A Framework for Exploratory Data Analysis with Python

By Tony Ojeda and Sasan Bahadaran

PORTLAND, OREGON
PYCON 2017

IDENTIFY
REVIEW

CATEGORY AGGREGATIONS

Transmission
Automatic 3--spd
Automatic 4--spd
Manual 5--spd
Automatic (S5)
Manual 6--spd
Automatic 5--spd
Auto(AM3)
Auto(AV--S7)
Automatic (S6)
Automatic (S9)
Manual 4--spd+
32 more

Transmission Type
Automatic
Manual

Vehicle Class
Special Purpose Vehicle 2WD
Midsize Cars
Subcompact Cars
Compact Cars
Sport Utility Vehicle - 4WD
Small Sport Utility Vehicle 2WD
Small Sport Utility Vehicle 4WD
Two Seaters
Small Station Wagons
Minicompact Cars
Minivan - 4WD
+ 23 more

Vehicle Category
Small Cars
Midsize Cars
Large Cars
Station Wagons
Pickup Trucks
Special Purpose
Sport Utility
Vans & Minivans

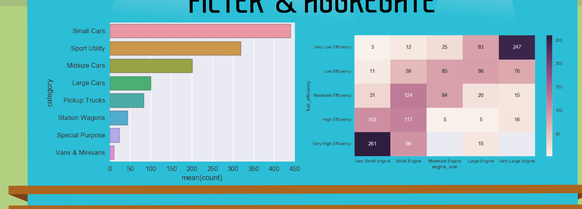
DistrictDataLabs
pycon.districtdatalabs.com

CONTINUOUS BINNING

VERY LOW
LOW
MODERATE
HIGH
VERY HIGH

PREP PHASE

FILTER & AGGREGATE



CLUSTER CATEGORIES

Takes multiple fields into consideration together

Automatically creates new categories (saves time).

Number of clusters? Looking for relatively clear boundaries.

Groups things in ways you may not have thought of.

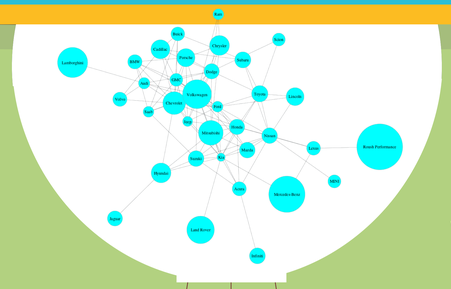
Come up with descriptive names for clusters.

EXPLORE PHASE

matplotlib
pandas
learn

FIELD RELATIONSHIPS

ENTITY RELATIONSHIPS



NOW ENTERING
INSIGHT CITY

Exploratory data analysis (EDA) is an important pillar of data science, a critical step required to complete every project regardless of the domain or the type of data you are working with. Yet, analysts and developers often resort to grasping at straws when it comes to their process for exploring their data and trying to find insights.

Having witnessed the lack of structure in conventional approaches, we decided to document our own processes and come up with a formal framework for data exploration with Python. This poster features both the resulting framework and the Python tools and libraries one can use with it.