1.5 Time Invariance

Code hiding script from [Damian Kao (http://blog.nextgenetics.net/?e=102)](http://blog.nextgenetics.net/?e=102).

```
Out[1]:   Code hidden for easier reading: toggle on/off.
```

1.5.1 New Train/Test Split

Ensure that both train and test sets have a full year of data, covering each month.

```
Read: 18712 records
Merged: 18712 records

Train: 2012-04-02 to 2013-03-29
Test:  2013-04-01 to 2014-03-31

Date Range Contains: 13642 records

Train: N = 5788, P(critical) = 0.138
Test:  N = 7854, P(critical) = 0.146
```

Figure 1. Number of Inspections Conducted By Month

- The test set includes a greater total number of inspections, both overall and per month.
- In the train set, most inspections were conducted in March, with January as a close second.
- In the test set, most inspections were conducted in October, with May as a close second.
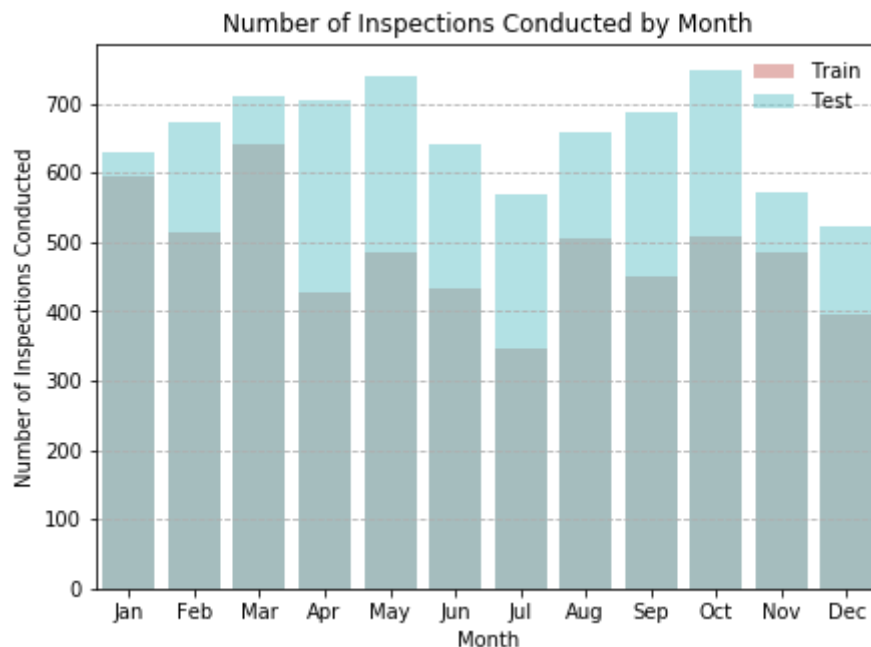

Figure 2. Proportion of All Inspections Conducted by Month

- Each series will sum to one.
- January, March, and April appear to be the months with the greatest difference in proportion of all inspections conducted between the train and test sets.

Proportion of All Inspections Conducted by Month
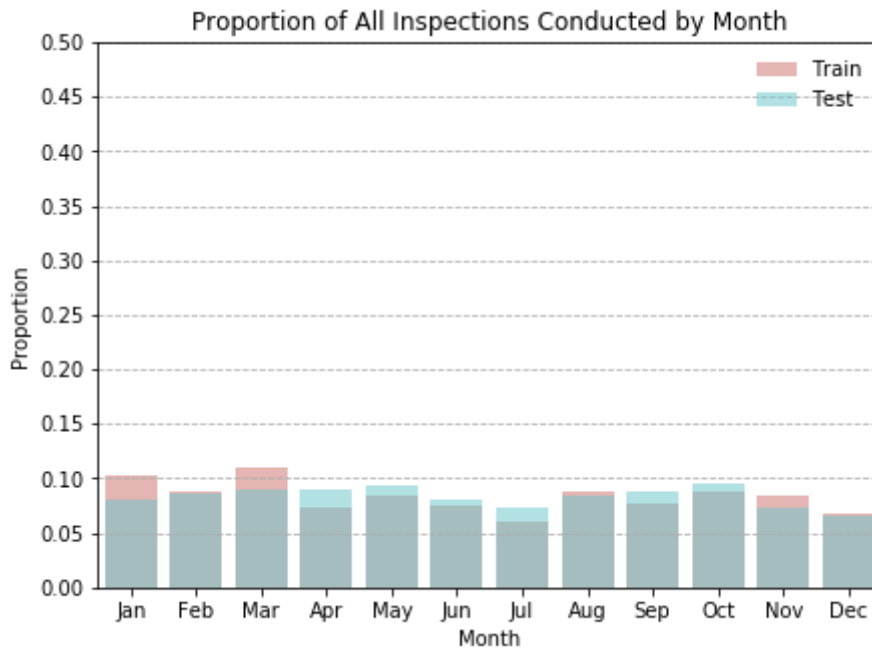
Table 1. Summary of Violation Codes in Train and Test

- Sorted by violation code.
- V5 does not appear in the train set at all.
- V7 accounts for less than 0.5% of all violation codes issued in the train set.

Critical violation code descriptions referenced from the City of Chicago Food Protection Division website (https://webapps1.cityofchicago.org/healthinspection/Code_Violations.jsp).

