Profiling Analytics Template

User Guide

Introduction

This profiling analytics template has been created to scan the results of multiple profile runs of a single dataset over time and identify the anomalies of the current run against the aggregated statistics derived from the previous runs. This template can be applied to any dataset that has columns with numeric data to be profiled and analyzed.

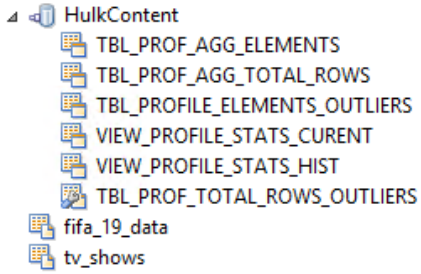
The template can be designed to pick data, of the runs from the last 20 days, and build aggregations and baseline statistics from it. The underlying data of the PDO can be altered, profiled and can be compared with the baseline statistics created (from history). The latest profile run is checked for changes to min and max values of a field below or above the 5-percentile or 95-percentile thresholds respectively, sudden hike of null counts, or drop of distinct value counts, etc.

1. Template Installation Steps

1.1. Create the PDO and Profile

- Choose a physical data object which has been profiled multiple times, over a period
- If this is not available, create a PDO and profile it. Keep changing the underlying data, and profile it at every data change
- We now have a history of runs, on which we can build the aggregated baseline statistics

1.2. Prepare the database



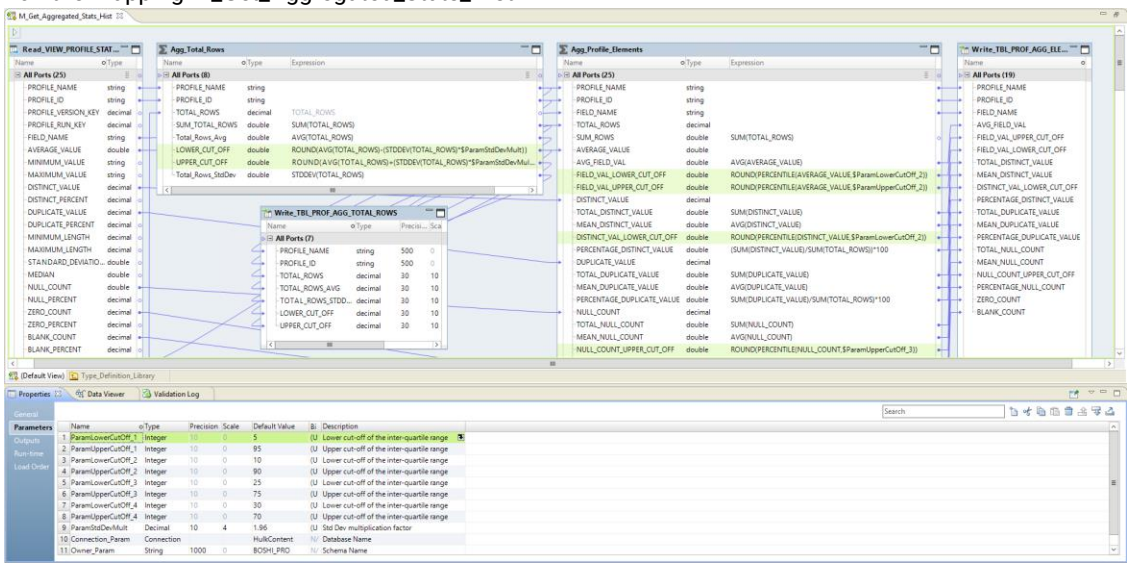
- Create the required views and tables in the database from the Profile_Analytics_Template_DB_Components.sql
- It is recommended to create the views and output tables in a separate schema (eg: reporting schema) that has access to the PWH
- Ensure that relevant permissions (select, update, insert) are given to the tables and views created, in the database, if created outside the profiling schema.

