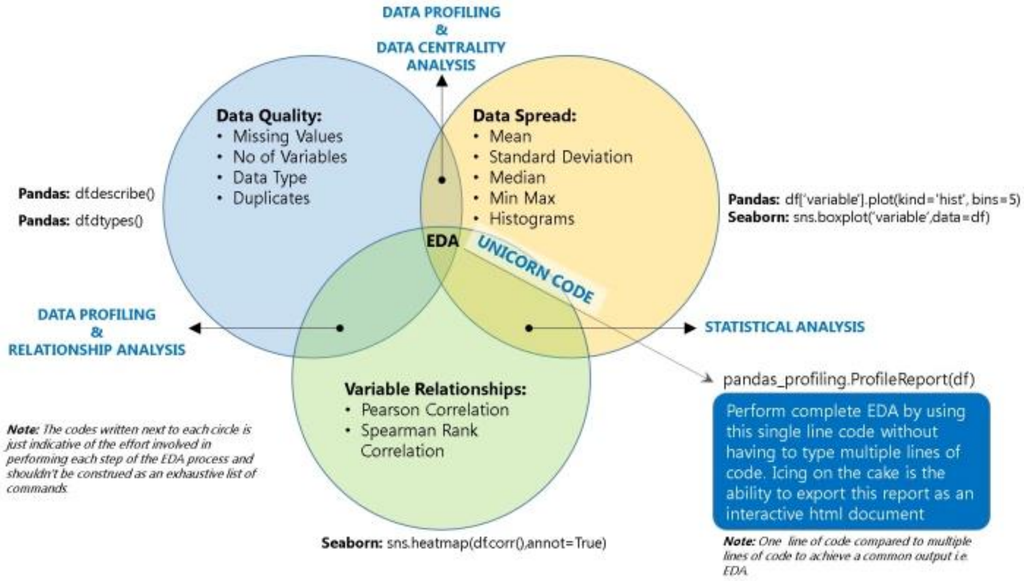
PANDAS PROFILING (PYTHON)

Python module for Exploratory Data Analysis (EDA))



DATA PROFILING & DATA CENTRALITY ANALYSIS

Data Quality:
- Missing Values
- No of Variables
- Data Type
- Duplicates

Pandas: df.describe()
Pandas: df.dtypes()

Data Spread:
- Mean
- Standard Deviation
- Median
- Min Max
- Histograms

Pandas: df['variable'].plot(kind='hist', bins=5)
Seaborn: sns.boxplot('variable',data=df)

EDA

UNICORN CODE

DATA PROFILING & RELATIONSHIP ANALYSIS

STATISTICAL ANALYSIS

Variable Relationships:
- Pearson Correlation
- Spearman Rank Correlation

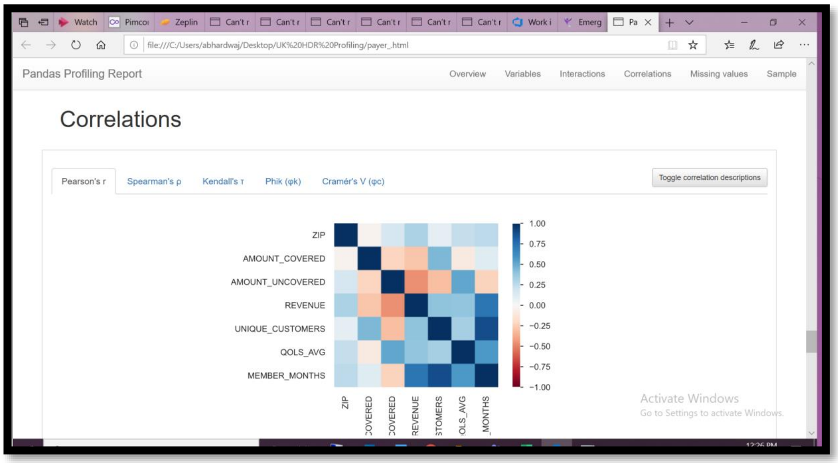
Seaborn: sns.heatmap(df.corr(),annot=True)

pandas_profiling.ProfileReport(df)

Perform complete EDA by using this single line code without having to type multiple lines of code. Icing on the cake is the ability to export this report as an interactive html document

Note: *The codes written next to each circle is just indicative of the effort involved in performing each step of the EDA process and shouldn't be construed as an exhaustive list of commands*

Note: *One line of code compared to multiple lines of code to achieve a common output i.e EDA.*



Links

- ✓ https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/
- ✓ https://pypi.org/project/pandas-profiling/#modal-close
- ✓ https://www.kaggle.com/nulldata/intro-to-pandas-profiling-simple-fast-eda
- ✓ https://github.com/pandas-profiling/pandas-profiling

License	MIT
Version	2.6.0
Last Update	4/14/2020

OS Linux, macOS, Windows

System Requirements JRE 11, Python 3.8

Description Pandas profiling is an open source Python module with which we can quickly do an exploratory data analysis with just a few lines of code. It generates profile reports from a pandas DataFrame. The pandas df.describe() function is great but a little basic for serious exploratory data analysis. pandas profiling extends the pandas DataFrame with df.profile_report() for quick data analysis.