V5: Likelihood of Infected Food Handler

> No person affected with or carrying any disease in a communicable form or af
> flicted with boils, infected wounds, sores, acute respiratory infection, or
> intestinal disorder shall work in any area of a food establishment in any c
> apacity where there is a likelihood of that person contaminating food or foo
> d contact surfaces.

V7: Adequate Washing of Utensils

> Hand washing of all tableware and drinking utensils shall be accomplished by
> the use of warm water at a temperature of 110°F to 120°F containing an adequa
> te amount of detergent effective to remove grease and solids.

```
Out[9]:
```

	Code	N(Train)	N(Test)	P(Train)	P(Test)	Rank(Train)	Rank(Test)
0	V1	31	33	0.027	0.020	8	8
1	V2	189	282	0.164	0.170	2	2
2	V3	444	606	0.385	0.365	1	1
3	V4	28	23	0.024	0.014	9	10
4	V5	0	1	0.000	0.001	14	14
5	V6	81	79	0.070	0.048	5	6
6	V7	3	9	0.003	0.005	13	13
7	V8	142	192	0.123	0.116	3	3
8	V9	33	70	0.029	0.042	7	7
9	V10	26	28	0.023	0.017	10	9
10	V11	56	140	0.049	0.084	6	5
11	V12	97	169	0.084	0.102	4	4
12	V13	14	17	0.012	0.010	11	11
13	V14	8	13	0.007	0.008	12	12

Table 2. Frequency Ranking of Violation Codes in Train and Test

- Sorted by frequency in train set.
- Ranks for the top four violation codes match between train and test.
- Ranks for the bottom four violation codes match between train and test.

```
Out[10]:
```

	Code	N(Train)	N(Test)	P(Train)	P(Test)	Rank(Train)	Rank(Test)
2	V3	444	606	0.385	0.365	1	1
1	V2	189	282	0.164	0.170	2	2
7	V8	142	192	0.123	0.116	3	3
11	V12	97	169	0.084	0.102	4	4
5	V6	81	79	0.070	0.048	5	6
10	V11	56	140	0.049	0.084	6	5
8	V9	33	70	0.029	0.042	7	7
0	V1	31	33	0.027	0.020	8	8
3	V4	28	23	0.024	0.014	9	10
9	V10	26	28	0.023	0.017	10	9
12	V13	14	17	0.012	0.010	11	11
13	V14	8	13	0.007	0.008	12	12
6	V7	3	9	0.003	0.005	13	13
4	V5	0	1	0.000	0.001	14	14

1.5.2 Frequency of Violations by Month

Are certain violation codes more frequent in different months?

- Figure 3 will answer: Which month had the **highest fraction of all inspections** resulting in at least one critical violation found?
- Figure 4 will answer: Which month had the **highest rate of monthly inspections** resulting in at least one critical violation found?

Figure 3. Proportion of All Inspections with Critical Found by Month

- Each series will sum to one.
- In the train set, October had the highest fraction of all inspections resulting in at least one critical violation found.
- In the test set, May and March had the highest fraction of all inspections resulting in at least one critical violation found.

Proportion of All Inspections with Critical Found by Month
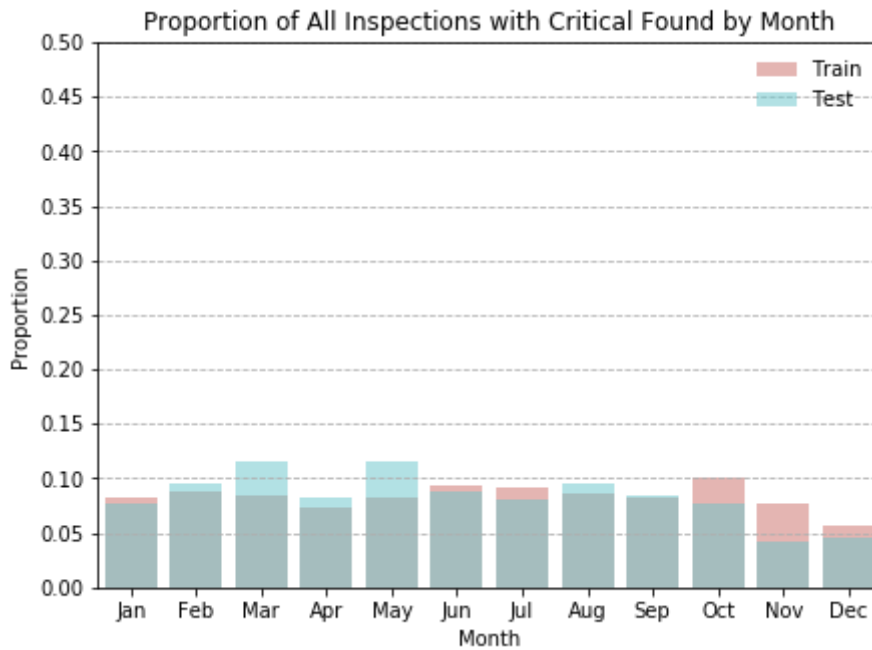
Figure 4. Proportion of Monthly Inspections with Critical Found by Month

