# Houston Police Department (HPD) Crime (Part 1) Data, 2009-2017

## Description

Houston Police Department (HPD) Crime (Part 1) Data are in alignment with the FBI's nationwide Uniform Crime Reporting (UCR) Program to collect local crime statistics. Crime (Part 1) offense types include murder, rape, robbery, aggravated burglary, theft, and auto theft.

## Source Data

**Original files:**

* HPD\_Crime\_2009\_06\_to\_2015\_12.xlsx
  + Created 11/7/2016
  + Source: <http://www.houstontx.gov/police/cs/crime-stats-archives.htm>
  + June 2009 – December 2015
  + See “HPD\_Crime\_RawDataAcquisition.docx” for notes on raw dataset creation
* [MMMYY].xls files
  + Where MMM=”jan” to “dec”, and YY=”16 or “17”
  + Downloaded 3/12/2018
  + Source: <http://www.houstontx.gov/police/cs/crime-stats-archives.htm>
  + 24 files to cover years 2016 and 2017
* HPD\_Crime\_Documentation
  + Contains record counts by month/year of files, to give an annual and sum total of all records from June 2009 – December 2015
* Premise\_Code\_Dictionary.xlsx
  + Created 10/25/2016
  + Source: <http://www.houstontx.gov/police/cs/beatpages/premise.htm>
  + See Appendix A for notes on dataset creation
* Premise\_Code\_Dictionary\_v02.xlsx
  + Created 5/2/2018
  + Source: Updated from Premise\_Code\_Dictionary.xlsx
  + See Appendix A for notes on dataset creation
* HPD\_beat\_map.pdf
  + A map of HPD’s beats, districts, store fronts, and stations
  + Source: <http://www.houstontx.gov/police/contact/images/hpd_beat>\_map.pdf
  + Not used in data processing – this is for reference only

**Notes:**

2009-2015 data were also downloaded on 3/12/2018 and compared to the 2009-2015 data acquired in 2016. It was determined that the source files are stagnant, and had not been updated since 2016. Thus, 2016-2017 data were processed and then combined with the previously processed 2009-2015 data.

## Processing

The raw data file for 2009-2015 contains 851,787 records and 14 variables. The raw data file for 2016-2017 contains 242,420 records, for a total of 1,094,207 records. Data were minimally processed to remove invalid values from select fields, remove extraneous dashes and apostrophes, rename and format fields, and create new constructed variables. A file was then generated for use in geocoding. When geocoding was completed, the data were combined with the results of geocoding to create a first level processed and GIS-enabled file.

**Unique ID for each record (RecordID variable)** – There is no unique ID supplied with the raw data. The field RecordID was created to uniquely identify each record. This RecordID is used for geocoding as well as subsequent processing.

**Hour** – The Hour variable was converted from character to numeric. This effectively standardized values with and without leading zeros to have the same value. For example, “2” and “02” are standardized to the numeric value of 2.

**Premise (Location) Code and Description** – The premise information (where the crime was committed) was reported as a coded value through May, 2014. Starting in August 2014, the premise information is supplied in the form of a long description. For records where the code is supplied in the raw data, it is used to attach the standard Code and Description from the *Premise\_Code\_Dictionary.xlsx* file. Similarly, where the description exists, it is used to attach the standard Code and Description from the *Premise\_Code\_Dictionary.xlsx* file for 2009-2015 source records, and the updated *Premise\_Code\_Dictionary\_v02.xlsx* file for 2016-2017 source records.

Notes on anomalies/issues with Premise (Location) information:

* Missing Location information
  + Premise is missing entirely for the source months of June and July 2014. This is consistent with the original data files on the HPD website.
  + Premise is missing for 40% of the records for source months August and September 2014, and 26% in October 2014.
  + With so much missingness in 2014, be cautious when using location information during that time period.
* “APT” – There is one record with premise = “APT”. It is assumed that this corresponds to the premise Description of “Apartment”.
* “RES” – There are two records with premise = “RES”. It is assumed that this corresponds to the premise Description of “Residence or House”.
* “23C” – There are 6 records with premise = “23C”. This Code does not exist in the list on the HPD website. Thus, these are not assigned a Description.
* Six premise descriptions in the 2016-2017 source data have no corresponding code in the Premise Code Dictionary. These were checked against the online source, and no code was available.

