DATA QUALITY – TEST SUMMARY REPORT

Document Overview

This report describes the data quality test scenarios executed with its corresponding test outputs, to uncover and spot the various quality issues introduced in the parquet input data files of the **"scrape_appearances"** dataset and the **"competitor_appearances"** dataset.

Test Summary

Below is the summary of the various tests and its corresponding Python modular source code.

#	Test Name	Python Modular Source Code
General	Data Extraction from AWS S3 Buckets	"s3_bucket_list.py"
1	Data Quality Test 1 - Schema Validations	"parquet_file_schema_generator.py"
2	Data Quality Test 2 - DataFrame Validations	"parquet_file_dataframe_converter.py"
3	Data Quality Test 3 - Null (NaN and None) Validations	"parquet_null_values_checks.py"
4	Data Quality Test 4 - Non-ASCII Characters Validations	"parquet_non_ascii_characters_checks.py"
5	Data Quality Test 5 - Unique Primary Key Combination - Duplicate Data Validations	"parquet_primary_key_duplicate_data_instances.py"
6	Data Quality Test 6 - Date Format Validations	"parquet_date_format_checks.py"
7	Data Quality Test 7 - Device Supported Values Validations	"parquet_device_supported_values_checks.py"
8a	Data Quality Test 8a - Search Term Maximum Allowable Characters Deviation Validations	"parquet_max_chars_deviation_checks.py"
8b	Data Quality Test 8b - Domain Maximum Allowable Characters Deviation Validations	"parquet_max_chars_deviation_checks.py"
9	Data Quality Test 9 - Scrape Count Minimum Acceptable Values Deviation Validations	"parquet_scrape_count_min_int_checks.py"
10a	Data Quality Test 10a - Sponsored Appearances - Minimum and Maximum Values Deviation Validations	"parquet_min_max_deviation_checks.py"
10b	Data Quality Test 10b - Natural Appearances - Minimum and Maximum Values Deviation Validations	"parquet_min_max_deviation_checks.py"
10c	Data Quality Test 10c - Pla Appearances - Minimum and Maximum Values Deviation Validations	"parquet_min_max_deviation_checks.py"
10d	Data Quality Test 10d - ctr Probability - Minimum and Maximum Values Deviation Validations	"parquet_min_max_deviation_checks.py"

Below is the summary of the various tests executed to identify the quality issues in the given two datasets.

#	Test Name	"scrape_appearances" Dataset	"competitor_appearances" Dataset
General	Data Extraction from AWS S3 Buckets	TEST APPLICABLE	TEST APPLICABLE
1	Data Quality Test 1 - Schema Validations	TEST APPLICABLE	TEST APPLICABLE
2	Data Quality Test 2 - DataFrame Validations	TEST APPLICABLE	TEST APPLICABLE
3	Data Quality Test 3 - Null (NaN and None) Validations	TEST APPLICABLE	TEST APPLICABLE
4	Data Quality Test 4 - Non-ASCII Characters Validations	TEST APPLICABLE	TEST APPLICABLE
5	Data Quality Test 5 - Unique Primary Key Combination - Duplicate Data Validations	TEST APPLICABLE	TEST APPLICABLE
6	Data Quality Test 6 - Date Format Validations	TEST APPLICABLE	TEST APPLICABLE
7	Data Quality Test 7 - Device Supported Values Validations	TEST APPLICABLE	TEST APPLICABLE
8a	Data Quality Test 8a - Search Term Maximum Allowable Characters Deviation Validations	TEST APPLICABLE	TEST APPLICABLE
8b	Data Quality Test 8b - Domain Maximum Allowable Characters Deviation Validations	TEST NOT APPLICABLE	TEST APPLICABLE
9	Data Quality Test 9 - Scrape Count Minimum Acceptable Values Deviation Validations	TEST APPLICABLE	TEST NOT APPLICABLE
10a	Data Quality Test 10a - Sponsored Appearances - Minimum and Maximum Values Deviation Validations	TEST NOT APPLICABLE	TEST APPLICABLE
10b	Data Quality Test 10b - Natural Appearances - Minimum and Maximum Values Deviation Validations	TEST NOT APPLICABLE	TEST APPLICABLE
10c	Data Quality Test 10c - Pla Appearances - Minimum and Maximum Values Deviation Validations	TEST NOT APPLICABLE	TEST APPLICABLE
10d	Data Quality Test 10d - ctr Probability - Minimum and Maximum Values Deviation Validations	TEST NOT APPLICABLE	TEST APPLICABLE

General - Data Extraction from AWS S3 Buckets

(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:

This **Generic Data Extraction from AWS S3 Buckets** test will fetch the below information from the given AWS S3 Buckets via AWS API:

- Total Number of parquet input files in each S3 Bucket.
- File Name of the individual parquet input files in each of the S3 Bucket.

