Streamline your process of building a data pipeline and modelling effectively

Emphasis on importance of EDA and statistical tests for Data Science modelling

Laisha Wadhwa

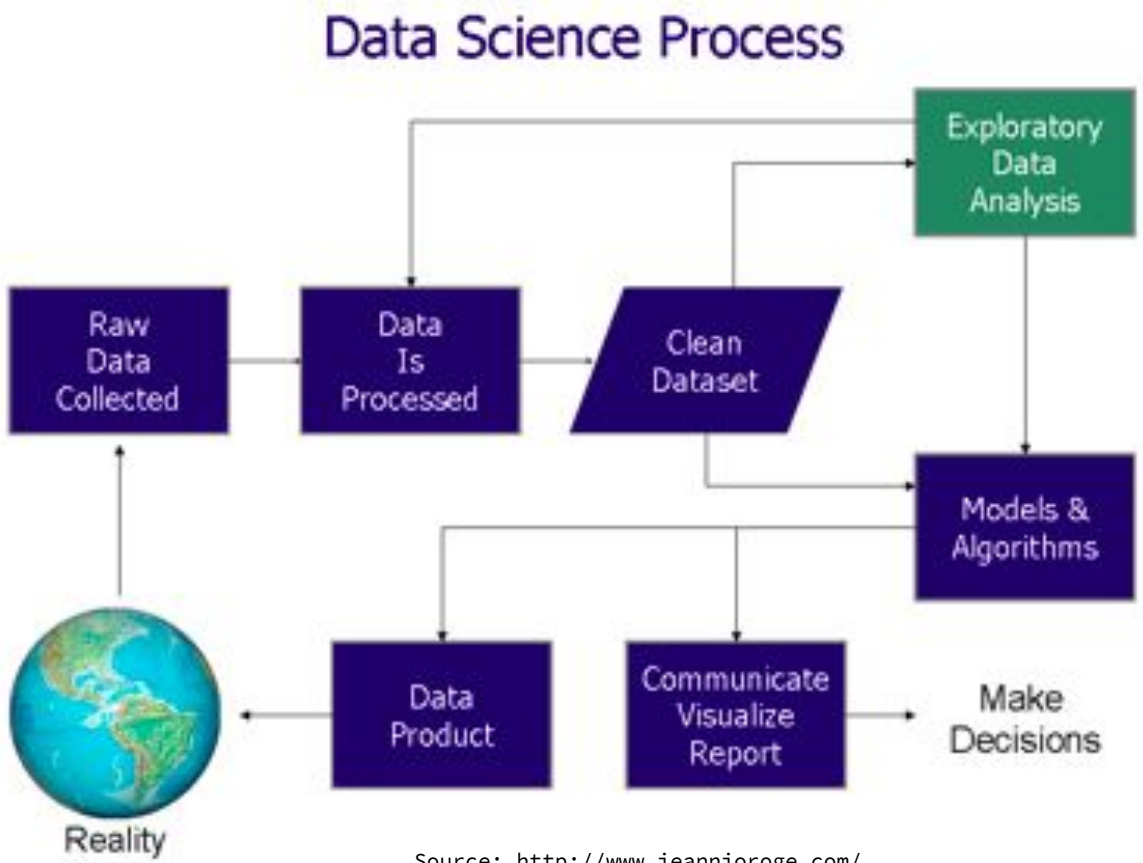
- ★ Data Engineer, Couture.ai
- ★ Microsoft AI hackathon 2018 winner
- ★ Sabre Hack RU
- ★ Amex AI hackathon(Techgig Geek Goddess winner
- ★ Icertis Blockchain and AIML hackathon RU
- ★ Mercedes Benz Digital challenge winner
- ★ Podcast Host - Co-Learning Lounge
- ★ Global Ambassador- Women.Tech Network

in laisha-wadhwa
laishawadhwa

Outline

- Data pipelines
- Importance of EDA
- EDA tools - Bokeh
- Need for statistical tests
- Saving time with lazypredict
- Easy conversion to HTML pages - Voila
- Streamlit - serve your models easily