- In the train set, July had the highest rate of monthly inspections resulting in at least one critical violation found.
- In the test set, March and May had the highest rate of monthly inspections resulting in at least one critical violation found.
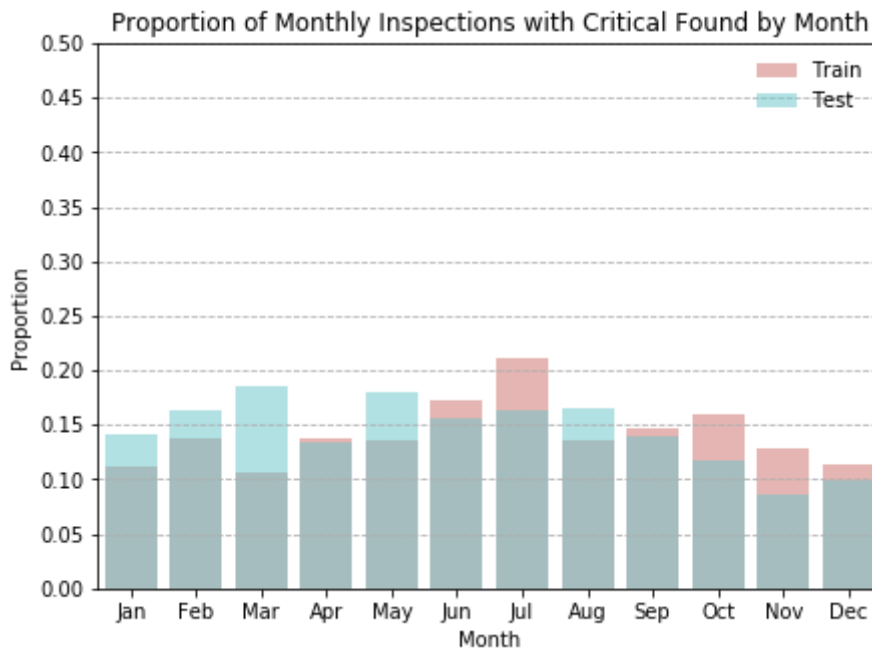

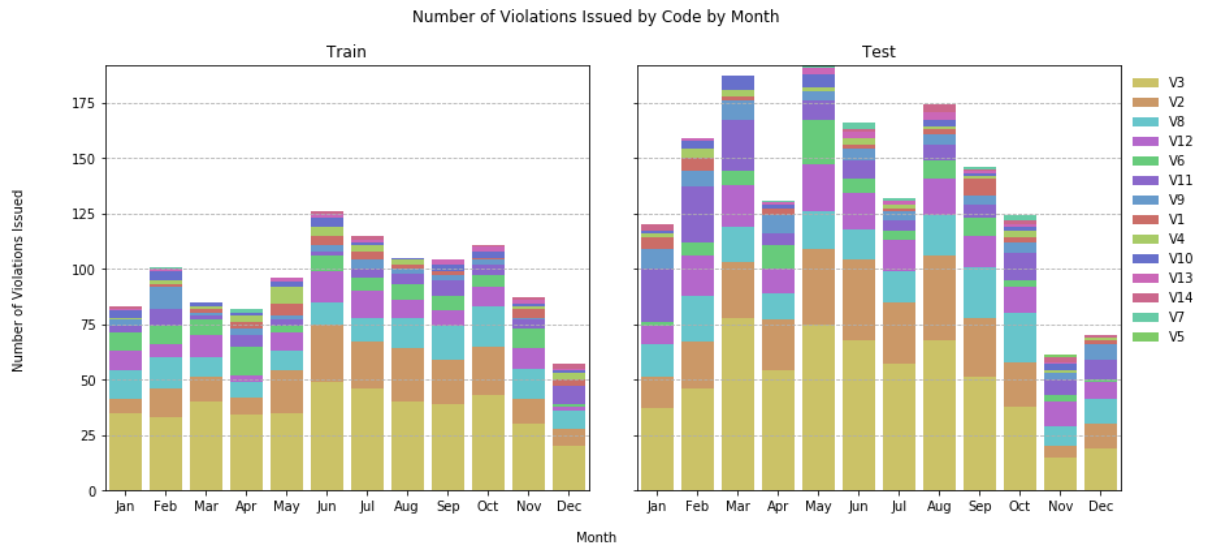Proportion of Monthly Inspections with Critical Found by Month

Figure 5. Number of Violations Issued by Month

- Monthly totals for each code are stacked in order of decreasing frequency of the code in the train set.

Number of Violations Issued by Code by Month

- Figure 6 will answer: Which month had the **highest fraction of all inspections** issuing the given code?
- Figures 7 and 8 will answer: Which month had the **highest rate of monthly inspections** issuing the given code?

