

## 1 APPENDICES

### 1.1 Appendix 1 - Summary of Well-known Data Quality Issues and the Data Quality Dimension Violations Associated

Table 1 - Data Quality Issues and Data Quality Dimension Violations [Adapted from Visengeriyeva and Abedjan, 2020]

#	Data Quality Issue	Description	Data Quality Dimensions
1	Missing data	Comprises missing tuples and missing values. Tuple completeness requires that all tuples are present in the table. Missing value issue consists of either null values or disguised values. Value completeness requires that all values are present in the table, while null values indicate that the value is unknown or non-existent.	Accuracy, Completeness
2	Incorrect data	Data that differ from the values of the real-world entity (e.g., wrong date of birth).	Accuracy
3	Misspellings	Syntactic deviation of the data value from its ground truth (e.g., “Smiht” instead of “Smith”).	Accuracy
4	Ambiguous data	Data values which might be interpreted in several ways (e.g., abbreviations or cryptic values).	Accuracy, Consistency
5	Extraneous data	Presence of additional data in the attribute value (e.g., the address column contains a person’s name in addition to the address).	Consistency, Uniqueness
6	Outdated temporal data	Values that are obsolete or outdated.	Timeliness
7	Misfielded values	Values that are placed inside the wrong attribute.	Accuracy, Consistency, Completeness
8	Incorrect references	Entities that contain wrong information concerning the reference relation (e.g., the employee is associated with a wrong department).	Accuracy
9	Duplicates	Tuples/values that represent the same real-world entity.	Uniqueness
10	Structural conflicts	Conflicting duplicates in different sources.	Consistency, Uniqueness
11	Different word orderings	Values that violate the expected word order (e.g., first name precedes last name).	Consistency, Uniqueness
12	Different aggregation levels	Entities produced by applying different aggregation methods (e.g., entries per quartal vs. entries per year).	Accuracy, Consistency
13	Temporal mismatch	Refers to erroneous data that arise due to non-enforcement of integrity constraints for temporal data.	Accuracy, Timeliness
14	Different units/representations	Occurrence of multiple representations for the same concept (e.g., Price in different currencies).	Consistency
15	Domain violation	Values that violate semantic rules defined on the specific attribute.	Accuracy
16	FD violation	Values that violate previously specified functional dependencies.	Accuracy, Consistency
17	Wrong data type	Values that violate the data type specification of the corresponding attribute, i.e., data type constraint violation.	Consistency
18	Referential integrity violation	Tuples that violate the referential integrity constraints defined on multiple relations (e.g., missing foreign key).	Accuracy, Consistency, Completeness
19	Uniqueness violation	Duplication of values under the uniqueness constraint.	Uniqueness
20	Use of synonyms	Occurrence of synonymous representations for the same concept inside the same column (e.g., “lecturer” and “professor”).	Uniqueness

21	Use of special characters (space, no space, dash, parentheses)	Refers to different representations of compound data, such as Social Security Number or phone number.	Consistency
22	Different encoding formats	Inconsistent usage of encodings for values within a dataset (e.g., ASCII or EBCDIC).	Consistency

## 1.2 Appendix 2 - Spreadsheet with first selection of ten datasets from the UCI Catalogue

The decision to choose these ten datasets was the following:

- The top two datasets from each of the areas: Life, Social, Physical, Computer and Financial.
- This is how the top two datasets were chosen for each area:
  - 1) Sorting from top to bottom of the 'web\_hits' column that showed how many web hits each dataset obtained.
  - 2) Sorting from top to bottom of the 'num\_papers' column, which showed the number of research papers that cited each dataset.
  - 3) Ranking from 1 to the lowest amount for the 'web\_hits' column. The top one is number 1
  - 4) Ranking from 1 to the lowest amount for the 'num\_papers' column. The top one is number 1
  - 5) Addition of the two rankings.
  - 6) Ranking from 1 to the lowest amount for the sum of rankings.
  - 7) Only the first 2 top in each area were chosen, and they are presented in green below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	index	name	url	instances	attributes	year	area	date_donat	web_hits	#webhits	dataset	attribute_info	papers_that_cite_this_data	num_papers	#numpapers	2#	FinalRank							
2	52	Iris	http	150	4	1988	Life	1/07/1988	5319836	1	data	1. sepal length in	100	1	2	1								
3	45	Heart Disease	http	303	75	1988	Life	1/07/1988	2184270	4	data	Only 14 attributes	58	3	7	2								
4	2	Adult	http	48842	14	1996	Social	1/05/1996	2769887	2	data	Listing of	51	8	10	3								
5	107	Wine	http	178	13	1991	Physical	1/07/1991	2167533	5	data	All attributes are	40	13	18	4								
6	17	Breast Cancer Wisconsin (Diagn	http	569	32	1995	Life	1/11/1995	1972842	8	data	1) ID number	40	14	22	5								
7	14	Breast Cancer	http	286	9	1988	Life	11/07/1988	705622	29	data	1. Class: no-	91	2	31	6								
8	34	Diabetes	https://archive	20			Life	N/A	741148	26	Z	Diabetes files	53	5	31	7								
9	15	Breast Cancer Wisconsin (Origir	http	699	10	1992	Life	15/07/1992	875007	17	data	1. Sample code	40	15	32	8								
10	72	Mushroom	http	8124	22	1987	Life	27/04/1987	821573	21	data	1. cap-shape:	46	12	33	9								
11	1	Abalone	http	4177	8	1995	Life	1/12/1995	1431163	12	data	Given is the	29	22	34	10								
12	42	Glass Identification	http	214	10	1987	Physical	1/09/1987	471902	36	data	1. Id number: 1 to	52	7	43	11								
13	19	Car Evaluation	http	1728	6	1997	N/A	1/06/1997	1748706	10	data	Class Values:	16	38	48	12								
14	20	Census Income	http	48842	14	1996	Social	1/05/1996	768711	22	data	Listing of	24	29	51	13								
15	51	Ionosphere	http	351	34	1989	Physical	1/01/1989	311140	62	data	-- All 34 are	55	4	66	14								
16	9	Auto MPG	http	398	8	1993	N/A	7/07/1993	856989	19	data	1. mpg:	12	48	67	15								
17	58	Letter Recognition	http	20000	16	1991	Computer	1/01/1991	489555	33	data	1. lettrcapital	16	39	72	16								
18	46	Hepatitis	http	155	19	1988	Life	1/11/1988	359983	56	data	1. Class: DIE, LIVE	33	20	76	17								
19	142	Statlog (German Credit Data)	http	1000	20	1994	Financial	17/11/1994	890119	16	data	Attribute 1:	7	63	79	18								
20	109	Zoo	http	101	17	1990	Life	15/05/1990	448279	40	data	1. animal name:	16	40	80	19								
21	100	Thyroid Disease	http	7200	21	1987	Life	1/01/1987	351663	57	data	N/A	26	24	81	20								
22	149	Connectionist Bench (Sonar, Mi	http	208	60		Physical	N/A	263556	82	https://	N/A	53	6	88	21								
23	16	Breast Cancer Wisconsin (Progn	http	198	34	1995	Life	1/12/1995	276394	76	data	1) ID number	40	16	92	22								
24	99	Tic-Tac-Toe Endgame	http	958	9	1991	Game	19/08/1991	310422	63	data	1. top-left-square:	19	33	96	23								
25	92	Spambase	http	4601	57	1999	Computer	1/07/1999	744313	25	data	The last column of	4	75	100	24								
26	10	Automobile	http	205	26	1987	N/A	19/05/1987	873273	18	data	Attribute:	3	84	102	25								
27	103	Congressional Voting Records	http	435	16	1987	Social	27/04/1987	290644	71	data	1. Class Name: 2	20	32	103	26								
28	108	Yeast	http	1484	8	1996	Life	1/09/1996	374003	54	data	1. Sequence	11	52	106	27								
29	31	Coverttype	http	581012	54	1998	Life	1/08/1998	444144	41	data	Given is the	6	66	107	28								
30	27	Credit Approval	http	690	15		Financial	N/A	666861	31	data	1. b, a,	4	76	107	29								

Figure 1 – Definition of ten datasets chosen from 5 areas

### 1.3 Appendix 3 – Definition of datasets for the second selection

In the second selection 40 more datasets from the UCI Catalogue were analysed. The definition was determined based on a proportional representation of each domain, considering the total number of datasets available in each. This was executed while excluding the datasets adopted in the first iteration from the initial five chosen domains. This iteration introduced two additional domains: the "Business" domain and an untitled category encompassing areas not specified in the UCI Catalogue.