Datasets Applicable:

This generic test is applicable to both the **”scrape_appearances”** dataset and the **“competitor_appearances”** dataset.

Data Extraction Test Results:

- There are no issues noticed while extracting the individual parquet input data files from both the **”scrape_appearances”** dataset and the **“competitor_appearances”** dataset.

Data Quality Test 1 - Schema Validations

(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:

This **Schema Validations** test will verify and uncover the following issues:

- Missing Data Attributes deviating the pre-defined data schema.
- Presence of New Data Attributes deviating the pre-defined data schema.
- Incorrect Data Types of the Data Attributes deviating the pre-defined data schema.

Datasets Applicable:

This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- There are no issues noticed in the parquet input data files in both the **”scrape_appearances”** dataset and the **“competitor_appearances”** dataset, that deviates the pre-defined data schema.

Data Quality Test 2 - DataFrame Validations

(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:

This **DataFrame Validations** test will verify and uncover the following issues:

- Missing Data Attributes in the pandas dataframe converted from the parquet input data file, deviating the pre-defined data schema.
- Presence of New Data Attributes in the pandas dataframe converted from the parquet input data file, deviating the pre-defined data schema.
- Incorrect Data Types of the Data Attributes in the pandas dataframe converted from the parquet input data file, deviating the pre-defined data schema.

Datasets Applicable:

This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- There are no issues noticed in the pandas dataframe generated for each of the parquet input data files in both the **”scrape_appearances”** dataset and the **“competitor_appearances”** dataset, that deviates the pre-defined data schema.

Data Quality Test 3 - Null (NaN and None) Validations
(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:
This **Null (NaN and None) Validations** test will verify and uncover the following issues:

- Data Instances for which the given data attribute has Null Values, which deviates the described data constraint for that attribute. This test is verified individually for each of the 4 Data Attributes of the **“scrape_appearances”** dataset and each of the 8 Data Attributes of the **“competitor_appearances”** dataset.

Datasets Applicable:
This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- Below are the Data Instances with the Null Values identified for the various parquet input data files, which deviates the given described data constraints.

3.1. “scrape_appearances” Dataset - '0012_part_00.parquet' File – ‘date’ Attribute

As per the data constraints defined for the **“scrape_appearances”** dataset, ‘date’ attribute is described as a Primary Key (Unique) and hence it cannot be NULL. Below are the 2 data instances for which ‘date’ attribute is ‘None’. This is a data quality issue where Primary Key value is Null/None/NaN without a valid date value in the expected format as ‘YYYY-MM-DD’, thereby deviating the given described data constraint.

```
Attribute 'date':

Total Number of Null Values: 2

Below are the Filtered Data Instances of the 2 Null Values found for the Attribute 'date' in the Parquet File '0012_part_00.parquet'.

             search_term  date   device  scrape_count
356000  date is bad here  None  desktop             1
356001  date is bad here  None   mobile             2
```

3.2. “scrape_appearances” Dataset - '0018_part_00.parquet' File – ‘scrape_count’ Attribute

As per the data constraints defined for the **“scrape_appearances”** dataset, ‘scrape_count’ attribute must be a non-NULL value greater than 0. Below is a single data instance for which ‘scrape_count’ attribute is ‘NaN’. This is a data quality issue where ‘scrape_count’ attribute is Null/None/NaN without a valid integer value, thereby deviating the given described data constraint.

```
Attribute 'scrape_count':

Total Number of Null Values: 1

Below are the Filtered Data Instances of the 1 Null Values found for the Attribute 'scrape_count' in the
Parquet File '0018_part_00.parquet'.

                search_term        date  device  scrape_count
356000  null in the scrape_count  2022-05-14  mobile           NaN
```

3.3. “competitor_appearances” dataset – ‘0004_part_01.parquet' File – ‘domain’ Attribute

As per the data constraints defined for the **“competitor_appearances”** dataset, ‘domain’ attribute must be a non-NULL string type value with a maximum of 100 characters long. Below is a single data instance for which ‘domain’ attribute is ‘None’. This is a data quality issue where ‘domain’ attribute is Null/None/NaN without a valid string value, thereby deviating the given described data constraint.

```
Attribute 'domain':

Total Number of Null Values: 1

Below are the Filtered Data Instances of the 1 Null Values found for the Attribute 'domain' in the Parquet File '0004_part_01.parquet'.

             search_term        date   device domain  sponsored_appearances  natural_appearances  pla_appearances       ctr
390605  electric bike for sale  2022-05-14  desktop   None                      0                    2                1  0.129663
```