**Additional Notes**

* Per conversation with HPD on 11/3/16, the “Date” field (renamed to Date\_Reported) is the date the crime was reported. Older dates in the variable reflect updates to cases (i.e. new supplemental information like lab results coming in, etc…).
* Per conversation with HPD on 11/3/16, the “Suffix” field is really the address prefix. This field was renamed to “Street\_Prefix”.
* The system for data collection was upgraded sometime around 2013. Changes in consistency across years may be a result of this system change.
* The type of offense variable (Offense\_Type) contains the invalid value of “1” for 8 records. These records are found in the 2015 source file months of January, March, April, May, July, October, November, and December (one in each month).
* Premise information changed from “codes” to “descriptions in 2014. Variable created that indicates whether the original record has a code, a description, or is missing.

## Geo-Coding, other GIS Processing

The address information provided in the raw data file includes only the street (prefix, name, and type) and a block range (for example, 5600-5699). The address information was prepped for geocoding as follows:

* Add City column, with all values = “Houston”
* Add State column, with all values = “TX”
* Add Zipcode column, with all values left empty
* Add Mid\_block column, where Mid\_block = lower block number + 50
* Add Address column, where Address  = Mid\_Block + Prefix + StreetName + StreetType)

***NOTE:*** *It is important to note that the geocoding was done using the mid-block address, which was always xxx50. Thus, the coordinates are an approximation of the actual location of the crime. Similarly, the census block attached is based on this approximate address. Analysts should account for this when using the data.*

### 2009-2015

The 2009-2015 addresses were processed using the “AutomaticGeocodingTool\_V1.97\_UDP”[[1]](#footnote-1) automatic geocoder.

The process took 14 hours to run, with the following results:

* Records with geocode score >= 80:  838,263 or 98.4%
* Unmatched records (below score 80):  13,524  or 1.6%
* Records with 100% match: 62,2857 or 74.3%

Only records with score >=80 were accepted as matches. For matched addresses, the following information was attached to the record:

* X/Y coordinates (GCS\_WGS\_1984)
* 2000 and 2010 census block
* Police Division
* Community Tabulation Area (CTA)
* Zip code

### 2016-2017

The 2016-2017 addresses were geocoded using CEHI's AutomaticGeocodingTool\_V1.99 [[2]](#footnote-2). Batch 1 records were matched with a minimum match score of 100% to one of three address locators: StreetMap Premium 2017’s building-level locator (SMP\_Pt17), StreetMap Premium 2017’s street-level locator (SMP\_St17), or Business Analyst 2013’s street-level locator (BA\_St), listed in order of decreasing priority. Batch 2 records were matched with a minimum match score of at least 90% to one of two address locators: StreetMap Premium 2017’s street-level locator (SMP\_St17), or Business Analyst 2013’s street-level locator (BA\_St), listed in order of decreasing priority. Batch 3 records were manually reviewed and matched to StreetMap Premium 2017’s street-level locator (SMP\_St17).

X and Y coordinates are in GCS\_WGS\_1984 format, and were set to 0 for un-geocoded records. Census block identifiers for 2000 and 2010, were attached to each geocoded address. Note that it is possible for an address to be assigned to a different county in 2000 than in 2010. This is due to the precision of the 2000 v. 2010 geography files used to assign the Census block. There are 6 addresses with county assignment discrepancies, affecting 6 records.

### Combined 2009-2017 Results

Notes:

* Since zip code, city and state are not included in the source data, these fields are dropped after the geocoding is combined with the base data.
* The NGC (Not GeoCodable) indicator was not created during the geocoding process for the 2009-2015 source data.

Geocoding results by year for the combined 2009-2017 data are as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year Report Date[[3]](#footnote-3)** | **Total** | **Geocodable[[4]](#footnote-4)** | | **Geocoded** | | |
| **#** | **%** | **#** | **% of Geocodable** | **% of Total** |
| **Missing/Invalid\*** | 10 | - | - | 10 | - | 100 |
| **Pre-2009** | 1,402 | - | - | 1,374 | - | 98.00 |
| **2009** | 89,050 | - | - | 87,801 | - | 98.60 |
| **2010** | 136,357 | - | - | 134,576 | - | 98.69 |
| **2011** | 127,168 | - | - | 125,467 | - | 98.66 |
| **2012** | 127,486 | - | - | 125,711 | - | 98.61 |
| **2013** | 130,200 | - | - | 128,337 | - | 98.57 |
| **2014** | 121,451 | - | - | 119,058 | - | 98.03 |
| **2015** | 120,079 | - | - | 117,264 | - | 97.66 |
| **2016** | 122,423 | 121,093 | 98.91 | 115,353 | 95.26 | 94.22 |
| **2017** | 118,581 | 116,809 | 98.51 | 108,309 | 92.72 | 91.34 |

\*Year of the date crime reported was missing or greater than 2017.