Area	Count	Frequency (%)	Count for the second iteration	New frequency	Distribution for 40 datasets	Final count for 40 datasets
Computer	232	37.3%	230	37.6%	15.03	15
Life	145	23.3%	143	23.4%	9.35	9
Physical	57	9.2%	55	9.0%	3.59	4
Social	41	6.6%	39	6.4%	2.55	3
Business	40	6.4%	40	6.5%	2.61	3
Game	12	1.9%	12	2.0%	0.78	0
Financial	5	0.8%	3	0.5%	0.20	0
Computer Security	1	0.2%	1	0.2%	0.07	0
(Missing)	89	14.3%	89	14.5%	5.82	6
Total	622	1	612	1	40	40

Figure 2 – Determination of 40 new datasets to be analysed.

The datasets were chosen based on their number of web hits, as a significant portion of the datasets in the UCI Catalog do not provide information on the number of papers that used them.

See below the final list of the datasets chosen according to the rule from above:

index	name	area	instances	attribute	year	#webhits	Order	date_donat	web_hits	dataset_file_format
186	Wine Quality	Business	4898	12	2009	6	1	7/10/2009	2160922	csv
222	Bank Marketing	Business	45211	17	2012	7	2	14/02/2012	2056977	zip
352	Online Retail	Business	541909	8	2015	20	3	6/11/2015	832562	xls
602	Dry Bean Dataset	Computer	13611	17	2020	3	1	14/09/2020	2213938	zip
545	Rice (Cammeo and Osm	Computer	3810	8	2019	9	2	6/10/2019	1969142	
232	Human Activity Recogn	Computer	10299	561	2012	14		10/12/2012	1314591	zip
360	Air quality	Computer	9358	15	2016	27	3	23/03/2016	725672	zip
346	Air Quality	Computer	9358	15	2016	28		23/03/2016	725671	zip
228	SMS Spam Collection	Computer	5574		2012	34		22/06/2012	479991	zip
242	Energy efficiency	Computer	768	8	2012	38	4	30/11/2012	462717	xls
267	banknote authentication	Computer	1372	5	2013	39	5	16/04/2013	460121	txt
29	Computer Hardware	Computer	209	9	1987	46	6	1/10/1987	414194	data
80	Optical Recognition of	Computer	5620	64	1998	52		1/07/1998	389787	z
6	Artificial Characters	Computer	6000	7	1992	61		1/07/1992	313031	z
294	Combined Cycle Power	Computer	9568	4	2014	69	7	26/03/2014	292540	zip
81	Pen-Based Recognition	Computer	10992	16	1998	73		1/07/1998	283022	z
342	Detect Malicious Execu	Computer	373	513	2016	77		3/03/2016	273969	rar
229	Skin Segmentation	Computer	245057	4	2012	86	8	17/07/2012	255159	txt
256	Daily and Sports Activi	Computer	9120	5625	2013	88		8/07/2013	251575	zip
246	3D Road Network (North	Computer	434874	4	2013	98	9	16/04/2013	235945	txt
51	Internet Advertisement	Computer	3279	1558	1998	45		1/07/1998	416162	zip
4	Anonymous Microsoft V	Computer	37711	294	1998	108		1/11/1998	221962	data
374	Appliances energy pred	Computer	19735	29	2017	109	10	15/02/2017	221573	csv
248	Buzz in social media	Computer	140000	77	2013	115	11	27/05/2013	206018	gz
343	Occupancy Detection	Computer	20560	7	2016	121	12	29/02/2016	200819	zip
128	KDD Cup 1999 Data	Computer	4000000	42	1999	127	13	1/01/1999	193028	gz
303	Perfume Data	Computer	560	2	2014	131	14	22/07/2014	189180	xls
225	Restaurant & consumer	Computer	138	47	2012	132	15	4/08/2012	186722	zip

This sheet is ordered by area. Here are the 3 datasets from Business area and 15 from Computer area as previously defined. It is easy to see that the order in each area is done by #webhits. There are many cases in red that were not chosen. Some cases are due to high quantity of attributes (561 in ID 232, 513 in ID 342, 5625 in ID 256, 1558 in ID 51 and 294 in ID 4); or repeated datasets (ID 360 and 346); or just have one or two columns (ID 228); or are basically just numbers in character recognition datasets (IDs 80, 6 and 81).

Below you see the continuation.

index	name	area	instances	attribute	year	#webhits	Order	date_donat	web_hits	dataset_file_format
17	Breast Cancer Wisconsin	Life	569	32	1995	8	1	1/11/1995	1972842	data
850	Raisin Dataset	Life	900	8	2021	11	2	1/04/2021	1563266	zip
1	Abalone	Life	4177	8	1995	12	3	1/12/1995	1431163	data
15	Breast Cancer Wisconsin	Life	699	10	1992	17	4	15/07/1992	875007	data
73	Mushroom	Life	8124	22	1987	21	5	27/04/1987	821573	data
34	Diabetes	Life	20			26		N/A	741148	Z
14	Breast Cancer	Life	286	9	1988	29	6	11/07/1988	705622	data
236	seeds	Life	210	7	2012	37	7	29/09/2012	465163	txt
111	Zoo	Life	101	17	1990	40	8	15/05/1990	448279	data
31	Covertypes	Life	581012	54	1998	41	9	1/08/1998	444144	data
19	Car Evaluation	N/A	1728	6	1997	10	1	1/06/1997	1748706	data
10	Automobile	N/A	205	26	1987	18	2	19/05/1987	873273	data
9	Auto MPG	N/A	398	8	1993	19	3	7/07/1993	856989	data
40	Flags	N/A	194	30	1990	49	4	15/05/1990	395601	data
164	Bag of Words	N/A	8000000	100000	2008	55		12/03/2008	363665	txt
104	University UISP-readable	N/A	285	17	1988	70		1/07/1988	292193	data
132	Movie	N/A	10000		1999	80		7/07/1999	269872	data
336	Chronic_Kidney_Disease	N/A	400	25	2015	81	5	3/07/2015	268517	rar
331	Sentiment Labelled Sentences	N/A	3000		2015	84		30/05/2015	255655	zip
50	Image Segmentation	N/A	2310	19	1990	89	6	1/11/1990	250837	data
162	Forest Fires	Physical	517	13	2008	15	1	29/02/2008	1184000	csv
235	Individual household electricity consumption	Physical	2075259	9	2012	32	2	30/08/2012	538670	zip
165	Concrete Compressive Strength	Physical	1030	9	2007	59	3	3/08/2007	328451	xls
138	Robot Execution Failure	Physical	463	90	1999	60		23/04/1999	324835	data
52	Ionosphere	Physical	351	34	1989	62		1/01/1989	311140	data
381	Beijing PM2.5 Data	Physical	43824	13	2017	78	4	19/01/2017	270894	csv
320	Student Performance	Social	649	33	2014	13	5	27/11/2014	1328168	zip
275	Bike Sharing Dataset	Social	17389	16	2013	24	6	20/12/2013	763587	zip
13	Balloons	Social	16	4		51	7	N/A	392085	data

And here are the 9 datasets from Life, 6 from N/A, 4 from Physical and 3 from Social areas. The datasets in red, that had to be discarded were due to: being in a Z compacted file that the code cannot access the contents (ID 34); a dataset with 100000 attributes (ID 164); a dataset that is not text (ID 104); a dataset in HTML (ID 132); a dataset in two columns (ID 331); a dataset where the data are not in one line, making it impossible to read it automatically (ID 138) and a dataset which does not exhibit names/labels of columns/attributes (ID 52).

