Think About Data Visualization

Communicating With Data
Language of Eye

About Me

Praneet Nigam

Trainee Data Scientist at
Spectral Tech

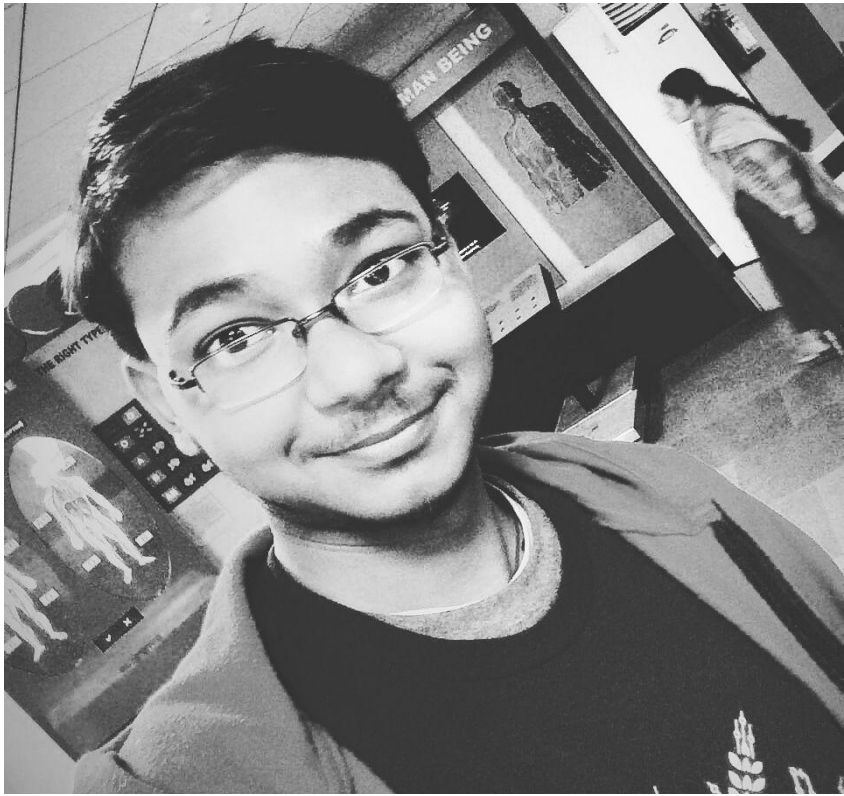
praneet.nigam@spectraltech.ai

linkedin.com/in/praneet460/

@praneetnigam

Praneet460

praneetnigam



Small Exercise : Sales of CNG in India from 2012 to 2017

States : [Gujarat, Delhi, Rajasthan, Maharashtra, Andhra Pradesh / Telangana, Uttar Pradesh, Tripura, Madhya Pradesh, Haryana, West Bengal]

CNG Sales (TMT) - 2012-13: [441.8, 695.1, 0.8, 425.1, 24.7, 137.7, 4.3, 14.5, 73.2, 0.6]

CNG Sales (TMT) - 2013-14 : [463.5, 697.6, 1.6, 476, 24.6, 162.6, 6.8, 15.9, 78.2, 1.154]

CNG Sales (TMT) - 2014-15 : [475.9, 717.1, 2.6, 531.4, 25.8, 184.8, 9.5, 16.6, 72.3, 1.244]

CNG Sales (TMT) - 2015-16 : [503.05, 738.3, 3.68, 565.01, 27.45, 211.61, 11.19, 19.19, 74.59, 1.36]

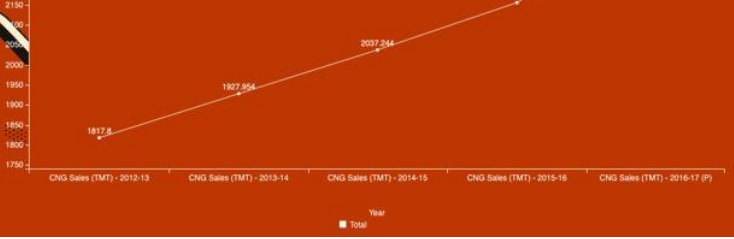
CNG Sales (TMT) - 2016-17 : [546.31, 803.84, 4.27, 592.59, 28.53, 245.42, 12.3, 21.59, 109, 1.64]

