POLITECNICO
MILANO 1863

DATA QUALITY ANALYZER

COMPUTER SCIENCE
AND ENGINEER

DATA AND INFORMATION QUALITY PROJECT

A.Y. 2018-2019

Leonardo Febbo

Giacomo Astolfi

INTRODUCTION

PRESENTATION & DEVELOPMENT ENVIRONMENT

DOMAIN PROBLEM

We have a large amount of data generated from a data stream of a transport company. In it there are informations regarding each validated ticket, on which bus and on which line.

GOAL

The goal is to analyze the quality of this dataset and make some analysis on it and comparisons on portions of it.

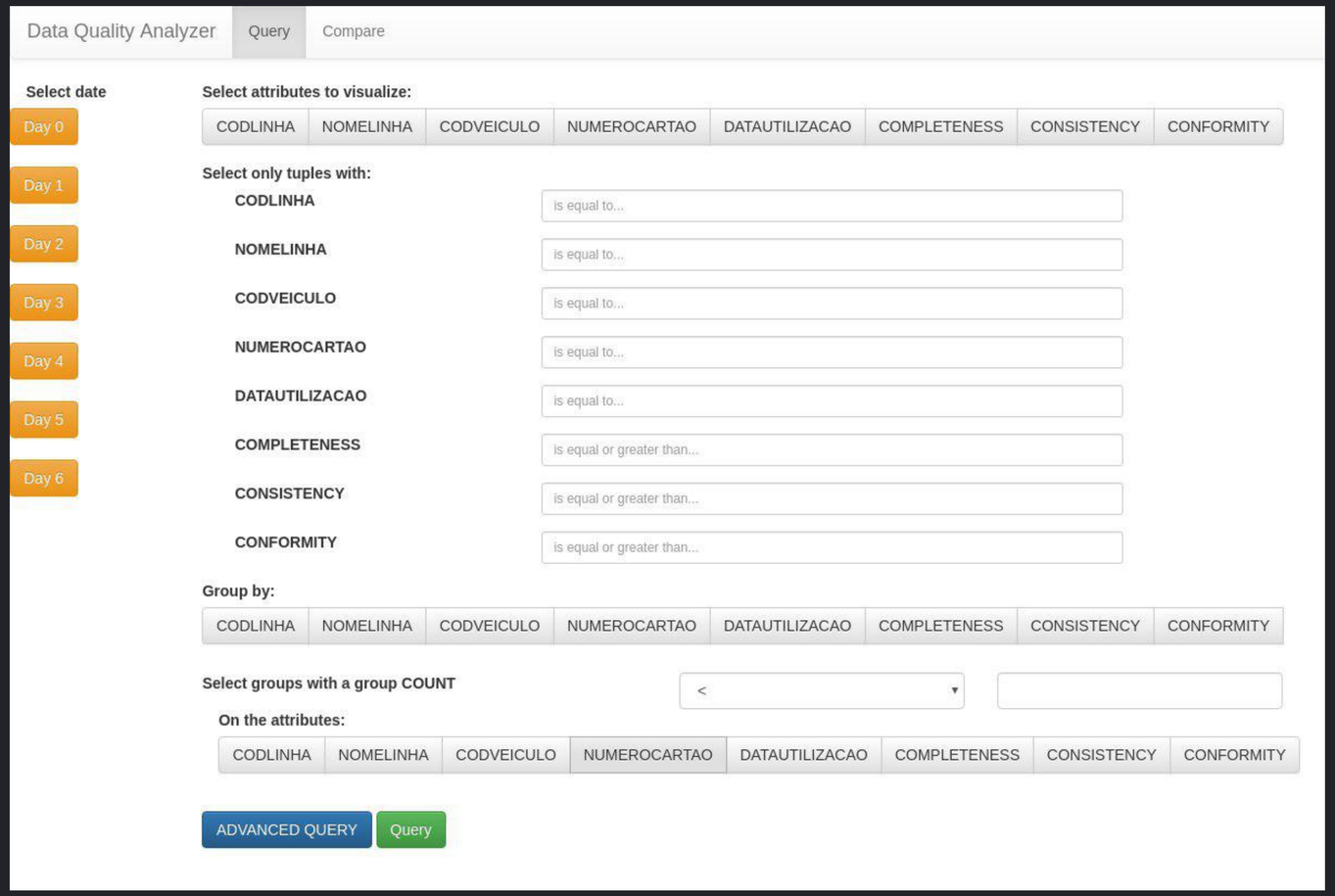
We want to give users the opportunity to analyze these datasets based on their data quality dimensions and to make queries based on dataset attributes. The entire amount of data will be grouped in different datasets per day or week.