Figure 3 – Definition of forty datasets from UCI for second iteration.

## 1.4 Appendix 4 – Algorithm and Results of Attribute-Based Semantic Type Detection

Here is the Algorithm for this section:

Attribute-Based Semantic Type Detection.	
1.	Read Dataset Information: Load dataset information, such as dataset names and column descriptions, from an external source (e.g., an Excel file).
2.	Clean Columns:
a.	Extract ID, column names, and descriptions from the 'Original Column' data.
b.	Standardize column names by converting them to lowercase, removing special characters, and separating descriptive information.
c.	Split descriptive information into multiple parts for further analysis.
3.	Preprocess Columns:
a.	Further preprocess the standardized column names to replace abbreviations with their full forms using an abbreviations dictionary.
b.	Clean and prepare the column names for semantic analysis.
4.	Open Formats and Abbreviations Dictionaries: Load dictionaries that contain mappings between abbreviations and their full forms, as well as mappings between column name keywords and their associated data formats.
5.	Replace Abbreviations in Column Names: For each column name, replace any abbreviations found with their full forms based on the abbreviations dictionary.
6.	Apply Semantic Analysis:
a.	For each column, identify the target word (keyword) from the column name that matches an entry in the formats dictionary.
b.	Assign the associated data format to the column based on the identified target word.
c.	For descriptions associated with each column, perform a similar analysis to identify additional keywords and associated data formats.
d.	Special handling for identifying 'ID' columns and handling cases where the column name matches specific patterns or criteria.
7.	Analysis of Column: Determine the final format for each column by comparing the findings from the column name and description analyses. Resolve any discrepancies or special cases according to predefined rules.
8.	Identify Origin: For each column, identify whether the final format determination came from the analysis of the column name or the description.
9.	Save Results: Output the analysis results to an external file (e.g., Excel) for further use or review.

This algorithm provides a structured approach to analyzing dataset attributes/columns based on their labels/names and descriptions, utilizing predefined dictionaries for semantic analysis and format identification. The goal is to standardize column names and identify potential data formats for further investigation.

To illustrate the practical application of this algorithm, let us consider its execution on the 'Heart Disease' dataset. Initially, the notebook reads the information from the 'AllColumnsFromTenDatasets.xlsx' file for a specific dataset (Table 3 from section 4.2.1 above) and initially breaks down each column's 'Original Column' data into its constituent parts: ID, Column, and Description. The description is formed from parts of text after the symbols '(' or ':' or '/'. There can be many Descriptions for a single column being analysed. See Table 7 for the result:

Table 2 – Breaking columns and descriptions for the 'Heart Disease' dataset

lex	name	area	Original Column	ID	Column	Description 1	Description 2
45	Heart Disease	Life	1. age	1	age		
45	Heart Disease	Life	2. sex	2	sex		
45	Heart Disease	Life	3. cp (chest pain type)	3	cp	chest pain type	
45	Heart Disease	Life	4. trestbps (resting blood pressure Integer)	4	trestbps	resting blood pressure Integer	
45	Heart Disease	Life	5. chol (serum cholestoral in mg/dl Integer)	5	chol	serum cholestoral in mg	dl Integer
45	Heart Disease	Life	6. fbs (fasting blood sugar > 120 mg/dl Categorical) (1 = true; 0 = false)	6	fbs	fasting blood sugar > 120 mg	dl Categorical
45	Heart Disease	Life	7. restecg (resting electrocardiographic results Categorical)	7	restecg	resting electrocardiographic results Categorical	
45	Heart Disease	Life	8. thalach (maximum heart rate achieved Integer)	8	thalach	maximum heart rate achieved Integer	
45	Heart Disease	Life	9. exang (exercise induced angina (1 = yes; 0 = no) Categorical)	9	exang	exercise induced angina	1 = yes; 0 = no
45	Heart Disease	Life	10. oldpeak (ST depression induced by exercise relative to rest Integer)	10	oldpeak	ST depression induced by exercise relative to rest Integer	
45	Heart Disease	Life	11. slope (the slope of the peak exercise ST segment Categorical)	11	slope	the slope of the peak exercise ST segment Categorical	
45	Heart Disease	Life	12. ca (number of major vessels (Categorical 0-3) colored by flourosopy)	12	ca	number of major vessels	Categorical 0-3
45	Heart Disease	Life	13. thal (3 = normal; 6 = fixed defect; 7 = reversible defect Categorical)	13	thal	3 = normal; 6 = fixed defect; 7 = reversible defect Categorical	
45	Heart Disease	Life	14. num (the predicted attribute)	14	num	the predicted attribute	

Notice above that the first two columns don't contain any words besides their own. All the other columns contain a first word and separate words after symbols '(' or '/'. So, description(s) were created.

After the Description is separated there are pre-processing activities that turn all words to lowercase and delete some symbols such as '-' and '\_'. The 'CleanedColumn' is then created. In this field the code searches for words that exist in the Formats dictionary, or maybe in the Abbreviations dictionary. The first one found goes to the 'ColumnKeyword'. The format associated with this target word goes to the column 'ColumnFormat'. See below in Table 8.

Observe below that the word 'age' brings the format 'age' which is numerical bounded and the word 'sex' brings the format 'categorical'. Then, all columns until 'slope' (format numerical), do not show any target word found, because the words that exist in the name of the columns do not exist in the formats nor in the abbreviations dictionary. But there are words in the Description parts that exist in the Formats Dictionary, and they are brought together with the format associated:

Table 3 – Obtaining words and formats for the 'Heart Disease' dataset.

ID	Column	Description 1	Description 2	CleanedColumn	ColumnKeyword	ColumnFormat	DescriptionKeyword	DescriptionFormat
1	age			age	age	age		
2	sex			sex	sex	categorical		
3	cp	chest pain type		cp			type	categorical
4	trestbps	resting blood pressure Integer		trestbps			integer	numerical
5	chol	serum cholestoral in mg	dl Integer	chol			integer	numerical
6	fbs	fasting blood sugar > 120 mg	dl Categorical	fbs			categorical	categorical
7	restecg	resting electrocardiographic results Categorical		restecg			categorical	categorical
8	thalach	maximum heart rate achieved Integer		thalach			rate	numerical
9	exang	exercise induced angina	1 = yes; 0 = no	exang			yes	categorical
10	oldpeak	ST depression induced by exercise relative to rest Integer		oldpeak			integer	numerical
11	slope	the slope of the peak exercise ST segment Categorical		slope	slope	numerical	categorical	categorical
12	ca	number of major vessels	Categorical 0-3	ca			categorical	categorical
13	thal	3 = normal; 6 = fixed defect; 7 = reversible defect Categorical		thal			categorical	categorical
14	num	the predicted attribute		num	num	numerical	predicted	categorical