Figure 6. Proportion of All Violations with Violation Code by Month

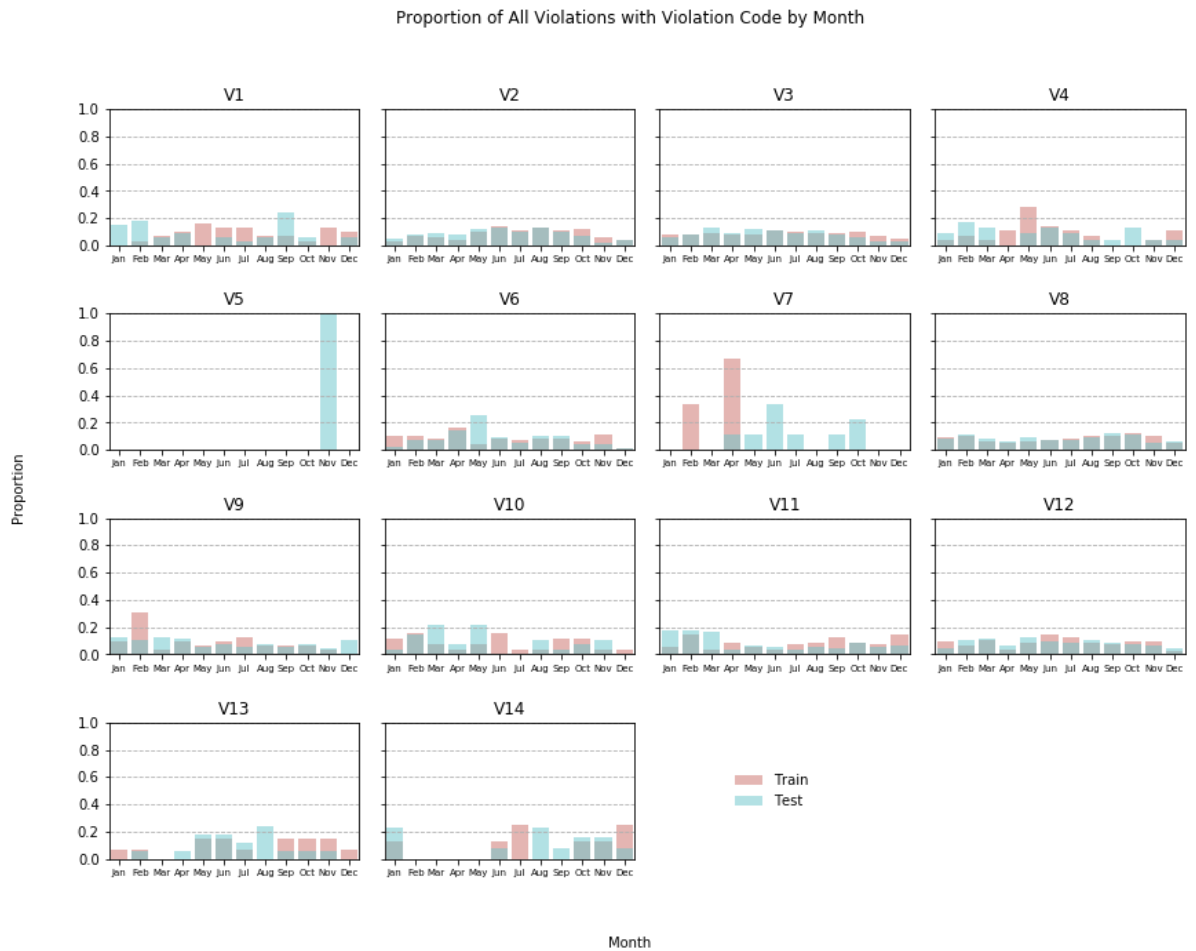- The series on each subplot will sum to one.

Figure 7. Proportion of Monthly Violations with Violation Code by Month

- The values for each month will sum to one.
- All subplot axes share the same range.