Data Source :
https://community.data.gov.in/sale-of-cng-in-india-during-2012-13-to-2016-17/

SALE OF

CNG

IN **INDIA**
DURING 2012-13 TO 2016-17



Sale of CNG in India during 2012-13 to 2016-17

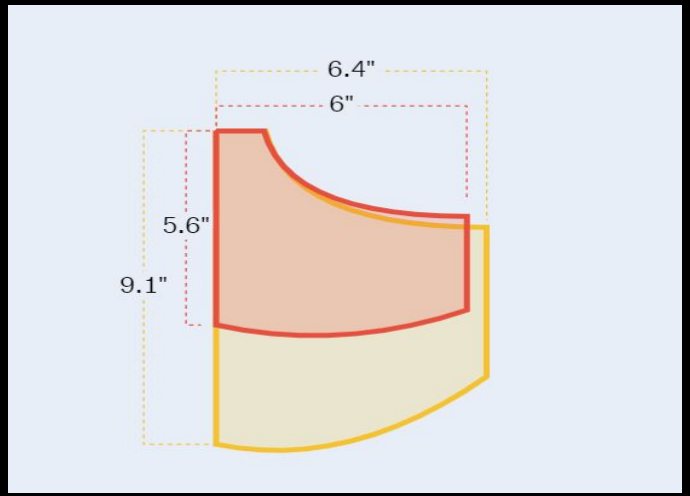
Fact :

Our Brain process visual data 60,000 times faster than the text data, and 90 percent of the information transmitted to the brain is visual.

Source : http://www.t-sciences.com/news/humans-process-visual-data-better

Men’s Vs Women’s Jeans Pockets

On average, the pockets in women’s jeans are 48% shorter and 6.5% narrower than men’s pockets.