- There are no Null/None/NaN values exists in any of the other 18 parquet input data files of the **“scrape_appearances”** dataset and in any of the other 39 parquet input data files of the **“competitor_appearances” dataset.**

Data Quality Test 4 - Non-ASCII Characters Validations
(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:

This **Non-ASCII Characters Validations** test will verify and uncover the following issues:

- Data Instances for which the given string type data attribute has values with Non-ASCII Characters. This test is verified individually for each of the string type Data Attributes namely ‘search_term’ and ‘device’ of the **“scrape_appearances”** dataset, and each of the string type Data Attributes namely ‘search_term’, ‘device’ and ‘domain’ of the **“competitor_appearances”** dataset.

NOTE: There is no explicit data constraint/s defined about the non-ASCII characters in the **“scrape_appearances”** dataset and **“competitor_appearances”** dataset. This test is verified, and the results are tracked as additional observations, which could potentially be a data quality issue to discuss and investigate based on the business requirements.

Datasets Applicable:

This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

Based on the test outcomes, we noticed the following:

- Presence of non-ASCII Characters in the data attribute ‘search_term’ in all the 20 parquet input data files of the **“scrape_appearances”** dataset
- There are no non-ASCII Characters present in the data attribute ‘device’ in all the 20 parquet input data files of the **“scrape_appearances”** dataset
- Presence of non-ASCII Characters in the data attributes namely ‘search_term’ and ‘domain’ in all the 40 parquet input data files of the “competitor**_appearances”** dataset
- There are no non-ASCII Characters present in the data attribute ‘device’ in all the 40 parquet input data files of the “competitor**_appearances”** dataset

Below tables represents the total data instances across various parquet input data files where non-ASCII characters are present and identified.

“scrape_appearances” Dataset

#	”scrape_appearances” Dataset	Total Data Instances with Non-ASCII Characters	
		'search_term'	'device'
1	'0000_part_00.parquet'	172	0
2	'0001_part_00.parquet'	197	0
3	'0002_part_00.parquet'	212	0
4	'0003_part_00.parquet'	174	0
5	'0004_part_00.parquet'	197	0
6	'0005_part_00.parquet'	198	0
7	'0006_part_00.parquet'	195	0
8	'0007_part_00.parquet'	191	0
9	'0008_part_00.parquet'	177	0
10	'0009_part_00.parquet'	178	0
11	'0010_part_00.parquet'	187	0
12	'0011_part_00.parquet'	193	0
13	'0012_part_00.parquet'	186	0
14	'0013_part_00.parquet'	203	0
15	'0014_part_00.parquet'	218	0
16	'0015_part_00.parquet'	196	0
17	'0016_part_00.parquet'	192	0
18	'0017_part_00.parquet'	180	0
19	'0018_part_00.parquet'	194	0
20	0019_part_00.parquet'	178	0

“competitor_appearances” Dataset

#	”competitor_appearances” Dataset	Total Data Instances with Non-ASCII Characters		
		'search_term'	'device'	'domain'
1	'0000_part_00.parquet'	1628	0	15
2	'0000_part_01.parquet'	148	0	2
3	'0001_part_00.parquet'	1654	0	11
4	'0001_part_01.parquet'	154	0	0
5	'0002_part_00.parquet'	1593	0	23
6	'0002_part_01.parquet'	124	0	2
7	'0003_part_00.parquet'	1552	0	13
8	'0003_part_01.parquet'	153	0	0
9	'0004_part_00.parquet'	1579	0	15
10	'0004_part_01.parquet'	152	0	2
11	'0005_part_00.parquet'	1584	0	14
12	'0005_part_01.parquet'	139	0	1
13	'0006_part_00.parquet'	1654	0	9
14	'0006_part_01.parquet'	152	0	0
15	'0007_part_00.parquet'	1605	0	23
16	'0007_part_01.parquet'	165	0	0
17	'0008_part_00.parquet'	1661	0	17
18	'0008_part_01.parquet'	141	0	1
19	'0009_part_00.parquet'	1613	0	9
20	'0009_part_01.parquet'	129	0	4
21	'0010_part_00.parquet'	1536	0	14
22	'0010_part_01.parquet'	136	0	1
23	'0011_part_00.parquet'	1501	0	15
24	'0011_part_01.parquet'	156	0	1
25	'0012_part_00.parquet'	1614	0	10
26	'0012_part_01.parquet'	178	0	1
27	'0013_part_00.parquet'	1572	0	20
28	'0013_part_01.parquet'	144	0	0
29	'0014_part_00.parquet'	1544	0	14
30	'0014_part_01.parquet'	136	0	1
31	'0015_part_00.parquet'	1635	0	10
32	'0015_part_01.parquet'	135	0	2
33	'0016_part_00.parquet'	1587	0	19
34	'0016_part_01.parquet'	172	0	0
35	'0017_part_00.parquet'	1601	0	16
36	'0017_part_01.parquet'	147	0	5
37	'0018_part_00.parquet'	1604	0	16
38	'0018_part_01.parquet'	132	0	1
39	'0019_part_00.parquet'	1596	0	11
40	0019_part_01.parquet'	135	0	2

