Austin Datasets

Contents

[Notes 1](#_Toc499558362)

[Merge 1](#_Toc499558363)

[Socioeconomic Data 3](#_Toc499558364)

[Old 4](#_Toc499558365)

# Notes

11/27, Mika

## Merge

* Merge.py
* What has been done?

1. We put the AUF2015 and AAnn datasets on top of each other
   1. This puts all of the columns from both datasets into one big pandas dataframe
   2. Where there is no data for a given cell (e.g. if that row came from AUF2015 and the column is from AAnn), the cell has a NaN
2. We replaced all NaNs in the following columns w/data from their corresponding columns in the other dataset:
   1. PrimaryKey – GO Primary Key
   2. CouncilDistrict – Council District
   3. X-Coordinate – GO X Coordinate
   4. Y-Coordinate – GO Y Coordinate
3. We removed spaces from all column names and renamed some columns so that it’s easier to navigate the dataset
   1. 'PrimaryKey': 'Key'
   2. EffectonOfficer': 'OfficerEffects'
   3. 'NatureofContact':'NatureOfContact'
   4. 'HighestNIBRS/UCROffenseDescription':'NIBRS'
   5. 'OfficerYrsofService': 'OfficerYrsServ'
   6. ‘X-Coordinate’: ‘XCoord’
   7. ‘Y-Coordinate’: ‘YCoord’
4. We created a column of True/False values for Use of Force incidents (‘UF’)
   1. Don’t know why the length and output in pandas is different. When you open the .csv in excel, it’s the full dataset.
   2. There do appear to be repeating values in the first column – like it starts to repeat somewhere in the 3000s, however the Key associated with these is independent and non-repeating, and the data itself doesn’t seem to be repeating. As long as we reference Key instead of that very first index column, I think we should be good. Therefore, we:
5. We set the primary key to ‘Key’
6. We re-ordered the table so that the important stuff we care about @ first are in the first columns, and then columns are sorted according to topic e.g. officer-related, subject-related, location etc.

Order of columns

(Important stuff): 'Key','CouncilDistrict','UF','XCoord','YCoord','RIN','DateOccurred', 'TimeOccurred','ClearanceDate', 'GOReportDate','R2RLevel','NIBRS'

(Location-related): 'AreaCommand', 'Location', 'GOCensusTract', 'GODistrict','GOLocation','GOLocationZip',

(Officer-related): 'OfficerEffects','OfficerCommissionDate', 'OfficerYrsServ', 'OfficerOrganizationDesc', 'ReasonDesc',

(Subject-related): 'SubjectConductDesc', 'SubjectEffects','SubjectEthnicity', 'SubjectRace', 'SubjectResistance', 'SubjectSex',

(Nature of contact-related): 'NatureOfContact', 'GOHighestOffenseDesc', 'NumberShots', 'WeaponUsed1', 'WeaponUsed2', 'WeaponUsed3', 'WeaponUsed4', 'WeaponUsed5',

(Miscellaneous): 'ClearanceStatus','MasterSubjectID'

1. We sorted based on CouncilDistrict (all council district 1 is @ top, all 9 is at bottom)
2. We saved to .csv: ‘stack.csv’

Now I think we have to figure out how to get the SES and other neighborhood data into here.

11/26, Mika

Merge

* SQL script: Merge.sql
* Database name: Austin.sqlite
  + 2 tables
    - AAnn2015 – the annual crime data from Austin\_Annual\_Crime\_Dataset\_2015.csv
    - AUF2015 – the Use of Force report data from Austin\_UF\_R2R\_2015.csv
* Merged dataset on council district
  + Dataset too big to post to git. Will need to run it yourself to generate.
  + If want to view what it will look like ahead of time without waiting two million years, look at “testAustinMerge\_CD.csv” or run that part of the Merge.sql script.
* Merging UF on *PrimaryKey*…Problem: the merge is keeping rows only where we have a match between the AAnn2015.*GOPrimaryKey* and the AUF2015.*PrimaryKey* (UoF incidents accounted for in overall crime dataset. Also, there’s 300-sum entries - quite a bit more than the 192 I got from Pandas...).
  + So, what we need to figure out is how to get all of the incidents to show up - for this query to not exclude them – while not creating duplicate rows for UF incidents that have a match in the annual crime set. This might be tangential to our new questions, though, which as I understand it are looking just at council district.

## Socioeconomic Data

* Located in <https://github.com/MikaArmenta/IPPP_Abbott_Lovins_Armenta/tree/master/Final/Austin> > Districts10\_Socioeconomics.csv

# Old

* Primary Key
* 11% of UoF incidents are accounted for in the annual crime stats.
* 192 duplicate keys (Crime IDs) between Austin Annual Crime 2015 and Austin UoF 2015
  + Total Austin Annual Crime: 38,573
  + Total Austin UoF 2015: 1,681
* *DupKey* = dictionary of duplicate keys
* Cleaning
* Removed all white spaces from indexes in both datasets
  + e.g. “Primary Key” now = “PrimaryKey”, “GP Primary Key” = “GOPrimaryKey”
* NaNs
  + *AAnn\_nan\_rows* and *AUF2015\_nan\_rows* = dataframe of rows with NaNs in them
* Drop rows with NaNs for *CouncilDistrict*