APPLIED BUSINESS ANALYTICS

Introduction

INTRODUCTION

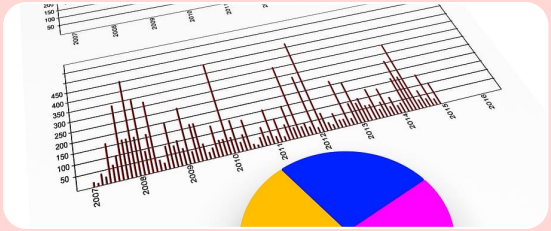
Introduction to Business Analytics

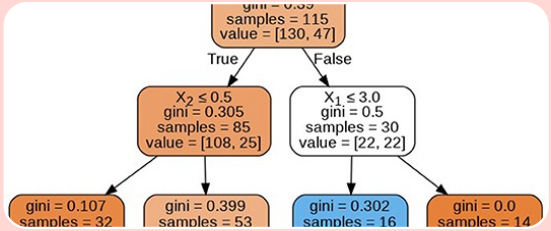
Business Analytics

… refers to the use of methodologies such as data mining, predictive analytics, and statistical analysis in order to analyse and transform data into useful information, identify and anticipate trends and outcomes, and ultimately make smarter, data-driven business decisions.

https://www.omnisci.com/technical-glossary/business-analytics

Business Analytics





Data Visualization and Reporting

Statistical Models

Data Mining Algorithms

Data mining

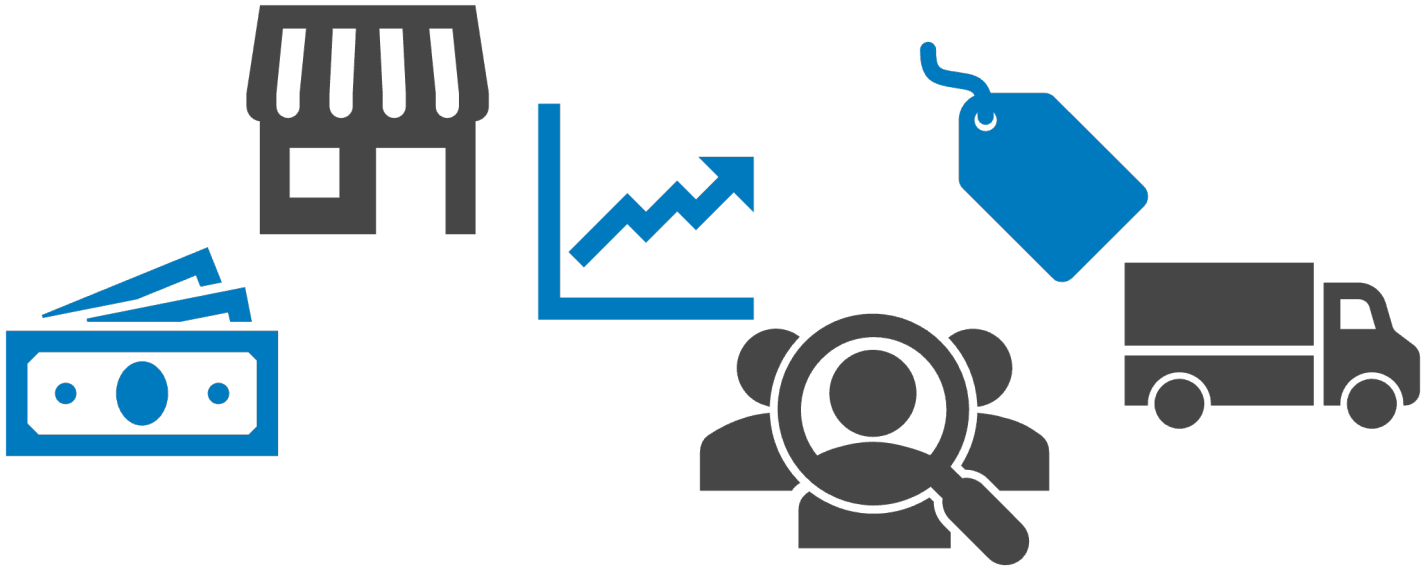
statistical and machine-learning methods that inform decision-making, often in an automated fashion

Big Data

The 4 V's

Volume
Velocity
Variety
Veracity

Applications



Steps involved into a BA Project

Develop an **understanding of the purpose** of the data mining project

Obtain the dataset to be used in the analysis

Explore, **clean and preprocess** the data

Reduce the data dimension, if necessary

Partition the data and choose the **data mining techniques** to be used

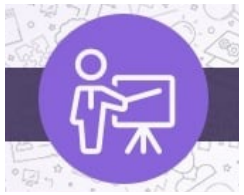
Use **algorithms** to perform the task and interpret the **results**

Deploy the model

APPLIED BUSINESS ANALYTICS

Overview of the Data Mining Process

Core Ideas in Data Mining



SUPERVISED LEARNING

The process of providing an algorithm with records in which an output variable of interest is known and the algorithm learns how to predict this value with new records where the output is unknown.

