In [1]:

```
%load_ext watermark
%watermark
```

```
2019-05-17T16:12:02+02:00

CPython 3.6.5
IPython 6.4.0

compiler   : GCC 7.2.0
system     : Linux
release    : 5.0.13-arch1-1-ARCH
machine    : x86_64
processor  :
CPU cores  : 4
interpreter: 64bit
```

Analisis Exploratorio de Datos - Herramientas adicionales

Aquí incluyo unas herramientas que son bastante útiles a la hora de hacer EDA

Ingesta de datos

In [2]:

```
import pandas as pd

vehiculos = pd.read_csv("../data/vehiculos.1.procesado_inicial.csv")
```

Pandas-profiling

https://github.com/JosPolfliet/pandas-profiling (https://github.com/JosPolfliet/pandas-profiling)

In [3]:

```
!conda install -y pandas-profiling
```

```
Collecting package metadata: done
Solving environment: done

## Package Plan ##

  environment location: /anaconda3

  added / updated specs:
    - pandas-profiling


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    certifi-2019.3.9           |           py37_0         155 KB
    pandas-profiling-1.4.1     |           py37_0          39 KB
    ------------------------------------------------------------
                                           Total:         194 KB

The following NEW packages will be INSTALLED:

  pandas-profiling   pkgs/main/osx-64::pandas-profiling-1.4.1-py37_0

The following packages will be SUPERSEDED by a higher-priority channel:

  ca-certificates    conda-forge::ca-certificates-2019.3.9~ --> pkgs/main::ca-certificates-2019
.1.23-0
  certifi                                       conda-forge --> pkgs/main
  conda                    conda-forge::conda-4.6.12-py37_2 --> pkgs/main::conda-4.6.12-py37_1
  openssl            conda-forge::openssl-1.1.1b-h01d97ff_2 --> pkgs/main::openssl-1.1.1b-h1de3
5cc_1



Downloading and Extracting Packages
certifi-2019.3.9     | 155 KB    | ##################################### | 100%
pandas-profiling-1.4 | 39 KB     | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

In [4]:

```
import pandas_profiling

pandas_profiling.ProfileReport(vehiculos)
```

Out[4]:

Overview

Dataset info

Number of variables	11
Number of observations	38436
Total Missing (%)	0.3%
Total size in memory	3.2 MiB
Average record size in memory	88.0 B

Variables types

Numeric	4
Categorical	6
Boolean	0
Date	0
Text (Unique)	0
Rejected	1
Unsupported	0

Warnings

- cilindros is highly correlated with desplazamiento (ρ = 0.90304) **Rejected**
- fabricante has a high cardinality: 133 distinct values **Warning**

- modelo has a high cardinality: 3791 distinct values **Warning**
- traccion has 1189 / 3.1% missing values **Missing**
- Dataset has 1506 duplicate rows **Warning**

Variables

~~cilindros~~
Highly correlated

This variable is highly correlated with desplazamiento and should be ignored for analysis

Correlation 0.90304

clase
Categorical

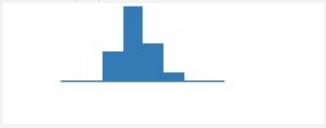
Distinct count	34
Unique (%)	0.1%
Missing (%)	0.0%
Missing (n)	0

Compact Cars	5562
Subcompact Cars	4906
Midsize Cars	4441
Other values (31)	23527

Toggle details

co2
Numeric

Distinct count	597
Unique (%)	1.6%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	472.09
Minimum	0
Maximum	1269.6
Zeros (%)	0.4%



Toggle details

combustible
Categorical

Distinct count	14
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

Regular	25356
Premium	10334
Gasoline or E85	1227
Other values (11)	1519

Toggle details

consumo
Numeric

Distinct count	84
Unique (%)	0.2%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	20.252
Minimum	7
Maximum	136
Zeros (%)	0.0%

Toggle details

desplazamiento
Numeric

Distinct count	67
Unique (%)	0.2%
Missing (%)	0.4%
Missing (n)	140
Infinite (%)	0.0%
Infinite (n)	0
Mean	3.3143
Minimum	0
Maximum	8.4