Data Science process



Source: http://www.jeannjoroge.com/

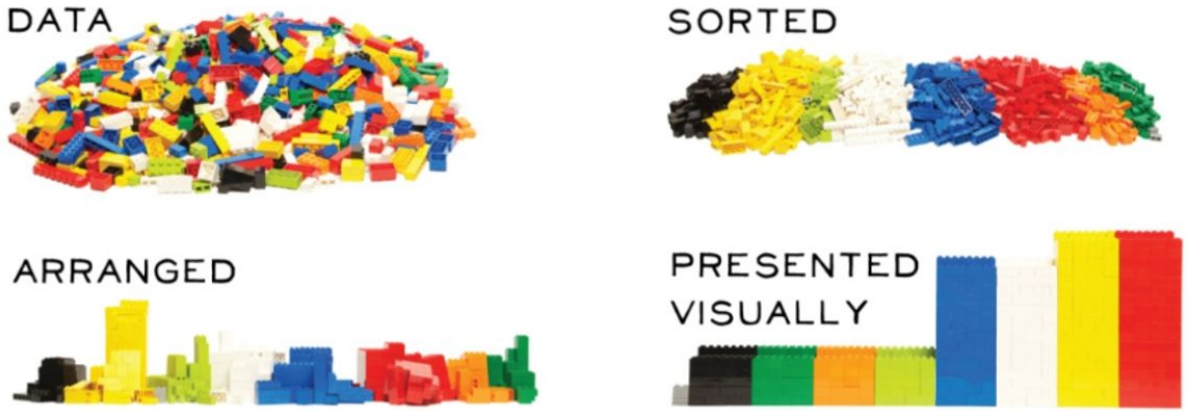
What is EDA?

By definition

“EDA is the practice of describing the data by means of statistical and visualization techniques to bring important aspects of that data into focus for further analysis.”

EDA methods

- Visualisation
- Descriptive statistics



EDA Methods

Descriptive statistics

Central tendency of a distribution

Mean	Sum of all values / Total number of values
Median	Middle value(when data are arranged in order
Mode	Most common value

Measure of Variation

Variance	how far a set of numbers are spread out from mean
Interquartile range	divides a data set into quartiles.
Standard deviation	dispersion of a set of data from mean

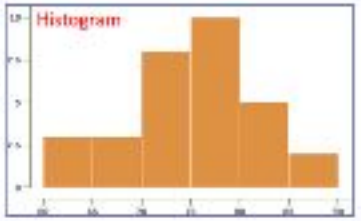
Skewness & Kurtosis

Skewness	Measure of symmetry
Kurtosis	Kurtosis is a measure of "peakedness" relative to a Gaussian shape

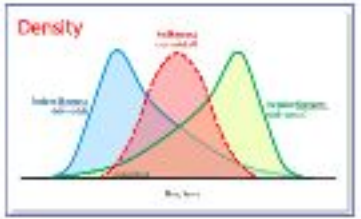
Visualizations

1-dimension

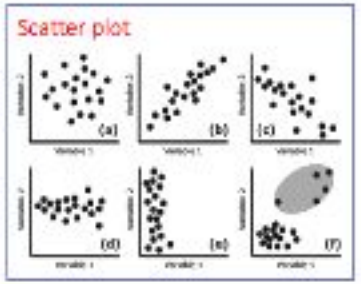
Few data points



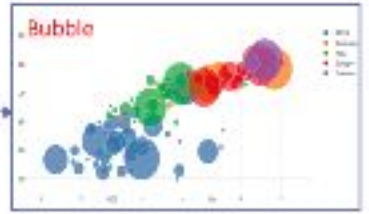
Many data points



2-dimension



3-dimension



Source: http://www.jeannjoroge.com/

Hands on EDA

Let’s look at visualisation:

- Visualisation
- Measure of variance
- Skewness and kurtosis
- Central tendency of distribution
- Heteroscedasticity

Assumptions for regression

- A linear relationship exists between the independent variable (X) and dependent variable (y)
- Little or no multicollinearity between the different features
- Residuals should be normally distributed (multi-variate normality)
- Little or no autocorrelation among residues
- Homoscedasticity of the errors

Fit the data for every model, apply metrics to find which model has better accuracy/metric being used,then choose best model.

Use Lazypredict

Lazy predict

```
In [11]:  1  # to check which model did better on the fetch_california_housing dataset
          2  models_r
```

Out[11]:

Model	R-Squared	RMSE	Time Taken
LGBMRegressor	0.84	0.46	0.42
HistGradientBoostingRegressor	0.84	0.46	2.80
XGBRegressor	0.84	0.47	1.13
ExtraTreesRegressor	0.82	0.50	3.32
RandomForestRegressor	0.81	0.50	7.87
BaggingRegressor	0.79	0.53	0.79
GradientBoostingRegressor	0.79	0.53	3.12
NuSVR	0.77	0.55	6.97
SVR	0.77	0.55	6.56
KNeighborsRegressor	0.75	0.58	0.47
DecisionTreeRegressor	0.64	0.69	0.21
ExtraTreeRegressor	0.58	0.75	0.07
OrthogonalMatchingPursuit	0.47	0.84	0.02
PoissonRegressor	0.46	0.85	0.05

What about time to fit so many models?