Data Quality Test 5 - Unique Primary Key Combination - Duplicate Data Validations (Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:
This **Unique Primary Key Combination - Duplicate Data Validations** test will verify and uncover the following issues:

- Data Instances for which the unique combination of the Primary Key Data Attributes has duplicate values, which deviates the described data constraint for that Primary Key (Unique) Combination of the Data Attributes. This test is verified individually for the Unique Combination of the Primary Key Data Attributes of the **“scrape_appearances”** dataset and the **“competitor_appearances”** dataset.

Datasets Applicable:
This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- Below are the Data Instances with the duplicate values identified for the Unique Combination of the Primary Key Data Attributes, which deviates the given described data constraint for that Primary Key (Unique) Combination of the Data Attributes.

5.1. “scrape_appearances” Dataset - '0000_part_00.parquet' File – Primary Key (Unique) Combination of the Data Attributes ‘date’, ‘device’ and ‘search_term’

As per the data constraints defined for the **“scrape_appearances”** dataset, the data attributes ‘date’, ‘device’ and ‘search_term’ are the Unique Primary Keys such that their combination cannot have duplicate values or NULL values. Below are the 2 data instances for which their combination has duplicate values. This is a data quality issue where the Unique Combination of the Primary Key values are duplicated between the 2 data instances, thereby deviating the given described data constraint.

```
--------------------------------------------------------------------------------------------------------------------------
***** DATA QUALITY TEST 5 – UNIQUE PRIMARY KEY COMBINATION – DUPLICATE DATA CHECKS AND VALIDATIONS *****
--------------------------------------------------------------------------------------------------------------------------
Total Number of Duplicate Data Instances: 2

Below are the 2 Duplicate Data Instances found for the Combination of the Unique Primary Key Data Attributes
'['search_term', 'date', 'device']' in the Parquet File '0000_part_00.parquet'.

       search_term        date   device  scrape_count
360000        dupe  2022-05-13  desktop             2
360001        dupe  2022-05-13  desktop             2
```

- There are no duplicate data instances exists in any of the other 19 parquet input data files of the **“scrape_appearances”** dataset, where the given described data constraint related to the Unique Combination of the Primary Key Data Attributes namely date’, ‘device’ and ‘search_term’ are deviated.
- There are no duplicate data instances exists in any of the 40 parquet input data files of the **“competitor_appearances”** dataset, where the given described data constraint related to the Unique Combination of the Primary Key Data Attributes namely ‘date’, ‘device’, ‘search_term’ and ‘domain’ are deviated.

Data Quality Test 6 - Date Format Validations
(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:
This **Date Format Validations** test will verify and uncover the following issues:

- Data Instances for which the format of the ‘date’ attribute is invalid and does not meet the given described data constraint about the expected date format as ‘YYYY-MM-DD’. This test is verified individually for the ‘date’ attribute in the **“scrape_appearances”** dataset and the **“competitor_appearances”** dataset.

Datasets Applicable:
This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- Below are the Data Instances with invalid date format or invalid data values identified for the ’date’ attribute, which deviates the given described data constraint about the expected date format as ‘YYYY-MM-DD’.

6.1. “scrape_appearances” Dataset - '0012_part_00.parquet' File – ‘date’ Attribute

As per the data constraints defined for the ‘date’ attribute in the **“scrape_appearances”** dataset and also for the **“competitor_appearances”** dataset, ‘date’ attribute is described as one of the Primary Key (Unique) attribute with the expected date format as ‘YYYY-MM-DD’ and hence it cannot hold any other date format (E.G. ‘YYYY-DD-MM’, ‘DD-MM-YYYY’, ‘MM-DD-YYYY’, etc) or NaN/None/NULL values. Below are the 2 data instances for which ‘date’ attribute is ‘None’. This is a data quality issue where Primary Key value is Null/None/NaN without a valid date value in the expected format as ‘YYYY-MM-DD’, thereby deviating the given described data constraint.

```
Attribute 'date':

Total Number of Null Values: 2

Below are the Filtered Data Instances of the 2 Null Values found for the Attribute 'date' in the Parquet File '0012_part_00.parquet'.

            search_term  date   device  scrape_count
356000  date is bad here  None  desktop             1
356001  date is bad here  None   mobile             2
```