In the end, the column 'FinalFormat' receives the final data type detection either from the Column Format or from the Description Format. And the word associated with the final format goes to the last column, 'SourceKeyword'. See Table9:

Table 4 – Final results for data type detection for the 'Heart Disease' dataset

ID	CleanedColumn	ColumnKeyword	ColumnFormat	DescriptionKeyword	DescriptionFormat	FinalFormat	SourceKeyword
1	age	age	age			age	age
2	sex	sex	categorical			categorical	sex
3	cp			type	categorical	categorical	type
4	trestbps			integer	numerical	numerical	integer
5	chol			integer	numerical	numerical	integer
6	fbs			categorical	categorical	categorical	categorical
7	restecg			categorical	categorical	categorical	categorical
8	thalach			rate	numerical	numerical	rate
9	exang			yes	categorical	categorical	yes
10	oldpeak			integer	numerical	numerical	integer
11	slope	slope	numerical	categorical	categorical	categorical	categorical
12	ca			categorical	categorical	categorical	categorical
13	thal			categorical	categorical	categorical	categorical
14	num	num	numerical	predicted	categorical	categorical	predicted

Above, only the first two 'FinalFormat' results were the same as 'ColumnFormat', and all the others were obtained from 'DescriptionFormat'.

The results of applying this code to all fifty datasets, covering 922 columns/attributes, are compiled in 'AnalysedColumns.xlsx' [Silva, 2024]. It's important to note that during the analysis of each dataset, the dictionaries were iteratively refined to enhance their effectiveness in generating the desired outputs.

**Observation: The Python version adopted was 3.9.12.**

## 1.5 Appendix 5 - Algorithm and Results of Attribute-Based Data Quality Assessment

Here is the Algorithm for this section:

Attribute-Based Data Quality Assessment	
1.	Read Ten/Forty Datasets file (depending if first or second set of datasets) <ul style="list-style-type: none"><li>a. Import necessary libraries.</li><li>b. Define the path to the Excel file containing ten/forty datasets' details.</li><li>c. Load the Excel file into a pandas DataFrame and display its first few rows for verification.</li></ul>
2.	Read AnalysedColumns file from previous code and Define Dataset Index <ul style="list-style-type: none"><li>a. Load the Excel sheet that lists AnalysedColumns for 50 datasets.</li><li>b. Define the dataset index for which the user wants to find the dataset file URL and name.</li><li>c. Print the selected dataset index and the current date-time.</li></ul>
3.	Step 3: Get Dataset File URL <ul style="list-style-type: none"><li>a. Define a function get_dataset_file_url that loads dataset details from the first Excel file above and returns the dataset file URL and name for a specific dataset index.</li><li>b. Call the function with the previously defined Excel file path and dataset index.</li><li>c. Print the obtained dataset file URL and name of file, or an error message if not found.</li></ul>
4.	Step 4: Load Dataset <ul style="list-style-type: none"><li>a. Implement several functions to load datasets:<ul style="list-style-type: none"><li>i. is_header_for_csv: Determines if a line is likely a header based on the presence of numeric values.</li><li>ii. load_csv: Loads a CSV file into a pandas DataFrame, with adjustments for delimiter detection and header presence.</li><li>iii. is_header_for_excel: Determines if the first row in an Excel file is likely a header.</li><li>iv. load_excel: Loads an Excel file into a pandas DataFrame, with adjustments for header presence.</li><li>v. download_and_extract: Downloads and extracts an archive file from a URL.</li><li>vi. select_file_from_extracted: Allows the user to select a file from an extracted directory.</li><li>vii. fetch_file_content: Fetches the content of a file from a URL.</li><li>viii. load_dataset: Determines the file type and loads it accordingly, supporting local and remote files and handling archives.</li></ul></li></ul>
5.	Assign Column Names <ul style="list-style-type: none"><li>a. Import necessary libraries.</li><li>b. Define a function assign_column_names that assigns column names to a DataFrame based on a given dataset index from an "AnalysedColumns" DataFrame. It checks if required columns exist in the DataFrame and assigns extracted column names to the target DataFrame.</li><li>c. Call the function with the analysed columns DataFrame, desired dataset index, and the previously loaded DataFrame to assign column names.</li></ul>
6.	Data Quality Issues <ul style="list-style-type: none"><li>a. Define a class "DataQualityIssues" containing static methods to handle various data quality issues. These methods cover a wide range of issues, including missing data, ambiguous data, extraneous data, outdated temporal data, duplicates, structural conflicts, domain violations, wrong data type, uniqueness violation, and the use of special characters.</li><li>b. Each method is designed to handle a specific type of data quality issue identified by DQI (Data Quality Issue) numbers or not directly associated with a DQI number.</li><li>c. The methods perform checks and return information about the presence of data quality issues, including the indices of problematic data points, error messages, and the specific data quality issue addressed.</li></ul>



- d. This step involves identifying and addressing various data quality issues using the DataQualityIssues class's methods. The specific implementation details of handling each data quality issue are encapsulated within the class's methods.
- 

Below are functions created to evaluate specific formats defined in the previous code.

- a. Each task involves:
    - i. Handling missing or invalid entries (blank, empty, null, NaN)
    - ii. Applying specific validations relevant to the data type or format being checked.
    - iii. Using predefined lists of exceptions where applicable (e.g., linking words for names, acceptable abbreviations for states).
    - iv. Reporting errors and summarizing the results, including value distributions and ranges where applicable.
    - v. Generate and report frequency distributions in some cases to provide insights into the data's distribution.
    - vi. For tasks involving dates, times, or specific formats, the algorithm may deduce the most likely format based on sample values before applying validations.
- 
7. Check Numerical Greater or Equal to Zero
- 
- a. Handle missing or blank values in the specified column.
  - b. Validate that all numerical values are greater than or equal to zero.
  - c. Report on non-numeric values.
  - d. Summarize the findings, including any errors and the range of numeric values.
- 
8. Check Numerical
- 
- a. Similar to item 07 but focused on ensuring all values are numerical without the greater than zero condition.
- 
9. Check Numerical Between
- 
- a. Ensure all numerical values fall within a specified range.
  - b. Similar error checking as previous tasks, with the addition of validating the numeric range.
- 
10. Check if ID
- 
- a. Determine if column values are suitable for use as a Primary Key by checking for uniqueness, non-negative values, and other ID-specific criteria.
- 
11. Check String Content
- 
- a. Ensure all values are non-empty strings and report on any values that do not meet this criterion.
- 
12. Check if Categorical
- 
- a. Validate if a column can be considered categorical based on the number of unique values and the presence of predefined unacceptable values.
- 
13. Check Month
- 
- a. Verify that all column values are valid representations of months.
- 
14. Check Weekday
- 
- a. Ensure all values correctly represent weekdays.
- 
15. Check Date
- 
- a. Validate date values, ensuring they adhere to a consistent format and fall within a reasonable range.
-

---

16. Check DateTime

---

- a. Similar to the date check but for datetime values, ensuring both the date and time components are valid.
- 

17. Check Time

---

- a. Validate time values, focusing on the format and range of the time component.
- 

18. Check Model Name

---

- a. Ensure model names meet certain criteria, such as being non-empty and potentially following a specific format.
- 

Check Name (Task 18.5)

---

- b. Validate names, ensuring they do not contain numbers or special characters and adhere to capitalization norms.
- 

19. Check Street

---

- a. Validate street names for standard conventions (e.g., capitalization, avoiding special characters).
  - b. Find the range of street names.
  - c. Return any identified errors along with the range of street names.
- 