UNSUPERVISED LEARNING

An analysis in which one attempts to learn patterns in the data other than predicting an output value of interest.

Supervised Learning

- **CLASSIFICATION** — To assign a class label to each instance in a dataset based on its features. The goal of classification is to build a model that accurately predicts the class labels of new instances based on their features
- **PREDICTION** — Similar to classification, except that we are trying to predict the value of a numerical variable rather than a class
- **FORECASTING** — Fitting a model to historical, time-stamped data in order to predict future values

Unsupervised Learning

- **ASSOCIATION RULES** — **Find general associations patterns between items in large databases**
- **SEGMENTATION** — **The process of dividing a dataset into distinct groups or segments based on certain characteristics or features. The goal is to create segments that share common traits or behaviors, making it easier to tailor strategies or interventions for each segment**

DATA ANALYSIS

Loading and Looking at the Data in Python

A **library** is a collection of precompiled codes that can be used later on in a program for some specific well-defined operations.

Pandas is a Python library used for working with data sets.
It has functions for analyzing, cleaning, exploring, and manipulating data.

Dataset

WestRoxbury

The data in WestRoxbury.csv includes information on single family owner-occupied homes in West Roxbury, a neighborhood in southwest Boston, MA, in 2014. The data include values for various predictor variables, and for an outcome—assessed home value ("total value"). This dataset has 14 variables and includes 5802 homes.

TOTAL VALUE	Total assessed value for property, in thousands of USD
TAX	Tax bill amount based on total assessed value multiplied by the tax rate, in USD
LOT SQ FT	Total lot size of parcel (ft^2)
YR BUILT	Year the property was built
GROSS AREA	Gross floor area
LIVING AREA	Total living area for residential properties (ft^2)
FLOORS	Number of floors
ROOMS	Total number of rooms
BEDROOMS	Total number of bedrooms
FULL BATH	Total number of full baths
HALF BATH	Total number of half baths
KITCHEN	Total number of kitchens
FIREPLACE	Total number of fireplaces
REMODEL	When the house was remodeled (recent/old/none)

In [1]:
```
# Import required packages
import pandas as pd
```

Load Data

In [2]:
```
# Load data
housing_df = pd.read_csv(r'C:\Users\pxufre\OneDrive - Nova SBE\Ambiente de Trabalho\2957 - ABA\Datasets Examples\WestRoxbury.
print(housing_df.shape)  # find the dimension of data frame
housing_df.head()  # show the first five rows
```

observations (5802, 14) variables

Out[2]:

	TOTAL VALUE	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS	FULL BATH	HALF BATH	KITCHEN	FIREPLACE	REMODEL
0	344.2	4330	9965	1880	2436	1352	2.0	6	3	1	1	1	0	None
1	412.6	5190	6590	1945	3108	1976	2.0	10	4	2	1	1	0	Recent
2	330.1	4152	7500	1890	2294	1371	2.0	8	4	1	1	1	0	None
3	498.6	6272	13773	1957	5032	2608	1.0	9	5	1	1	1	1	None
4	331.5	4170	5000	1910	2370	1438	2.0	7	3	2	0	1	0	None

Rename columns and showing slices of data

A slice returns an object usually containing a portion of a sequence, such as a subset of rows and columns from a data frame.

In [3]:
```
# Rename columns: replace spaces with '_' to allow dot notation
housing_df = housing_df.rename(columns={'TOTAL VALUE ': 'TOTAL_VALUE'}) # explicitly: one column
housing_df.columns = [s.strip().replace(' ', '_') for s in housing_df.columns] # all columns
```

In [4]:
```
# Practice showing the first four rows of the data
housing_df.loc[0:3]  # loc[a:b] gives rows a to b, inclusive
housing_df.iloc[0:4]  # iloc[a:b] gives rows a to b-1
```