1.3. Import the mappings

- Import the mappings from the Analytics_Mappings.xml file
- There are 2 mappings present: one for getting the aggregated stats for all the historical runs, and the other to compare the current run, with the previous run
- Each mapping has 2 outputs. M_Get_Aggregated_Stats_Hist contains one output for calculating the statistics only for the total rows of the profile runs and the other output for calculating the metrics of all the other fields of the profile runs
- M_Compare_Current_Run_To_Agg_Hist_Stats compares the current profile run's data to the historical stats generated in the previous mapping to give two outputs – one for the outliers based on the row count and the other for calculating the outliers based on the metrics of all the other fields

2. **Execution Steps:**

- Ensure that there is a history of profile runs for the PDO
- Run the mapping M_Get_Aggregated_Stats_Hist:



- You will be able to see the outputs in the tables
- The database and schema names are parameterized
- The cut-offs used in the mapping are parameterized as well. One can use custom values for the standard deviation and the lower and upper cut-off values

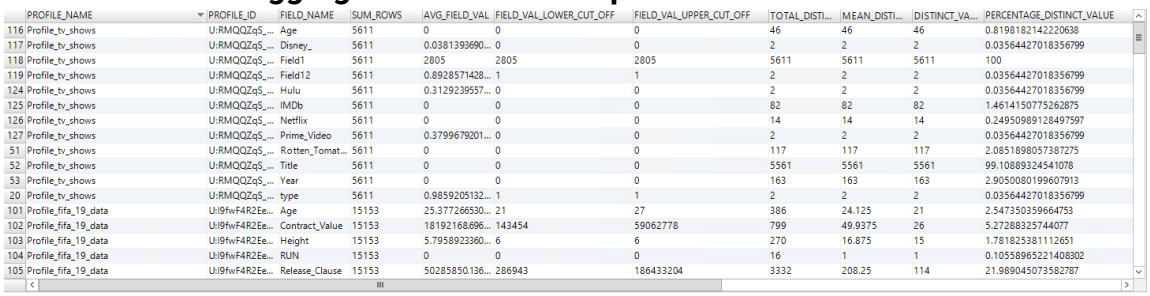
2.1. **Total Rows aggregated statistics output:**

	PROFILE_NAME	PROFILE_ID	TOTAL_ROWS	SUM_TOTAL_ROWS	Total_Rows_Avg	LOWER_CUT_OFF	UPPER_CUT_OFF	Total_Rows_StdDev
1	Profile_Copy_of_Profile_DD_Test	U:9PQCmabyEeqJhnosuWPKAQ	6	48	6	6	6	0
2	Cals_Prof_Test	U:FuV9PacdEeq_eHxJbgLMwg	8	32	8	8	8	0
3	Profile_SSN_Test	U:LxAdtpUyEeqd7UvtLs2JqQ	86	258	86	86	86	0
4	Profile_drizzle_ssn_classic	U:Q7cRCZXtEeqd7UvtLs2JqQ	20020	20020	20020	20020	20020	0
5	Profile_DD_test	U:Q8Q7iZOiEeqkuebFU742_A	13	195	13	13	13	0
6	Profile_O_SSN_Out	U:QDKYcJUZEeqONRBWhLr_1A	20021	80084	20021	20021	20021	0
7	Profile_tv_shows	U:RMQQZqS_Eeqr-PMXEns3Fg	5611	67332	5611	5611	5611	0
8	Profile_drizzle_out	U:SXbjcZXsEeqd7UvtLs2JqQ	20020	100100	20020	20020	20020	0
9	Profile_drizzle_ssn_classic	U:_u6OFJTwEeqCbzwYGIYIkA	20020	40040	20020	20020	20020	0
10	Profile_UK_AV_Demo_Contact	U:boInSqcFEeqCm4FiyTRkRg	298	14304	298	298	298	0
11	Profile_fifa_19_data	U:I9fwF4R2EeqoSWhBCjJwow	1000	106071	947.0625	506	1388	225.0253751957365
12	Calvin_DD_Profile_Test	U:nw0gpKb0Eeq_eHxJbgLMwg	6	56	7	5	9	1.0690449676496976
13	Profile_DD_test1	U:odE5QZRMEeqONRBWhLr_1A	13	195	13	13	13	0
14	Profile_OFAC_consolidated_list_3_6_2020	U:q1KjUWRnEeqPdrWjCWh9ZQ	11324	317072	11324	11324	11324	0

- This table is to record the statistics of the total number of rows, that are processed across the profile runs of all the profiles present in the MRS, over a period of 90 days

