CPSC 340

Assignment 1

Friday, September 23, 2016

Brian Taylor

Student Number: 52311859

CS Account: z2b0b

* 1. Summary Statistics

The following table shows the minimum, maximum, mean, median and mode for each of the regions (columns).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 'NE' | 'MidAtl' | 'ENCentral' | 'WNCentral' | 'SAtl' | 'ESCentral' | 'WSCentral' | 'Mtn' | 'Pac' | 'WtdILI' |
| min | 0.4280 | 0.4830 | 0.4520 | 0.4640 | 0.4680 | 0.5540 | 0.4560 | 0.3520 | 0.3770 | 0.6060 |
| max | 2.3100 | 2.2050 | 2.5150 | 3.1150 | 2.7140 | 3.8590 | 3.2190 | 4.8620 | 2.6600 | 3.2600 |
| mean | 1.2233 | 1.2335 | 1.2753 | 1.4602 | 1.2988 | 1.5625 | 1.2923 | 1.2700 | 1.0632 | 1.5670 |
| median | 1.1295 | 1.1160 | 1.2650 | 1.2775 | 1.1025 | 1.4165 | 1.1075 | 0.9785 | 0.9570 | 1.3035 |
| mode | 0.4280 | 0.4900 | 0.4520 | 0.5050 | 0.5420 | 0.5540 | 0.4990 | 0.3520 | 0.3770 | 0.7150 |

The following table shows the minimum, maximum, mean, median and mode for the entire dataset.

|  |  |
| --- | --- |
|  | X |
| min | 0.3520 |
| max | 4.8620 |
| mean | 1.3246 |
| median | 1.1590 |
| mode | 0.7700 |

The following table shows the 10%, 25%, 50%, 75%, and 90% quantiles for each of the regions (columns).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 'NE' | 'MidAtl' | 'ENCentral' | 'WNCentral' | 'SAtl' | 'ESCentral' | 'WSCentral' | 'Mtn' | 'Pac' | 'WtdILI' |
| 10% | 0.4587 | 0.5095 | 0.4894 | 0.5050 | 0.5420 | 0.5985 | 0.4978 | 0.4363 | 0.4196 | 0.6358 |
| 25% | 0.7135 | 0.7105 | 0.6965 | 0.6765 | 0.7275 | 0.9125 | 0.6830 | 0.6055 | 0.5510 | 0.7905 |
| 50% | 1.1295 | 1.1160 | 1.2650 | 1.2775 | 1.1025 | 1.4165 | 1.1075 | 0.9785 | 0.9570 | 1.3035 |
| 75% | 1.6940 | 1.7265 | 1.6745 | 1.9980 | 1.7610 | 2.1090 | 1.7765 | 1.6975 | 1.4630 | 2.3460 |
| 90% | 2.1449 | 2.0492 | 2.0869 | 2.6998 | 2.2954 | 2.7197 | 2.3714 | 1.9997 | 1.6162 | 3.0899 |

The following table shows the 10%, 25%, 50%, 75%, and 90% quantiles for the entire dataset.

|  |  |
| --- | --- |
|  | X |
| 10% | 0.5015 |
| 25% | 0.7170 |
| 50% | 1.1590 |
| 75% | 1.8135 |
| 90% | 2.3170 |

The following table shows the mean and variance for each of the regions (columns).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 'NE' | 'MidAtl' | 'ENCentral' | 'WNCentral' | 'SAtl' | 'ESCentral' | 'WSCentral' | 'Mtn' | 'Pac' | 'WtdILI' |
| mean | 1.2233 | 1.2335 | 1.2753 | 1.4602 | 1.2988 | 1.5625 | 1.2923 | 1.2700 | 1.0632 | 1.5670 |
| variance | 0.3603 | 0.3222 | 0.3653 | 0.6632 | 0.4244 | 0.6856 | 0.5181 | 0.7988 | 0.3220 | 0.7467 |

The region with the highest mean is 'WtdILI' and the region with the lowest mean is 'Pac'. The region with the highest variance is 'Mtn' and the region with the lowest variance is 'Pac'.

The following table is the correlation matrix for the different regions.