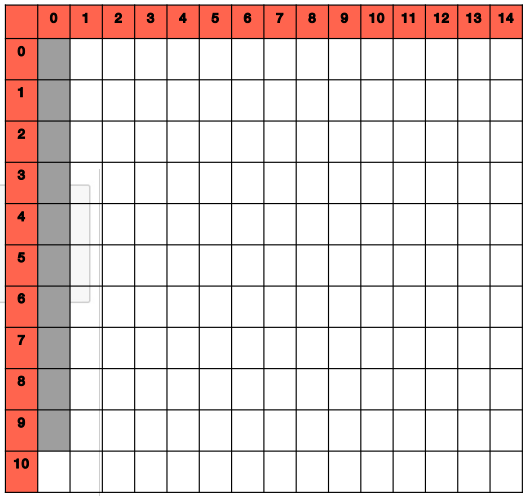
Out[4]:

	TOTAL_VALUE	TAX	LOT_SQFT	YR_BUILT	GROSS_AREA	LIVING_AREA	FLOORS	ROOMS	BEDROOMS	FULL_BATH	HALF_BATH	KITCHEN	FIREP
0	344.2	4330	9965	1880	2436	1352	2.0	6	3	1	1	1	
1	412.6	5190	6590	1945	3108	1976	2.0	10	4	2	1	1	
2	330.1	4152	7500	1890	2294	1371	2.0	8	4	1	1	1	
3	498.6	6272	13773	1957	5032	2608	1.0	9	5	1	1	1	

Showing slices of data

In [5]:
```
# Different ways of showing the first 10 values in column TOTAL_VALUE
housing_df['TOTAL_VALUE'].iloc[0:10]
housing_df.iloc[0:10]['TOTAL_VALUE']
housing_df.iloc[0:10].TOTAL_VALUE  # use dot notation if the column name has no spaces
```

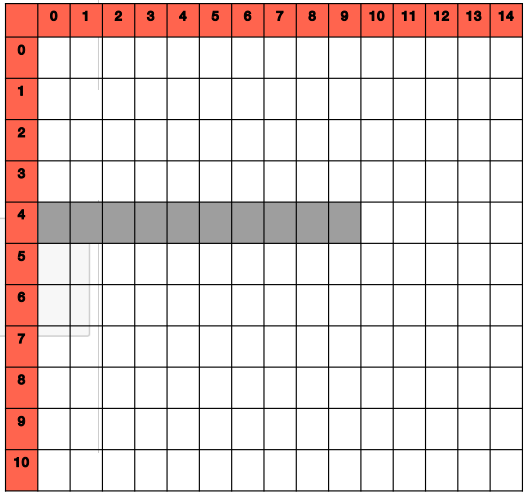
Out[5]:
```
0    344.2
1    412.6
2    330.1
3    498.6
4    331.5
5    337.4
6    359.4
7    320.4
8    333.5
9    409.4
Name: TOTAL_VALUE, dtype: float64
```



In [6]:
```
# Show the fifth row of the first 10 columns
housing_df.iloc[4][0:10]
housing_df.iloc[4, 0:10]
housing_df.iloc[4:5, 0:10]  # use a slice to return a data frame
```

Out[6]:

	TOTAL_VALUE	TAX	LOT_SQFT	YR_BUILT	GROSS_AREA	LIVING_AREA	FLOORS	ROOMS	BEDROOMS	FULL_BATH
4	331.5	4170	5000	1910	2370	1438	2.0	7	3	2

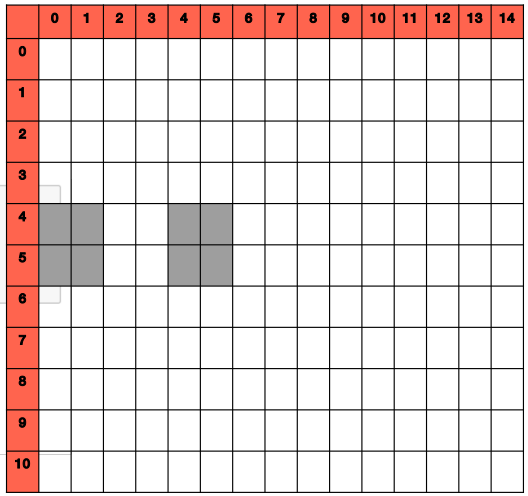


Showing slices of data

In [7]:
```
# Use pd.concat to combine non-consecutive columns into a new data frame.
# The axis argument specifies the dimension along which the
# concatenation happens, 0=rows, 1=columns.
pd.concat([housing_df.iloc[4:6,0:2], housing_df.iloc[4:6,4:6]], axis=1)
```