Additionaly, since the dataset is almost perfect, it will be "dirtied" in different ways on the different groups in order to analyze more cases of data quality problems.

FEATURES

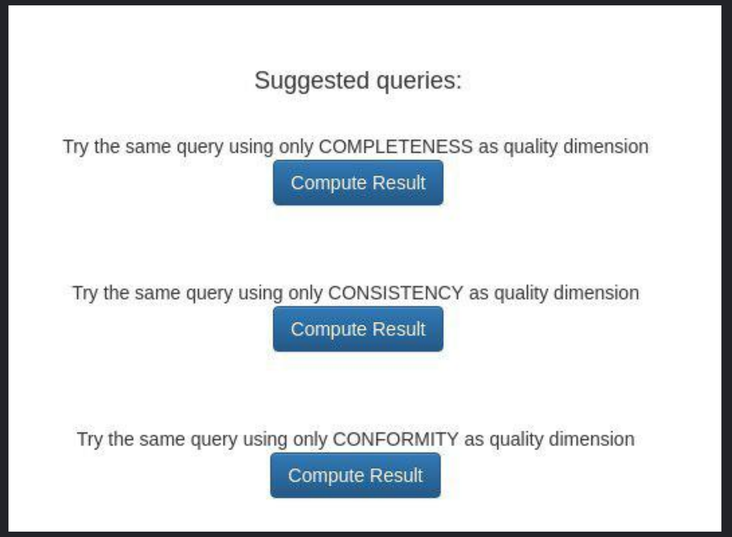
1. The user can choose one of the provided dataset he wants to analyze and build a query for it.

When provided with the query, the application will return the portion of the dataset which respects the requirements, and the user will be able to download it in CSV format. Some of the data will also be printed in order to provide feedback to the user.



FEATURES

2. In addition, the user will also receive as suggestions other portions of the dataset that perhaps only respect some of the requirements imposed by him and he will be able to decide whether to download even those or not.



