# Data Exploration

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

# Things to remember for today

- Teams and collaboration
- (Group) Assignment due this Thursday
- Project Proposal due next Thursday

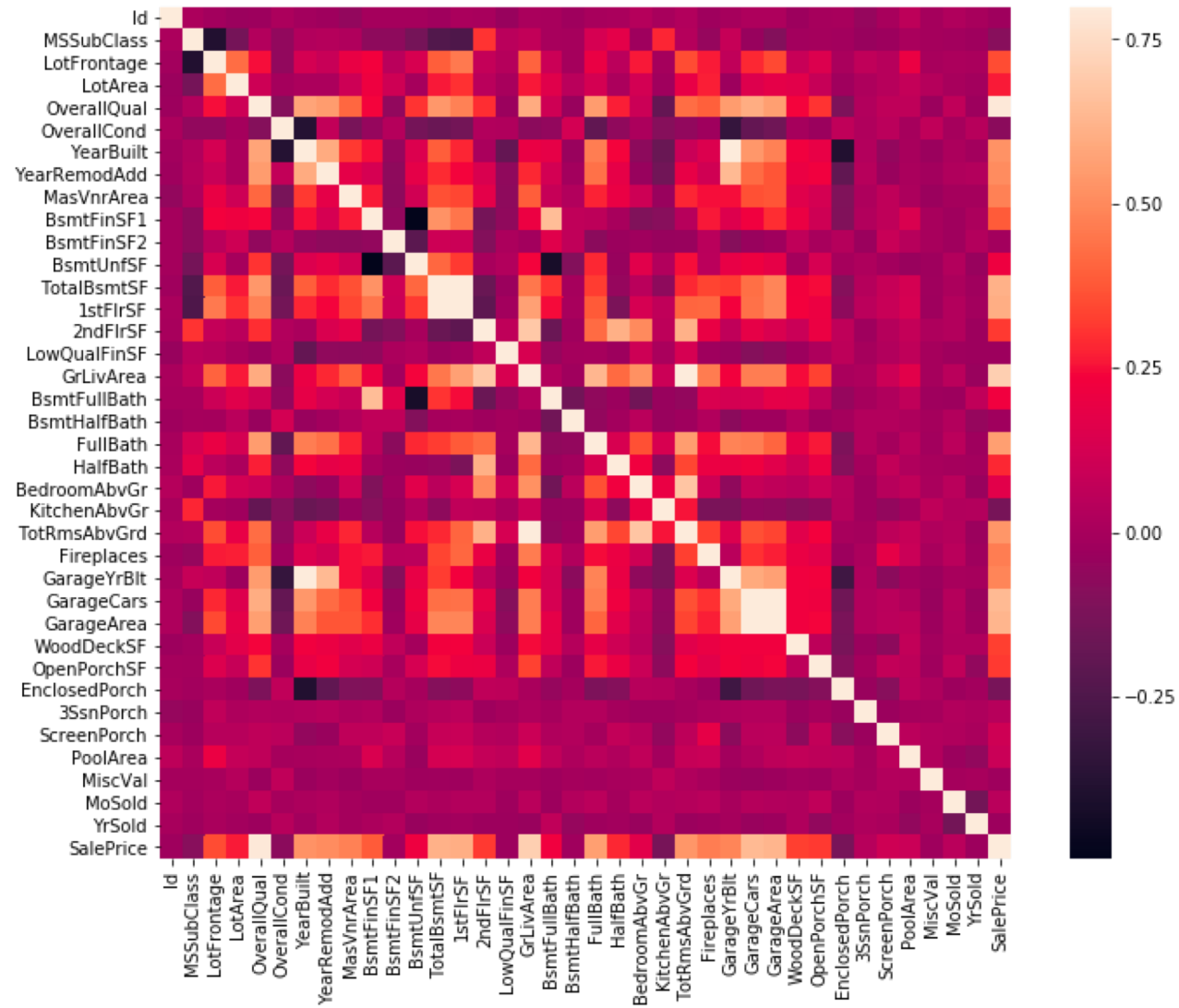
# How does record linkage affect fairness and bias?