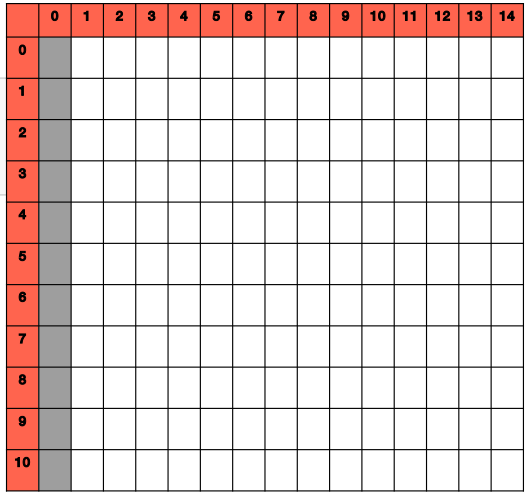
Out[7]:

	TOTAL_VALUE	TAX	GROSS_AREA	LIVING_AREA
4	331.5	4170	2370	1438
5	337.4	4244	2124	1060



In [8]:
```
# To specify a full column, use:
housing_df.iloc[:,0:1]
housing_df.TOTAL_VALUE
housing_df['TOTAL_VALUE'][0:10]  # show the first 10 rows of the first column
```

Out[8]:
```
0    344.2
1    412.6
2    330.1
3    498.6
4    331.5
5    337.4
6    359.4
7    320.4
8    333.5
9    409.4
Name: TOTAL_VALUE, dtype: float64
```



Sampling from a Database

sampling and over/under-sampling

In [10]:
```
# random sample of 5 observations
housing_df.sample(5)

# oversample houses with over 10 rooms
weights = [0.9 if rooms > 10 else 0.01 for rooms in housing_df.ROOMS]
housing_df.sample(5, weights=weights)
```

Out[10]:

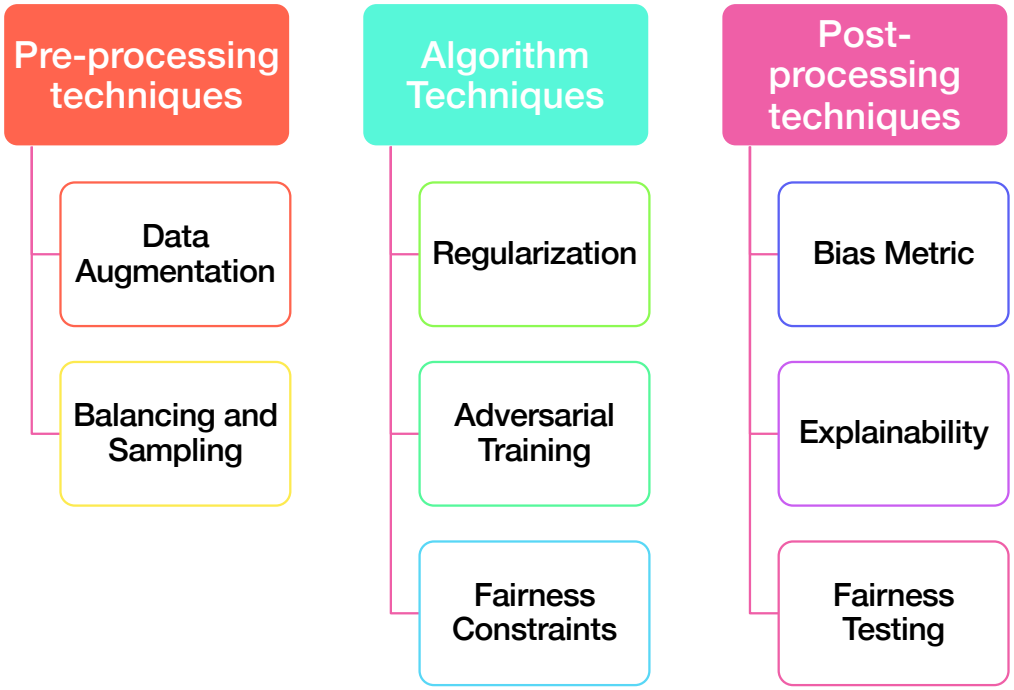
	TOTAL_VALUE	TAX	LOT_SQFT	YR_BUILT	GROSS_AREA	LIVING_AREA	FLOORS	ROOMS	BEDROOMS	FULL_BATH	HALF_BATH	KITCHEN	FI
2118	935.1	11763	25200	1954	6840	5289	1.0	13	9	2	1	2	
4739	666.4	8383	12137	1915	5600	3462	2.5	11	6	2	0	1	
4578	430.6	5416	6894	1965	2771	2187	1.0	11	3	2	0	2	
2455	597.3	7514	6900	1919	5243	2926	2.0	14	6	3	1	1	
5061	384.2	4833	5500	1928	2325	1380	2.0	7	3	1	0	1	

Bias and Discrimination

If data contains human bias

Then the algorithms learn the bias

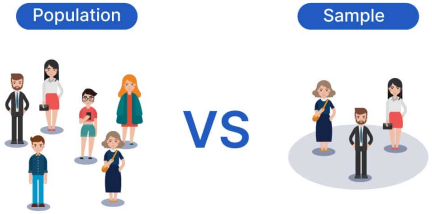
Techniques for Mitigating AI Bias



APPLIED BUSINESS ANALYTICS

Data Analysis

Basic Concepts



A **population** includes all of the entities of interest in a study.