FEATURES

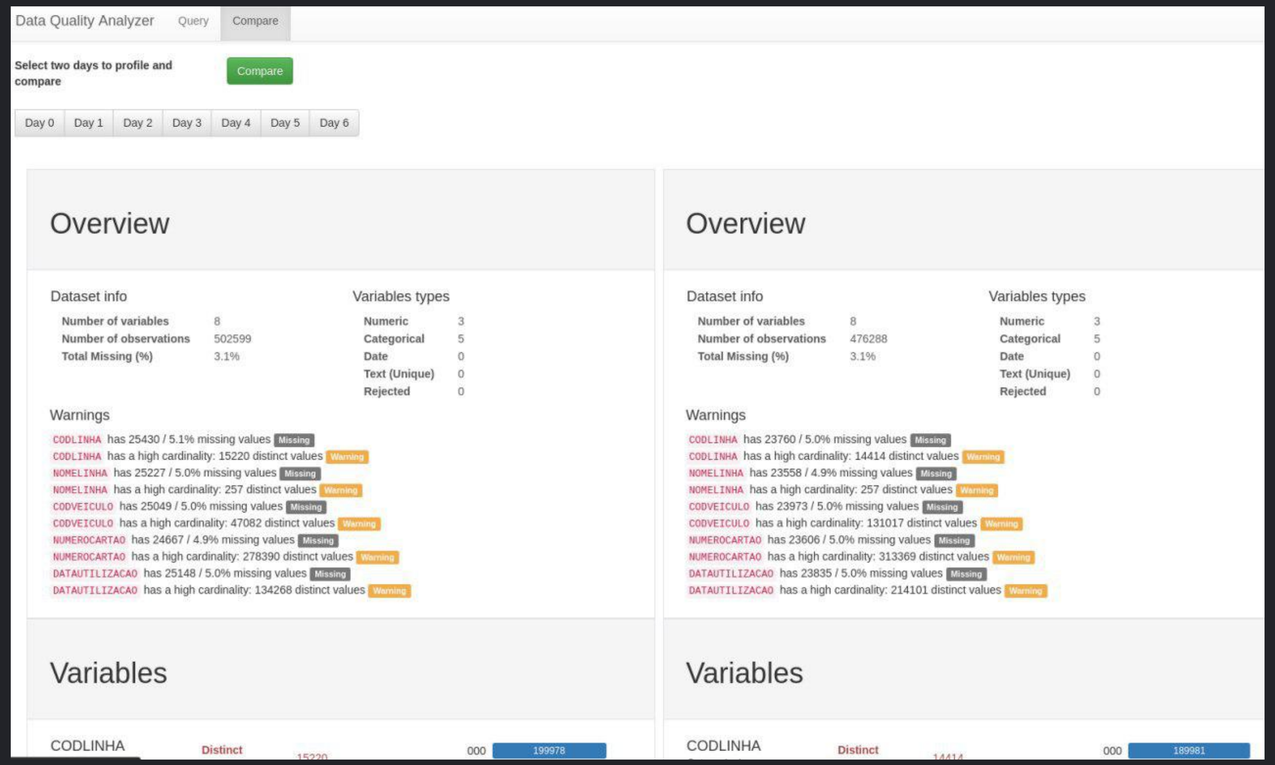
3. We also give the user the opportunity to see further details regarding each portion of the dataset obtained from the query (a sort of profiling) before downloading it.

4. The user can also compare two dataset using the profiling informations or compare two sections of them based on a query of his choice when possible.

FEATURES

3. We also give the user the opportunity to see further details regarding each portion of the dataset obtained from the query (a sort of profiling) before downloading it.

4. The user can also compare two dataset using the profiling informations or compare two sections of them based on a query of his choice when possible.



QUERY REQUIREMENTS

`Completeness > value` **: the user can search for data that has a completeness greater than the value entered**

`Consistency > value` **: the user can search for data that has a consistency greater than the value entered**

`Timeliness > value` **: the user can search for data that has a timeliness greater than the value entered**

`Group by attributes` **: the user can group the data based on a selected attributes**

`Attributes: value` **: the user can search for data that has a certain v alue for the attributes chosen**

PROFILING REQUIREMENTS

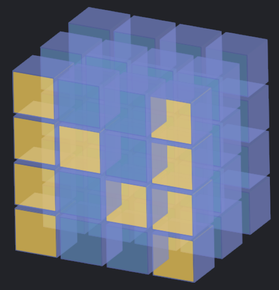
Upon selecting a dataset, the application will show metadata about each column, including:
- The type (categorical, numerical, datetime etc.)
- The number and percentage of missing values
- The number and percentage of distinct values - A plot of the values (when possible)

And other information about the dataset itself, including:
- Simple metadata of the dataset (number of variables, size in memory, number of tuples, duplicate rows etc.)
- Mean and standard deviation of the completeness, timeliness and consistency
- Matrix plot of the missing values
- A random sample of the data

COMPARE REQUIREMENTS

- The profiling information will be visible for both selected datasets
- The stats of the dataset with higher quality will be highlited in some way (either by color or direct comparison)

DEVELOPMENT ENVIRONMENT



NumPy

DIRTY DATASET

HOW WE DUMPED THE DATASET

DIRTY DATAFRAME

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

COMPLETENESS

```
df = df.mask(np.random.choice([True, False], size=df.shape, p=[.2, .8]))
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	280	N. SRA.DE NAZARÉ	BC911	1430250	07/10/15 07:37:02
1	280	N. SRA.DE NAZARÉ	BC911	NaN	07/10/15 07:51:25
2	NaN	N. SRA.DE NAZARÉ	BC911	NaN	NaN

DIRTY DATAFRAME

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

CODLINHA