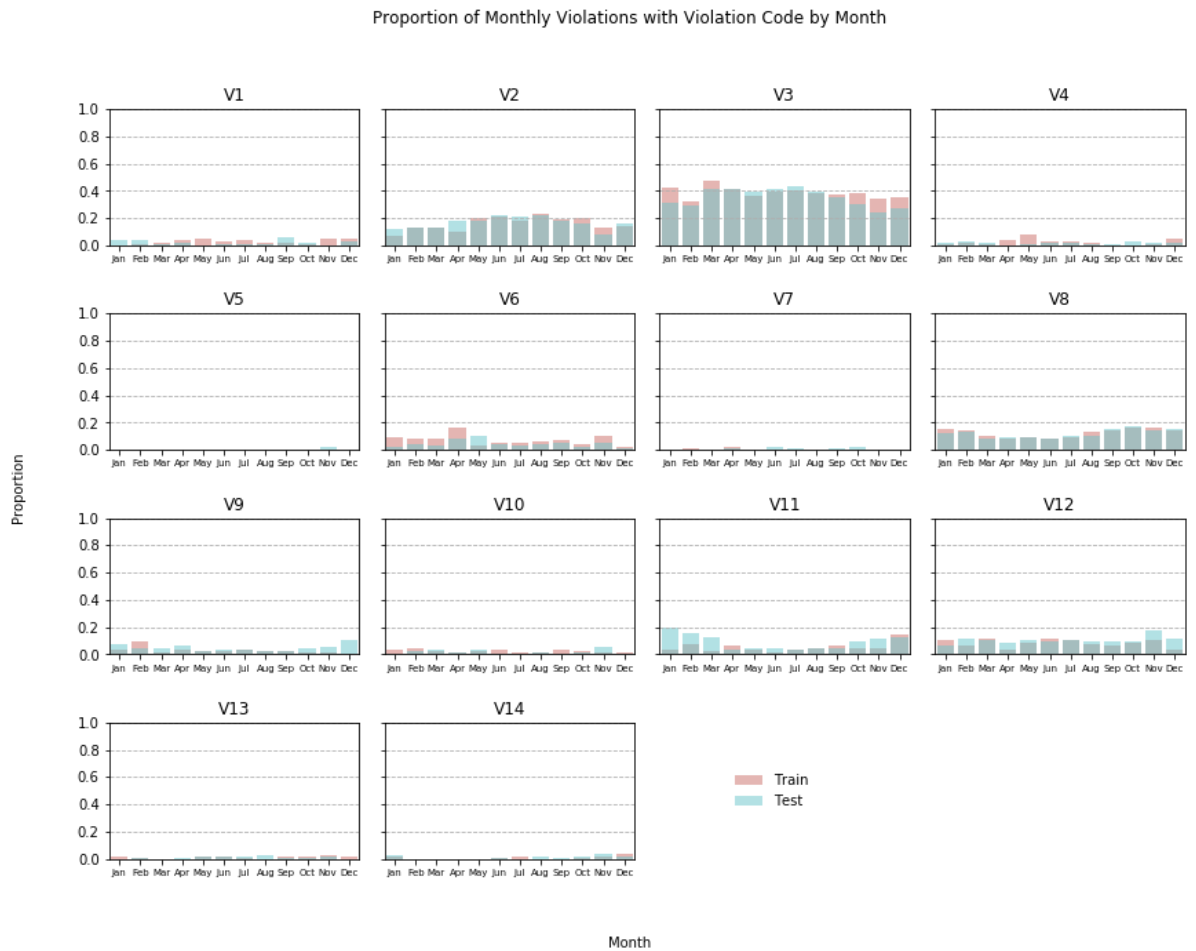Proportion of Monthly Violations with Violation Code by Month
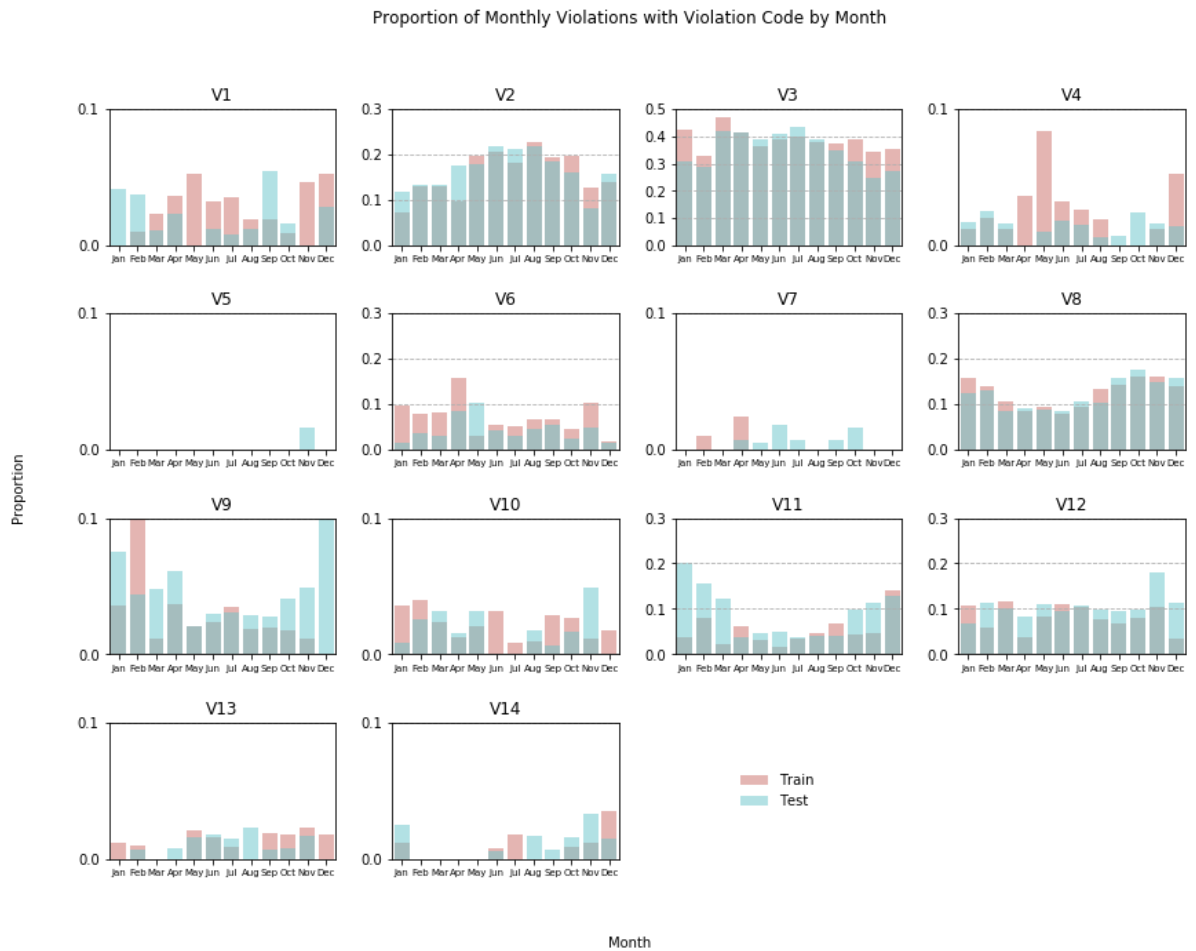
Figure 8. Proportion of Monthly Violations with Violation Code by Month (Different Axes)

- Same data as Figure 7.
- The values for each month will sum to one.
- Each subplot axis has a different range, to emphasize less frequent violation codes.