Column Name	Description
PROFILE_NAME	Name of the profile for which the statistics are derived
PROFILE_ID	Unique identifier for a profile
TOTAL_ROWS	The total number of rows across the profile runs
TOTAL_ROWS_AVG	Indicates the average number of rows processed across the profile runs
LOWER_CUT_OFF	Indicates the lower threshold of the number of rows across profile runs
UPPER_CUT_OFF	Indicates the upper threshold of the number of rows across profile runs

2.2. **Profile elements aggregated statistics output:**



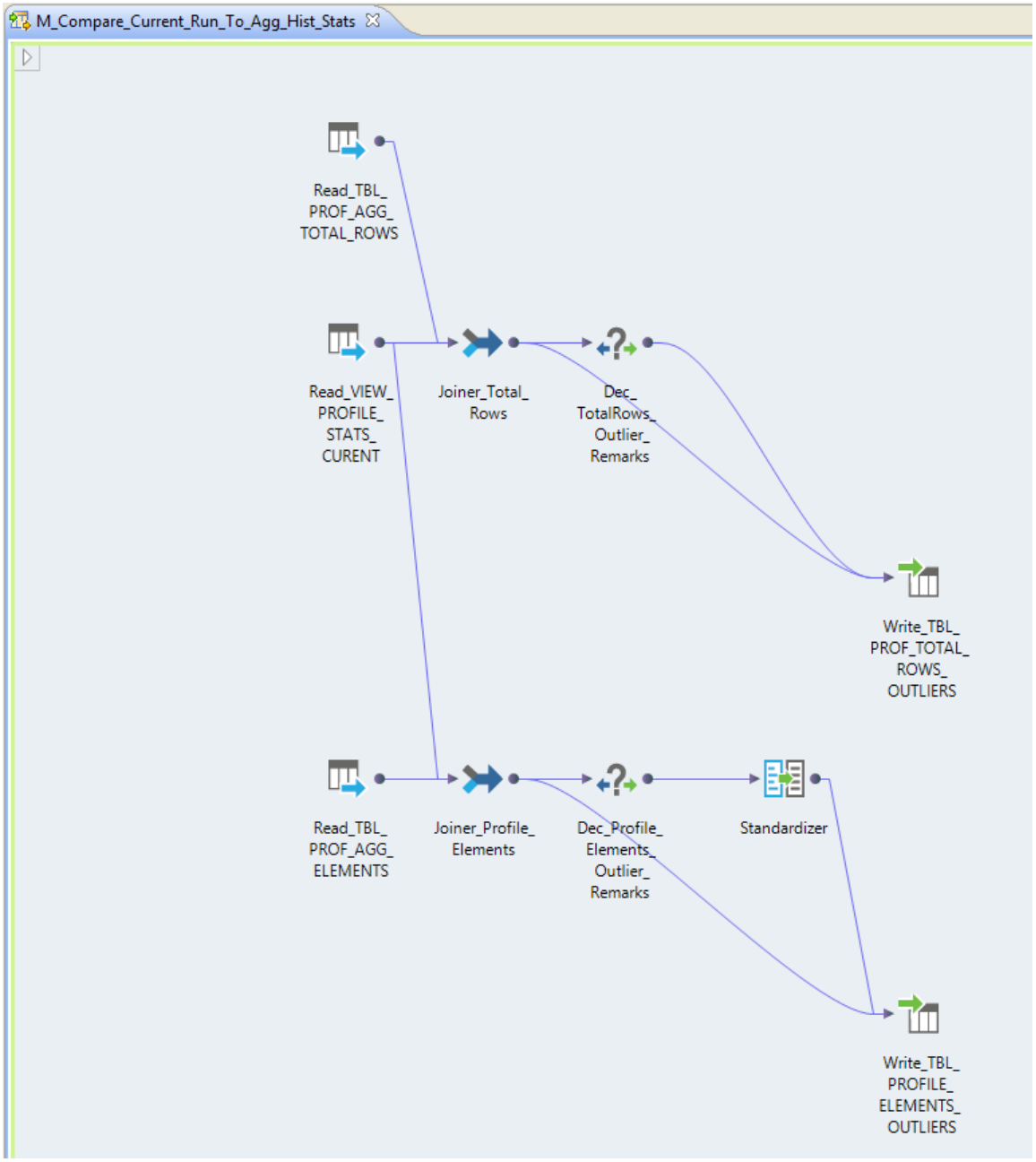
	PROFILE_NAME	PROFILE_ID	FIELD_NAME	SUM_ROWS	AVG_FIELD_VAL	FIELD_VAL_LOWER_CUT_OFF	FIELD_VAL_UPPER_CUT_OFF	TOTAL_DISTI...	MEAN_DISTI...	DISTINCT_VA...	PERCENTAGE_DISTINCT_VALUE
116	Profile_tv_shows	U:RMQQZqS_...	Age	5611	0	0	0	46	46	46	0.8198182142220638
117	Profile_tv_shows	U:RMQQZqS_...	Disney_	5611	0.0381393690...	0	0	2	2	2	0.03564427018356799
118	Profile_tv_shows	U:RMQQZqS_...	Field1	5611	2805	2805	2805	5611	5611	5611	100
119	Profile_tv_shows	U:RMQQZqS_...	Field12	5611	0.8928571428...	1	1	2	2	2	0.03564427018356799
124	Profile_tv_shows	U:RMQQZqS_...	Hulu	5611	0.3129239557...	0	0	2	2	2	0.03564427018356799