```
x = df[“CODLINHA”].to_numpy().copy()
random.shuffle(x)

import rstr
for i in range(size_df):
    new = rstr.xeger(r’(\d|[A-Z]){3}')
    x.extend([new])

df["CODLINHA"] = df[“CODLINHA”].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	OPC	N. SRA.DE NAZARÉ	BC911	1430250	07/10/15 07:37:02
1	000	N. SRA.DE NAZARÉ	BC911	2470195	07/10/15 07:51:25
2	280	N. SRA.DE NAZARÉ	BC911	3234514	07/10/15 18:49:49

NOMELINHA

```
x = df[“NOMELINHA”].to_numpy().copy()
random.shuffle(x)

df["NOMELINHA"] = df[“NOMELINHA”].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	280	OPER S/LINHA	BC911	1430250	07/10/15 07:37:02
1	280	OPER S/LINHA	BC911	2470195	07/10/15 07:51:25
2	280	VITÓRIA RÉGIA	BC911	3234514	07/10/15 18:49:49
3	280	V. IZABEL	BC911	3234514	07/10/15 18:49:51

CODVEICULO

```
import rstr
x=[]
for i in range(size_df):
    new = rstr.xeger(r'(\d\d|[A-Z][A-Z])(\d){3}')
    x.extend([new])

df["CODVEICULO"] = df[“CODVEICULO”].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	280	N. SRA.DE NAZARÉ	BC911	1430250	07/10/15 07:37:02
1	280	N. SRA.DE NAZARÉ	EA332	2470195	07/10/15 07:51:25
2	280	N. SRA.DE NAZARÉ	BC911	3234514	07/10/15 18:49:49
3	280	N. SRA.DE NAZARÉ	TL272	3234514	07/10/15 18:49:51

DATAUTILIZACAO

```
x=[]
for i in range(size_df):
    new = rstr.xeger(r'(0[1-9]|[12][0-9]|3[01])/(0[1-9]|1[0-2])/\d\d
                     (00|0[0-9]|1[0-9]|2[0-3]):(0[0-9]|[0-5][0-9]):(0[0-9]|[0-5][0-9])')
    x.extend([new])

df["DATAUTILIZACAO"] = df["DATAUTILIZACAO"].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	280	N. SRA.DE NAZARÉ	BC911	1430250	07/10/15 07:37:02
1	280	N. SRA.DE NAZARÉ	EA332	2470195	07/10/15 07:51:25
2	280	N. SRA.DE NAZARÉ	BC911	3234514	31/01/33 00:50:10
3	280	N. SRA.DE NAZARÉ	TL272	3234514	20/09/11 15:12:50

DIRTY DATAFRAME

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

CODLINHA

...

```
x=[]
for i in range(size_df):
    new = rstr.xeger(r'(\d|[A-Z]){1,5}')
    x.extend([new])

df["CODLINHA"] = df[“CODLINHA”].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
3	280	N. SRA.DE NAZARÉ	BC911	0595898100	07/10/15 18:49:51
4	000	BAIRRO NOVO B	UV372	771305	07/05/16 00:04:09
5	DL4RG	OPER S/LINHA	08047	856665	07/10/15 16:50:18

CODVEICULO

```
import rstr
x=[]
for i in range(size_df):
    new = rstr.xeger(r'(\d|[A-Z]){0,1}(\d\d|[A-Z][A-Z]|\d[A-Z])(\d){3}(\d|[A-Z]){0,1}')
    x.extend([new])

df["CODVEICULO"] = df[“CODVEICULO”].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	07/10/15 07:51:25
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49

NUMEROCARTAO

```
x=[]
for i in range(size_df):
    new = rstr.xeger(r'(\d){1,13}(\w){0,1}')
    x.extend([new])

df["NUMEROCARTAO"] = df["NUMEROCARTAO"].mask(
        np.random.choice([True, False], size=size_df, p=probability) == True, x)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
12	814	MOSSUNGUÊ	LN404	3527951	07/10/15 17:07:31
13	542	MOSSUNGUÊ	LN404	7324I	07/10/15 17:46:36
18	653	SABARÁ	HA293	12O	07/10/15 16:15:46
19	653	SABARÁ	ND550	2471107	07/10/15 18:44:03

