## Version 1

* Crimes Against Category: removed Property and Society, leaving only People
  + Also deleted field to make dataset smaller
* Victim Sex Category: removed Male and Unknown, leaving only Female
  + Also deleted field to make dataset smaller
* Stolen Vehicles: removed field (not relevant to us)
* BCS Area: removed field (not relevant to us)
* Beat: removed field (not relevant to us)

## Version 2

* Victim Race: combined subcategories to create overall categories
  + Black + East African = Black
  + Other Asian + Filipino + Indian + Vietnamese + Chinese + Laotian + Asian Indian + Cambodian + Japanese + Korean = Asian
  + Pacific Islander + Hawaiian + Samoan + Guamanian = Pacific Islander

## Version 3

* City name:
  + Removed all ‘Sheriff’ values
  + Used ‘Zip Code’ column to fill in missing values for city names - so now the only missing values in the city name column also have missing values in the zip code column
  + Suggest putting BCS area and beat back in, they might help us with filling in more missing values for city name/zip code

## Version 4

* Added ‘Beat’ and ‘BCS’ columns back in to help with finding city names
* City column:
  + Grouped unincorporated areas together with their city names (e.g. 'ESCONDIDO UNINC' → 'ESCONDIDO')
  + Used BCS Areas and Police Beats (police beats data in shared drive) to fill in zip codes & city names. Sometimes, only city names could be filled in since the beats don’t get specific enough for zip codes. So the ‘City’ column should be the **primary column** for location, NOT zip code
  + Used Agency column to fill in some of the city names (e.g. Oceanside PD → “Oceanside” city name)
  + Based on trends for beats and BCS Areas = ‘DETENTION FACILITY”, filled in city names (see city\_data\_cleaning.ipynb in shared drive for details)

## Version 5

* Overall Race:
  + Filled null values with ‘Unknown’ since there were only 93
* Victim Age:
  + Calculated the median age for each racial group, then used these to fill null values.
* Incident Date:
  + Converted data type to DateTime to make it easier to perform date operations on
* Zip Code:
  + Fixed various typos in the data where the zip code and city name did not align
  + Used most frequent zip codes from each city, then used ‘City’ column to fill in null values of Zip Code
* Added Health and Human Service Agency (HHSA) Region Column:
  + Used info from [this website](https://www.sandiegocounty.gov/content/sdc/hhsa/programs/phs/community_health_statistics/regional-community-data.html) to group zip codes together in regions
  + Mapped zip codes to regions and double checked to make sure only the right city names were in each of the regions (see data\_cleaning.ipynb)
* Questions for future cleaning:
  + Do we want to use Census Tract data? Benefits of using this over zip code column?
* Need to fix:
  + Fallbrook and Bonsall should be classified as North Inland, not North Coastal

## Version 6

* HHSA Region:
  + Fixed instances where city names were in the wrong regions
  + Corrected typos in the original data (‘San Ysabel’ should be ‘Santa Ysabel’)
* Questions for future cleaning:
  + For the Age column, there is one row where age = -1 which needs fixing
  + For Age, do we want to remove crimes against children (Age < 18) since we want to focus on women? (This would remove 7600 rows)

## Version 7

* Victim Age: dropped all rows where victim age < 18
* Dropped rows where city/zip code/HHSA Region is null (25 rows)