**EXERCISE-5 DATA MANAGEMENT (ISTM 6212)**

The dataset which I used for this exercise contains the data of crimes committed in Kansas City which are reported in the year 2015. The dataset has around 1,22,000 records of transactional data spread across 24 columns with each record corresponding to a crime committed. The data could be found at the following website:

<https://data.kcmo.org/Crime/KCPD-Crime-Data-2015/kbzx-7ehe>

**Transformations:**

1. The column “Sex” has some mismatched values like “U” and also has some missing values. I transformed this column such that a record with value “U” is renamed to “Unknown” and the missing values are assigned value “Not Available”.
2. The column “Beat” has some mismatched values like “CPD” as 99% of the data in that column has numeric data ranged from 111 to 999 and hence I cleaned that column and assigned the value 99 to mismatched columns and the missing values are also replaced with 000.
3. The column “Reported\_date” is being reported in the format “mm/dd/yyyy hh:mm:ss”. All the records have the time part as “12:00:00 AM”. The time value is same for all the records as the day starts at 12 AM and hence I felt it is better to remove the time part from the “Reported\_date” column. Hence I applied a transformation to create a new column “dateformat\_Reported\_date” which stores only the date part of the column “Reported\_date” and later I dropped the original column. Have applied a similar transformations for columns “From\_Date” & “To\_Date” and deleted original columns. Also replaced missing values from “From\_Date” & “To\_Date” columns with “00/00/0000”.
4. The column “Age” has many missing values. Hence I applied a transformation to assign the value “999” to the records with missing value of age.
5. The column “Race” has many missing values and hence I have replaced them with “NA”.

**Script:**

splitrows *col*: column1 *on*: '\n' *quote*: '\"'  
split *col*: column1 *on*: ',' *limit*: 23 *quote*: '\"'

header

set *col*: Sex *value*: 'Unknown' *row*: mismatched(Sex, ['Gender'])

set *col*: Sex *value*: 'Not Available' *row*: empty([Sex])

set *col*: Beat *value*: 99 *row*: mismatched(Beat, ['Integer'])

set *col*: Beat *value*: 0 *row*: empty([Beat])

derive *value*: dateformat(Reported\_Date, 'MM\/dd\/yyyy') *as*: 'dateformat\_Reported\_Date'

drop *col*: Reported\_Date

derive *value*: dateformat(From\_Date, 'MM\/dd\/yyyy') *as*: 'dateformat\_From\_Date'

drop *col*: From\_Date

derive *value*: dateformat(To\_Date, 'MM\/dd\/yyyy') *as*: 'dateformat\_To\_Date'

drop *col*: To\_Date

set *col*: dateformat\_To\_Date *value*: '00\/00\/0000' *row*: empty([dateformat\_To\_Date])

set *col*: dateformat\_From\_Date *value*: '00\/00\/0000' *row*: empty([dateformat\_From\_Date])

set *col*: Age *value*: 999 *row*: empty([Age])

set *col*: Race *value*: 'NA' *row*: empty([Race])

**Results:**

The results show that the data is 97% valid now with 0.02% of the data being mismatched and 3% of data missing (data missing and mismatched is from the columns on which Transformations are not applied). After having a thorough check on the data I came to a conclusion that all the transformations worked without any error on the whole dataset. All the missing values and mismatched values are corrected in the columns on which transformations are applied. The full results showed me that I didn’t miss any important factors while writing the scripts for the selected columns and after seeing the results I learn that I can go ahead and clean the other columns on which transformations are not applied.

I liked this tool very much as this makes the job of an Analyst easier.

**THANK YOU**