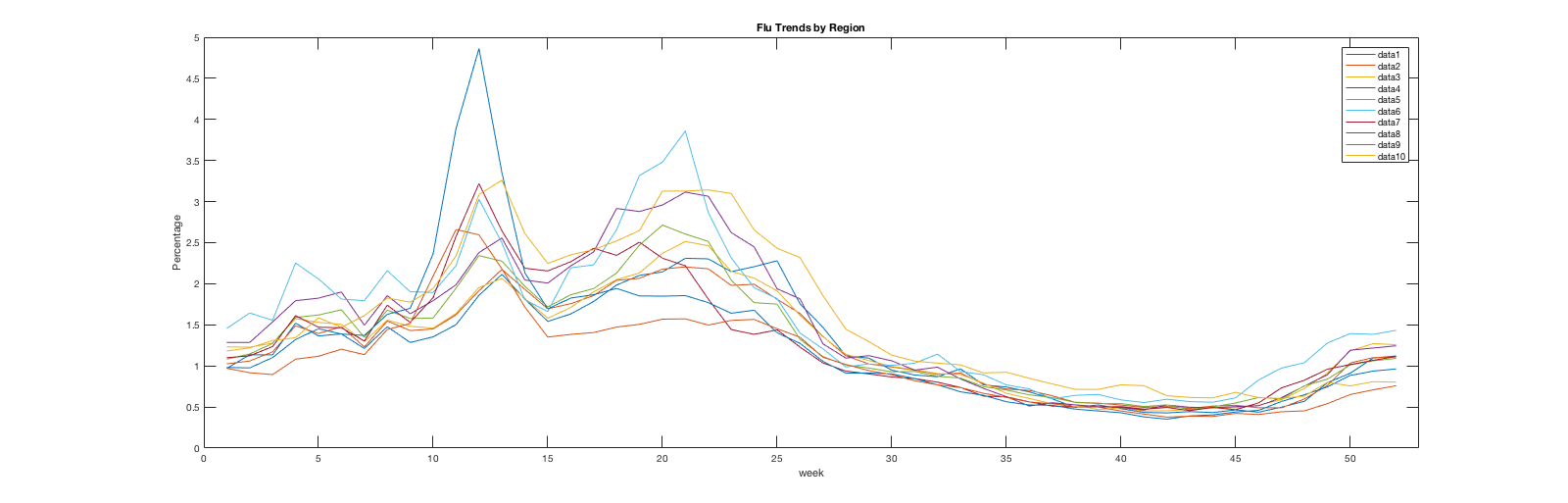
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 'NE' | 'MidAtl' | 'ENCentral' | 'WNCentral' | 'SAtl' | 'ESCentral' | 'WSCentral' | 'Mtn' | 'Pac' | 'WtdILI' |
| 'NE' | 1 | 0.9859 | 0.9826 | 0.9621 | 0.9467 | 0.8868 | 0.8330 | 0.7093 | 0.8208 | 0.9533 |
| 'MidAtl' | 0.9859 | 1 | 0.9879 | 0.9813 | 0.9742 | 0.9123 | 0.8936 | 0.7639 | 0.8591 | 0.9628 |
| 'ENCentral' | 0.9826 | 0.9879 | 1 | 0.9859 | 0.9744 | 0.9374 | 0.8678 | 0.7302 | 0.8249 | 0.9399 |
| 'WNCentral' | 0.9621 | 0.9813 | 0.9859 | 1 | 0.9844 | 0.9523 | 0.8920 | 0.7391 | 0.8207 | 0.9427 |
| 'SAtl' | 0.9467 | 0.9742 | 0.9744 | 0.9844 | 1 | 0.9680 | 0.9278 | 0.7970 | 0.8601 | 0.9460 |
| 'ESCentral' | 0.8868 | 0.9123 | 0.9374 | 0.9523 | 0.9680 | 1 | 0.8938 | 0.7485 | 0.7882 | 0.8714 |
| 'WSCentral' | 0.8330 | 0.8936 | 0.8678 | 0.8920 | 0.9278 | 0.8938 | 1 | 0.9107 | 0.9042 | 0.8722 |
| 'Mtn' | 0.7093 | 0.7639 | 0.7302 | 0.7391 | 0.7970 | 0.7485 | 0.9107 | 1 | 0.9538 | 0.7936 |
| 'Pac' | 0.8208 | 0.8591 | 0.8249 | 0.8207 | 0.8601 | 0.7882 | 0.9042 | 0.9538 | 1 | 0.8778 |
| 'WtdILI' | 0.9533 | 0.9628 | 0.9399 | 0.9427 | 0.9460 | 0.8714 | 0.8722 | 0.7936 | 0.8778 | 1 |

The pair of regions with the highest correlation is 'ENCentral' and 'MidAtl' with a correlation of 0.9879. The pair of regions with the lowest correlation is 'Mtn' and 'NE' with a correlation of 0.7093.

