Flexible Session #1
Exploratory Data Analysis

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

After this lesson, you should be able to:

- Review Step ❸ Parse the Data and more specifically
  - Descriptive Statistics and Exploratory Data Analysis
  - Apply *pandas* on a Kaggle dataset
- Have fun doing Data Science!

DS

Announcements and Exit Tickets

Review

Review

❸ *Parse the Data*

Introduction to pandas

Codealong | Tidying up (more) the SF housing dataset

	DataFrame	Series
Column subsetting		

by name
 	
(Columns names are stored in df.columns)
(df.columns.get_loc('X1') returns X1’s column index)
	`# New DataFrame with column named X1`
	`df[ ['X1'] ]`
	
	`# 2+ columns (in the order listed)`
	`df[ ['X1', 'X2', …] ]`
		`df['X1']`
		
		`df.X1`

by location	`# New DataFrame with column at location column_i (numbering starts at 0)`
	`df[ [column_i] ]`
	
	`# 2+ columns (in the order listed)`
	`df[ [column_i, column_j, …] ]`	

Row subsetting		

by index label	`df.loc[ [index_label_i] ]`
	`df.loc[ [index_label_i, index_label_j, …] ]`
	
	`# Can use a range if the index is made of numbers (rows “a” to “b” included)`
	`df.loc[ index_label_a : index_label_b ]`
		`df.loc[index_label_i]`

by location	`df.loc[ [row_i] ]`
	`df.loc[ [row_i, row_j, …] ]`
	
	`# (rows “a” to “b’ excluded)`
	`df.iloc[row_a : row_b ] or df[row_a : row_b ]`
		`df.iloc[location_i]`

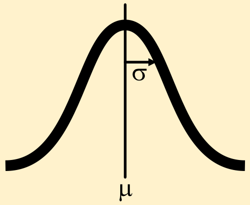
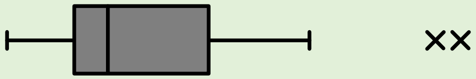
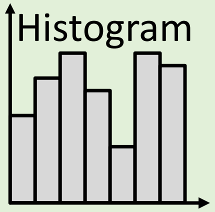
Cell subsetting/scalar lookup		
By index label/column name	`df.at[index_label, 'X1']`	
By location	`df.iat[row_i, column_j]`	

Review

❸ *Parse the Data*

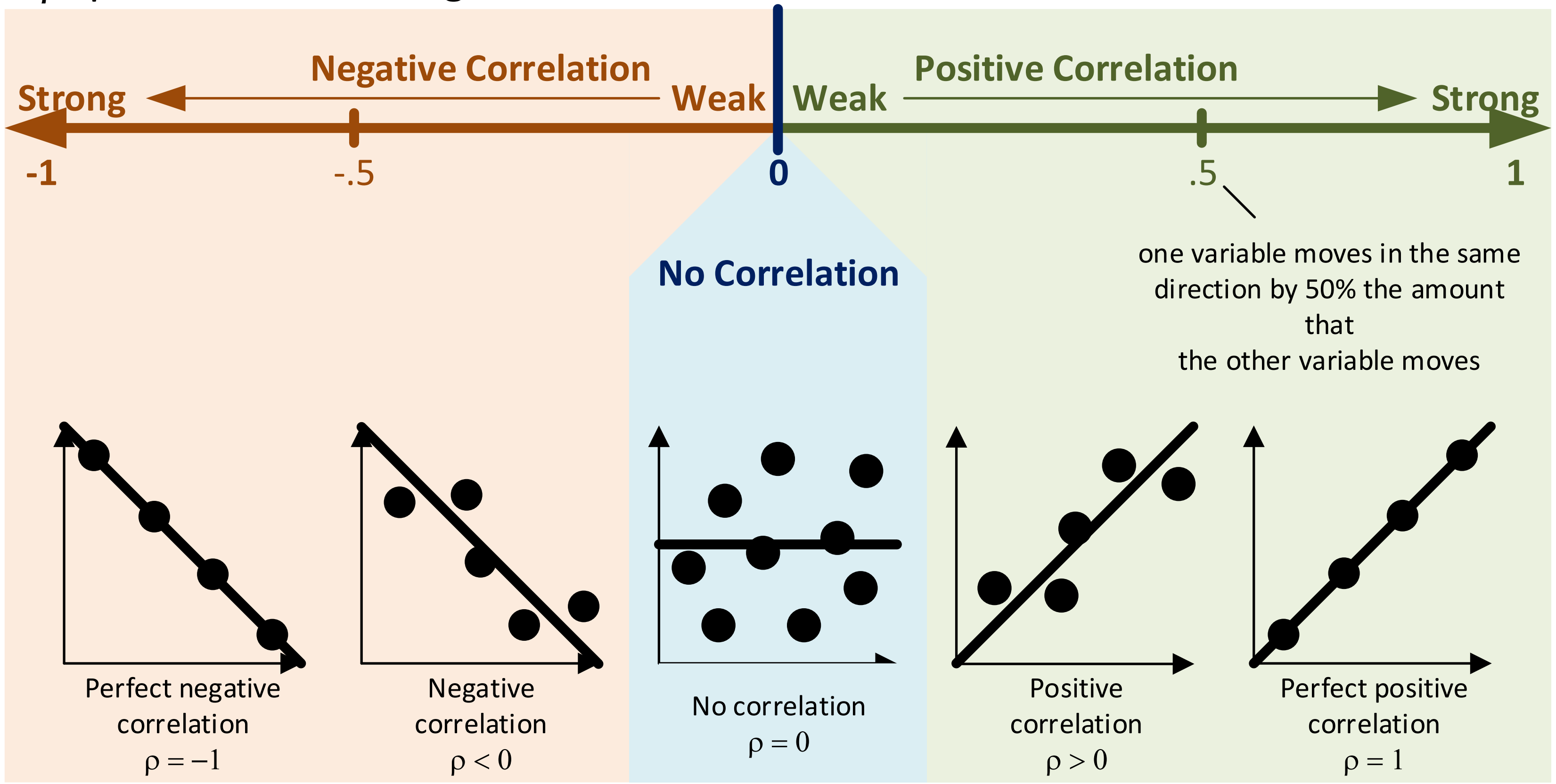
Descriptive Statistics for Exploratory Data Analysis