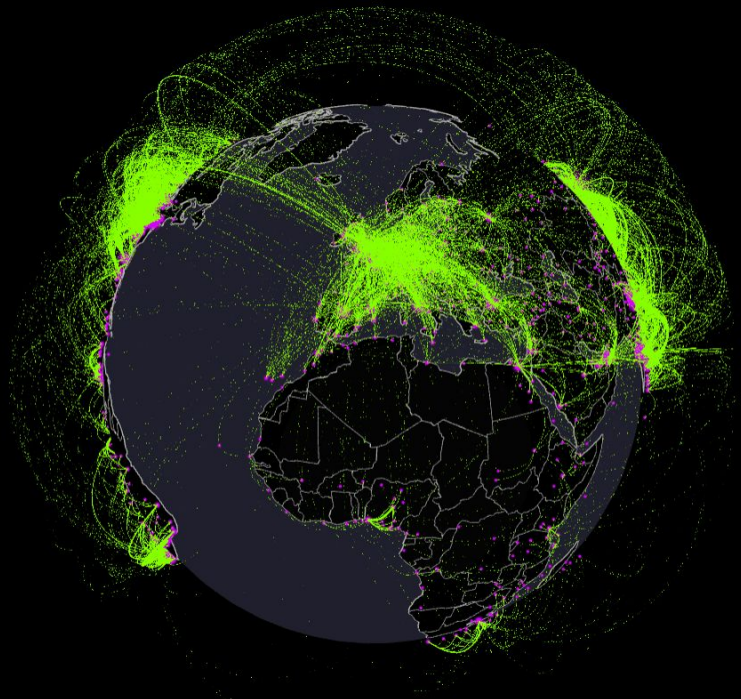


Source : https://pudding.cool/2018/08/pockets/

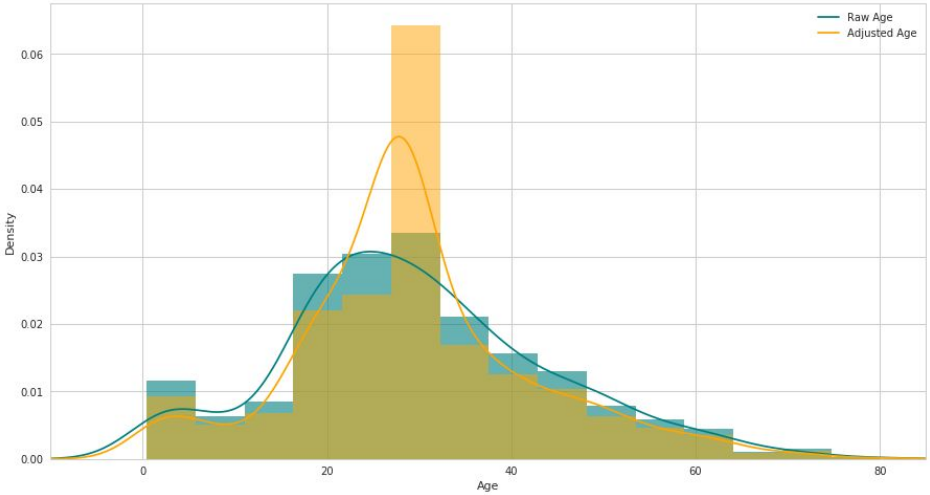
Visualization gives you answer to questions you didn’t know you had.

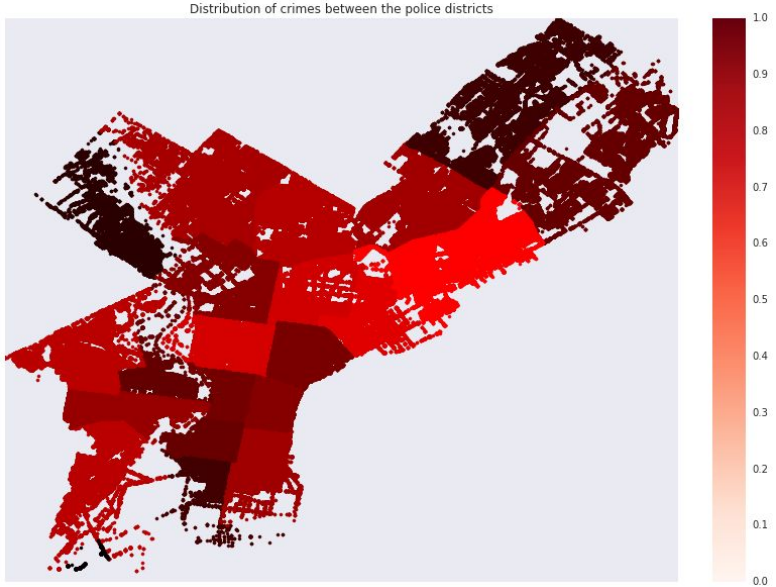
- Ben Schneiderman

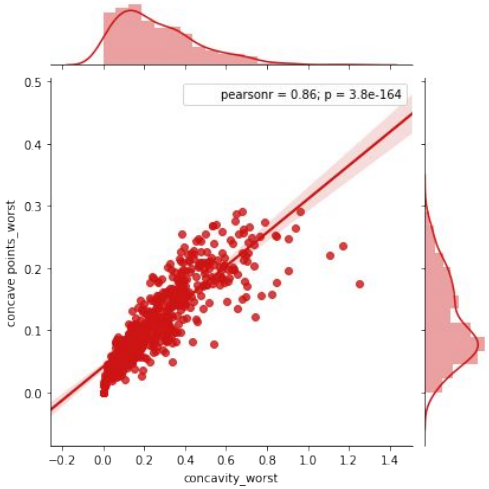
When most people think about DataViz, they think about this.