125	Profile_tv_shows	U:RMQQZqS_...	IMDb	5611	0	0	0	82	82	82	1.4614150775262875
126	Profile_tv_shows	U:RMQQZqS_...	Netflix	5611	0	0	0	14	14	14	0.24950989128497597
127	Profile_tv_shows	U:RMQQZqS_...	Prime_Video	5611	0.3799679201...	0	0	2	2	2	0.03564427018356799
51	Profile_tv_shows	U:RMQQZqS_...	Rotten_Tomat...	5611	0	0	0	117	117	117	2.0851898057387275
52	Profile_tv_shows	U:RMQQZqS_...	Title	5611	0	0	0	5561	5561	5561	99.10889324541078
53	Profile_tv_shows	U:RMQQZqS_...	Year	5611	0	0	0	163	163	163	2.9050080199607913
20	Profile_tv_shows	U:RMQQZqS_...	type	5611	0.9859205132...	1	1	2	2	2	0.03564427018356799
101	Profile_fifa_19_data	U:I9fwF4R2Ee...	Age	15153	25.377266530...	21	27	386	24.125	21	2.547350359664753
102	Profile_fifa_19_data	U:I9fwF4R2Ee...	Contract_Value	15153	18192168.696...	143454	59062778	799	49.9375	26	5.272883257440777
103	Profile_fifa_19_data	U:I9fwF4R2Ee...	Height	15153	5.7958923360...	6	6	270	16.875	15	1.781825381112651
104	Profile_fifa_19_data	U:I9fwF4R2Ee...	RUN	15153	0	0	0	16	1	1	0.10558965221408302
105	Profile_fifa_19_data	U:I9fwF4R2Ee...	Release_Clause	15153	50285850.136...	286943	186433204	3332	208.25	114	21.989045073582787

