Module 6: Data Sampling

DAT-375

Brandon Hobbs

April 3, 2023

This assignment had me analyze crime data from the Miami area for 2019. The original data set consist of 443 records each with 6 attributes. This data set was then sampled by taking every tenth record, i.e., records 1, 11, 21…441.

The larger data set consists of 51 unique categories, but the sampled set only has 18.

Upon pivoting each data set to ascertain which Crime type had the greatest occurrence Tables 1 and 2 may be generated – only the Top 12 categories are shown.

**Table 1: Count of Crime Type from Larger Data Set**

|  |  |
| --- | --- |
| Crime Type | Count of Crime Type |
| BURGLARY | 93 |
| THEFT - ALL OTHER | 52 |
| CRIMINAL MISCHIEF / VANDALISM | 46 |
| RETAIL THEFT | 27 |
| THEFT/LARCENY | 27 |
| BURGLARY - RESIDENCE | 26 |
| PETIT THEFT | 22 |
| THEFT - PICK POCKET | 19 |
| GRAND THEFT | 13 |
| BURGLARY - OTHER STRUCTURE | 12 |
| ROBBERY - STRONG ARM | 10 |
| PETIT THEFT - RETAIL / LESS THAN $100 | 9 |

|  |  |
| --- | --- |
| Crime Type | Count of Crime Type |
| BURGLARY | 8 |
| THEFT - ALL OTHER | 6 |
| BURGLARY - RESIDENCE | 5 |
| CRIMINAL MISCHIEF / VANDALISM | 4 |
| ROBBERY - STRONG ARM | 3 |
| RETAIL THEFT | 3 |
| GRAND THEFT - UNSPECIFIED - $300 TO LESS THAN $5000 | 2 |
| PETIT THEFT - RETAIL / LESS THAN $100 | 2 |
| THEFT/LARCENY | 2 |
| PETIT THEFT | 2 |
| BURGLARY / DWELLING OCCUPIED (NO ASSAULT / BATTERY - NO WEAPON) | 1 |
| THEFT - FROM BUILDING | 1 |

**Table 2: Count of Crime Type After Sampling Larger Data Set**

The count of individual Crime types is reduced due to the sampling, but more subtly the order and contents of these tables has changed.

To illustrate the relative changes between individual categories’ counts and their relative weights Figure 1 was created. Figure 1 has been sorted by the categorical count according to the count in the larger data set.

**Figure 1: Plot Showing the Count of Violations per Category and Their Relative Weights**

A few key changes occur in Figure 1, namely *Burglary Residence*, *Petite Theft – Retail / Less Than $100*, and *Robbery Strong Arm*, in the sampled set these categories now carry a much larger weight than they did in the original data. Moreover, *Theft - Pickpocket*, *Burglary – Other Structure* carry a smaller relative weight.

At a sampling rate of every tenth record the sampled data set no longer looks like the larger set. As shown in Figure 1 the relative weights have been skewed, but three categories in the top ten are missing (*theft - pick pocket, grand theft,* and *burglary - other structure*) – let alone the 33 other categories completely missing in the smaller data set.

While 443 is not unmanageable other data sets are on the order of millions of records with hundreds of attributes, e.g., the US Census data. At this scale it becomes unwieldy to use all the data and sampling is needed. Sampling tries to obtain a representative subset of a larger population or group of items that can be analyzed to draw conclusions about the entire population. Taking a sample allows for more efficient and effective analysis, as it enables researchers and professionals to draw conclusions about a larger population based on a smaller, more manageable subset