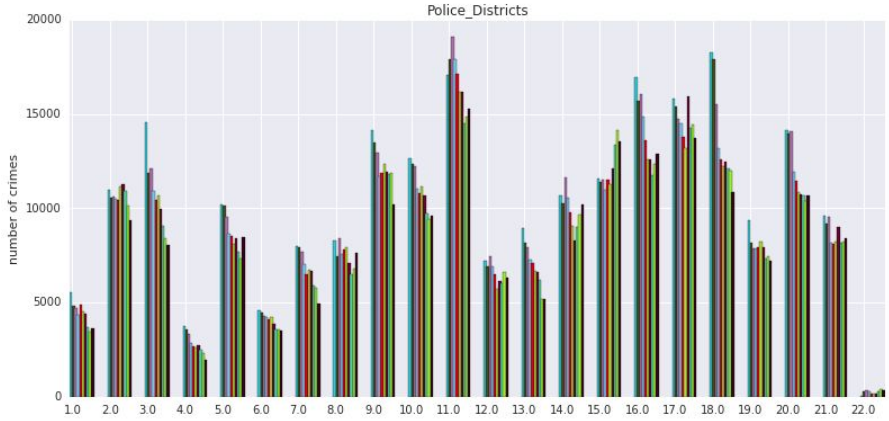


One Year of Air Traffic / One Million Particles


Raw Age
Adjusted Age
Density
Age


Distribution of crimes between the police districts


pearsonr = 0.86; p = 3.8e-164
concave points_worst
concavity_worst


Police_Districts
number of crimes

Data Visualization

- The visual representation of data.
- Primary goal is to communicate information clearly and efficiently to users via the statistical graphics, plots, information graphics, tables, and charts selected.
- Visualizing the data help you find the hidden patterns inside the data.
- The purpose of visualization is insight, not pictures.
- Show data variation, not design variation.

Steps to make your visualization successful

Purpose — Why this visualization?

Content — What to visualize?

Structure — How to visualize it?

Formatting — Everything else like text-font, axis

Basics of DataViz Starts With

- ★ Scatter plot
- ★ Line plot
- ★ Bar plot
- ★ Histogram
- ★ Box plot
- ★ Kernel Density Estimation plot i.e. KDE
- ★ Violin plot

Know more about these plots with examples - ***http://bit.ly/edanalysis***

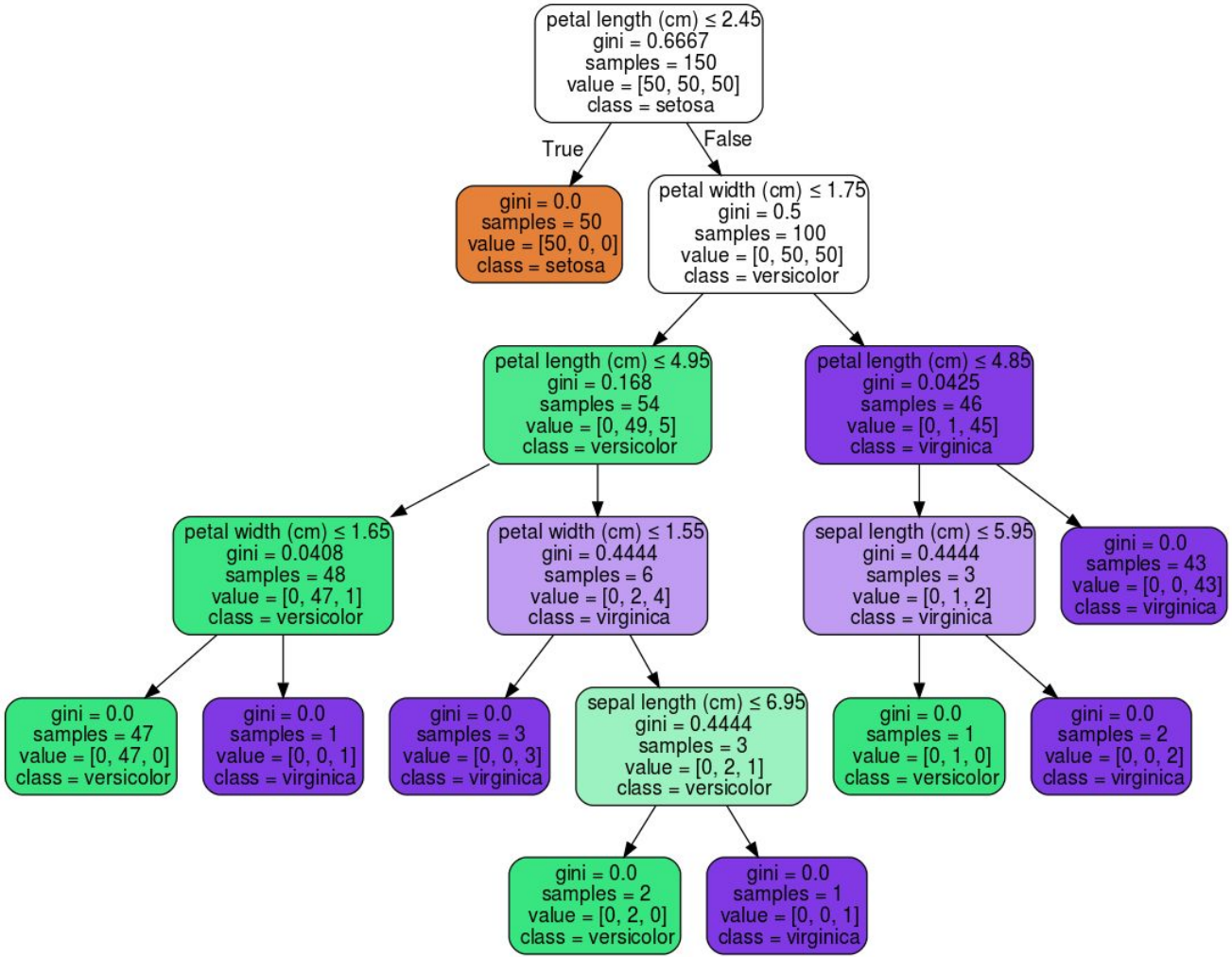
One of the most popular machine learning model for structured data…….