This issue is also identified as part of the **Data Quality Test 3 - Null (NaN and None) Validations** as mentioned in the issue named **3.1. “scrape_appearances” Dataset - '0012_part_00.parquet' File – ‘date’ Attribute.**

- There are no data instances with invalid date format exists in any of the other 19 parquet input data files of the **“scrape_appearances”** dataset, where the given described data constraint related to the expected date format as ‘YYYY-MM-DD’ is deviated.

- There are no data instances with invalid date format exists in any of the 40 parquet input data files of the **“competitor_appearances”** dataset, where the given described data constraint related to the expected date format as ‘YYYY-MM-DD’ is deviated.

Data Quality Test 7 - Device Supported Values Validations
(Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:
This **Device Supported Values Validations** test will verify and uncover the following issues:

- Data Instances for which the device values in the ‘device’ attribute are not in the pre-defined list of supported device values as per the given described data constraint. This test is verified individually for the ‘device’ attribute in the **“scrape_appearances”** dataset and the **“competitor_appearances”** dataset.

Datasets Applicable:
This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- Below are the Data Instances with non-supported device values or invalid device values identified for the ’device’ attribute, which deviates the given described data constraint about the pre-defined list of supported device values as either ‘desktop’ or ‘mobile’.

7.1. “scrape_appearances” Dataset - '0000_part_00.parquet' File – ‘device’ Attribute

As per the data constraints defined for the ‘device’ attribute in the **“scrape_appearances”** dataset and also for the **“competitor_appearances”** dataset, the pre-defined list of supported device values for the ‘device’ attribute is expected to be either ‘desktop’ or ‘mobile’ and hence it cannot hold any other device values. Below are the 2 data instances for which ‘device’ attribute has a value as ‘tablet’, which is not in the pre-defined list of supported device values. This is a data quality issue where the ‘device’ attribute holds an unsupported device value which does not match with the pre-defined set of supported device values, thereby deviating the given described data constraint.

```
--------------------------------------------------------------------------------------------
***** DATA QUALITY TEST 7 – UNIQUE VALUES SUPPORTED BY THE DEVICE DATA ATTRIBUTE – VALIDATIONS *****
--------------------------------------------------------------------------------------------

Total Number of New UnSupported Device Values: 1


List of New UnSupported Device Values: ['tablet']

Below are the 2 Filtered Data Instances for the 1 New UnSupported Device Values found for the Attribute 'device'
in the Parquet File '0000_part_00.parquet'.

        search_term        date  device  scrape_count
360002  tablet in device  2022-05-15  tablet             1
360003  tablet in device  2022-05-14  tablet             1
```

- There are no data instances with unsupported device values exists in any of the other 19 parquet input data files of the **“scrape_appearances”** dataset, where the given described data constraint related to the supported device values of the ‘device’ attribute is deviated.
- There are no data instances with unsupported device values exists in any of the 40 parquet input data files of the **“competitor_appearances”** dataset, where the given described data constraint related to the supported device values of the ‘device’ attribute is deviated.

Data Quality Test 8a - Search Term Maximum Allowable Characters Deviation Validations (Datasets Applicable: ”scrape_appearances” and “competitor_appearances”)

Test Description:

This **Search Term Maximum Allowable Characters Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the string values in the ‘search_term’ attribute has a length of about less than or equal to a maximum of 400 characters as per the given described data constraint. This test is verified individually for the ‘search_term’ attribute in the **“scrape_appearances”** dataset and the **“competitor_appearances”** dataset.

Datasets Applicable:

This test is applicable to all the 20 parquet input data files in the **”scrape_appearances”** dataset and all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

As per the tests executed, below are the outcomes:

- There are no data instances with the string length of the ‘search_term’ attribute greater than the maximum characters length of about 400 characters exists in any of the 20 parquet input data files of the **“scrape_appearances”** dataset, where the given described data constraint related to the maximum characters length of the ‘search_term’ attribute is deviated.
- There are no data instances with the string length of the ‘search_term’ attribute greater than the maximum characters length of about 400 characters exists in any of the 40 parquet input data files of the **“competitor_appearances”** dataset, where the given described data constraint related to the maximum characters length of the ‘search_term’ attribute is deviated.