Descriptive Statistics

Measure of Centrality	Mean	Median	Mode
Measurement Scales	Interval - Ratio	Interval - Ratio	Nominal - Ratio
• In the dataset?	☹	😐	☺
• Easy of compute	☺	😐	☹
• Resistant to outliers?	☹	☺	☺
Measure of Dispersion	☺ (Variance, Standard Deviation)	☺ (Interquartile Range)	☹
Extensive used in mathematical models?	☺	☹	☹
Graphical Methods		Boxplot 	Histogram 

Correlation

ρ quantifies the strength and direction of movements of two random variables



No Correlation

one variable moves in the same direction by 50% the amount that
the other variable moves

Perfect negative correlation
$\rho = -1$

Negative correlation
$\rho < 0$

No correlation
$\rho = 0$

Positive correlation
$\rho > 0$

Perfect positive correlation
$\rho = 1$

Python and *pandas*

Measure of Centrality	.mean()	.median()	.mode()

Measure of Dispersion	.var(), .std()	.min(), .max()
		.quantile()	

Summary	.describe()		
Graphical Methods		.plot(kind = 'box')	.plot(kind = 'hist')
Correlation Matrix	.corr()		
Scatter plot	*DataFrame*.plot(kind = 'scatter', x = '*SerieName*', y = '*SerieName*')		
Scatter matrix	pd.tools.plotting.scatter_matrix(*DataFrame*)		

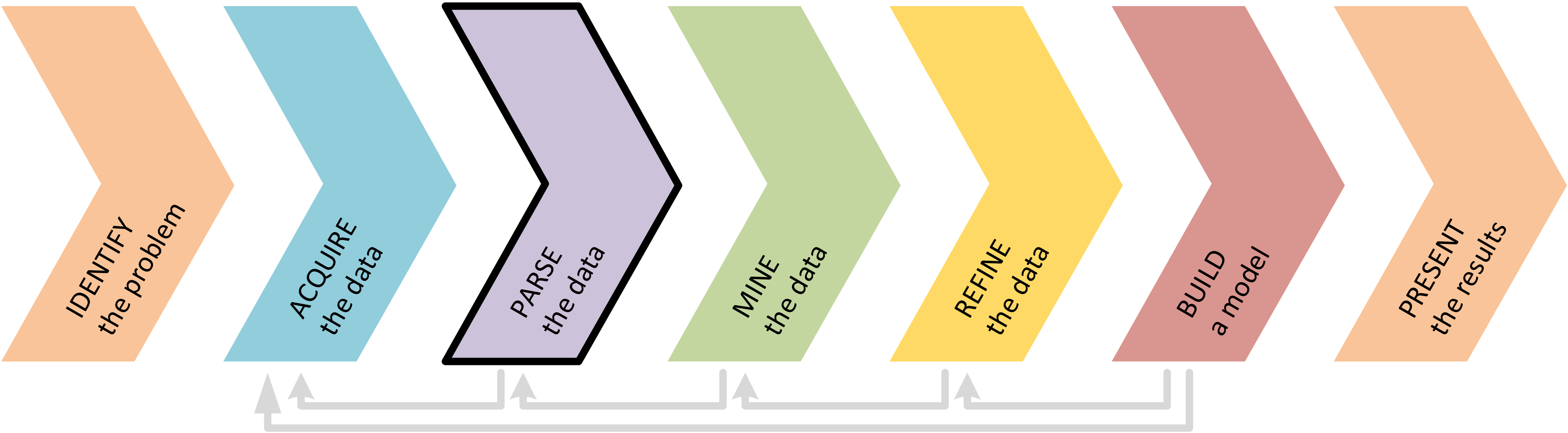
.columns, .set_index(), .drop()	len(), .count(), .sum(), .unique() .value_counts(), .isnull(), .notnul(), .dropna()	np.sort(), .apply()

Q & A

DS

Today

Today we'll keep our focus on PARSE the data



And more precisely on the Exploratory Data Analysis using the *pandas* library

Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Descriptive Statistics for Exploratory Data Analysis Inferential Statistics for Model Fit	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

Here’s what happening today:

- Announcements and Exit Tickets
- Review
- ❸ Parse the Data
  - Kaggle – Exploratory Data Analysis
- Unit Project 2 (due next session on 5/19)

Kaggle

Exploratory Data Analysis

Unit Project 2

Q & A

Before Next Class

Before Next Class

- Projects
  - Unit Project 2 (due next time on 5/19)

Next Class

Inferential Statistics for Model Fit

Learning Objectives

After this next lesson, you should be able to:

- Explain the difference between causation and correlation
- Identify a normal distribution within a dataset using summary statistics and visualization
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (t-tests, p-values, t-values, confidence intervals)

Exit Ticket

Don’t forget to fill out your exit ticket here