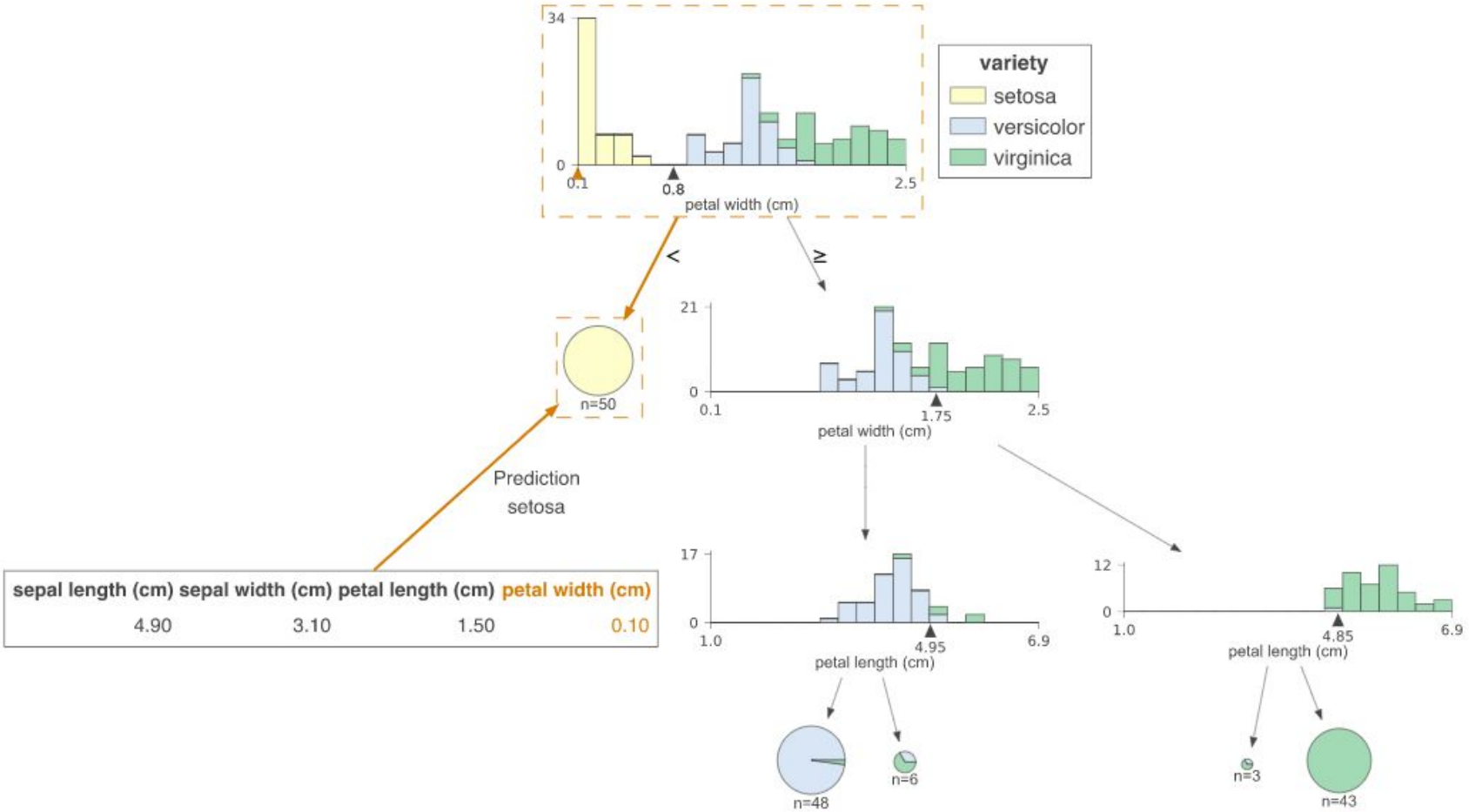
Decision Tree

IRIS Dataset

The Iris dataset is a classic and very easy multi-class classification dataset.

Classes	3
Samples per class	50
Samples total	150
Dimensionality	4
Features	Real, Positive