Data Quality Test 8b - Domain Maximum Allowable Characters Deviation Validations (Datasets Applicable: "competitor_appearances")

Test Description:
This **Domain Maximum Allowable Characters Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the string values in the 'domain' attribute has a length of about less than or equal to a maximum of 100 characters as per the given described data constraint. This test is verified for the 'domain' attribute in the **"competitor_appearances"** dataset.

Datasets Applicable:
This test is applicable to all the 40 parquet input data files in the **"competitor_appearances"** dataset.

Data Quality Test Results:
As per the tests executed, below are the outcomes:

- There are no data instances with the string length of the 'domain' attribute greater than the maximum characters length of about 100 characters exists in any of the 40 parquet input data files of the **"competitor_appearances"** dataset, where the given described data constraint related to the maximum characters length of the 'domain' attribute is deviated.

Data Quality Test 9 - Scrape Count Minimum Acceptable Values Deviation Validations (Datasets Applicable: “scrape_appearances”)

Test Description:
This **Scrape Count Minimum Acceptable Values Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the scrape count integer value in the ‘scrape_count’ attribute has a value which is less than or equal to 0 as per the given described data constraint. This test is verified for the ‘scrape_count’ attribute in the **“scrape_appearances”** dataset.

Datasets Applicable:
This test is applicable to all the 20 parquet input data files in the **“scrape_appearances”** dataset.

Data Quality Test Results:

- Below are the Data Instances with scrape count values as less than or equal to 0 identified for the ’scrape_count’ attribute, which deviates the given described data constraint about the minimum acceptable scrape count value as 1 in the ‘scrape_count’ data attribute.

9.1. “scrape_appearances” Dataset - '0010_part_00.parquet' File – ‘scrape_count’ Attribute

As per the data constraints defined for the ‘scrape_count’ attribute in the **“scrape_appearances”** dataset, the minimum acceptable scrape count value is expected to be 1 (any value greater than 0) in the ‘scrape_count’ data attribute and hence it cannot hold any other minimum values which is less than 1 (i.e., 0 or negative integer values). Below are the 2 data instances for which ‘scrape_count’ attribute has values less than 1 as 0 and -1, which are not the acceptable scrape count values. This is a data quality issue where the ‘scrape_count’ attribute holds values that are less than or equal to 0, thereby deviating the given described data constraint.

```
--------------------------------------------------------------------------------------------------------------------------------
***** DATA QUALITY TEST 9 – DEVIATIONS IN THE MINIMUM ALLOWABLE INTEGER COUNT OF THE SCRAPE COUNT DATA ATTRIBUTE – VALIDATIONS *****
--------------------------------------------------------------------------------------------------------------------------------
Total Number of Data Instances for the Data Attribute 'scrape_count' that Deviated the Minimum Allowable Scrape Count of 1:
2

Below are the 2 Data Instances that Deviated the Minimum Allowable Scrape Count of 1 in the Parquet File '0010_part_00.parquet'.

             search_term        date   device  scrape_count
356983  did not scrape this  2022-05-13  desktop             0
356984     scraped negative  2022-05-14   mobile            -1
```

- There are no data instances with the scrape count integer value in the ‘scrape_count’ attribute as less than or equal to 0 exists in any of the other 19 parquet input data files of the **“scrape_appearances”** dataset, where the given described data constraint related to the acceptable scrape count in the ‘scrape_count’ attribute is deviated.

Data Quality Test 10a - Sponsored Appearances - Minimum and Maximum Values Deviation Validations
(Datasets Applicable: "competitor_appearances")

Test Description:
This **Scrape Count Minimum Acceptable Values Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the minimum acceptable sponsored appearances integer value in the 'sponsored_appearances' attribute for a given combination of the 'date', 'device' and the 'search_term' attributes, has a value which is equal to 0 and the maximum acceptable value is less than or equal to the scrape count value retrieved from the **"scrape_appearances"** dataset for the same combination of the 'date', 'device' and the 'search_term' attributes as per the given described data constraint. This test is verified for the 'sponsored_appearances' attribute in the **"competitor_appearances"** dataset.

Datasets Applicable:
This test is applicable to all the 40 parquet input data files in the **"competitor_appearances"** dataset.

Data Quality Test Results:
Based on the test execution, below table represents the list of all the parquet input data files in the **"competitor_appearances"** dataset such that the minimum and the maximum acceptable sponsored appearances integer values in the 'sponsored_appearances' attribute are deviated from that defined in the data constraints.