Features

- Type inference: detect the types of columns in a dataframe.
- Essentials: type, unique values, missing values
- Quantile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range
- Descriptive statistics like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- Most frequent values
- Histogram
- Correlations highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
- Missing values matrix, count, heatmap and dendrogram of missing values
- Text analysis learn about categories (Uppercase, Space), scripts (Latin, Cyrillic) and blocks (ASCII) of text data.

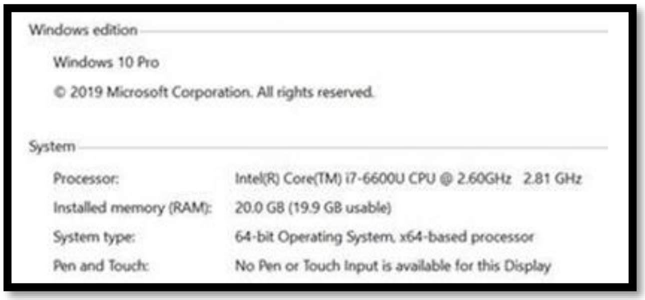
Connectivity / Supported Data Sources & Formats

- Text: - CSV, fixed-width test files, JSON, HTML, Clipboard, Excel
- Binary: OpenDocument, HDF5 Format, Feather Format, Parqeuet Format, ORC Format, Msgpak, Stata, SAS, SPSS, Python Pickle Format
- SQL, Google BigQuery

Limitations

With the increase in the size of the data the time to generate the report also increases a lot. This problem can be solved by generating the report from only a part of the data or by using "minimum mode" introduced in version 2.4 for a simplified report. The tool requires people with technical knowledge.

Performance



Windows edition

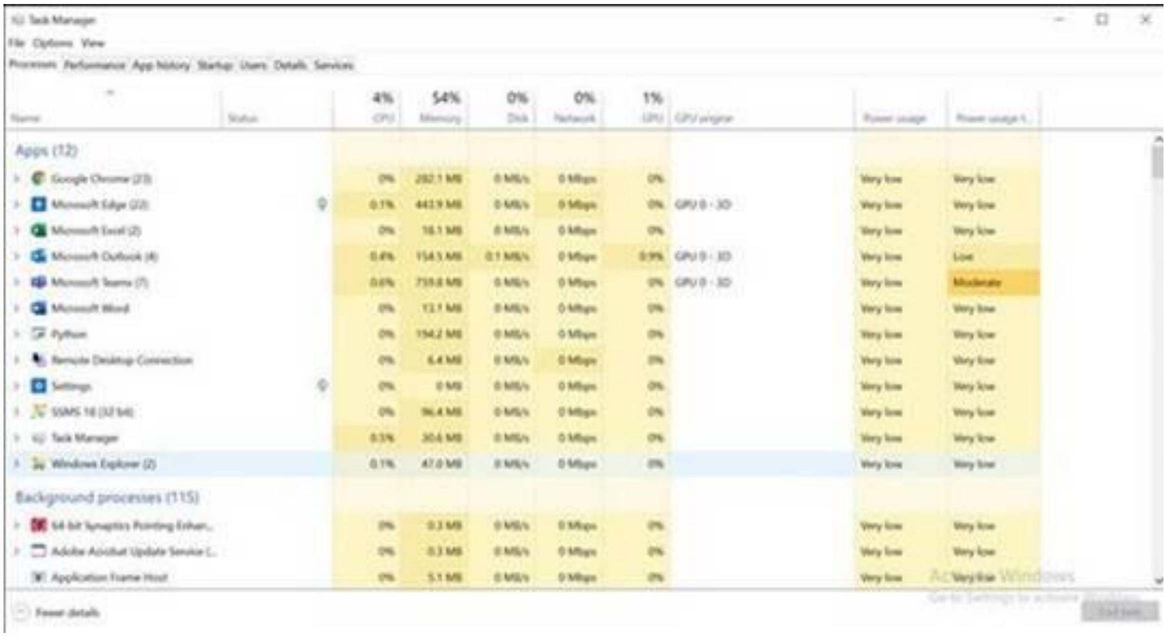
Windows 10 Pro

© 2019 Microsoft Corporation. All rights reserved.