Proportion of Monthly Violations with Violation Code by Month

1.5.3 City Model Scheduling

How does the City model schedule a full year of inspections?
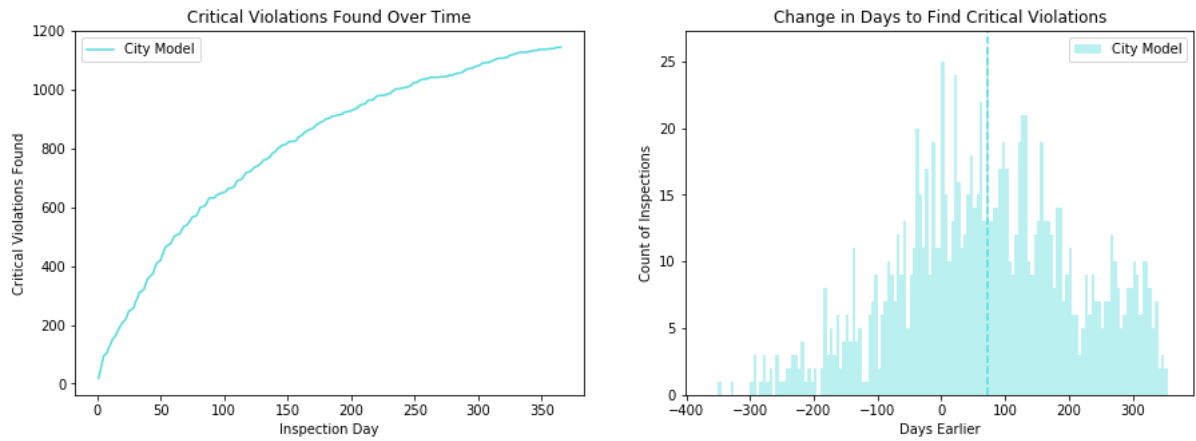
Table 3. Confusion Matrix for City Model Predictions on Test Set

```
F1 Score = 0.03532
Precision = 0.47727
Recall = 0.01834
```

Out[17]:

	Predicted +	Predicted -
Actual +	21	1124
Actual -	23	6686

Figure 9. City Model Performance According to City Metrics on Test Set

Out[18]:

	Model	First Half	Mean Change	Std. Change
0	City Model	0.784	72.08	140.254

Figure 10. Number of Inspections Conducted by Month

- City Model schedule should have exactly the same number of inspections per month as the original test set.
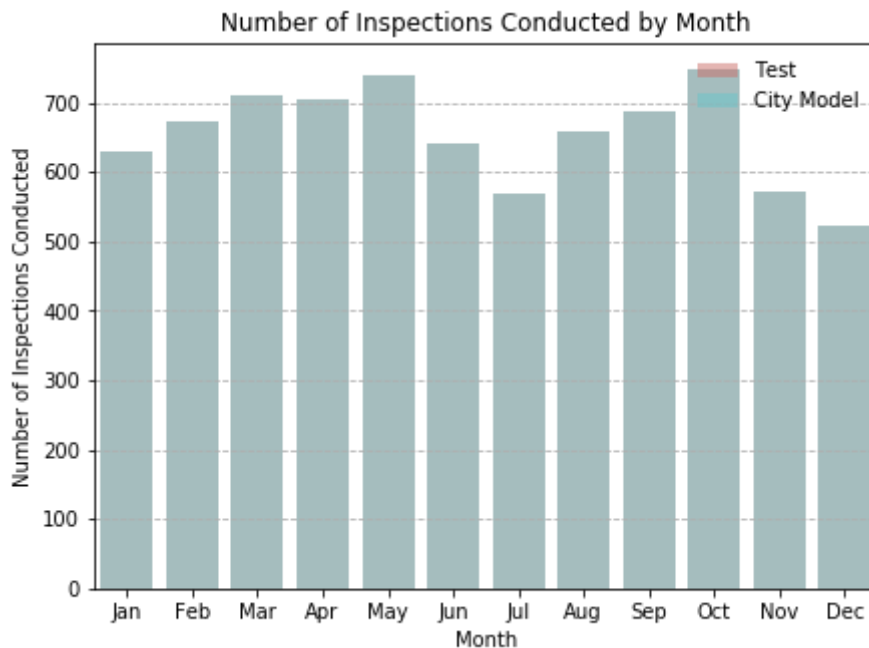
- Figure 11 will answer: Which month had the **highest fraction of all inspections** resulting in at least one critical violation found?
- Figure 12 will answer: Which month had the **highest rate of monthly inspections** resulting in at least one critical violation found?

Figure 11. Proportion of All Inspections with Critical Found by Month

- Remember that the test set starts in April and continues through the year to March.
- As expected, the first month of the model schedule has the highest fraction of all inspections resulting in at least one critical violation found and the fraction generally decreases with each successive month.