20. Check City

---

- a. Validate city names for proper capitalization and format.
  - b. Generate a frequency distribution of city names.
  - c. Report on non-standard city names and provide a frequency distribution.
- 

21. Check State

---

- a. Validate state names or abbreviations for capitalization and correct format.
  - b. Create a frequency distribution of state names.
  - c. Highlight and report non-standard state names along with the frequency distribution.
- 

22. Check Country

---

- a. Ensure country names meet expected standards of capitalization and correctness.
  - b. Generate and report a frequency distribution for country names.
- 

23. Check Postal Code

---

- a. Validate postal codes for standard formats (length, numeric/alphanumeric values).
  - b. Report on non-standard postal codes and provide the range of valid postal codes.
- 

24. Check Phone Numbers

---

- a. Validate phone numbers for standard formats (including international formats).
  - b. Identify and report non-standard phone numbers and provide the range of valid phone numbers.
- 

25. Check IP Format

---

- a. Ensure IP addresses are in valid formats (IPv4, IPv6).
  - b. Report on any non-standard IP address formats.
- 

26. Check URL Format

---

	<ul style="list-style-type: none"> <li>a. Validate URLs for standard formats.</li> <li>b. Identify and report non-standard URLs.</li> </ul>
27. Check Email Format	
	<ul style="list-style-type: none"> <li>a. Ensure email addresses meet standard email format criteria.</li> <li>b. Report on any non-standard email formats.</li> </ul>
28. Check Binary Values	
	<ul style="list-style-type: none"> <li>a. Validate if values in a column conform to binary values (e.g., '0', '1', 'true', 'false', and variations thereof).</li> <li>b. Provide a frequency distribution of binary values.</li> <li>c. Highlight and report non-standard binary values along with their frequency distribution.</li> </ul>
29. Analyse Data Quality	
<p>This function comprehensively evaluates the quality of data across various columns specified in a DataFrame that specifies the expected data format for each column. The process involves several key steps:</p>	
	<ul style="list-style-type: none"> <li>a. Initialization and Setup: Define any global parameters or thresholds needed for analysis, such as valid year ranges or thresholds for categorical uniqueness.</li> <li>b. Preparation: Determine the set of columns to be analyzed based on a provided <code>desired_dataset_index</code>, which helps to filter <code>analysed_columns_df</code> for relevant columns.</li> <li>c. Iteration over Columns: Loop through each column specified for the analysis, ensuring that each column exists in the primary DataFrame.</li> <li>d. Determination of Column Format: For each column, identify the desired data format based on the information in <code>analysed_columns_df</code>. This step involves mapping textual descriptions of formats to specific validation functions.</li> <li>e. Validation: Apply the appropriate validation function based on the determined format. This might include: <ul style="list-style-type: none"> <li>i. Checking for numerical ranges or specific conditions (e.g., greater than zero, within a specific range).</li> <li>ii. Verifying categorical data against a uniqueness threshold.</li> <li>iii. Validating string formats for names, addresses, etc.</li> <li>iv. Confirming the format of dates, times, URLs, email addresses, etc.</li> </ul> </li> <li>f. Frequency Distribution: For categorical data, generate and display a frequency distribution to provide insights into data diversity and potential anomalies.</li> <li>g. Result Compilation and Reporting: Aggregate the results of each column's analysis into a structured format, typically as a dictionary or a text summary, indicating any detected issues or confirming adherence to expected formats.</li> <li>h. Output Presentation: Print or return a comprehensive summary of the data quality analysis, highlighting any columns with issues and providing insights into the distribution of valid data.</li> <li>i. Handling Special Cases: Depending on the column's intended format, perform specialized checks (e.g., for IP addresses, geographical coordinates, or binary values) using tailored validation functions.</li> </ul>

Below we can see some of the output of the Data Quality Assessment of Dataset 45 – Heart Disease, that explains the results in easy-to-read descriptions. Observe that it shows the name of the column being analysed, followed by the ('SourceKeyword', when different) found from the previous code, then it exhibits the format being analysed and the respective output, confirming data integrity, with the number of items and the range for numerical values, and possible Data Quality Issues and their associated data quality dimensions, as defined in Appendix 1:

#### 45 Heart Disease

```
age:
Age format: All 303 values are numerical and valid in the range [0, 130].
Actual range of values: (29.0 : 77.0)

sex:
All 303 values are correctly categorical.
Categorical format with 2 unique values:
Category Frequency
1.0 206
0.0 97
...
trestbps (integer):
Numerical format: All 303 values are numerical in the range (94.0:200.0).
...
ca (categorical):
Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
Unacceptable value(s) at index(es): 166, '?', (192, '?'), (287, '?'), (302, '?')]]
Categorical format with 5 unique values:
Category Frequency
0.0 176
1.0 65
2.0 38
3.0 20
? 4
```

As seen above the 'age' column has been analysed as a numerical valid in the range [0,130], and the number of values and actual range were provided. The column sex was evaluated as categorical and as such its data integrity show that it had only two values, with the associated frequency. The column 'trestbps' had the word 'integer' in the description, so it is analysed as numerical. The last case provided regards the column 'ca', which due to having the word 'categorical' in the description was evaluated this way. Notice that our code found 4 cases of the content '?', which is an 'Unacceptable value' for categorical data. It was considered Data Quality Issue #4, related to 'Ambiguous Data' and the Data Quality Dimensions of 'Accuracy and Consistency'. Other Data Quality Assessment tools do not find this error.

The complete output from the latest analysis of 50 datasets and 922 columns/attributes is available on GitHub ('Discoveries on the 10 datasets.pdf' and 'Discoveries on the forty datasets.pdf' [Silva, 2024]).