System

Processor:	Intel(R) Core(TM) i7-6600U CPU @ 2.60GHz 2.81 GHz
Installed memory (RAM):	20.0 GB (19.9 GB usable)
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	No Pen or Touch Input is available for this Display

Successfully able to load the 1.3M rows from CSV in < 5 seconds and generation of profiling report took ~1 min



Task Manager
File Options View
Processes Performance App history Startup Users Details Services

		4%	54%	0%	0%	1%			
Name	Status	CPU	Memory	Disk	Network	GPU	GPU engine	Power usage	Power usage t...
Apps (12)									
Google Chrome (23)		0%	282.1 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Microsoft Edge (22)		0.1%	443.9 MB	0 MB/s	0 Mbps	0%	GPU 0 - 3D	Very low	Very low
Microsoft Excel (2)		0%	18.1 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Microsoft Outlook (4)		0.4%	154.5 MB	0.1 MB/s	0 Mbps	0.9%	GPU 0 - 3D	Very low	Low
Microsoft Teams (7)		0.6%	759.8 MB	0 MB/s	0 Mbps	0%	GPU 0 - 3D	Very low	Moderate
Microsoft Word		0%	13.1 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Python		0%	194.2 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Remote Desktop Connection		0%	6.4 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Settings		0%	0 MB	0 MB/s	0 Mbps	0%		Very low	Very low
SSMS 18 (32 bit)		0%	96.4 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Task Manager		0.3%	30.6 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Windows Explorer (2)		0.1%	47.0 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Background processes (115)									
64-bit Synaptics Pointing Enhan...		0%	0.3 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Adobe Acrobat Update Service (...		0%	0.3 MB	0 MB/s	0 Mbps	0%		Very low	Very low
Application Frame Host		0%	5.1 MB	0 MB/s	0 Mbps	0%		Very low	Very low

DATA PROFILING WITH PANDAS PROFILING (PYTHON)

PYTHON:

1. Download Python:
   a. Download the latest version of Python 3 by running Python Installer from https://www.python.org/downloads/.
   b. Download Anaconda from “https://www.anaconda.com/products/individual”.
2. Install pandas-profiling:
   a. the pip package manager by running “pip install pandas-profiling[notebook]”
   b. from Github: “pip install https://github.com/pandas-profiling/pandas-profiling/archive/master.zip”
   c. Using conda: conda install -c conda-forge pandas-profiling
3. The documentation for pandas_profiling can be found at https://pandas-profiling.github.io/pandas-profiling/docs/master/.
4. Run jupyter notebook or preferred Python IDE
5. Import required packages including pandas, pandas-profiling, ProfileReport as

```
import pandas as pd

import pyodbc

import pandas_profiling

import ProfileReport
```

6. Connect to database by:

```
pyodbc.connect('DRIVER={SQL Server}; SERVER=' + DB['servername'] + '; DATABASE=' +
DB['database name'] + '; Trusted_Connection=yes')
```

7. Import data from SQL database table:

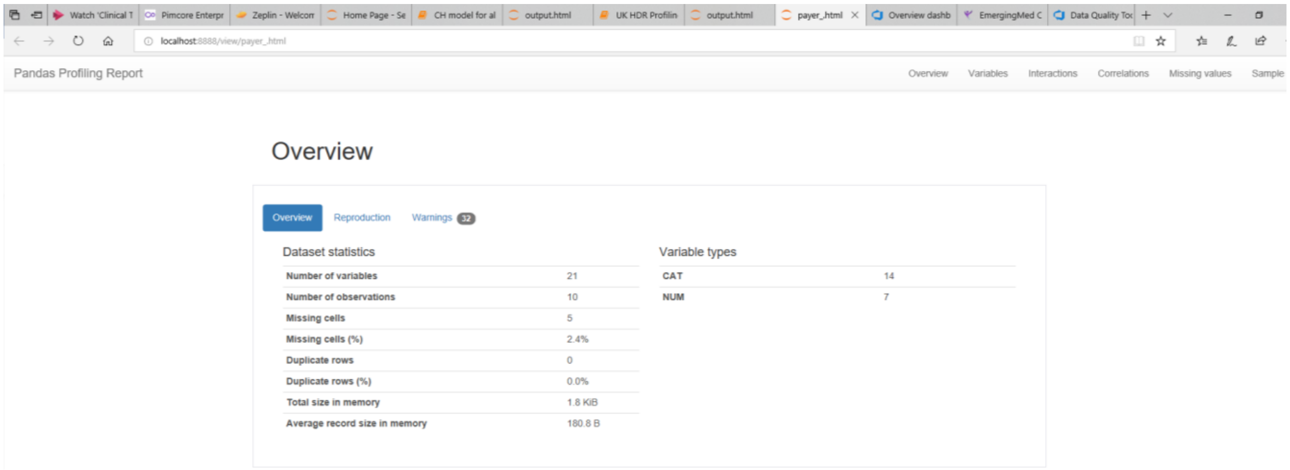
    df = pd.read_sql_query(f"""select * from table_name_from_database """, conn)

8. To generate the report, run: pandas_profiling.ProfileReport(df)

9. Saving the report: report can be save as HTML or json by using to_file() function:

    a. save as HTML file: profile.to_file("your_report.html")
    b. save as JSON file: profile.to_file("your_report.json")

PROFILING REPORT OVERVIEW: PROVIDES BASIC INFORMATION ABOUT DATA.