DATAUTILIZACAO

...

```
#2015-10-07 19:30:20
for i in range(size_df):
    new = rstr.xeger(r'[12][0-9][0-9][0-9]-(0?[1-9]|1[0-2])-(0?[1-9]|[12][0-9]|3[01])
                       (00|0[0-9]|1[0-9]|2[0-3]):(0[0-9]|[0-5][0-9]):(0[0-9]|[0-5][0-9])')
    x.extend([new])

#07-10-2015 19:30:20
for i in range(size_df):
    new = rstr.xeger(r'(0?[1-9]|[12][0-9]|3[01])-(0?[1-9]|1[0-2])-\d\d
                       (00|0[0-9]|1[0-9]|2[0-3]):(0[0-9]|[0-5][0-9]):(0[0-9]|[0-5][0-9])')
    x.extend([new])
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49

DATA QUALITY DIMENSIONS

CHECK DATA QUALITY DIMENSIONS

DATA QUALITY DIMENSIONS

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

COMPLETENESS

```
df['null-values'] = df.isnull().sum(axis=1)

conditions = [
    df['null-values'] == 0,
    df['null-values'] == 1,
    df['null-values'] == 2,
    df['null-values'] == 3,
    df['null-values'] == 4]
choices = [100, 80, 60, 40, 20]

df['COMPLETENESS'] = np.select(conditions, choices, default=0)

df = df.drop(['null-values'], axis=1)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	COMPLETENESS
0	280	NaN	NaN	1430250	NaN	40
1	280	N. SRA.DE NAZARÉ	BC911	2470195	07/10/15 07:51:25	100
2	280	NaN	NaN	NaN	NaN	20
3	280	N. SRA.DE NAZARÉ	BC911	NaN	07/10/15 18:49:51	80

DATA QUALITY DIMENSIONS

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

LIST LINHA

```
for i,row in df.iterrows():
    if (pd.isnull(row["CODLINHA"]) or pd.isnull(row["NOMELINHA"])):
        df.at[i, 'CHECK-LIST-LINHA'] = 0
    else:
        try:
            row_list_linha = list_linha.loc[row["CODLINHA"]]
            if row_list_linha["NOMELINHA"] == row["NOMELINHA"]:
                df.at[i, 'CHECK-LIST-LINHA'] = 1
            else:
                df.at[i, 'CHECK-LIST-LINHA'] = 0
        except:
            df.at[i, 'CHECK-LIST-LINHA'] = 0
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	COMPLETENESS	CHECK-LIST-LINHA
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02	100	0.0
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0	100	1.0
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49	100	0.0

CODLINHA

```
for i,row in df.iterrows():
    if pd.isnull(row["CODLINHA"]):
        df.at[i, 'CHECK-CODLINHA'] = 0
    else:
        try:
            row_list_linha = list_linha.loc[row["CODLINHA"]]
            df.at[i, 'CHECK-CODLINHA'] = 1
        except:
            df.at[i, 'CHECK-CODLINHA'] = 0
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	COMPLETENESS	CHECK-LIST-LINHA	CHECK-CODLINHA
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02	100	0.0	1.0
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0	100	1.0	1.0
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49	100	0.0	0.0

NOMELINHA

```
for i,row in df.iterrows():
    if pd.isnull(row["NOMELINHA"]):
        df.at[i, 'CHECK-NOMELINHA'] = 0
    else:
        try:
            row_list_linha = list_linha.loc[row["NOMELINHA"]]
            df.at[i, 'CHECK-NOMELINHA'] = 1
        except:
            df.at[i, ‘CHECK-NOMELINHA'] = 0
```

	CODLINHA	NOMELINHA	CHECK-LIST-LINHA	CHECK-CODLINHA	CHECK-NOMELINHA
0	280	ABAETÉ	0.0	1.0	1.0
1	280	N. SRA.DE NAZARÉ	1.0	1.0	1.0
2	23A	OPER S/LINHA	0.0	0.0	1.0