#	”competitor_appearances” Dataset	Total Data Instances where 'sponsored_appearances' Attribute Deviated the Maximum Values as described in the Data Constraints	
		Minimum Value Deviation	Maximum Value Deviation
1	'0000_part_00.parquet'	No Deviation	7
2	'0000_part_01.parquet'	No Deviation	4
3	'0001_part_00.parquet'	No Deviation	9
4	'0001_part_01.parquet'	No Deviation	1
5	'0002_part_00.parquet'	No Deviation	13
6	'0002_part_01.parquet'	No Deviation	No Deviation
7	'0003_part_00.parquet'	No Deviation	5
8	'0003_part_01.parquet'	No Deviation	1
9	'0004_part_00.parquet'	No Deviation	6
10	'0004_part_01.parquet'	No Deviation	1
11	'0005_part_00.parquet'	No Deviation	7
12	'0005_part_01.parquet'	No Deviation	No Deviation
13	'0006_part_00.parquet'	No Deviation	8
14	'0006_part_01.parquet'	No Deviation	No Deviation
15	'0007_part_00.parquet'	No Deviation	14
16	'0007_part_01.parquet'	No Deviation	No Deviation
17	'0008_part_00.parquet'	No Deviation	6
18	'0008_part_01.parquet'	No Deviation	No Deviation
19	'0009_part_00.parquet'	No Deviation	9
20	'0009_part_01.parquet'	No Deviation	No Deviation
21	'0010_part_00.parquet'	No Deviation	5
22	'0010_part_01.parquet'	No Deviation	No Deviation
23	'0011_part_00.parquet'	No Deviation	10
24	'0011_part_01.parquet'	No Deviation	No Deviation
25	'0012_part_00.parquet'	No Deviation	8
26	'0012_part_01.parquet'	No Deviation	No Deviation
27	'0013_part_00.parquet'	No Deviation	8
28	'0013_part_01.parquet'	No Deviation	No Deviation
29	'0014_part_00.parquet'	No Deviation	15
30	'0014_part_01.parquet'	No Deviation	No Deviation
31	'0015_part_00.parquet'	No Deviation	11
32	'0015_part_01.parquet'	No Deviation	2
33	'0016_part_00.parquet'	No Deviation	3
34	'0016_part_01.parquet'	No Deviation	No Deviation
35	'0017_part_00.parquet'	No Deviation	12
36	'0017_part_01.parquet'	No Deviation	No Deviation
37	'0018_part_00.parquet'	No Deviation	7
38	'0018_part_01.parquet'	No Deviation	3
39	'0019_part_00.parquet'	No Deviation	10
40	0019_part_01.parquet'	No Deviation	1

Data Quality Test 10b - Natural Appearances - Minimum and Maximum Values Deviation Validations

(Datasets Applicable: "competitor_appearances")

Test Description:

This **Natural Appearances - Minimum and Maximum Values Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the minimum acceptable natural appearances integer value in the 'natural_appearances' attribute has a value which is equal to 0 and the maximum acceptable value as infinite as per the given described data constraint. This test is verified for the 'natural_appearances' attribute in the **"competitor_appearances"** dataset.

Datasets Applicable:

This test is applicable to all the 40 parquet input data files in the **"competitor_appearances"** dataset.

Data Quality Test Results:

As per the tests executed, below are the outcomes:

- There are no data instances with the natural appearances integer value in the 'natural_appearances' attribute across the 40 parquet input data files of the **"competitor_appearances"** dataset that deviated the minimum and the maximum acceptable values as per the given defined data constraint.

Data Quality Test 10c - Pla Appearances - Minimum and Maximum Values Deviation Validations

(Datasets Applicable: "competitor_appearances")

Test Description:

This **Pla Appearances - Minimum and Maximum Values Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the minimum acceptable pla appearances integer value in the 'pla_appearances' attribute has a value which is equal to 0 and the maximum acceptable value as infinite as per the given described data constraint. This test is verified for the 'pla_appearances' attribute in the **"competitor_appearances"** dataset.

Datasets Applicable:

This test is applicable to all the 40 parquet input data files in the **"competitor_appearances"** dataset.

Data Quality Test Results:

As per the tests executed, below are the outcomes:

- There are no data instances with the pla appearances integer value in the 'pla_appearances' attribute across the 40 parquet input data files of the **"competitor_appearances"** dataset that deviated the minimum and the maximum acceptable values as per the given defined data constraint.