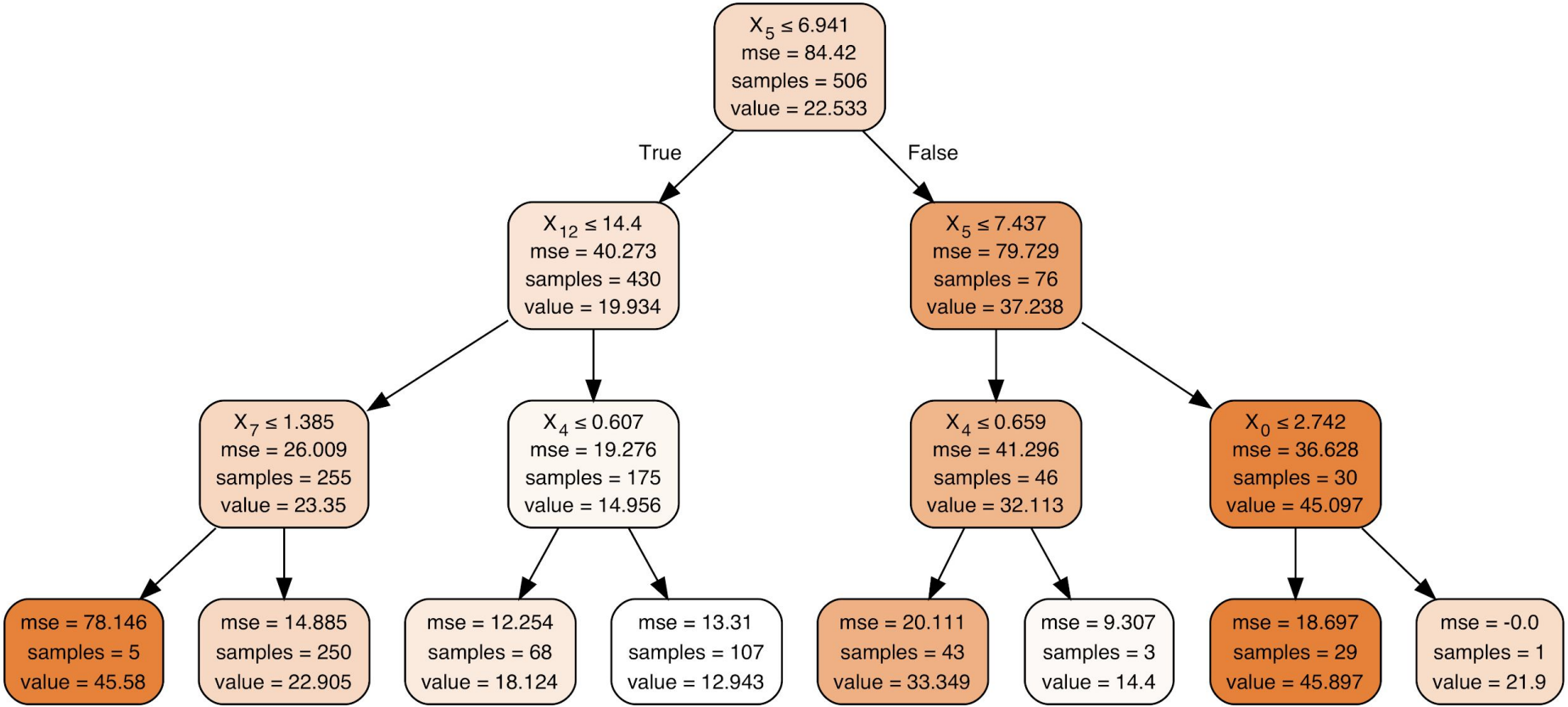

petal length (cm) ≤ 2.45
gini = 0.6667
samples = 150
value = [50, 50, 50]
class = setosa
True
False
gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa
petal width (cm) ≤ 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor
petal length (cm) ≤ 4.95
gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor
petal length (cm) ≤ 4.85
gini = 0.0425
samples = 46
value = [0, 1, 45]
class = virginica
petal width (cm) ≤ 1.65
gini = 0.0408
samples = 48
value = [0, 47, 1]
class = versicolor
petal width (cm) ≤ 1.55
gini = 0.4444
samples = 6
value = [0, 2, 4]
class = virginica
sepal length (cm) ≤ 5.95
gini = 0.4444
samples = 3
value = [0, 1, 2]
class = virginica
gini = 0.0
samples = 43
value = [0, 0, 43]
class = virginica
gini = 0.0
samples = 47
value = [0, 47, 0]
class = versicolor
gini = 0.0
samples = 1
value = [0, 0, 1]
class = virginica
gini = 0.0
samples = 3
value = [0, 0, 3]
class = virginica
sepal length (cm) ≤ 6.95
gini = 0.4444
samples = 3
value = [0, 2, 1]
class = versicolor
gini = 0.0
samples = 1
value = [0, 1, 0]
class = versicolor
gini = 0.0
samples = 2
value = [0, 0, 2]
class = virginica
gini = 0.0
samples = 2
value = [0, 2, 0]
class = versicolor
gini = 0.0
samples = 1
value = [0, 0, 1]
class = virginica