CODVEICULO

```
for i,row in df.iterrows()):
    if pd.isnull(row["CODVEICULO"]):
        df_dirt.at[i, 'CHECK-CODVEICULO'] = 0
    else:
        try:
            row_list_vehicle = list_vehicle.loc[row["CODVEICULO"]]
            df_dirt.at[i, 'CHECK-CODVEICULO'] = 1
        except:
            df_dirt.at[i, ‘CHECK-CODVEICULO'] = 0
```

	CODLINHA	NOMELINHA	CODVEICULO	CHECK-CODVEICULO
0	280	ABAETÉ	29073	0.0
1	280	N. SRA.DE NAZARÉ	HD128K	0.0
2	23A	OPER S/LINHA	BC911	1.0

DATAUTILIZACAO

```
def is_sameday(str_datetime): #consistency datetime
    a = re.search("^07/10/15 (00|0?[0-9]|1[0-9]|2[0-3]):(0?[0-9]|[0-5][0-9]):(0?[0-9]|[0-5][0-9])$", str_datetime)
    b = re.search("^07-10-15 (00|0?[0-9]|1[0-9]|2[0-3]):(0?[0-9]|[0-5][0-9]):(0?[0-9]|[0-5][0-9])$", str_datetime)
    c = re.search("^2015-10-07 (00|0?[0-9]|1[0-9]|2[0-3]):(0?[0-9]|[0-5][0-9]):(0?[0-9]|[0-5][0-9])$", str_datetime)
    return a,b,c

for i,row in tqdm(df.iterrows()):
    if pd.isnull(row["DATAUTILIZACAO"]):
        df.at[i, 'CHECK-DATA'] = 0
    else:
        a,b,c = is_sameday(row[“DATAUTILIZACAO"])
        if a or b or c:
            df.at[i, 'CHECK-DATA'] = 1
        except:
            df.at[i, ‘CHECK-DATA'] = 0
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	CHECK-DATA
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02	1.0
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0	0.0
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49	1.0

CONSISTENCY

```
for i,row in df.iterrows():
        df.at[i, 'TOT'] = row["CHECK-LIST-LINHA"] +
                          row["CHECK-CODLINHA"] + row["CHECK-NOMELINHA"] +
                          row["CHECK-CODVEICULO"] + row[“CHECK-DATA”]
conditions = [
    df['TOT'] == 5.0,
    df['TOT'] == 4.0,
    df['TOT'] == 3.0,
    df['TOT'] == 2.0,
    df['TOT'] == 1.0]
choices = [100, 80, 60, 40, 20]
df['CONSISTENCY'] = np.select(conditions, choices, default=0)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	COMPLETENESS	CONSISTENCY
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02	100	60
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0	100	60
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49	100	60
3	280	N. SRA.DE NAZARÉ	435304P	0595898100	07/10/15 18:49:51	100	80
4	000	BAIRRO NOVO B	70X707M	771305	2567-11-20 14:56:9	100	40

DATA QUALITY DIMENSIONS

Parte Uno

Completeness

Parte Due

Consistency

Parte Tre

Conformity

CODLINHA

```
def is_codlinha(str_datetime):
    return re.search("^(\d|[A-Z]){3}$", str_datetime)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	CONF-CODLINHA
4	000	BAIRRO NOVO B	70X707M	771305	2567-11-20 14:56:9	1.0
5	DL4RG	OPER S/LINHA	08047	NaN	2619-4-31 09:21:35	0.0
6	NaN	ALTO BOQUEIRÃO	30918	2115218	07/10/15 13:18:26	0.0

CODVEICULO

```
def is_codveiculo(str_datetime):
    return re.search("^(\d\d|[A-Z][A-Z])(\d){3}$", str_datetime)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	CONF-CODVEICULO
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02	1.0
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0	0.0
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49	1.0

NUMEROCARTAO

...

```
def is_numerocartao(str_datetime):
    return re.search("^(\d){1,10}$", str_datetime)
```