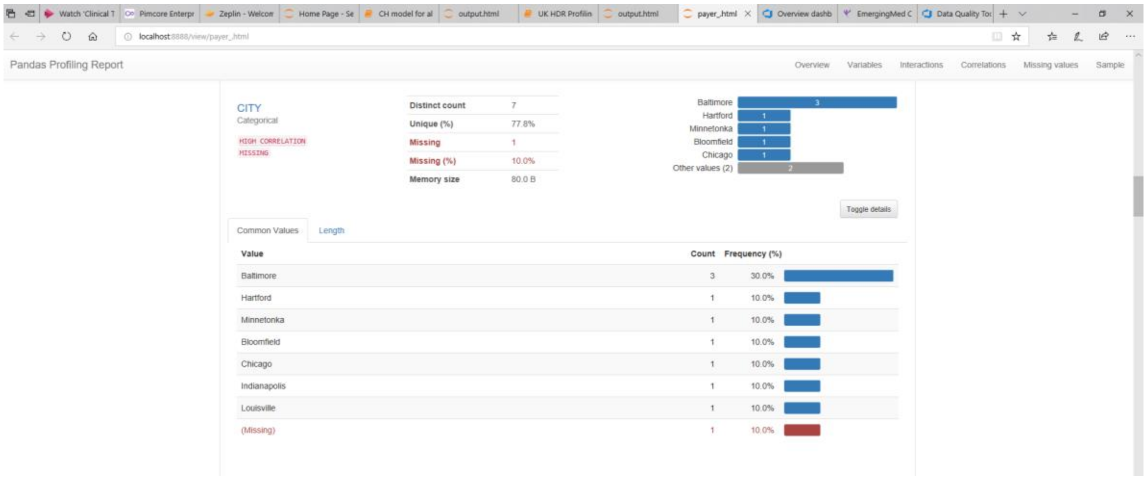
Dataset statistics

Number of variables	21
Number of observations	10
Missing cells	5
Missing cells (%)	2.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.8 KiB
Average record size in memory	180.8 B

Variable types

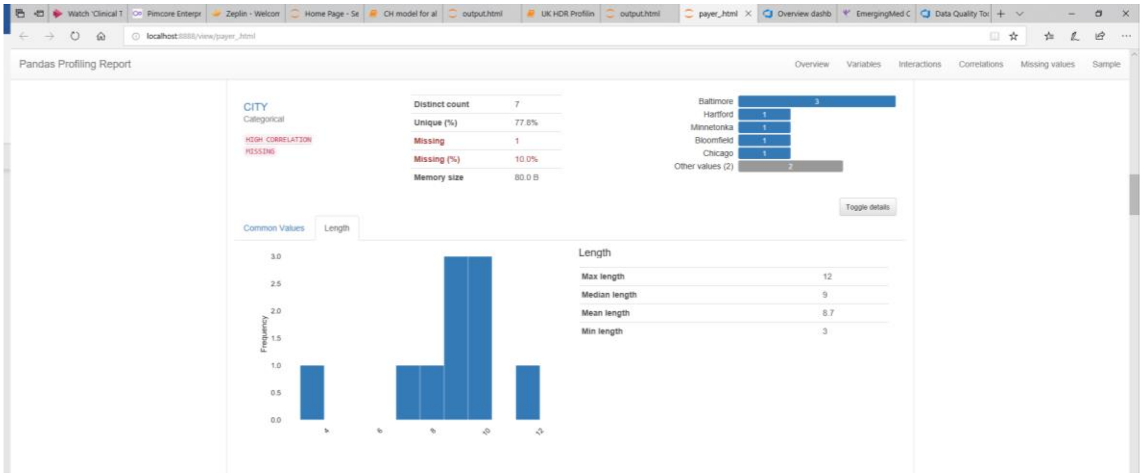
CAT	14
NUM	7

Profiling report categorical variables provides distinct values, unique (%), missing values, missing (%), histogram of count and frequency (%), etc.