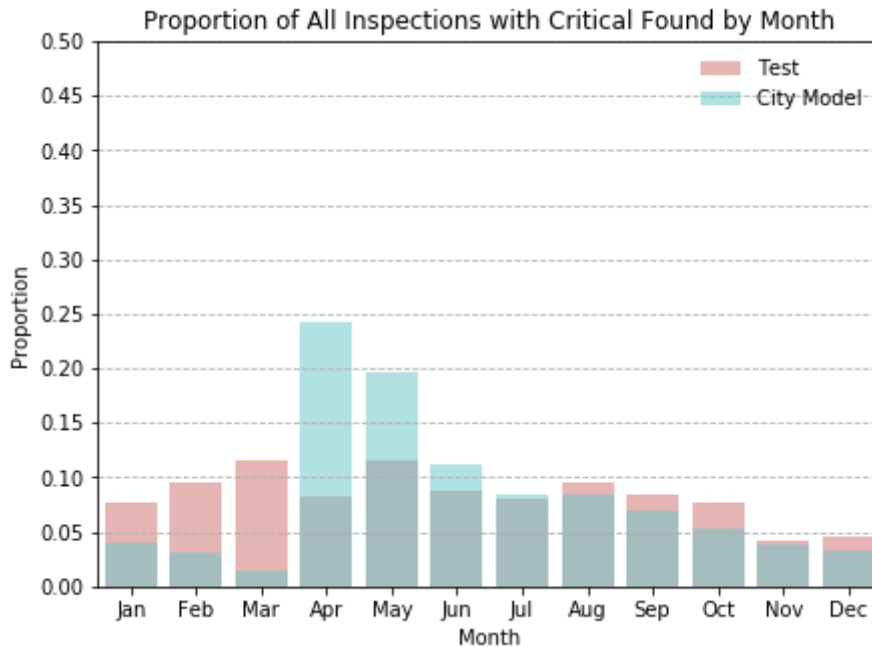Proportion of All Inspections with Critical Found by Month

Figure 12. Proportion of Monthly Inspections with Critical Found by Month

- Remember that the test set starts in April and continues through the year to March.
- As expected, the first month of the model schedule has the highest rate of monthly inspections resulting in at least one critical violation found and the rate generally decreases with each successive month.

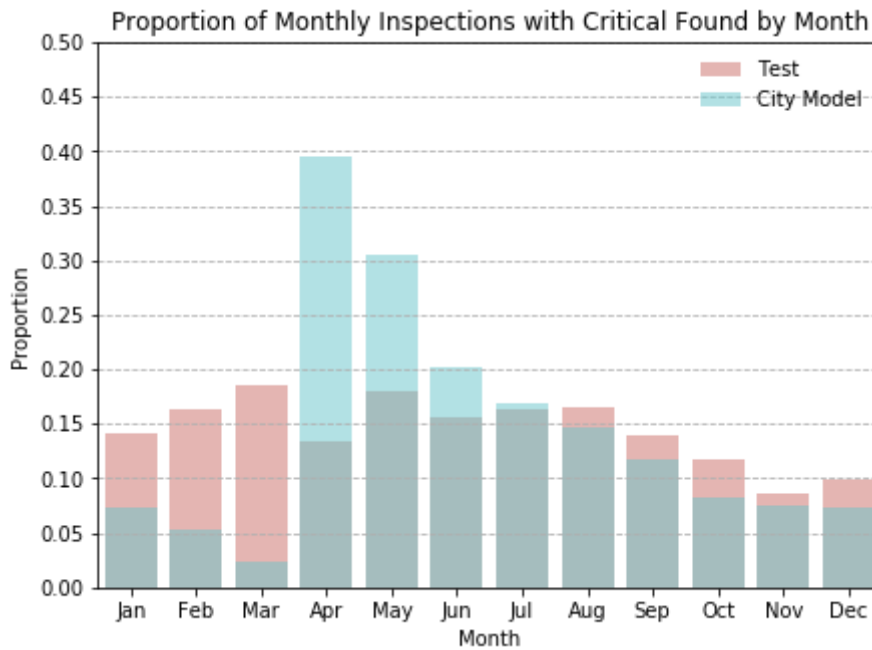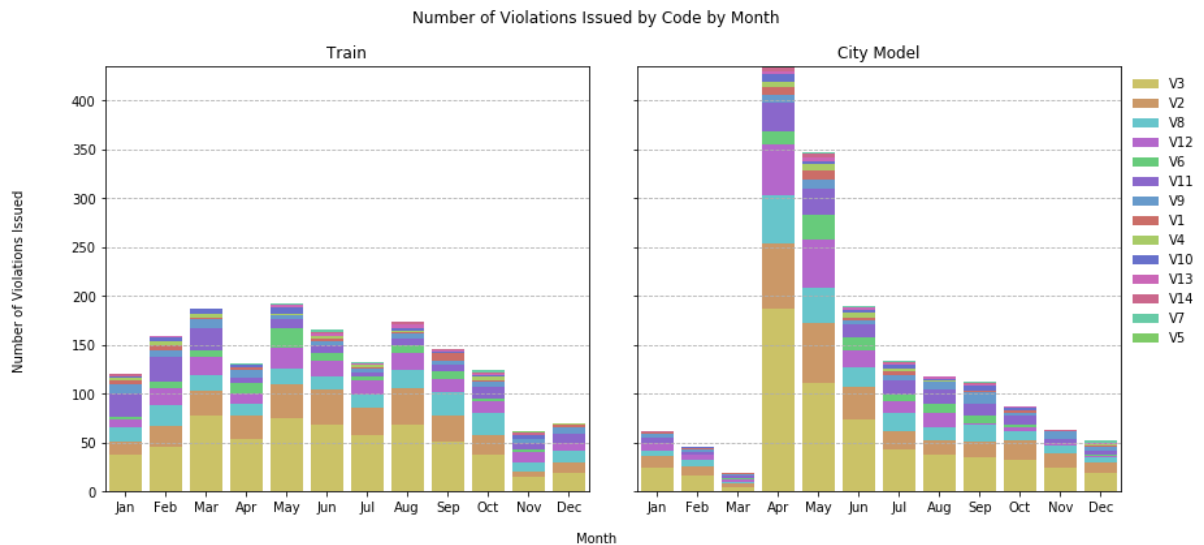Proportion of Monthly Inspections with Critical Found by Month