DATAUTILIZACAO

```
#07/10/15 19:30:20
def is_strdate(str_datetime):
    return re.search("^(0?[1-9]|[12][0-9]|3[01])/(0?[1-9]|1[0-2])/\d\d
                     (00|0?[0-9]|1[0-9]|2[0-3]):(0?[0-9]|[0-5][0-9]):(0?[0-9]|[0-5][0-9])$”,
                     str_datetime)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	CONF-DATAUTILIZACAO
4	000	BAIRRO NOVO B	70X707M	771305	2567-11-20 14:56:9	0.0
5	DL4RG	OPER S/LINHA	08047	NaN	2619-4-31 09:21:35	0.0
6	NaN	ALTO BOQUEIRÃO	30918	2115218	07/10/15 13:18:26	1.0

CONFORMITY

```
for i,row in df.iterrows():
        df.at[i, 'TOT'] = row["CONF-CODLINHA"] + row["CONF-CODVEICULO"] +
                          row["CONF-NUMEROCARTAO"] + row["CONF-DATAUTILIZACAO"]
conditions = [
    df['TOT'] == 4.0,
    df['TOT'] == 3.0,
    df['TOT'] == 2.0,
    df['TOT'] == 1.0]
choices = [100, 75, 50, 25]
df['CONFORMITY'] = np.select(conditions, choices, default=0)
```

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	COMPLETENESS	CONSISTENCY	CONFORMITY
0	280	ABAETÉ	29073	1430250	07/10/15 07:37:02	100	60	100
1	280	N. SRA.DE NAZARÉ	HD128K	2470195	2948-02-4 00:0:0	100	60	50
2	23A	OPER S/LINHA	BC911	3234514	07/10/15 18:49:49	100	60	100
3	280	N. SRA.DE NAZARÉ	435304P	0595898100	07/10/15 18:49:51	100	80	75
4	000	BAIRRO NOVO B	70X707M	771305	2567-11-20 14:56:9	100	40	50

PROFILING

...

DQA

OVERVIEW

Overview

Dataset info

Number of variables	8
Number of observations	100342
Total Missing (%)	25.0%

Variables types

Numeric	3
Categorical	5
Date	0
Text (Unique)	0
Rejected	0

Warnings

- `CODLINHA` has 40324 / 40.2% missing values **Missing**
- `CODLINHA` has a high cardinality: 15072 distinct values **Warning**
- `NOMELINHA` has 40191 / 40.1% missing values **Missing**
- `NOMELINHA` has a high cardinality: 205 distinct values **Warning**
- `CODVEICULO` has 40262 / 40.1% missing values **Missing**
- `CODVEICULO` has a high cardinality: 41968 distinct values **Warning**
- `NUMEROCARTAO` has 39948 / 39.8% missing values **Missing**
- `NUMEROCARTAO` has a high cardinality: 50985 distinct values **Warning**
- `DATAUTILIZACAO` has 40055 / 39.9% missing values **Missing**