- This table is to record the various statistical measures of all the fields (other than total rows) of a profile, across its runs

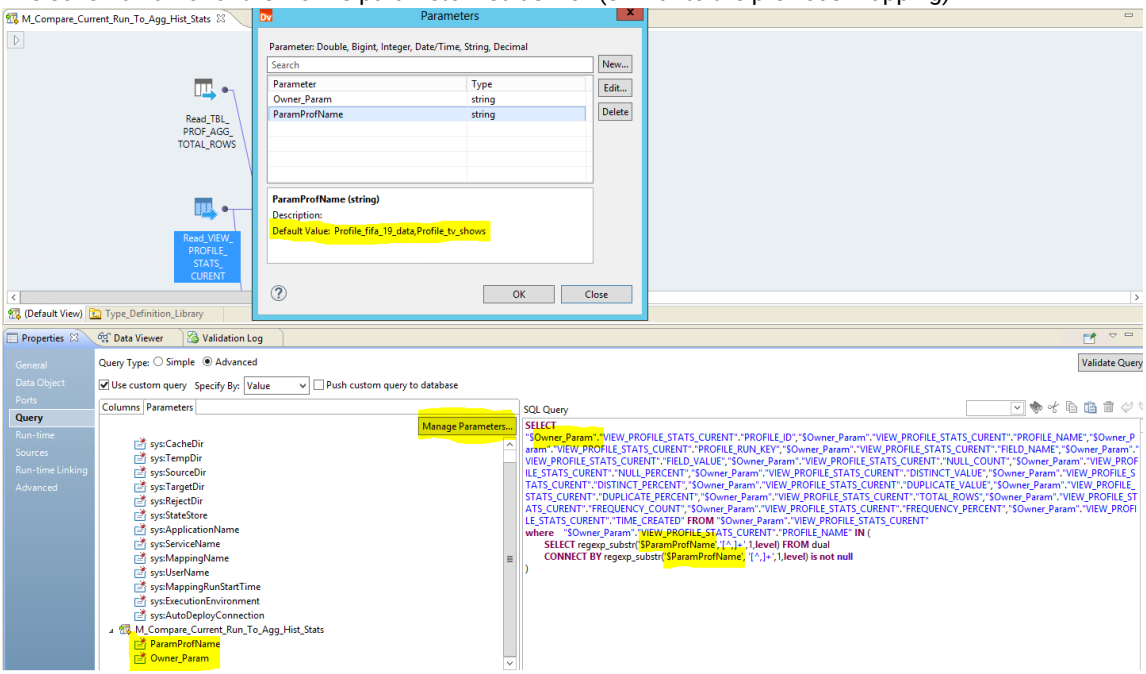
Column Name	Description
PROFILE_NAME	Name of the profile for which the statistics are derived
PROFILE_ID	Unique identifier for a profile
FIELD_NAME	Column name of the data
AVG_FIELD_VAL	indicates the average value across runs, of the respective field
FIELD_VAL_LOWER_CUT_OFF	indicates the lower threshold limit (5-percentile) calculated, across runs of the respective field
FIELD_VAL_UPPER_CUT_OFF	indicates the upper threshold limit (95-percentile) calculated, across runs of the respective field
TOTAL_DISTINCT_VALUE	indicates the sum of the distinct values of the respective field, across runs
MEAN_DISTINCT_VALUE	indicates the average distinct values present, across runs, of the respective field
DISTINCT_VAL_LOWER_CUT_OFF	indicates the lower threshold limit (5-percentile) of distinct values, across runs, of the respective field
PERCENTAGE_DISTINCT_VALUE	Indicates the percentage of distinct values of the respective field, across runs
TOTAL_DUPLICATE_VALUE	Indicates the sum of all duplicate values of the respective field, across runs
MEAN_DUPLICATE_VALUE	Indicates the average duplicate value count of the respective field, across runs
PERCENTAGE_DUPLICATE_VALUE	Indicates the percentage of duplicate values of the respective field, across runs
TOTAL_NULL_COUNT	indicates the sum of null counts of the respective field, across runs
MEAN_NULL_COUNT	indicates the average null count, across runs, of the respective field

NULL_COUNT_UPPER_CUT_OFF	indicates the upper threshold limit (95-percentile) of null count, across runs, of the respective field
PERCENTAGE_NULL_COUNT	Indicates the percentage of null count of the respective field, across runs

- Change the underlying data to have anomalies – increased null values in a certain column, reduce or increase the number of rows drastically
- Run the profile
- Then run the mapping M_Compare_Current_Run_To_Agg_Hist_Stats:

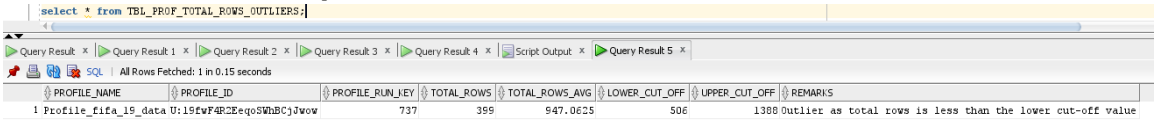


- The profile names that can be queried from the VIEW_PROFILE_STATS_CURRENT, is parameterized. The schema name for the view is parameterized as well (similar to the previous mapping):