ID	QC	Issue Description	Data Quality Dimension	Function Name	Check#	Format	Input	Error Explanation	Tested examples
1	13	Missing Data	Completeness	check_numerical_gz_zero	1	Numerical	= 0	Blank/Empty/Null/NaN values	None, '-', 'null', 'nan', '-', 'nan'
2	13	Missing Data	Completeness	check_numerical	2	Numerical		Blank/Empty/Null/NaN values	None, '-', 'null', '-', 'nan'
3	13	Missing Data	Completeness	check_numerical_between	3	Numerical	between	Blank/Empty/Null/NaN values	None, '-', 'null', '-', 'nan'
4	13	Missing Data	Completeness	check_id_attributes	4	ID		Blank/Empty/Null/NaN values	None, '-', 'null', '-', 'nan'
5	13	Missing Data	Completeness	check_string	5	String		Blank/Empty/Null/NaN values	None, 'nan', '-', 'nan'
6	13	Missing Data	Completeness	check_if_categorical	6	Categorical		Blank/Empty/Null/NaN values	-, 'null', 'None', '-', 'null'
7	13	Missing Data	Completeness	check_month	7	Month		Blank/Empty/Null/NaN values	None
8	13	Missing Data	Completeness	check_weekday	8	Weekday		Blank/Empty/Null/NaN values	None
9	13	Missing Data	Completeness	check_date	9	Date		Blank/Empty/Null/NaN values	-, 'nan', 'None', 'Null'
10	13	Missing Data	Completeness	check_datetime	10	DateTime		Blank/Empty/Null/NaN values	nan, 'None', 'Null', '-', 'nan'
11	13	Missing Data	Completeness	check_time	11	Time		Blank/Empty/Null/NaN values	-, 'None'
12	13	Missing Data	Completeness	check_name	12.5	Name		Blank/Empty/Null/NaN values	-, 'None'
13	13	Missing Data	Completeness	check_street	13	Street		Blank/Empty/Null/NaN values	None, '-', 'Null'
14	13	Missing Data	Completeness	check_city	14	City		Blank/Empty/Null/NaN values	None, '-', 'Null'
15	13	Missing Data	Completeness	check_state	15	State		Blank/Empty/Null/NaN values	None, '-', 'Null'
16	13	Missing Data	Completeness	check_country	16	Country		Blank/Empty/Null/NaN values	None, '-', 'Null'
17	13	Missing Data	Completeness	check_postal_code	17	Postal Code		Blank/Empty/Null/NaN values	None, '-', 'Null'
18	13	Missing Data	Completeness	check_phone_numbers	18	Phone		Blank/Empty/Null/NaN values	None, '-', 'Null'
19	13	Missing Data	Completeness	check_up_format	19	UP		Blank/Empty/Null/NaN values	None, '-', 'Null'
20	13	Missing Data	Completeness	check_url_format	20	URL		Blank/Empty/Null/NaN values	-, 'None', 'Null'
21	13	Missing Data	Completeness	check_email_format	21	Email		Blank/Empty/Null/NaN values	-, 'None', 'Null'
22	13	Missing Data	Completeness	check_binary_values	22	Binary		Blank/Empty/Null/NaN values	None, '-', 'P'
23	13	Extraneous Data	Accuracy, Consistency	check_if_categorical	6	Categorical		Extraneous content	-, 'P'
24	5	Extraneous Data	Consistency, Uniqueness	check_name	12.5	Name		Extraneous data	-, 'P', 'john3 Doe', 'Emily', '11'
25	5	Extraneous Data	Consistency, Uniqueness	check_street	13	Street		Extraneous data	-, 'P', 'Emily', '11'
26	5	Extraneous Data	Consistency, Uniqueness	check_city	14	City		Extraneous data	-, 'P', 'Dubai', '11'
27	5	Extraneous Data	Consistency, Uniqueness	check_state	15	State		Extraneous data	-, 'P', 'CA', 'P', 'California', '11'
28	5	Extraneous Data	Consistency, Uniqueness	check_country	16	Country		Extraneous data	-, 'P', 'Canada', '11'
29	6	Outdated Temporal Data	Timeliness	check_date	9	Date		Dates not in [1800-2020] period	3/4/2121, '13101720', '14/5/2222', '01/01/1500', '31/12/2121', '2121/12/25'
30	6	Outdated Temporal Data	Timeliness	check_datetime	10	DateTime		Dates not in [1800-2020] period	3/4/2121 13:00, '14/5/2222 13:05', '01/01/1500 13:00-10', '31/12/2121 13:00-20', '2121/12/25 13:00-12'
31	6	Duplicates	Uniqueness	check_id_attributes	4	ID		Duplicate values	'AB123CD456', 'Duplicate 1', '32/01/2021', '28/02/2021', '31/11/2021', '06/01/2021', '01/06/2021', '2021/13/01', 'not a date', '2021-02-30', '29022002'
32	13	Temporal mismatch	Accuracy, Timeliness	check_date	9	Date		Invalid date values	13/01/2021 12:00, '2021/16/01 14:00', '18/01/2021 25:00', '2021-01-1915:30', '21/01/2021 16:00:66', 'not a date time', '24/01/2021 26:30', '29/02/2021 15:20', '2021-01-30 15:20:05'
33	13	Temporal mismatch	Accuracy, Timeliness	check_time	11	Time		Invalid time values	'13:01', 'invalid', '02:30 PM', '25:05'
34	13	Temporal mismatch	Consistency	check_datetime	9	Date		Dates without format DDMYYYY in [1800-2020] period	'12/13/2021', '2021/12/25', '2021/04/7', '2021/08/15', '2021/11/03', '12/30/2021', '02/280202', '2021218'
35	13	Temporal mismatch	Consistency	check_datetime	10	DateTime		Dates without format DDMYYYY in [1800-2020] period	'01-11-23:45', '2021/01/12 23:40', 'January 14, 2021 1:00', '2021/01/23'

This sheet shows that, many cases of bad data such as contents: None, ", ' ', 'null', nan, ' ', '"" are related to Data Quality Issue #1, Missing Data, related to Data Quality Dimension 'Completeness', and are found in all different format functions, and their Error Explanation is Blank/Empty/Null/NaN values.

The same error related to content '?' is considered Data Quality Issue number 5, Extraneous Data, related to Consistency and Uniqueness, when the data is a geographic format. Other examples such as 'CA2', for state or 'Canada! ', for country are also shown. These geographic functions do not consider valid characters different than text in their content.

[illegible]

Continuing the analysis from the previous content, this Figure 5 shows many cases of Data Quality Issue #15, Domain Violation, related to Accuracy. Many different functions related to many formats are exhibited for this DQI.

13

For Format 'Numerical between', Values outside range are shown.

For ID attributes the content cannot be negative or floating-point numbers.

For 'Month' values cases such as numbers 0 or 13 or contents 'not a month' or 'mn' are flagged errors.

For 'Weekday' we don't accept 0, -1, 'Mn'.

But months from 1 to 12 or any variation on names such as 'December', 'FEB', and weekdays from 1 to 7, or 'Mon', 'TUE', and 'FR' are valid.

Geographic formats, such as Street, City, State and Country do not accept words which are not Capitalized or some cases such as 'Short length alphanumeric values' for Postal Code.

Phone, IP, URL and Email values are also quite restricted. See above many cases when the data are considered bad data for Accuracy.

Binary values also just accept the maximum of 2 distinct values, and usually cases such as 0 and 1 or 'Yes' and 'No', are accepted.

Another interesting DQ Issue is #17, 'Wrong Data Type', related to Dimension Consistency, and used in many functions such as numerical functions when non-numerical values are not accepted, or ID attributes when there is an “Inconsistent length or format in alphanumeric values”. This has to do with an analysis that is made from the first 10 items and if the format changes due to different length of content this is flagged. It also flags unacceptable formats in ID data, such as '?', '—' or 'aa'.

Besides Wrong Data Type there is another case for DQI 17, Non-String Data Type, when the format being tested is 'String', and the content is 'Non-String'. Values such as 11, 5.67, True, {'key': 'value'} are examples for this.

## 1.7 Appendix 7 – Summary of Discoveries with real dataset bad data

DQI#	DQ Issue Description	Data Quality Dimension	Function Name	Check#	Format Being Analyzed	Error Explanation	Data Issues	Dataset	Columns - Attributes
2	4 Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	?"	45 Heart Disease	ca, thal
3	4 Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	?"	2 Adult	workclass, occupation
4	5 Extraneous Data	Consistency, Uniqueness	check_country	15	Country	Extraneous data	?"	2 Adult	native-country
<div> <div>handicapped-infants, water-project-cost-sharing, adoption-of-the-budget-resolution, physician-fee-freeze, el-salvador-aid, religious-groups-in-schools, anti-satellite-test-ban, aid-to-nicaraguan-contras, mx-missile, immigration, synfuels-corporation-cutback, education-spending, superfund-right-to-sue, crime, duty-free-exports, export-administration-act-south-africa,</div> </div>									
5	4 Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	?"	103 Congressional Voting Records	
6	4 Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	?"	27 Credit Approval	A1, A4, A5, A6, A7
7	17 Wrong Data Type	Consistency	check_numerical_ge_zero	1	Numerical >= 0	Non-numeric values	?"	27 Credit Approval	A2, A14
8	17 Wrong Data Type	Consistency	check_numerical	2	Numerical	Non-numeric values	'C536379', 'C536383', etc.	352 - Online Retail	InvoiceNo
9	10 Structural Conflicts	Consistency, Uniqueness	check_if_categorical	6	Categorical	Data seems not categorical or has too many ca 4070 categories	"	352 - Online Retail	StockCode
10	1 Missing Data	Completeness	check_string	3	String	Blank/Empty/Null/NaN values	"	352 - Online Retail	Description
11	17 Non-String Data Type	Consistency	check_string	5	String	Non-string values	20713	352 - Online Retail	Description
12	1 Missing Data	Completeness	check_id_attributes	4	ID	Blank/Empty/Null/NaN values	"	352 - Online Retail	CustomerID
13	9 Duplicates	Uniqueness	check_id_attributes	4	ID	Duplicate values	17850, 13047, etc	352 - Online Retail	CustomerID
14	19 Uniqueness Violation	Uniqueness	check_id_attributes	4	ID	Uniqueness violation	17850, 13047, etc	352 - Online Retail	CustomerID

Figure 6 – First part of Summary of Discoveries with real bad data from UCI datasets.