Data Quality Test 10d - ctr Probability - Minimum and Maximum Values Deviation Validations

(Datasets Applicable: “competitor_appearances”)

Test Description:

This **ctr Probability - Minimum and Maximum Values Deviation Validations** test will verify and uncover the following issues:

- Data Instances for which the minimum acceptable probability float value in the ‘ctr’ attribute has a value which is equal to 0.0 and the maximum acceptable value as 1.0 as per the given described data constraint. This test is verified for the ‘ctr’ attribute in the **“competitor_appearances”** dataset.

Datasets Applicable:

This test is applicable to all the 40 parquet input data files in the **“competitor_appearances”** dataset.

Data Quality Test Results:

- Below are the Data Instances with the Null Values identified for the various parquet input data files, which deviates the given described data constraints.

10d.1. “competitor_appearances” Dataset - '0004_part_01.parquet' File – ‘ctr’ Attribute

As per the data constraints defined for the **“competitor_appearances”** dataset, the ‘ctr attribute is expected to have a minimum float value of 0.0 and a maximum float value of 1.0. Below are the 2 data instances for which ‘ctr attribute has a minimum value of -0.1 (which is less than the minimum acceptable value of 0.0) for one of the data instances and a maximum value of 1.01 (which is greater than the maximum acceptable value of 1.0) for the other data instance. This is a data quality issue where the minimum and the maximum float probability values deviated the given described data constraint.

```
----------------------------------------------------------------------------------------------------------------------------------
***** DATA QUALITY TEST 10 - DEVIATIONS IN THE MINIMUM AND MAXIMUM ALLOWABLE COUNT OF THE GIVEN DATA ATTRIBUTE - VALIDATIONS *****
----------------------------------------------------------------------------------------------------------------------------------
Attribute 'ctr':
Actual Minimum Value of the Data Attribute 'ctr' in the Parquet File '0004_part_01.parquet:'
-0.1

Actual Maximum Value of the Data Attribute 'ctr' in the Parquet File '0004_part_01.parquet:'
1.01

Total Number of Data Instances for the Data Attribute 'ctr' that Deviated the Minimum Allowable Count of 0.0
in the Parquet File '0004_part_01.parquet'.

Below are the ctr Data Instances that Deviated the Minimum Allowable Count of 0.0 in the Parquet File '0004_part_01.parquet'.

                search_term        date   device     domain  sponsored_appearances  natural_appearances  pla_appearances  ctr
390603  click through rate negative  2022-05-13  desktop  click.com                      1                    0                1 -0.1

Total Number of Data Instances for the Data Attribute 'ctr' that Deviated the Maximum Allowable Count of 1.0
in the Parquet File '0004_part_01.parquet'.

Below are the ctr Data Instances that Deviated the Maximum Allowable Count of 1.0 in the Parquet File '0004_part_01.parquet'.

        search_term        date  device    domain  sponsored_appearances  natural_appearances  pla_appearances   ctr
390604  ctr greater  2022-05-14  mobile  rate.com                      0                    2                3  1.01
```

- There are no data instances with the probability float value in the ‘ctr’ attribute across the other 39 parquet input data files of the **“competitor_appearances”** dataset that deviated the minimum and the maximum acceptable values as per the given defined data constraint.

Other Observations/Issues

Below is the summary of the other issues observed during the execution of the above-mentioned data quality tests.

1. ‘date’ Attribute Issue During Data Ingestion:

Issue Description:
Data Instances ingested into each of the parquet input data file, includes a mix of different dates in the random order and the records are not injected in the increasing order of the dates. This applies to all the parquet input data file for both the “scrape_appearances” dataset and the “competitor_appearances” dataset.

2. **Dependency on Multiple Historic Parquet Input Data Files to retrieve the scrape count:**

All the historic “scrape_appearances” parquet input data files for a given date must be referred to extract the maximum acceptable value of the ‘sponsored_appearances’ attribute in the “competitor_appearances” dataset.

Impact Due to the Above Two Observations/Issues

- Due to the ‘date’ attribute data ingestion issue 1 mentioned above; there is a limitation to extract the data instances of the “competitor_appearances” dataset with the same combination of ‘date’, ‘device’, ‘search_term’ and ‘domain’ attributes appearing in a single parquet input data file for a given date (E.G: 2022-05-13).
- The above said two issues causes challenges in the effective calculation of the maximum acceptable value of the sponsored_appearances’ attribute for the same combination of ‘date’, ‘device’ and ‘search_term’ attributes from the “scrape_appearances” parquet input data files. Hence, we will be able to calculate the maximum acceptable value of the sponsored_appearances’ attribute by considering only the data instances from a single parquet input data file for a given date (E.G: 2022-05-13) and not by consolidating the different parquet input data files which could possibly include the data instances with the same date (E.G: 2022-05-13).