Distinct count	7
Unique (%)	77.8%
Missing	1
Missing (%)	10.0%
Memory size	80.0 B

Value	Count	Frequency (%)
Baltimore	3	30.0%
Hartford	1	10.0%
Minnetonka	1	10.0%
Bloomfield	1	10.0%
Chicago	1	10.0%
Indianapolis	1	10.0%
Louisville	1	10.0%
(Missing)	1	10.0%



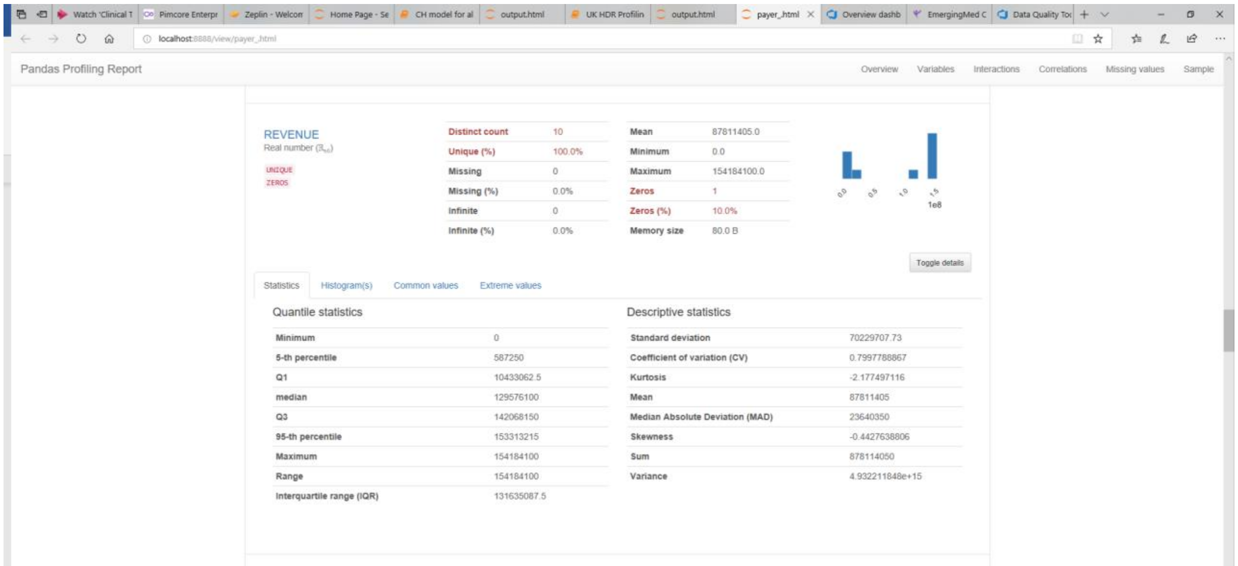
CITY
Categorical
HIGH CORRELATION
MISSING

Distinct count	7
Unique (%)	77.8%
Missing	1
Missing (%)	10.0%
Memory size	80.0 B

Length

Max length	12
Median length	9
Mean length	8.7
Min length	3

Profiling report numerical variables distinct values, unique (%), missing values, missing (%), histogram of count and frequency (%), mean, min, max, zeros, Quantile statistics (5th percentile, Q1, median, Q3, Max, Range, IQR), Quantile statistics (SD, CV, Kurtosis, Mean, MAD, Sum, Skewness, Variance) etc.



REVENUE
Real number ($\mathbb{R}_{\geq 0}$)
UNIQUE
ZEROS

Distinct count	10	Mean	87811405.0
Unique (%)	100.0%	Minimum	0.0
Missing	0	Maximum	154184100.0
Missing (%)	0.0%	Zeros	1
Infinite	0	Zeros (%)	10.0%
Infinite (%)	0.0%	Memory size	80.0 B

