Intro to Data Science with Python – Part 1

IBM **Developer**

Saishruthi Swaminathan| Developer Advocate | IBM San Francisco
Gabriela de Queiroz | Sr. Developer Advocate & Manager| IBM San Francisco

What can you take away from this workshop?

- Know the process on how to make data speak right

- Importance and common steps in Data Cleaning and Data Visualization

Photo by Flickr user CyperHades, used under Creative Commons Licensing.

Agenda

1. What is Data Science? Why is it important?
2. The Data Science Process | Pipeline
3. Why Data cleaning and visualization is important?
4. Common Data cleaning steps
5. Common Data Visualization steps
6. Recap
7. Workshop

What is Data Science?

Breaking down data science definition

Goal	• Find solution to the business problem
How?	• Transforming problems to well-posed questions
Using?	• Mathematics, programming and scientific method
Finally?	• Communicate results and its business impact

Data Science Pipeline

Terms

Dataset

Feature, field, variable, attribute or characteristics

	x0	x1	Dealer	Type
0	5	1	AA	Table
1	8	3	AA	Tab.
2	9	1	AA	Table
3	5	7	AA	Chair
4	6	9	AA	Chair
5	10	9	AA	Chair
6	20	40	AA	Dining
7	25	45	AA	Dining

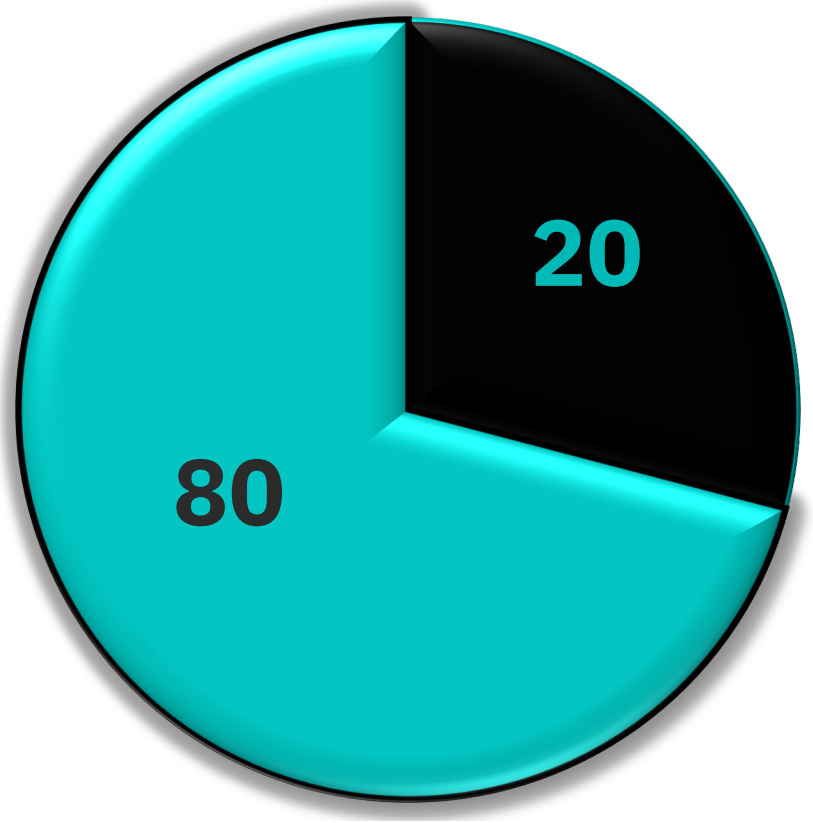
Data point, record, sample, entity or instance

Here,

- 'Type' and 'Dealer' are **categorical variables** as they have finite number of distinct groups.
- 'X0' and 'x1' are **continuous variables**.

Why Data Cleaning and Visualization is important?

The 80/20 Rule



What our data scientist has to say about data cleaning and visualization

Key takeaways:

- Work with domain experts to get more insights about the data.
- Data cleaning prevents you from building a faulty model.
- Visualizations help in understanding the data and convey analyzed information effectively to stakeholders.

“ **No one is going to collect data that can fit well with machine learning models**”

Let’s explore common data cleaning steps

- Import and export dataset.
- Renaming features.
- Changing type of features. (e.g. float -> int).
- Selecting features from the dataset.
- Filter rows and extract only records needed.
- Append / Join tables.
- Fill missing values.
- Summarize data.
- Normalizing and scaling.
- Date formatting.
- String variables to numeric.
- Binning.

Data Cleaning - demo

Problem statement: Predict if the furniture is either table or chair given its length and breadth.