Zeros (%)	0.0%

Toggle details

fabricante
Categorical

Distinct count	133
Unique (%)	0.3%
Missing (%)	0.0%
Missing (n)	0

Chevrolet	3835	
Ford	3164	
Dodge	2531	
Other values (130)		28906

Toggle details

modelo
Categorical

Distinct count	3791
Unique (%)	9.9%
Missing (%)	0.0%
Missing (n)	0

F150 Pickup 2WD	210
F150 Pickup 4WD	188

Truck 2WD	187	
Other values (3788)		37851

Toggle details

traccion
Categorical

Distinct count	8
Unique (%)	0.0%
Missing (%)	3.1%
Missing (n)	1189

Front-Wheel Drive		13437
Rear-Wheel Drive		13104
4-Wheel or All-Wheel Drive	6648	
Other values (4)	4058	

Toggle details

transmision
Categorical

Distinct count	38
Unique (%)	0.1%
Missing (%)	0.0%
Missing (n)	11

Automatic 4-spd	11043	
Manual 5-spd	8325	
Automatic 3-spd	3151	
Other values (34)		15906

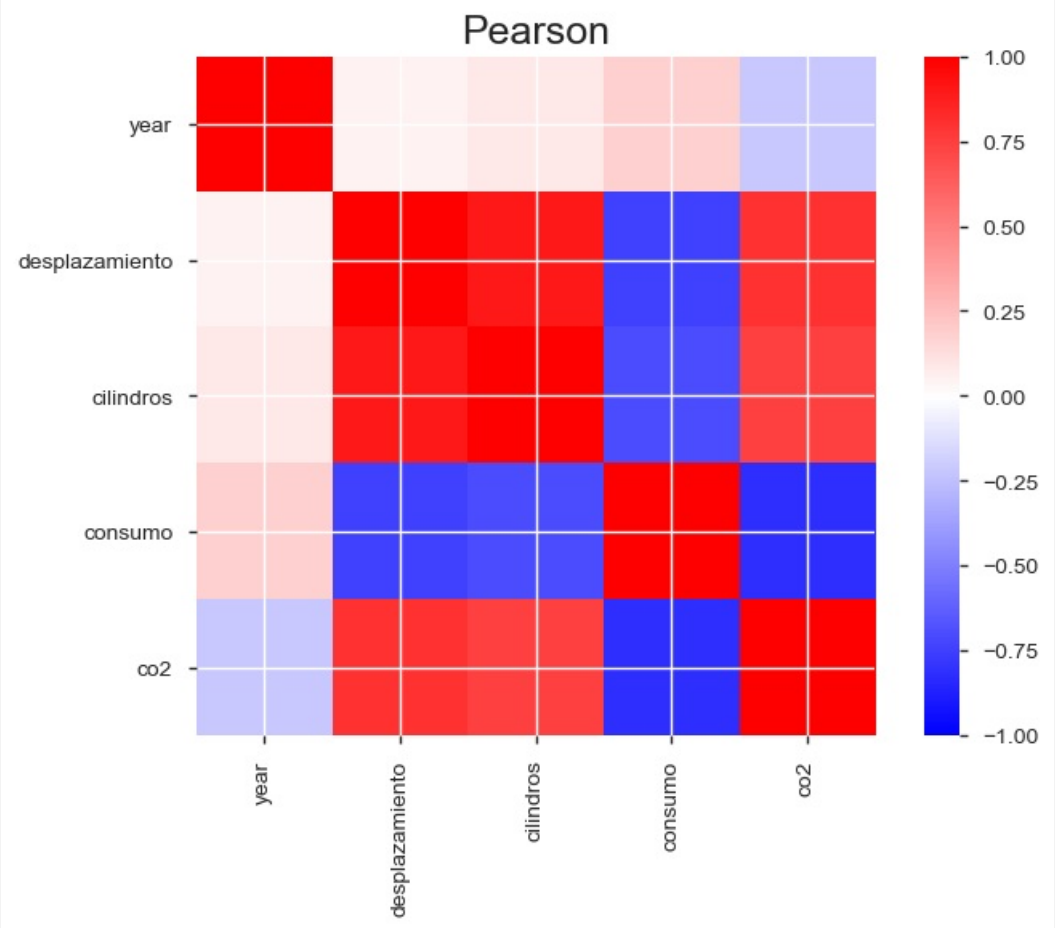
Toggle details

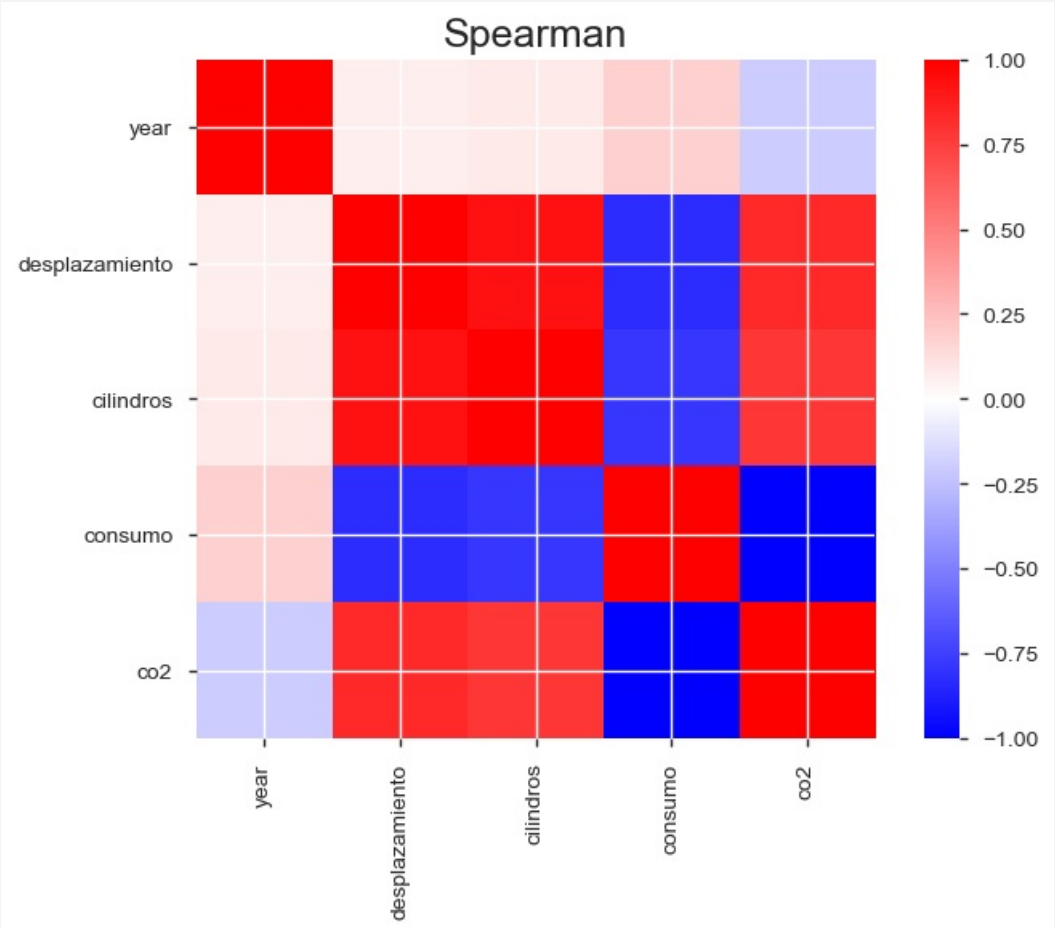
year
Numeric

Distinct count	35
Unique (%)	0.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	2000.3
Minimum	1984
Maximum	2018
Zeros (%)	0.0%

Toggle details

Correlations


Pearson
year
desplazamiento
cilindros
consumo
co2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Spearman
year
desplazamiento
cilindros
consumo
co2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Sample

	fabricante	modelo	year	desplazamiento	cilindros	transmision	traccion	c
0	AM General	DJ Po Vehicle 2WD	1984	2.5	4.0	Automatic 3-spd	2-Wheel Drive	S
1	AM General	DJ Po Vehicle 2WD	1984	2.5	4.0	Automatic 3-spd	2-Wheel Drive	S
2	AM General	FJ8c Post Office	1984	4.2	6.0	Automatic 3-spd	2-Wheel Drive	S
3	AM General	FJ8c Post Office	1984	4.2	6.0	Automatic 3-spd	2-Wheel Drive	S
4	AM General	Post Office DJ5 2WD	1985	2.5	4.0	Automatic 3-spd	Rear-Wheel Drive	S

In [5]:

```
%matplotlib inline
```

Missigno

https://github.com/ResidentMario/missingno (https://github.com/ResidentMario/missingno)

In [7]:

```
!conda install -c conda-forge missingno
```

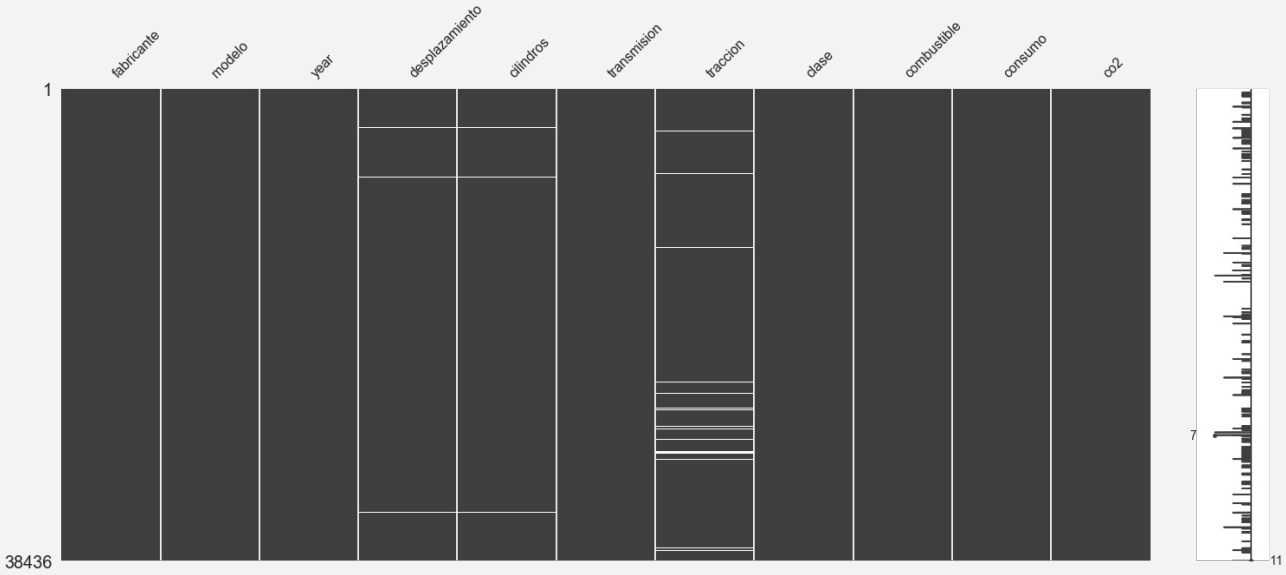
In [8]:

```
import missingno as msno

msno.matrix(vehiculos)
```

Out[8]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1c3e3630>
```



In [9]:

```
msno.heatmap(vehiculos)
```

Out[9]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1d69cb38>
```