- `DATAUTILIZACAO` has a high cardinality: 53977 distinct values **Warning**
- `COMPLETENESS` has 1011 / 1.0% zeros
- `CONSISTENCY` has 12096 / 12.1% zeros
- `CONFORMITY` has 12978 / 12.9% zeros

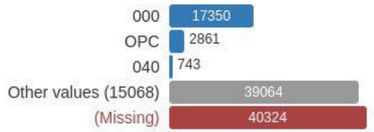
DQA

VARIABLES

Variables

CODLINHA
Categorical

Distinct count	15072
Unique (%)	25.1%
Missing (%)	40.2%
Missing (n)	40324
Unparsed Json Types Found	int, float



Toggle details

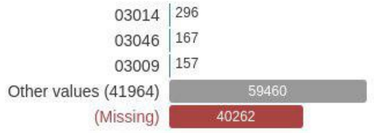
NOMELINHA
Categorical

Distinct count	205
Unique (%)	0.3%
Missing (%)	40.1%
Missing (n)	40191
Unparsed Json Types Found	[]

Toggle details

CODVEICULO
Categorical

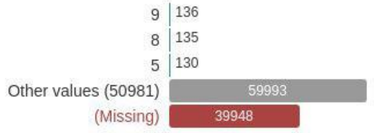
Distinct count	41968
Unique (%)	69.9%
Missing (%)	40.1%
Missing (n)	40262
Unparsed Json Types Found	int, float



Toggle details

NUMEROCARTAO
Categorical

Distinct count	50985
Unique (%)	84.4%
Missing (%)	39.8%
Missing (n)	39948
Unparsed	



DQA

DETAILS

NOMELINHA
Categorical

Distinct count	205
Unique (%)	0.3%
Missing (%)	40.1%
Missing (n)	40191
Unparsed Json Types Found	[]

Toggle details

Value	Count	Frequency (%)
OPER S/LINHA	28860	28.8%
OP. CONTIGENCIA	4835	4.8%
INTERBAIRROS IV	1254	1.2%
SISTEMA ARAUCARIA	828	0.8%
INTERBAIRR II H	787	0.8%
INTERBAIRROS III	683	0.7%
INTERB II ANTI H	620	0.6%
INTERBAIRROS V	558	0.6%
RIO BONITO	514	0.5%
S.RITA/PINHEIRIN	503	0.5%
CABRAL / PORTÃO	479	0.5%
C.RASO/CAIUÁ	439	0.4%
CENTENÁRIO/HAUER	404	0.4%
TERMINAL PINHEIRINHO	354	0.4%
ALTO BOQUEIRÃO	317	0.3%
C.COMP/C.RASO	297	0.3%
BARREIRINHA	259	0.3%

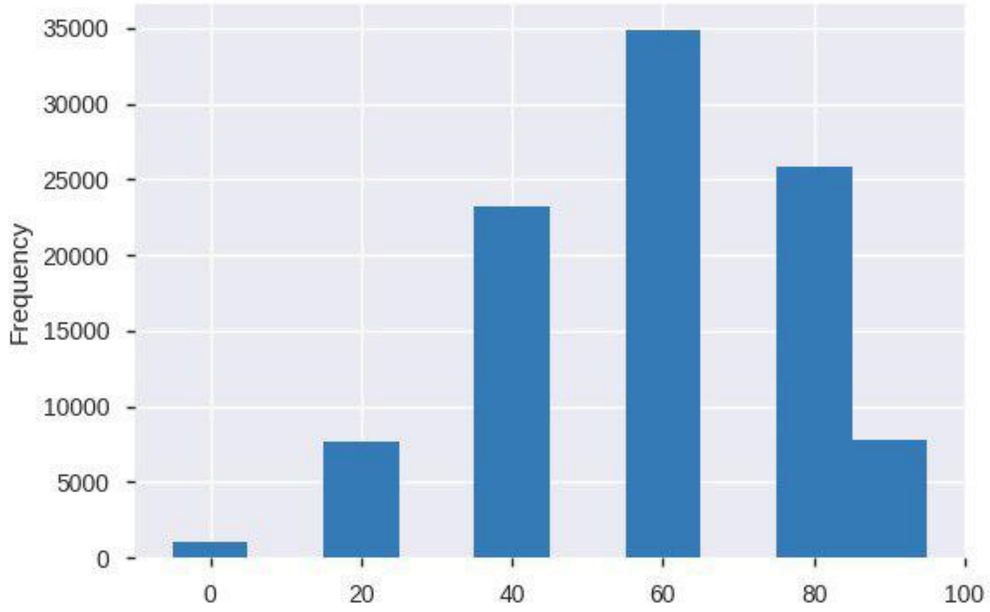
DETAILS

Quantile statistics

Minimum	0
5-th percentile	20
Q1	40
Median	60
Q3	80
95-th percentile	100
Maximum	100
Range	100
Interquartile range	40

Descriptive statistics

Standard deviation	21.876
Coef of variation	0.36472
Kurtosis	-0.36384
Mean	59.981
MAD	16.551
Skewness	-0.17633
Sum	6018600
Variance	478.57
Memory size	0.0 B



DQA

SAMPLES

Sample

	CODLINHA	NOMELINHA	CODVEICULO	NUMEROCARTAO	DATAUTILIZACAO	COMPL
0	170	None	7D4642	None	2833-12-30 10:0:52	60
1	000	None	00186	None	31/02/23 00:00:02	60
2	P26I	INTERBAIRROS V	None	None	None	40
3	None	SISTEMA ARAUCARIA	None	None	None	20
4	000	None	None	None	25/10/15 16:31:19	40

DEMO