This sheet shows all bad data from the first 10 datasets analysed and also from one dataset in the second set of datasets. The datasets with real bad data are only four: 45 – Heart Disease, 2 – Adult, 103 – Congressional Voting Records and 27 – Credit Approval. The dataset 352 – Online Retail is the first from the second set of 40 datasets from the UCI Catalogue.

Here we can see many cases where '?' exist in the datasets and are considered Data Issues. In fact, in the first set of 10 datasets, the only problems found were the appearance of '?'. According to the format being analysed, from the words that appeared in the attribute label, an appropriate output was presented. In Datasets 45, 2, 103 and 27 the DQI# was 4, related to Ambiguous Data, according to the table presented in Appendix 1. They were all associated with attributes that the dictionaries considered to be Categorical information. Line 2, presenting the attributes 'ca' and 'thal', were considered categorical because in the descriptions on the 'attributes information' column obtained in the original procedure where the information from all 622 datasets from UCI Catalogue came, there is the word 'Categorical' for these two attributes. Line 3, from Dataset 2, has '?' in attributes 'workclass' and 'occupation'. These are words that are automatically considered Categorical from the formats dictionary. Line 4 has another format being analysed, also from Dataset 2. The attribute is titled 'native-country', so, the format being analysed is Country. Country has some specific rules, not accepting lower-case names, for example, but it also does not accept Extraneous data, and '?' is considered so. Extraneous Data is Data Quality Issue #5, associated to Dimensions 'Consistency and Uniqueness'. Line 5, from dataset 103, contains many attributes that are considered Categorical because in the description the word 'yes' do appear. 'Yes' is automatically considered Categorical. Then, in all the attributes presented there are cases where '?' appear. Line 6 shows 5 attributes that are too generic, titled A1 to A7, but in the description of these attributes there is the word 'Categorical'. Therefore, they follow the same rule as the previous case. Line 7 shows a case where the description has the word 'Continuous'. Then, the attributes A2 and A14 are considered Numerical >= 0. And for this format '?' is considered 'Non-numeric values', leading to the Data Quality Issue (DQI) #17, 'Wrong Data Type', and Data Quality Dimension 'Consistency'.

Lines 8 to 14 are from Dataset 352, and they contain many different situations. The first one is similar to the previous one. It is also DQI# 17, but now the attribute is 'InvoiceNo', with the word 'number' in the description, associated with format numerical, and the content has a letter in the beginning. Instead of being a number the bad data are: 'C536379', and 'C536383', etc., so these show a case different than the usual '?' values presented earlier. Line 9 shows the first case of DQI# 10, 'Structural Conflicts', Dimensions 'Consistency and Uniqueness'. The attribute is 'StockCode', associated with the word code, which is considered 'Categorical'. But there are 4070 unique values in this attribute, which is not common to Categorical cases. Lines 10 and 11 are for Attribute 'Description', which is considered format 'string'. But there are many cases of content "" and one case of content 20713. The first is a case for Data Quality Issue #1, Missing Data, dimension Completeness. And the second case is again for DQI 17, due to 'Non-String Data Type'.

The next 3 cases are related to the attribute 'CustomerID'. Due to having the letters 'ID' in the name it is considered an 'ID column format'. The first problem is DQI#1 again, 'Missing Data', due to the value ". The second problem is related to the existence of 'Duplicates', associated with 'Dimension Uniqueness'. Some values that appear more than once are 17850 and 13047. This causes DQI#19 'Uniqueness'.

1	DQI#	DQ Issue Description	Data Quality Dimension	Function Name	Check#	Format Being Analyzed	Error Explanation	Data Issues	Dataset	Columns - Attributes
15	15	Domain Violation	Accuracy	check_numerical_ge_zero	1	Numerical >= 0	Negative values	-200	360 - Air quality	Relative Humidity, AH
16	Below are lines 524 to 526. Observe that many other values have the content -200, besides the two last Humidity values:									
17										Absolute Humidity
18										
19										
20										
21										
22										
23										
24										
25	9	Duplicates	Uniqueness	check_id_attributes	4	ID	Duplicate values	144552912, 93323205, etc	246 - 3D Road Network (North Jutland OSM_ID	
26	17	Wrong Data Type	Consistency	check_id_attributes	4	ID	Inconsistent length in alphanumeric value(s)	42991631, 42991632, etc	246 - 3D Road Network (North Jutland OSM_ID	
27	19	Uniqueness Violation	Uniqueness	check_id_attributes	4	ID	Uniqueness violation	144552912, 93323205, etc	246 - 3D Road Network (North Jutland OSM_ID	
28	10	Structural Conflicts	Consistency, Uniqueness	check_if_categorical	6	Categorical	Data seems not categorical or has too many categories	248 - Buzz in social media	Feature to predict	
29	5	Extraneous Data	Consistency, Uniqueness	check_name	12.5	Name	Extraneous data	'constrected2'	303 - Perfume Data	Perfume_name
30	10	Structural Conflicts	Consistency, Uniqueness	check_if_categorical	6	Categorical	Data seems not categorical or has too many categories	225 - Restaurant & consumer data	the_geom_meter	
31	5	Extraneous Data	Consistency, Uniqueness	check_name	12.5	Name	Extraneous data	'Restaurante 75', 'Cenaduria El 725 - Restaurant & consumer data	name	
32	15	Domain Violation	Accuracy	check_name	12.5	Name	Capitalization/Format issues	'puesto de tacos', 'little pizza Er 225 - Restaurant & consumer data	name	
33	5	Extraneous Data	Consistency, Uniqueness	check_street	13	Street	Extraneous data	'?'	225 - Restaurant & consumer data	address
34	15	Domain Violation	Accuracy	check_street	13	Street	Capitalization/Format issues	'esquina santos degollado y leo 225 - Restaurant & consumer data	address	
35	5	Extraneous Data	Consistency, Uniqueness	check_city	14	City	Extraneous data	'?'	225 - Restaurant & consumer data	city
36	15	Domain Violation	Accuracy	check_city	14	City	Capitalization/Format issues	's.l.p.', 'Victoria', etc	225 - Restaurant & consumer data	city
37	5	Extraneous Data	Consistency, Uniqueness	check_state	15	State	Extraneous data	'?'	225 - Restaurant & consumer data	state
38	15	Domain Violation	Accuracy	check_state	15	State	Capitalization/Format issues	's.l.p.', 'tamaulipas', etc	225 - Restaurant & consumer data	state
39	5	Extraneous Data	Consistency, Uniqueness	check_country	16	Country	Extraneous data	'?'	225 - Restaurant & consumer data	country
40	15	Domain Violation	Accuracy	check_country	16	Country	Capitalization/Format issues	'mexico'	225 - Restaurant & consumer data	country
41	15	Domain Violation	Accuracy	check_phone_numbers	18	Phone	Incorrect telephone number format	'?'	225 - Restaurant & consumer data	fax
42	17	Wrong Data Type	Consistency	check_postal_code	17	Postal Code	Non-alphanumeric values	'?'	225 - Restaurant & consumer data	zip
43	15	Domain Violation	Accuracy	check_url_format	20	URL	Invalid URL format	'?'	225 - Restaurant & consumer data	url

Figure 7 – Second part of Summary of Discoveries with real bad data.