Figure 13. Number of Violations Issued by Month

- Remember that the test set starts in April and continues through the year to March.
- Monthly totals for each code are stacked in order of decreasing frequency of the code in the train set.

Number of Violations Issued by Code by Month

- Figure 14 will answer: Which month had the **highest fraction of all inspections** issuing the given code?
- Figure 15 will answer: Which month had the **highest rate of monthly inspections** issuing the given code?

Figure 14. Proportion of All Violations with Violation Code by Month

- Remember that the test set starts in April and continues through the year to March.
- The series on each subplot will sum to one.

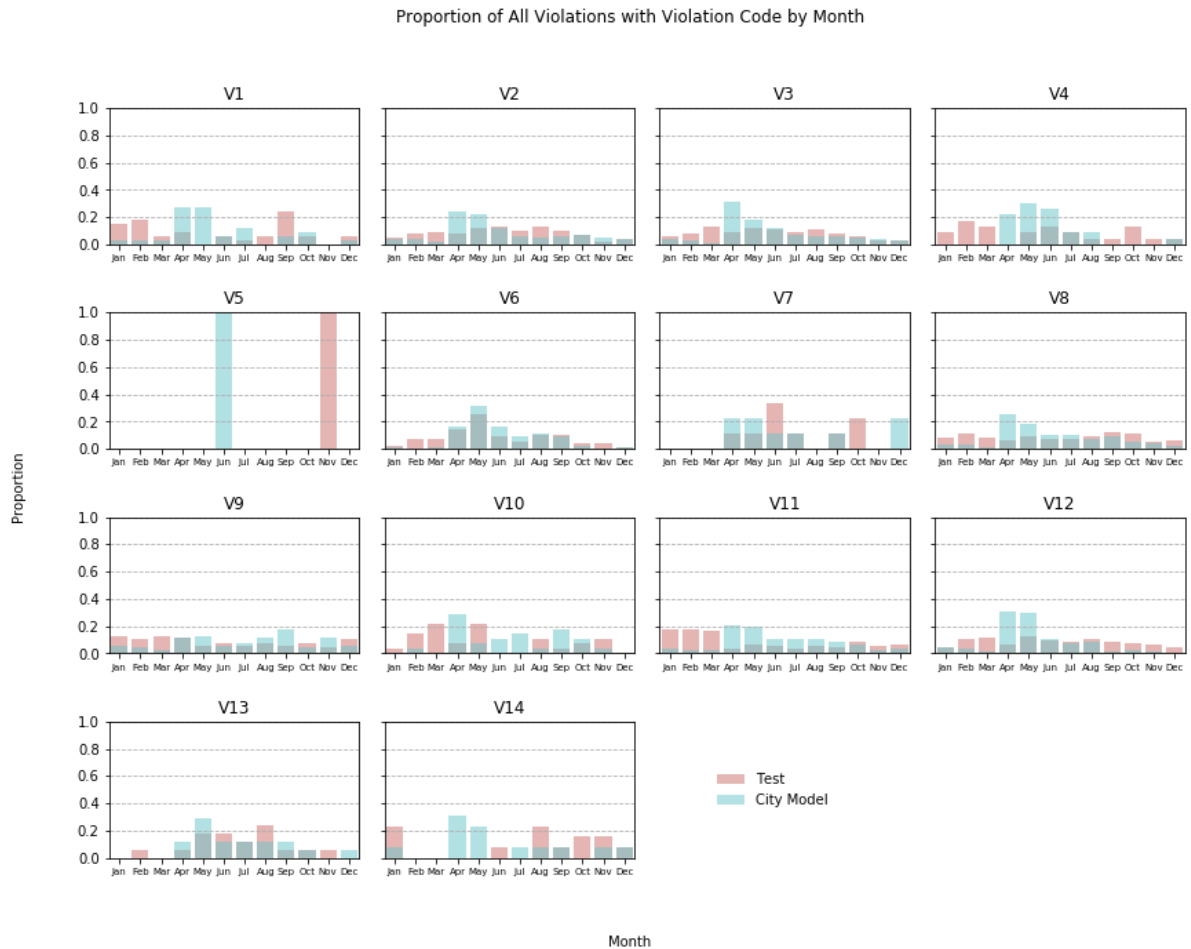Proportion of All Violations with Violation Code by Month

Figure 15. Proportion of Monthly Violations with Violation Code by Month (Different Axes)

- Different data from Figure 14.
- Remember that the test set starts in April and continues through the year to March.
- The values for each month will sum to one.
- Each subplot axis has a different range, to emphasize less frequent violation codes.

Proportion of Monthly Violations with Violation Code by Month