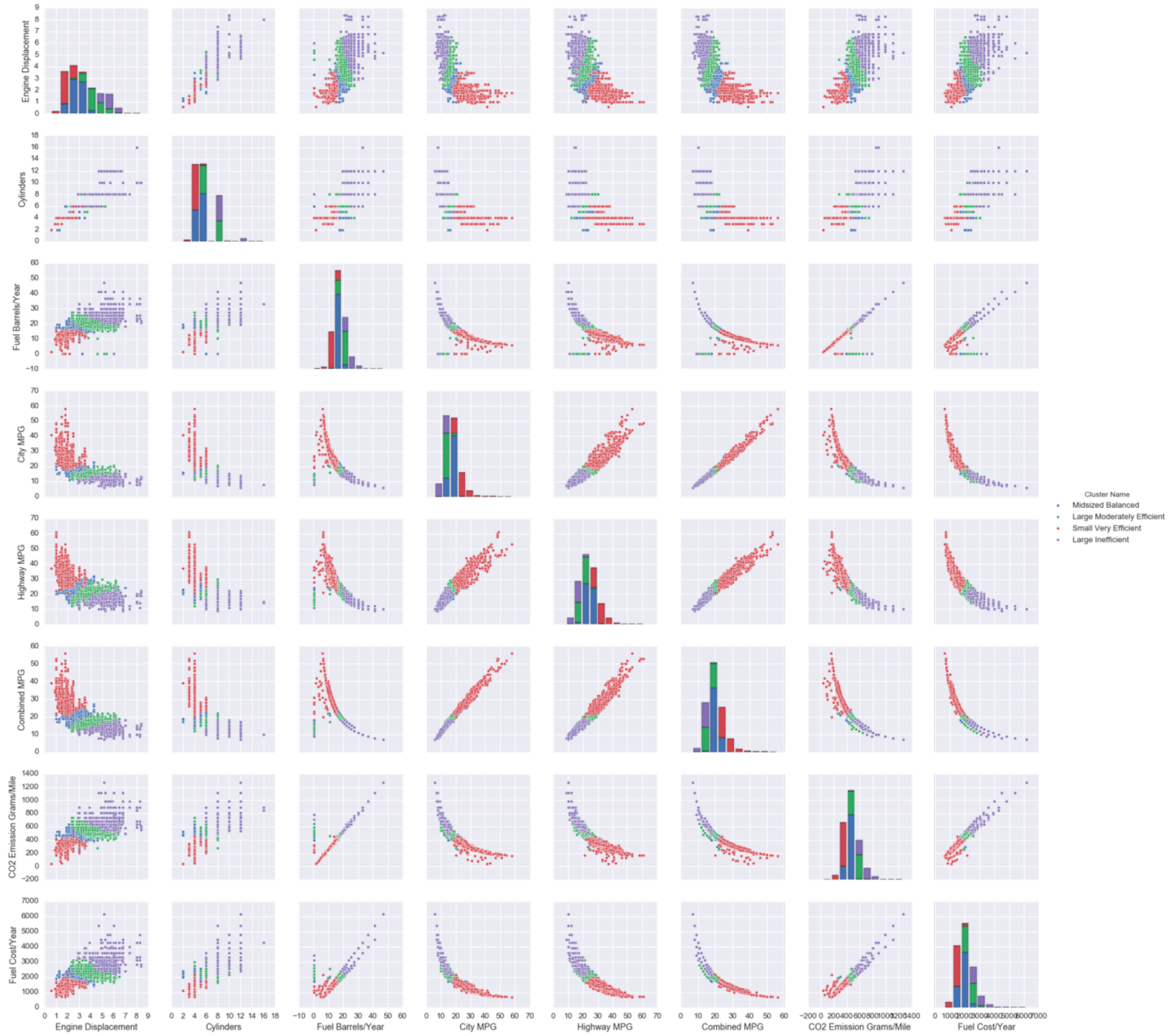
# Why is data exploration important?

1. Sanity Check
2. Understanding the domain
3. Problem Formulation
4. Debugging
5. Feature Generation/Selection
6. Interpretation of results

# Typical data exploration tasks

- Distributions of different variables
- Correlations between variables - correlation matrix and turn it into a heatmap
- Changes and trends over time - how does the data and the entities in the data change over time
- Missing values - are there lots of missing values? is there any pattern there? How/why are they missing?
- Outliers- this can be done using clustering but also by plotting distributions.
- Cross-tabs (if you're looking at multiple classes/labels), describing how the positive and negative classes are different without doing any machine learning.


Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
0.75
0.50
0.25
0.00
−0.25


Engine Displacement
Cylinders
Fuel Barrels/Year
City MPG
Highway MPG
Combined MPG
CO2 Emission Grams/Mile
Fuel Cost/Year
Cluster Name
Midsized Balanced
Large Moderately Efficient
Small Very Efficient
Large Inefficient

# Tools

- SQL (directly and through python – psycopg2)
- Python (matplotlib, seaborn, altair, …)
- Pandas (if you have to)
- Tableau

# Simple task for data exploration

- Write a SQL query that takes a “person id” and gives you everything you know about that person.

- Add a date parameter to it

# Things to remember for today

- Teams and collaboration
- (Group) Assignment due this Thursday
- Project Proposal due next Thursday