Continuing the analysis, the next Dataset is 360 - Air Quality. Our code provided the information that there are Negative values in the attributes 'Relative Humidity' and 'AH Absolute Humidity'. Checking the output, it was found the value '-200', which is surely a not expected value for these attributes. Observing the content, it was noticed that many other columns also contain this value '-200'. The other columns did not output the Data Quality Issue because they were considered only numerical, not numerical greater or equal to 0. Humidity is the word that determined that this analysis should be for values  $\geq 0$ . This is something unique in our research.

In Lines 25 to 27, attribute OSM\_ID also contain issues associated with an ID column. Besides Duplicates and Uniqueness as before, now there is the DQI# 17, due to 'Inconsistent length in alphanumeric values'. Notice that the values shown in the Data Issues have a smaller size than the ones shown in the other cases.

Line 28 shows again a case of DQI# 10, where the Attribute 'Feature to predict' is considered a Categorical format, due to the word 'feature' on it, but it contains 8895 different values, which might be a problem.

Line 29 shows a case of DQI# 5 for the format 'Name', in the attribute titled 'Perfume\_name' on dataset 303. The Data Issue is in the content 'constrected2'. The check name function considers an error when a number appears on it.

The next lines are all for Dataset 225 – Restaurant & consumer data. It contains many geographical columns that contain data issues. '?' appears in the 'address' attribute (the system uses check street function to analyse address information), in the city attribute, in the state, country, fax, zip, and url attributes. Each one of these attributes is analysed by a specific function.

Besides that, the first line analyses a possible categorical attribute (the \_geom\_meter), because in its description there is the word Nominal, which is classified as Categorical. Unfortunately, there are 130 different values, and it may be an issue for DQI# 10.

The 'name' attribute is associated with two different DQI's. DQI 5 for Extraneous Data, because there are numbers in the name ('Restaurante 75'), and DQI# 15, for Domain Violation, due to the Problem in 'Capitalization/Format Issues' because there are words that are not Capitalized ('puesto de tacos').

The address attribute also contains Capitalization issues ('esquina santos degollado y leon guzman'), as well as the city attribute ('s.l.p.'), the state attribute (also 's.l.p.'), and the country attribute ('mexico').



1	DQI#	DQ Issue Description	Data Quality Dimension	Function Name	Check#	Format Being Analyzed	Error Explanation	Data Issues	Dataset	Columns - Attributes
44	9	Duplicates	Uniqueness	check_id_attributes	4	ID	Duplicate values	1017023, 1033078, etc	15 - Breast Cancer Wisconsin (Original)	Sample code number
45	17	Wrong Data Type	Consistency	check_id_attributes	4	ID	Inconsistent length in alphanumeric value(s)	128059, 144888, etc	15 - Breast Cancer Wisconsin (Original)	Sample code number
46	19	Uniqueness Violation	Uniqueness	check_id_attributes	4	ID	Uniqueness violation	1033078, 1070935, etc	15 - Breast Cancer Wisconsin (Original)	Sample code number
47	17	Wrong Data Type	Consistency	check_numerical	2	Numerical	Non-numeric values	"?"	15 - Breast Cancer Wisconsin (Original)	Bare Nuclei
48	4	Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	"?"	73 - Mushroom	node-caps, breast-quad
49	17	Non-String Data Type	Consistency	check_string	5	String	Non-string values	3, 1, etc	10 - Automobile	symboling
50	17	Wrong Data Type	Consistency	check_numerical	2	Numerical	Non-numeric values	"?"	10 - Automobile	normalized-losses, price
51	4	Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	"?"	10 - Automobile	num-of-doors
52	17	Wrong Data Type	Consistency	check_numerical_ge_zero	1	Numerical >= 0	Non-numeric values	"?"	10 - Automobile	bore, stroke, horsepower, peak-rpm
53	17	Wrong Data Type	Consistency	check_numerical_between	3	Numerical between [1800, 2100]	Value(s) outside range [1800, 2100]	70, 82, etc	10 - Automobile	model year
54	5	Extraneous Data	Consistency, Uniqueness	check_name	12.5	Name	Extraneous data	buick skylark 320', ford galaxie	10 - Automobile	car name
55	17	Wrong Data Type	Consistency	check_numerical_ge_zero	1	Numerical >= 0	Non-numeric values	"?"	9 - Auto MPG	horsepower
56	17	Wrong Data Type	Consistency	check_numerical_between	3	Numerical between [1800, 2100]	Value(s) outside range [1800, 2100]	70, 82, etc	9 - Auto MPG	model year
57	5	Extraneous Data	Consistency, Uniqueness	check_name	12.5	Name	Extraneous data	buick skylark 320', ford galaxie	9 - Auto MPG	
58	17	Wrong Data Type	Consistency	check_numerical_between	3	Numerical between [0, 130]	Non-numeric values	"?"	336 - Chronic_Kidney_Disease	Age
59	17	Wrong Data Type	Consistency	check_numerical	2	Numerical	Non-numeric values	"?"	336 - Chronic_Kidney_Disease	Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell clumps, Bacteria, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count
60	4	Ambiguous Data	Accuracy, Consistency	check_if_categorical	6	Categorical	Unacceptable content	"?"	336 - Chronic_Kidney_Disease	
61	17	Wrong Data Type	Consistency	check_numerical_ge_zero	1	Numerical >= 0	Non-numeric values	"?"	336 - Chronic_Kidney_Disease	global_active_power, global_reactive_power, voltage, global_intensity, sub_metering_1, sub_metering_2, sub_metering_3
62	17	Wrong Data Type	Consistency	check_numerical	2	Numerical	Non-numeric values	"?"	235 - Individual household electric po	sub_metering_3
63	17	Wrong Data Type	Consistency	check_numerical_ge_zero	1	Numerical >= 0	Non-numeric values	"NA"	381 - Beijing PM2.5 Data	pm2.5

Figure 8 – Final part of Summary of Discoveries with real bad data.

Finalising this analysis, we observe Dataset 15, with two different Attributes. 'Sample code number' has in its description the information that it is an 'id number'. Therefore, it is being analysed as an ID. Unfortunately, there are Duplicates and Uniqueness violations, and it also has an 'Inconsistent length in alphanumeric values' situation. And the second attribute, 'Bare Nuclei', being considered a Numerical, because in its description there is the text: '1 – 10', contains some '?' data issues.

The next dataset, 73 shows an Ambiguous Data DQI# 4, also due to '?'.

The next two datasets, 10 and 9 are for Automobile related information. They contain many attributes that are the same. '?' appears in some cases for numerical and categorical attributes, but also there are two new cases related to DQI# 17, Wrong Data Type, for Values outside range [1800-2100], for the same attribute 'model year', where there are values 70 and 82 for example. This is surely a case of year with two digits only, but it is an interesting output. Other two cases are related to the 'car name' attribute, where the names of cars are not Capitalized.

The dataset 336 contains only numerical and categorical cases due to '?'.

The same happens with dataset 235.

The final dataset, 381, contains once again DQI# 17, but now the data issue is not '?', but 'NA'.