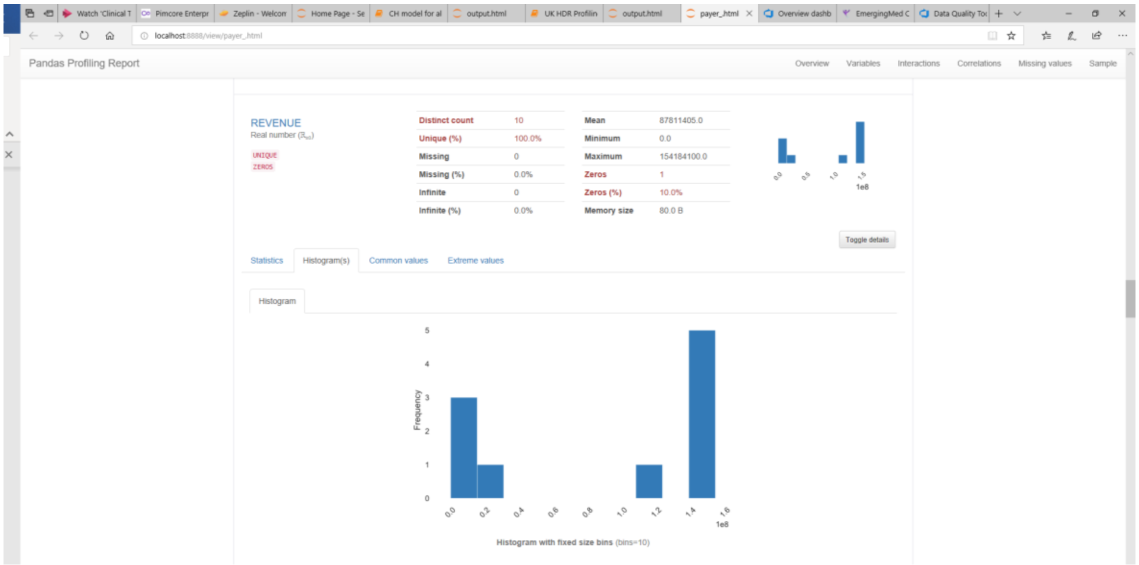
Quantile statistics

Minimum	0
5-th percentile	587250
Q1	10433062.5
median	129576100
Q3	142068150
95-th percentile	153313215
Maximum	154184100
Range	154184100
Interquartile range (IQR)	131635087.5

Descriptive statistics

Standard deviation	70229707.73
Coefficient of variation (CV)	0.7997788867
Kurtosis	-2.177497116
Mean	87811405
Median Absolute Deviation (MAD)	23640350
Skewness	-0.4427638806
Sum	878114050
Variance	4.932211848e+15

FREQUENCY HISTOGRAM:

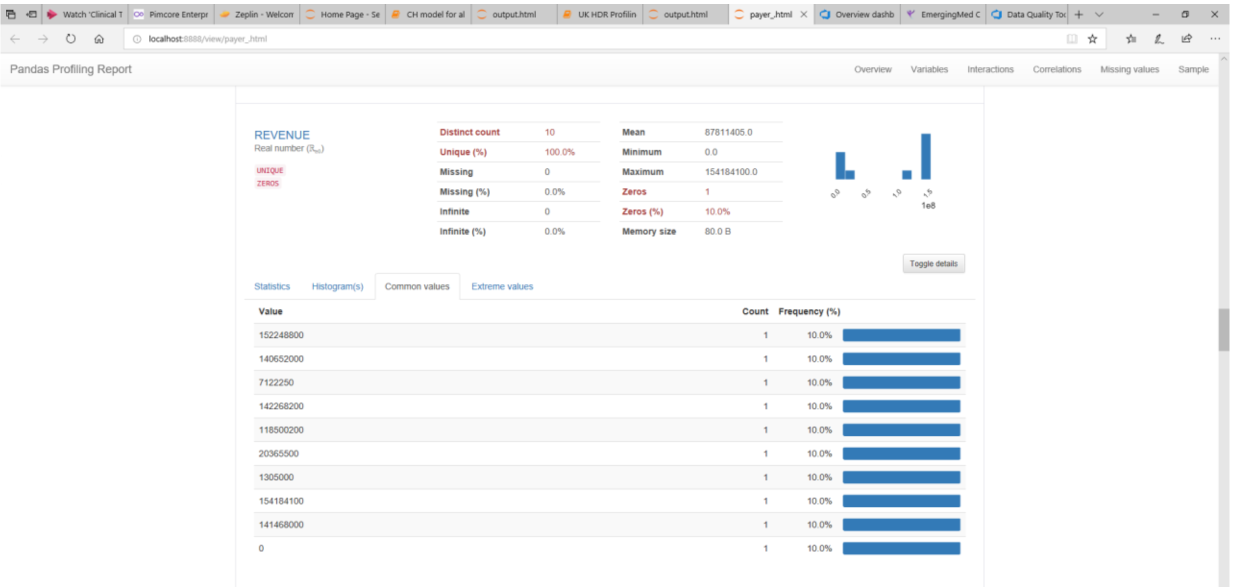


REVENUE
Real number ($\mathbb{R}_{\geq 0}$)
UNIQUE
ZEROS

Distinct count	10	Mean	87811405.0
Unique (%)	100.0%	Minimum	0.0
Missing	0	Maximum	154184100.0
Missing (%)	0.0%	Zeros	1
Infinite	0	Zeros (%)	10.0%
Infinite (%)	0.0%	Memory size	80.0 B

Histogram with fixed size bins (bins=10)

COMMON VALUES COUNT AND FREQUENCY:

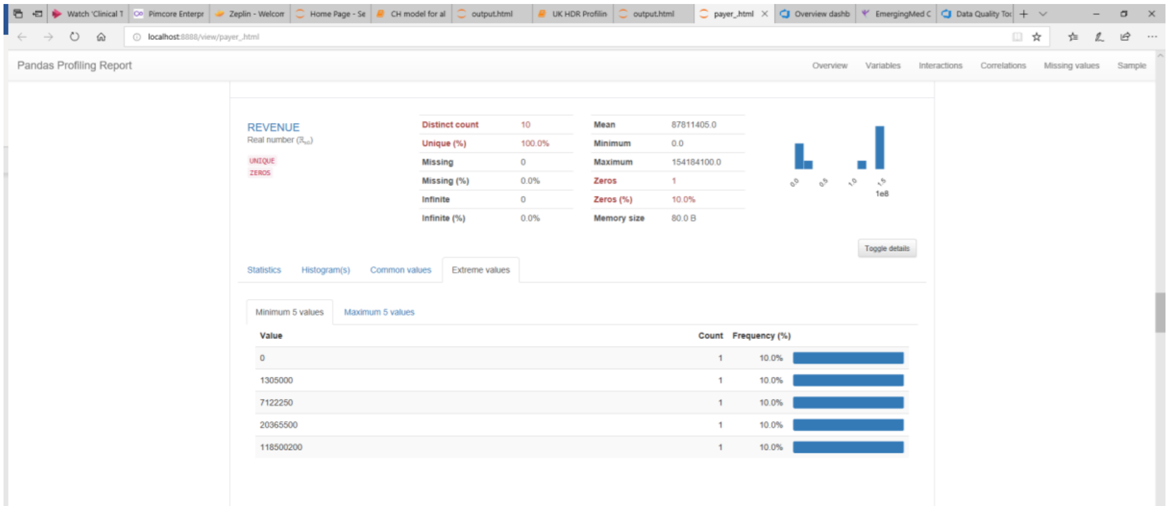


REVENUE
Real number ($\mathbb{R}_{\geq 0}$)
UNIQUE
ZEROS

Distinct count	10	Mean	87811405.0
Unique (%)	100.0%	Minimum	0.0
Missing	0	Maximum	154184100.0
Missing (%)	0.0%	Zeros	1
Infinite	0	Zeros (%)	10.0%
Infinite (%)	0.0%	Memory size	80.0 B

Value	Count	Frequency (%)
152248800	1	10.0%
140652000	1	10.0%
7122250	1	10.0%
142268200	1	10.0%
118500200	1	10.0%
20365500	1	10.0%
1305000	1	10.0%
154184100	1	10.0%
141468000	1	10.0%
0	1	10.0%