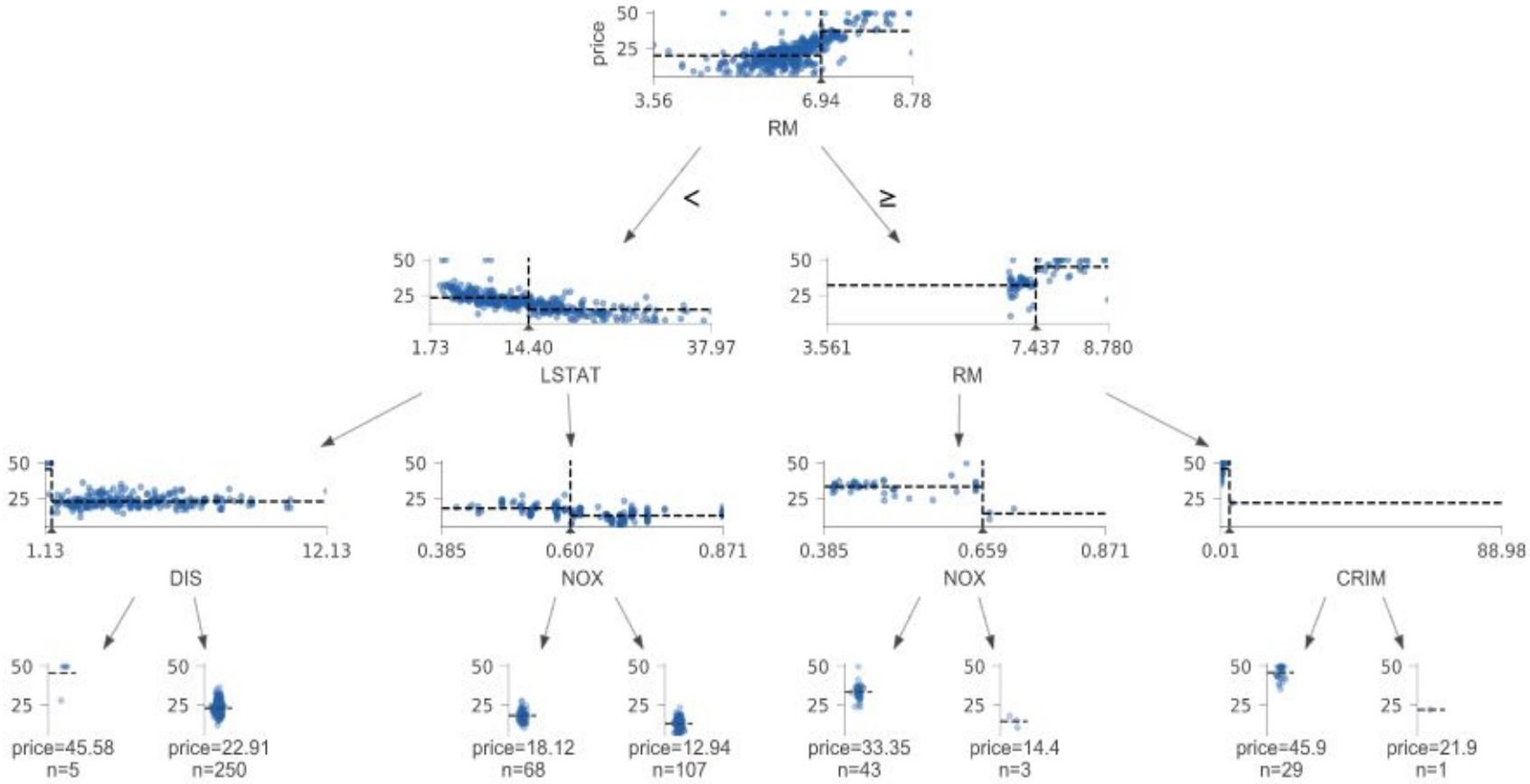

34
0
0.1
0.8
2.5
petal width (cm)
variety
setosa
versicolor
virginica
<
≥
n=50
21
0
0.1
1.75
2.5
petal width (cm)
Prediction
setosa
sepal length (cm) sepal width (cm) petal length (cm) petal width (cm)
4.90 3.10 1.50 0.10
17
0
1.0
4.95
6.9
petal length (cm)
12
0
1.0
4.85
6.9
petal length (cm)
n=48
n=6
n=3
n=43

Code : IRIS Classification

```
from sklearn.datasets import *
from dtreeviz.trees import *

clas = tree.DecisionTreeClassifier(max_depth = 2)
iris = load_iris()

X_train = iris.data
y_train = iris.target

viz = dtreeviz(clas, X_train, y_train, target_name = ‘price’, feature_names =
iris.feature_names, class_names = [‘setosa’, ‘versicolor’, ‘virginica’])

viz
```

Boston Housing


X5 ≤ 6.941
mse = 84.42
samples = 506
value = 22.533
True
False
X12 ≤ 14.4
mse = 40.273
samples = 430
value = 19.934
X5 ≤ 7.437
mse = 79.729
samples = 76
value = 37.238
X7 ≤ 1.385
mse = 26.009
samples = 255
value = 23.35
X4 ≤ 0.607
mse = 19.276
samples = 175
value = 14.956
X4 ≤ 0.659