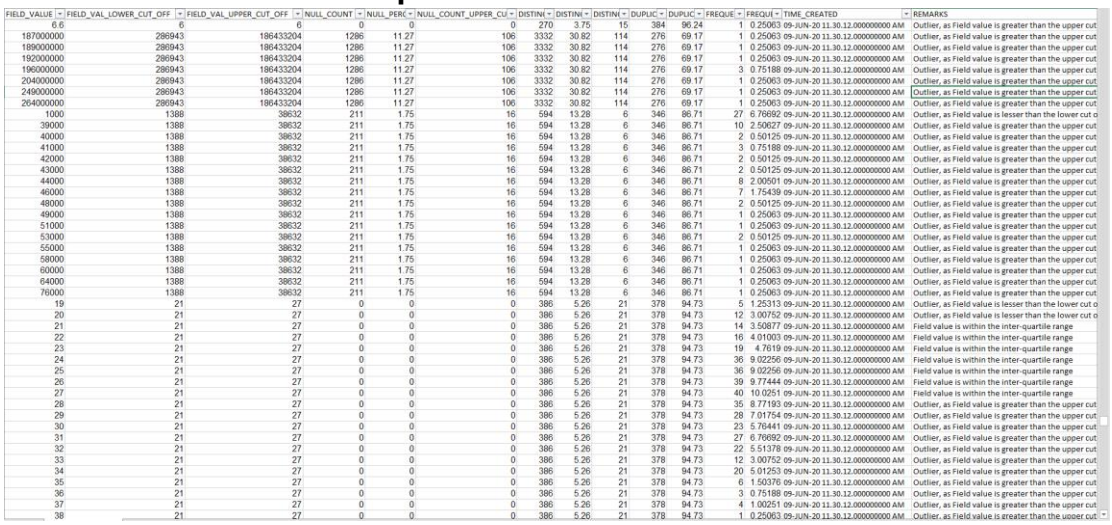
- The values given to the parameter MUST be comma separated as shown
- The M_Compare_Current_Run_To_Agg_Hist_Stats, compares this latest run with the aggregated statistics gathered in the previous mapping
- Repeat the above 3 steps (introducing anomalies to the underlying data, running the profile, running the 2nd mapping) a few times

2.3. Total rows outlier output:



- We can see in the 'Remarks' column, the reason that the particular run is tagged as an outlier – the total rows of a particular run, is less than the 5percentile value or greater than the 95-percentile value
- The conditions are as follows:
- TOTAL_ROWS < LOWER_CUT_OFF – If the total rows in this run, are lesser than the lower threshold
- OR TOTAL_ROWS > UPPER_CUT_OFF – If the total rows in this run, are greater than the upper threshold
- This output is only for detecting the anomalies of the total rows

2.4 Profile elements outlier output:



FIELD_VALUE	FIELD_VAL_LOWER_CUT_OFF	FIELD_VAL_UPPER_CUT_OFF	NULL_COUNT	NULL_PERC	NULL_COUNT_UPPER_CU	DISTINC	DISTINC	DISTINC	DUPLIC	DUPLIC	FREQUE	FREQUE	TIME_CREATED	REMARKS
6.6	6	6	0	0	0	270	3.75	15	384	96.24	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
187000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
189000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
192000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
196000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	3	0.75188	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
204000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
249000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
264000000	286943	186433204	1286	11.27	106	3332	30.82	114	276	69.17	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
1000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	27	6.76692	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is lesser than the lower cut o
39000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	10	2.50627	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
40000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	2	0.50125	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
41000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	3	0.75188	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
42000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	2	0.50125	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
43000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	2	0.50125	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
44000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	8	2.00501	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
46000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	7	1.75439	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
48000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	2	0.50125	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
49000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
51000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
53000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	2	0.50125	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
55000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
58000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
60000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
64000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
76000	1388	38632	211	1.75	16	594	13.28	6	346	86.71	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
19	21	27	0	0	0	386	5.26	21	378	94.73	5	1.25313	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is lesser than the lower cut o
20	21	27	0	0	0	386	5.26	21	378	94.73	12	3.00752	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is lesser than the lower cut o
21	21	27	0	0	0	386	5.26	21	378	94.73	14	3.50877	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
22	21	27	0	0	0	386	5.26	21	378	94.73	16	4.01003	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
23	21	27	0	0	0	386	5.26	21	378	94.73	19	4.7619	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
24	21	27	0	0	0	386	5.26	21	378	94.73	36	9.02256	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
25	21	27	0	0	0	386	5.26	21	378	94.73	36	9.02256	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
26	21	27	0	0	0	386	5.26	21	378	94.73	39	9.77444	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
27	21	27	0	0	0	386	5.26	21	378	94.73	40	10.0251	09-JUN-20 11.30.12.000000000 AM	Field value is within the inter-quartile range
28	21	27	0	0	0	386	5.26	21	378	94.73	35	8.77193	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
29	21	27	0	0	0	386	5.26	21	378	94.73	28	7.01754	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
30	21	27	0	0	0	386	5.26	21	378	94.73	23	5.76441	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
31	21	27	0	0	0	386	5.26	21	378	94.73	27	6.76692	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
32	21	27	0	0	0	386	5.26	21	378	94.73	22	5.51378	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
33	21	27	0	0	0	386	5.26	21	378	94.73	12	3.00752	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
34	21	27	0	0	0	386	5.26	21	378	94.73	20	5.01253	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
35	21	27	0	0	0	386	5.26	21	378	94.73	6	1.50376	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
36	21	27	0	0	0	386	5.26	21	378	94.73	3	0.75188	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
37	21	27	0	0	0	386	5.26	21	378	94.73	4	1.00251	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut
38	21	27	0	0	0	386	5.26	21	378	94.73	1	0.25063	09-JUN-20 11.30.12.000000000 AM	Outlier, as Field value is greater than the upper cut