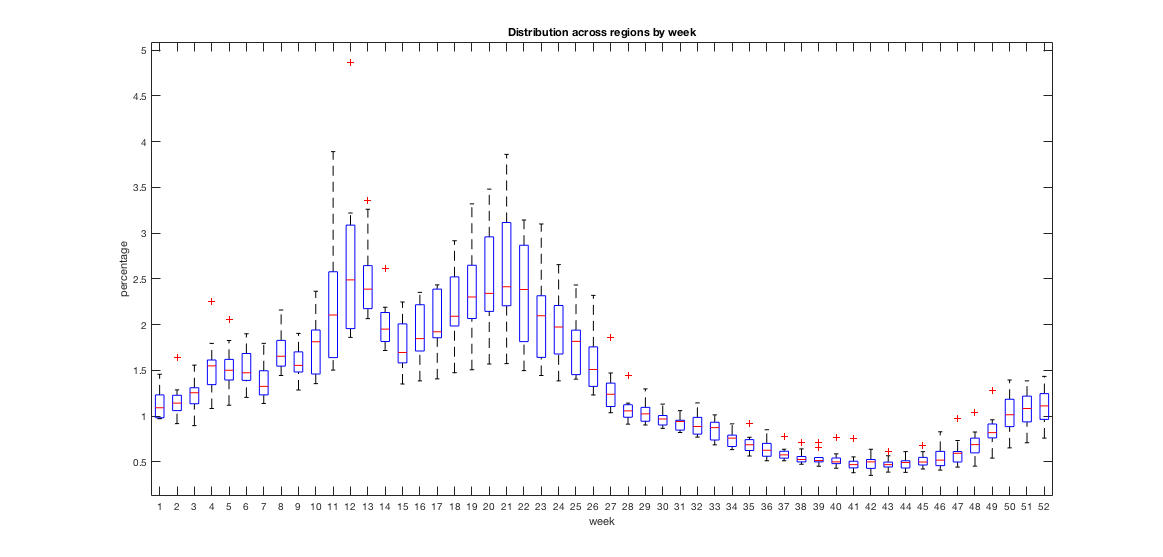
Mode is not a particularly useful summary statistic for this sort of data. A more meaningful ‘mode’ measurement for continuous data would require binning the data first into a number of discrete bins and then looking at the bin with the largest number of data points.

* 1. Summary Statistics

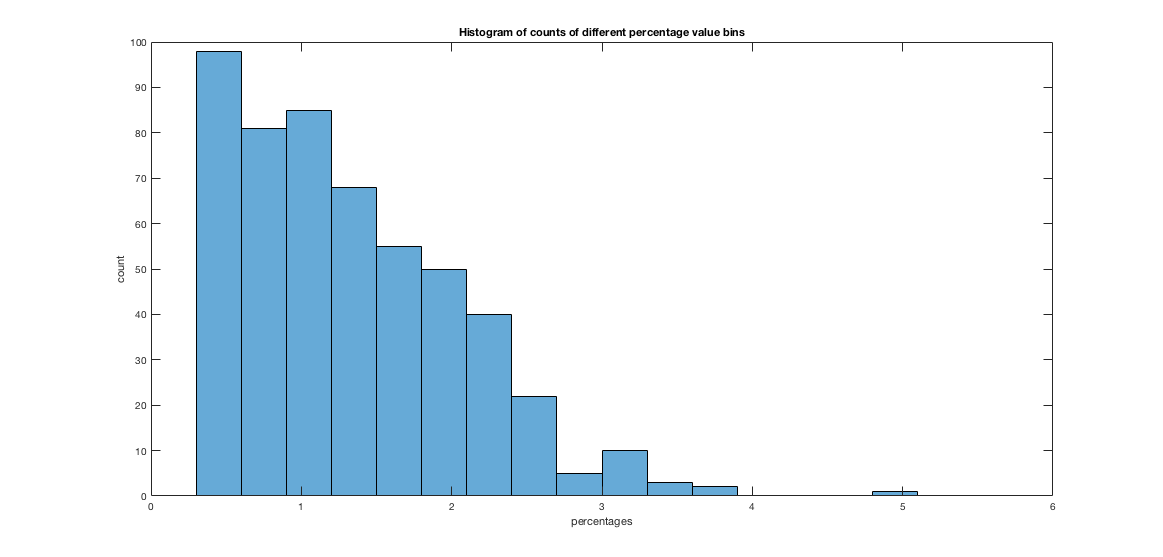
A plot containing the weeks on the x-axis and the percentages for each region on the y-axis.