EXTREME VALUES: COUNTS AND FREQUENCIES OF TOP 5 MIN AND MAX VALUES

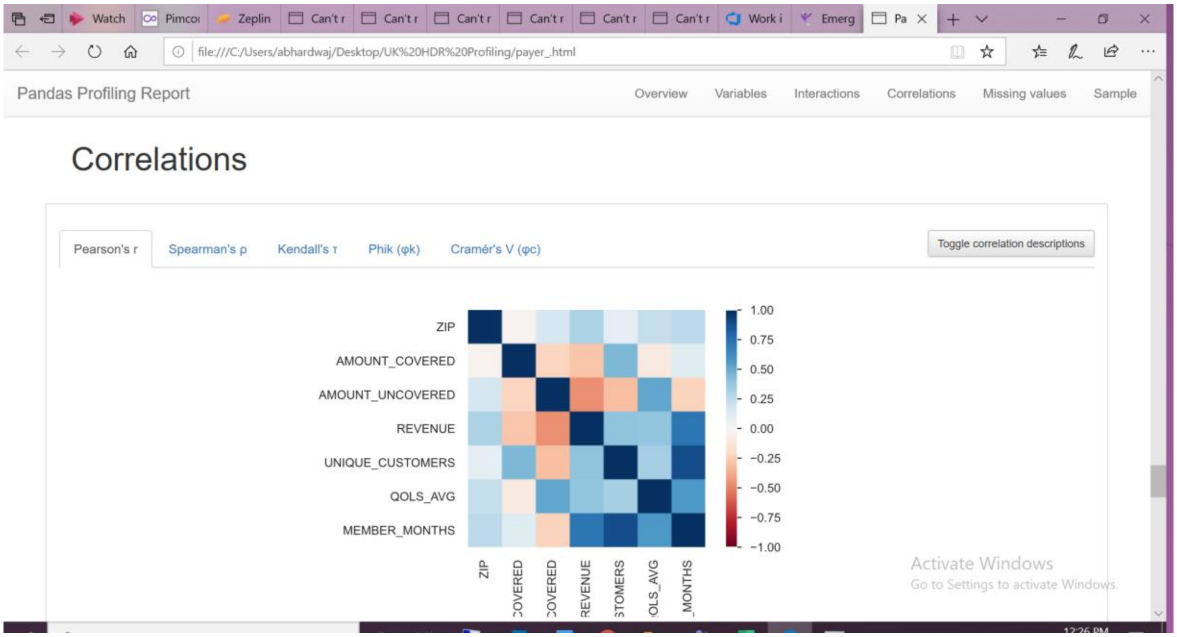


REVENUE
Real number ($\mathbb{R}_{\geq 0}$)
UNIQUE
ZEROS

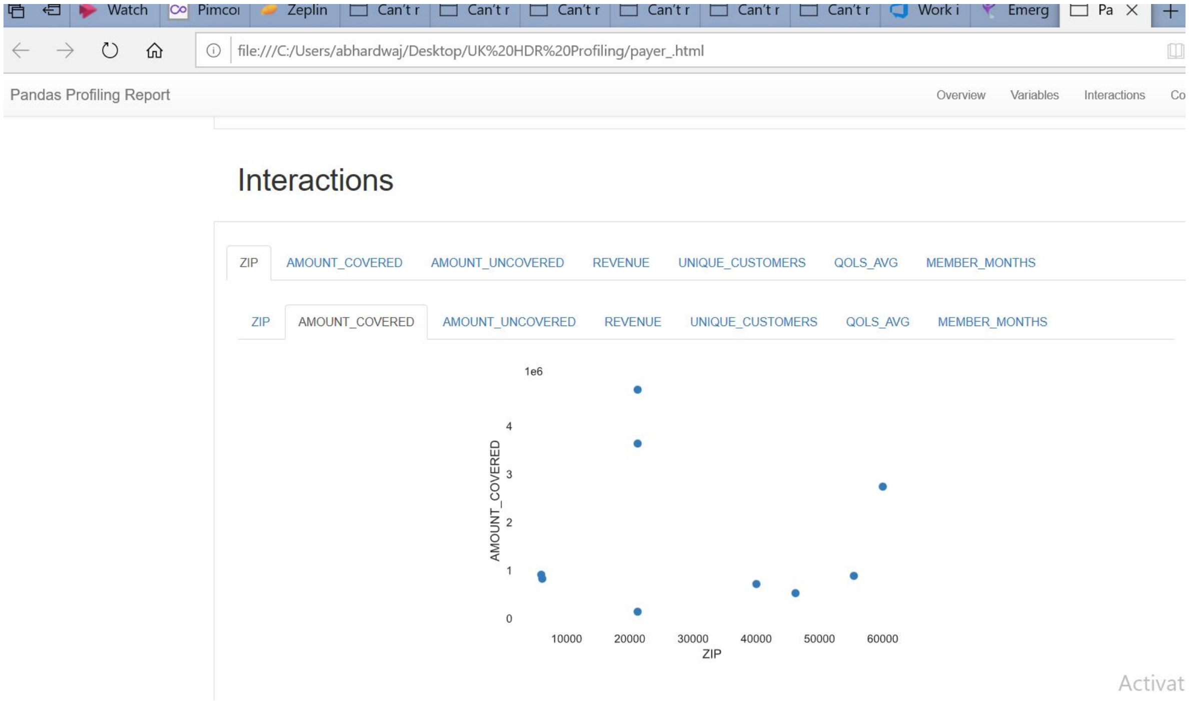
Distinct count	10	Mean	87811405.0
Unique (%)	100.0%	Minimum	0.0
Missing	0	Maximum	154184100.0
Missing (%)	0.0%	Zeros	1
Infinite	0	Zeros (%)	10.0%
Infinite (%)	0.0%	Memory size	80.0 B

Value	Count	Frequency (%)
0	1	10.0%
1305000	1	10.0%
7122250	1	10.0%
20365500	1	10.0%
118500200	1	10.0%

CORRELATIONS:



INTERACTIONS BETWEEN VARIABLES:


Watch
Pimcoi
Zeplin
Can't r
Can't r
Can't r
Can't r
Can't r
Can't r
Work i
Emerg
Pa
file:///C:/Users/abhardwaj/Desktop/UK%20HDR%20Profiling/payer_.html
Pandas Profiling Report
Overview
Variables
Interactions
Co
Interactions
ZIP
AMOUNT_COVERED
AMOUNT_UNCOVERED
REVENUE
UNIQUE_CUSTOMERS
QOLS_AVG
MEMBER_MONTHS
ZIP
AMOUNT_COVERED
AMOUNT_UNCOVERED
REVENUE
UNIQUE_CUSTOMERS
QOLS_AVG
MEMBER_MONTHS
1e6
4
3
2
1
0
AMOUNT_COVERED
10000
20000
30000
40000
50000
60000
ZIP
Activat