Data from location 'A'

	x0	x1	Dealer	Type
0	5	1	AA	Table
1	8	3	AA	Tab.
2	9	1	AA	Table
3	5	7	AA	Chair
4	6	9	AA	Chair
5	10	9	AA	Chair
6	20	40	AA	Dining
7	25	45	AA	Dining

Data from location 'B'

	x0	x1	Dealer	Type
0	2	2.0	AA	Table
1	4	3.0	AA	Table
2	7	2.0	AA	Table
3	2	9.0	AA	Chair
4	2	8.0	AA	Chair
5	7	8.0	AA	Chair
6	6	NaN	AA	Table

Handling missing values

General steps for handling missing values:

1. See how many missing data points are there in the dataset
2. Find answer for the below question:

   Is this value missing because it wasn't recorded or because it doesn't exist?

3. Think how to handle the missing values:
   * Removing data points having missing values (not recommended as you may lose important information).
   * Try filling values.
     * Fill with 0.
     * Imputation techniques.
       1. Continuous variable: Mean, Median, Mode, Linear Regression and Mixed Imputation.
       2. Categorical variables: Use 'NA' as a separate level, Logistic Regression, K-Nearest Neighbor, etc.

Label and one-hot encoding

	x0	x1	Dealer	Type
0	5	1	AA	Table
1	8	3	AA	Tab.
2	9	1	AA	Table
3	5	7	AA	Chair
4	6	9	AA	Chair
5	10	9	AA	Chair
6	20	40	AA	Dining
7	25	45	AA	Dining

One-hot Encoding

Length	Breadth	Table	Chair	Dining
5	1	1	0	0
5	7	0	1	0
20	40	0	0	1

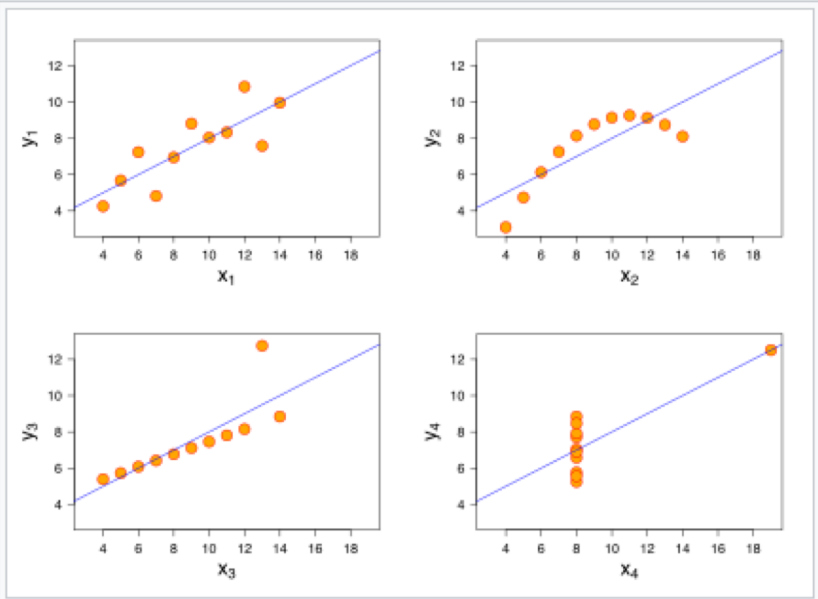
Label Encoding

Length	Breadth	Type
5	1	0
5	7	1
20	40	2

Stepping into the world of data visualization!!

Anscombe's quartet

Property	Value
Mean of *x*	9
Sample variance of *x*	11
Mean of *y*	7.50
Sample variance of *y*	4.125
Correlation between *x* and *y*	0.816
Linear regression line	*y* = 3.00 + 0.500*x*
Coefficient of determination of the linear regression	0.67



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

How Data Visualization is used in SF Gov?