- The outliers for the rest of the columns are determined based on the below logic:
- FIELD_VALUE <= FIELD_VAL_LOWER_CUT_OFF – If min value of a field in this run, is lesser or equal to the lower threshold
- OR FIELD_VALUE >= FIELD_VAL_UPPER_CUT_OFF – if the maximum value of a field in this run, is greater or equal to the upper threshold
- OR NULL_COUNT >= NULL_COUNT_UPPER_CUT_OFF – if the null count of a field in this run, is greater or equal to the upper threshold
- OR DISTINCT_VALUE <= DISTINCT_VAL_LOWER_CUT_OFF – if the distinct value count of a field in this run, is lesser or equal to the lower threshold
- These reasons if any are captured in the ‘Remarks’ column of the table
- This output records the all the values of the current profile run and identifies the outliers

3. Example

The data used in this example is that of football players (Fifa 2019), from Kaggle. There are multiple columns present in this dataset. Of those, the ones that have been selected are Age, Contract_Value, Wage, Height, Weight and Release_Clause. There are 18k rows of data in this dataset. These have been split into 18 sets, one for each profile run, to simulate profile runs of the dataset over a period.

The anomalies have been introduced in the last 3 runs. Data is said to be an outlier when it is either above or below the min or max threshold. For example, a player should be at least middle school age and cannot be too old (say 50 years) or young (say 10 years) and he would not be ludicrously tall (say 8 feet) or short (say 4 feet), weigh 300 pounds and so on. Also, if the count of null data of a field in a particular run, is more than the max threshold, or the number of distinct values of a field in a particular run, is less than the min threshold.

3.1. Create the PDO and Mapping

PDO.zip

- The PDO needs to be created first, its profile run. Then, the underlying data needs to be changed and profiled at every instance.
- The fifa_19_data.csv is the main file
- The split_files folder, contains the data for the 18 profile runs. Files fifa_run1,2,4,5,6,7,8,9,11,12,13,15,16,17 are to be used for creating the profile history (to simulate everyday profile runs over a period)
- Runs 3,10,18 have anomalies in them, and they can be used to compare the current run vs historical run, to bring out the outliers
- Profile the fifa_19_data.csv file
- Replace the underlying data of the fifa_19_data.csv with contents of the files in the split files folder - fifa_run1,2,4,5,6,7,8,9,11,12,13,15,16,17. Right click on the profile, and click run profile after every dataset change.
- You will now have a profile history of 15 runs

3.2. Prepare the database

- Follow the steps in 2.2

3.3. Import the mappings

- Follow the steps in 2.3

3.4. Execution steps

- Now that we have a profile history, run the first mapping
- Observe the results as explained in 3.1 and 3.2
- Use the files from split_files folder – fifa_run3,10,18, which have anomalies in them, and profile for each dataset
- After each profile run, execute the second mapping
- Observe the outliers as explained in 3.3 and 3.4

4. References

https://www.kaggle.com/karangadiya/fifa19