A **sample** is a subset of the population, often randomly chosen and it should be representative of the population as a whole.

	ids	bday	Gender	Rank
1	23643	22NOV1990	0	.
2	30953	23AUG1995	1	1
3	20531	29DEC1994	0	1
4	22416	.	0	1
5	41227	19APR1994	1	2
6	37301	06JUN1993	1	2
7	39181	17MAY1992	0	3
8	22652	04DEC1989	1	3
9	35684	29JUN1991	0	4
10	43344	26MAR1993	0	.

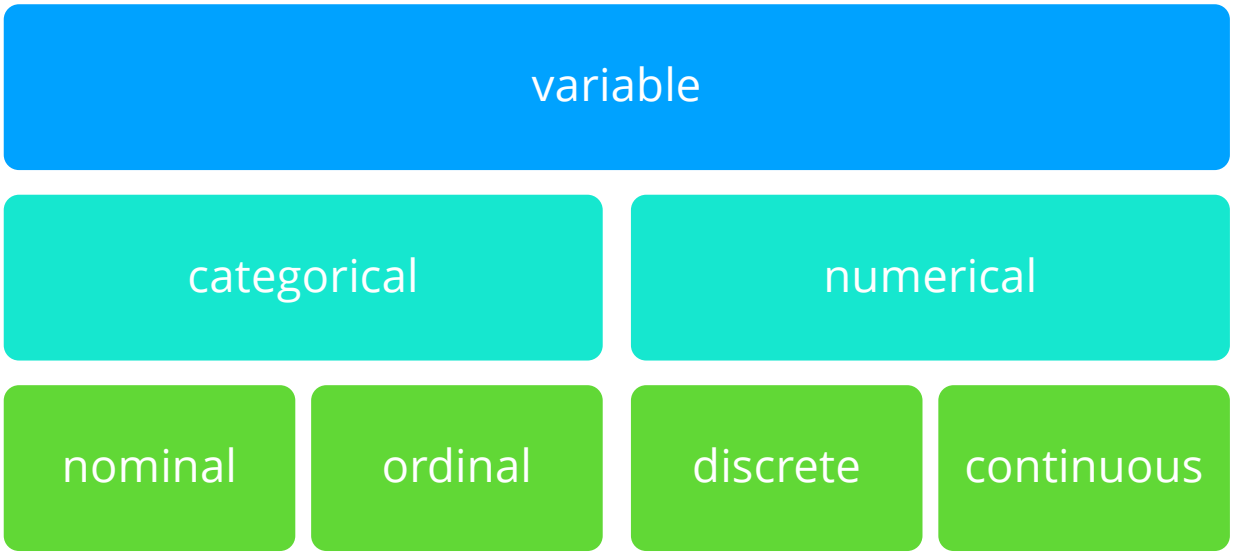
A **data set** is usually a rectangular table of data, with variables in columns and observations in rows.

A **variable** is a characteristic of members of a population. An **observation** is a list of all variable values for a single member of a population.

Cross-sectional data are data on a cross section of a population at a distinct point in time.

Time series data are data collected over time.

Data Types


variable
categorical
numerical
nominal
ordinal
discrete
continuous

Variables' Classification

TOTAL VALUE	Total assessed value for property, in thousands of USD	
TAX	Tax bill amount based on total assessed value multiplied by the tax rate, in USD	**Numerical - continuous**
LOT SQ FT	Total lot size of parcel (ft^2)	
YR BUILT	Year the property was built	**Numerical - discrete**
GROSS AREA	Gross floor area	
LIVING AREA	Total living area for residential properties (ft^2)	
FLOORS	Number of floors	
ROOMS	Total number of rooms	
BEDROOMS	Total number of bedrooms	
FULL BATH	Total number of full baths	
HALF BATH	Total number of half baths	
KITCHEN	Total number of kitchens	
FIREPLACE	Total number of fireplaces	
REMODEL	When the house was remodeled (recent/old/none)	**Categorical - nominal**