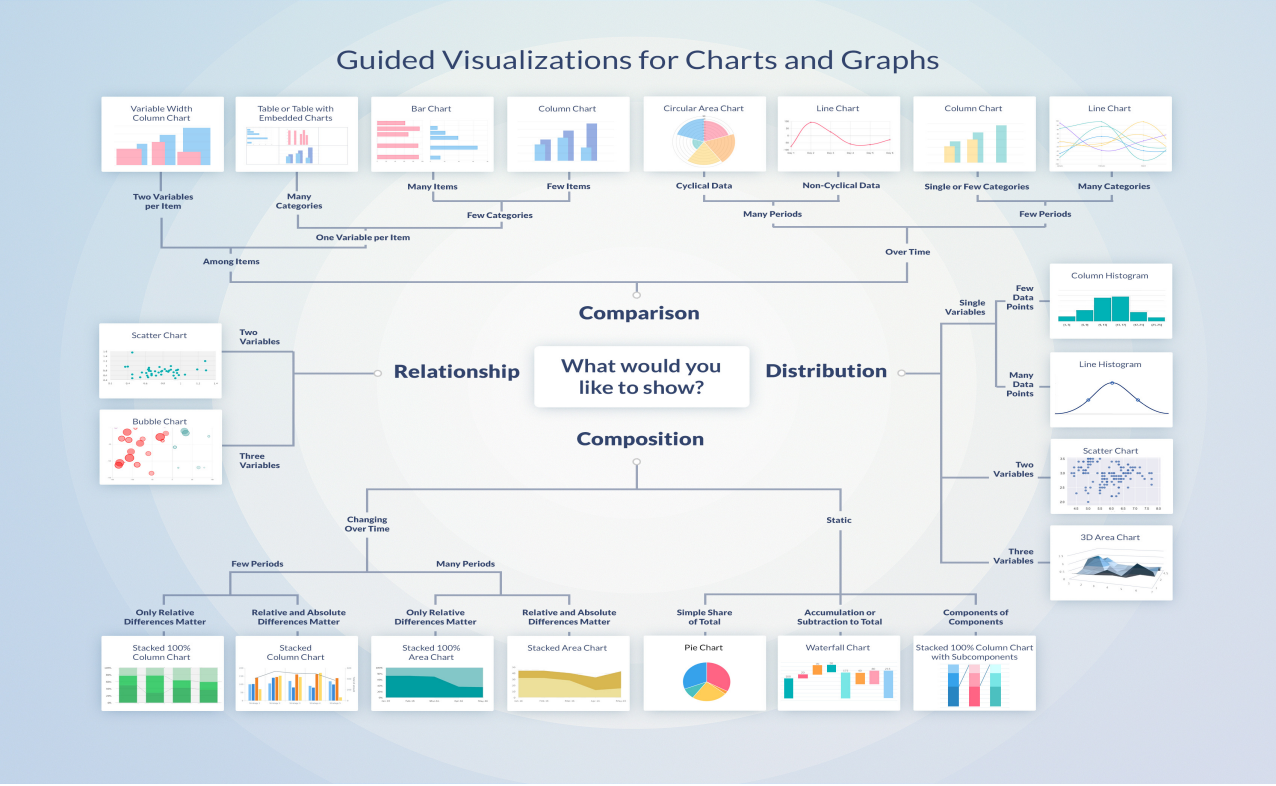
"Visualizing data is all about effective communication. In the public sector, we need to clearly communicate to very broad and diverse audiences. Data tables or traditional statistical output are simply unreadable for most people. In order to pursue clarity, transparency, and accountability, we work to make our analytical output as readable as possible. We emphasize clear graphs and maps of the City in all our reports to ensure readers can understand our results and then leverage that knowledge to push for the changes they want to see in their community."

- Emily Vontsolos

Source:

https://sfcontroller.org/sites/default/files/Documents/Auditing/FY18%20Annual%20Park%20Maintenance%20Standards%20Report.pdf

Which graph to use?



Source : https://www.tapclicks.com/guided-visualization/

Popular Data Visualization tools

1. Matplotlib
2. Seaborn
3. ggplot
4. Bokeh
5. pygal
6. Plotly
7. geoplotlib
8. Gleam
9. missingno
10. Leather

Business Problem : Employee Attrition

Is it a true statement?

Is there other hidden reasons ?

And|Or, ...

Press Center / Press Releases / 2018-01-10

GLASSDOOR SURVEY FINDS MORE EMPLOYEES EXPECTED TO QUIT IN UPCOMING YEAR, WITH SALARY CITED AS TOP REASON

New Data Shows Employee Salary Expectation Key Motivation for Resignation; One-Third of Employers Anticipate More Employees Leaving in 2018

MILL VALLEY, CALIF. (January 10, 2018) - Glassdoor®, one of the world's largest job and recruiting sites, today released new data1 which reveals that 35 percent of hiring decision makers expect more employees to quit over the next 12 months. The survey, conducted among 750 hiring decision makers (those in recruitment, HR and responsible for hiring) in the U.S. and UK, also finds that nearly half (45 percent) note that salary is the top reason for employees changing jobs, followed by career advancement opportunities, benefits and location.

Questions!!

- What features are highly correlated with attrition?
- Which age group employees are quitting more?
- Can we get top three departments facing attrition ?
- Who is quitting more? Male or Female?
- Are they quitting because of travel time?
- Is low monthly income causing attrition to increase?
- Is it possible to predict if an employee is likely to quit ?

Let's explore our data and find answers !

Summary

1. Data science helps in resolving the business problem
2. To get answers for data science question, use data pipeline
3. Importance and steps involved in data cleaning
4. Data Visualization overview

Let’s practice

Resources and Credits

Github Resource : https://github.com/SSaishruthi/women_who_code

Video Credit : Stacey Ronaghan
Medium : http://medium.com/@srnghn/
LinkedIn : https://www.linkedin.com/in/staceyronaghan/

Data Visualization message credit : Emily Vontsolos
Twitter: @EmilyVontsolos

saishruthis

SSaishruthi

gdequeiroz

gdequeiroz