mse = 41.296
samples = 46
value = 32.113
X0 ≤ 2.742
mse = 36.628
samples = 30
value = 45.097
mse = 78.146
samples = 5
value = 45.58
mse = 14.885
samples = 250
value = 22.905
mse = 12.254
samples = 68
value = 18.124
mse = 13.31
samples = 107
value = 12.943
mse = 20.111
samples = 43
value = 33.349
mse = 9.307
samples = 3
value = 14.4
mse = 18.697
samples = 29
value = 45.897
mse = -0.0
samples = 1
value = 21.9


price
50
25
3.56
6.94
8.78
RM
<
≥
1.73
14.40
37.97
LSTAT
3.561
7.437
8.780
RM
1.13
12.13
DIS
0.385
0.607
0.871
NOX
0.385
0.659
0.871
NOX
0.01
88.98
CRIM
price=45.58
n=5
price=22.91
n=250
price=18.12
n=68
price=12.94
n=107
price=33.35
n=43
price=14.4
n=3
price=45.9
n=29
price=21.9
n=1

Code : Boston Housing

```
from sklearn.datasets import *
from dtreeviz.trees import *

regr = tree.DecisionTreeRegressor(max_depth = 3)
boston = load_boston()

X_train = boston.data
y_train = boston.target

viz = dtreeviz(regr, X_train, y_train, target_name = ‘price’, feature_names =
boston.feature_names)

viz
```