## Published File(s), Programs and Audit Trail

HPD\_Crime\_2009\_2015\_v01 851,787 records, 25 variables

HPD\_Crime\_2009\_2017\_v01 1,094,207 records, 23 variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Program Name** | **Input File(s)** | **Output File(s)** | **Date/Time Run** |
| Read\_hpd\_crime\_2009\_2015 | HPD\_Crime\_2009\_06\_to\_2015\_12 | HPD\_Crime\_2009\_2015\_Raw  HPD\_Crime\_2009\_2015\_ToGeo.dbf | 11/7/2016 11:07:12 am |
| GIS process – GeoCode | HPD\_Crime\_2009\_2015\_ToGeo.dbf | Crime\_Geocoded\_After\_Join.txt | 12/9/2016  02:58:18 pm |
| CrHPD\_Crime\_2009\_2015 | HPD\_Crime\_2009\_2015\_Raw  Crime\_Geocoded\_After\_Join.txt | HPD\_Crime\_2009\_2015\_v01  HPD\_Crime\_2009\_2015\_v01.csv  HPD\_Crime\_2009\_2015\_v01.dta | 12/19/2016  08:57:33 am |
| CrHPD\_Crime\_2009\_2017\_v01 | HPD\_Crime\_2016\_2017\_v00  HPD\_Crime\_2016\_2017\_v01\_ToGeo\_GEOID00\_10.dbf  HPD\_Crime\_2009\_2015\_v01  Premise\_Code\_Dictionary\_v02.xlsx | HPD\_Crime\_2009\_2017\_v01  HPD\_Crime\_2009\_2017\_v01.csv  HPD\_Crime\_2009\_2017\_v01.dta  Cr\_HPD\_Crime\_2009\_2017\_v01.rtf | 05/09/2018 10:41:37 am |

*Note: Input/Output files are SAS datasets (.sas7bdat) unless otherwise noted.*

Appendix A

## Creation of Premise Code Dictionary

The Premise (Location) code/description dictionary was created from a table found on the HPD website that was hand modified. The table <http://www.houstontx.gov/police/cs/beatpages/premise.htm> was downloaded and placed into an Excel file. The premise (location) codes and descriptions from the website are the “standard” values. The codes and descriptions found in the raw data were then matched and added to the dictionary. The primary differences occurred in the descriptions, where the standard values are all upper case, and the values in the raw data are mixed case. In addition, small formatting differences exist, such as “Park and Ride Terminal” v. “PARK & RIDE TERMINAL”.

12/1/2016 Created by Danielle Jean-Louis

12/1/2016 Visually verified by Hien Le

* Checked each column for duplicate values
* Checked that “Premise Type = “Standard Premise Type”
* Checked that “Premise Description” is essentially equal to “Standard Premise Description”
* Three anomalies found – for entries of ‘APT’, ‘RES’, and ‘N’.

12/2/2016 Modifications made by Claire Osgood to delete entries for ‘APT’, ‘RES’, and ‘N’.

12/2/2016 Visually re-verified by Hien Le, comparing “standard” values in the dictionary to the original table on the HPD website. All checks passed.

This dictionary is used to attach the standard location code and description to each record that contains premise information.

04/30/2018 Updated by Claire Osgood

Copied the original version to create the V02 version, and added 5 new rows, one for each new spelling of a location that already existed in the dictionary. These new entries have a missing value for the Premise Type (code) because the 2016-2017 data does not include that information. They are matched to existing standardized codes and descriptions.

05/02/2018 Verified, and updated by Mamta Singh

Added 6 new rows to the V02 version, one for each location that did not exist in the dictionary. These new entries have a missing value for the Premise Type (code) and also for the standardized code. They are assigned to a standardized description which is the UPCASE version of the Premise Description.

1. The geocoding tool was developed by Ruiyang Li, and is found at K:\Admin\IT\ArcGIS\AutomaticGeocodingTool\_V1.97\_UDP . [↑](#footnote-ref-1)
2. The geocoding tool was developed by Ruiyang Li, and is found at K:\Admin\IT\ArcGIS\CEHI\_AutomaticGeocodingTool\_V1.99\_UDP\_2017. [↑](#footnote-ref-2)
3. Year the crime was reported, which may be different from the year of the source file. [↑](#footnote-ref-3)
4. This information is only available for the 2016-2017 source files. Though there are some reports in years 2009-2015 in those source files, there are very few and thus the information is not reported here. [↑](#footnote-ref-4)