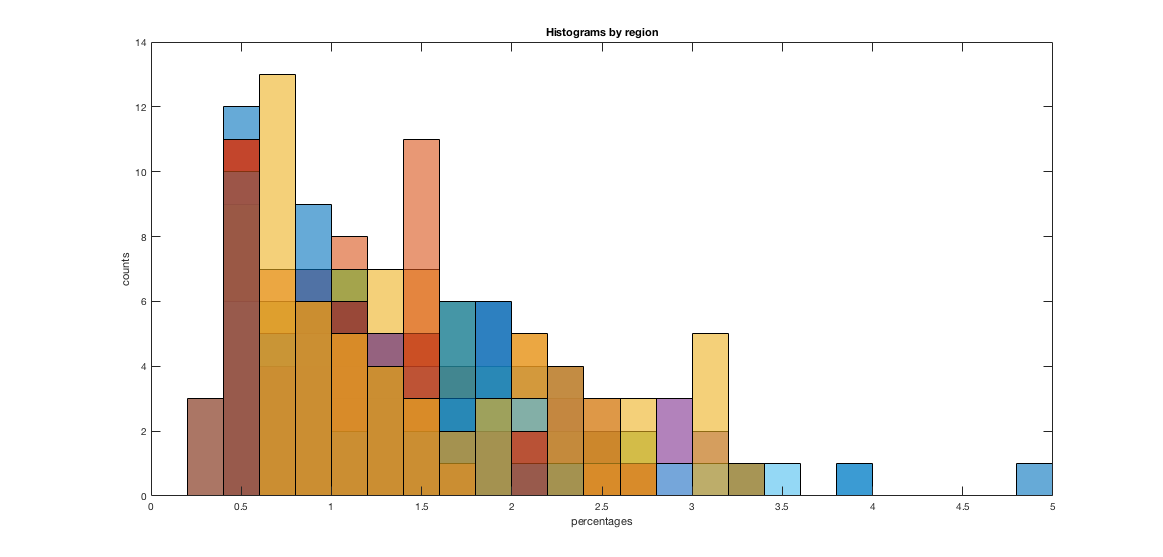
A boxplot grouping data by weeks, showing the distribution across the regions for each week.



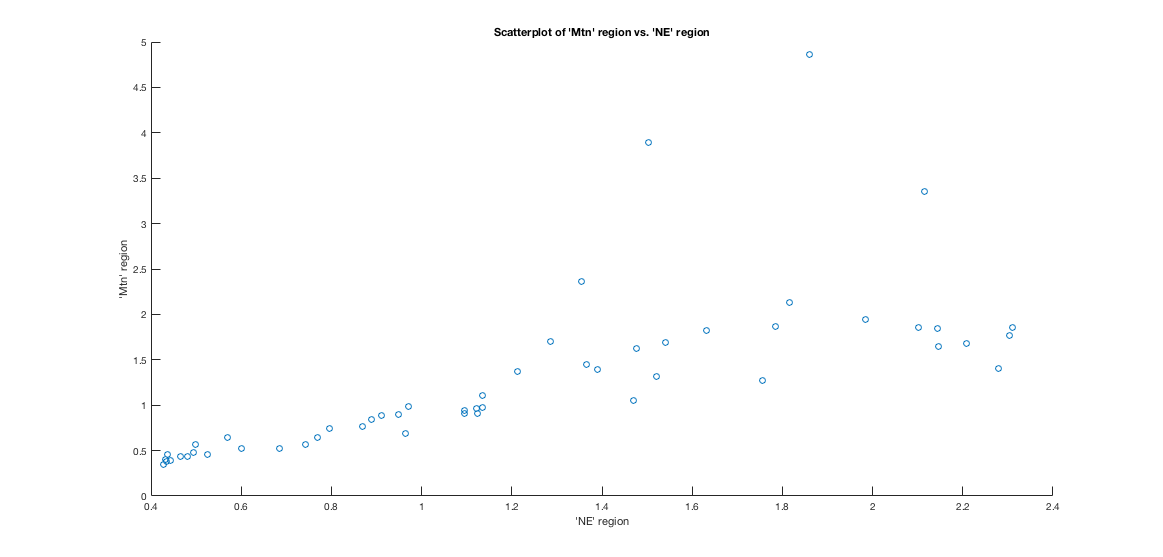
A histogram showing the distribution of all the values in the matrix.



A single histogram showing the distribution of all the columns in X.



A scatterplot between the two regions with lowest correlation.



A scatterplot between the two regions with highest correlation.

