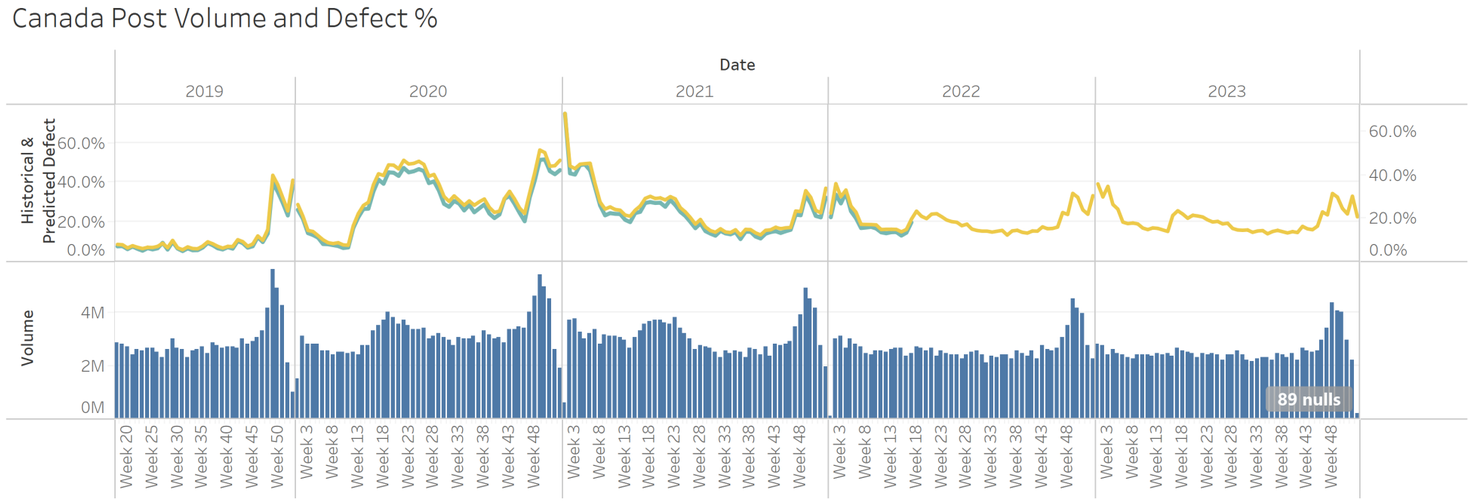




Year of Date
- (All)
- 2019
- 2020
- 2021
- 2022
- 2023

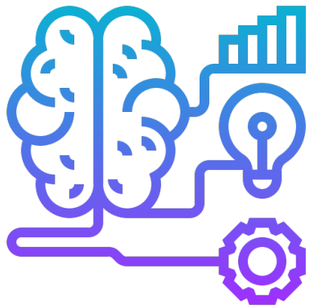
Measure Names
- Historical Defect
- Predicted Defect
- Volume

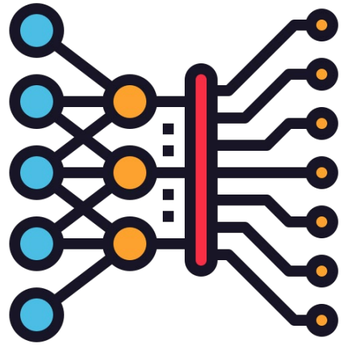
Site1
- 4
- 5
- 7
- 9
- 15
- 16
- 17



Phase 7: Presentation and Delivery

- **Presentation in next Client meeting- July 29**
- **Final delivery to Client**
  - Visualization – *Tableau File*
  - Predicted Data for every day – *Excel File*
  - Presentation slides – *PowerPoint File*
  - Python Codes for modeling - *ipynb File*
  - Dataset for modeling - *Excel File*





Thank you!