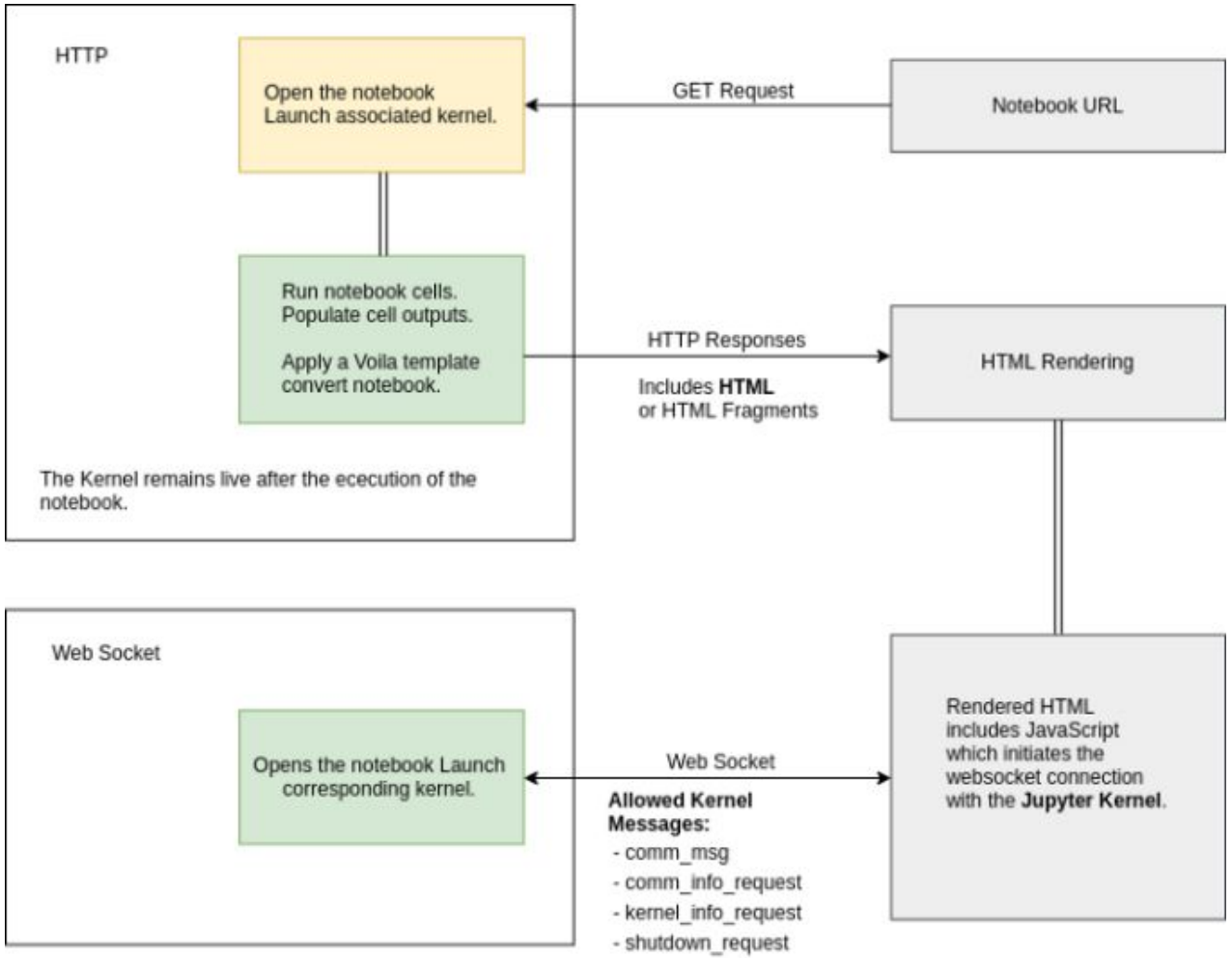
A take on time taken to fit upto 30 models at a time

Voila!
From Jupyter notebooks to standalone applications and dashboards

– – –

- From the exploratory phase of their work to the communication of the results in minutes.
- Rendering the EDA and results visualisation in a web application with no arbitrary code execution by the end user!
- More on voila-gridstack and voila-vuetify template: Dashboarding made easy

Voila execution model



Voila does it all!

Voilà can render custom Jupyter widget libraries, including (but not limited to) bqplot, ipyleafet, ipyvolume, ipympl, ipysheet, plotly, ipywebrtc, etc.

Beyond the voila command-line utility, the voilà package also include a Jupyter ***server extension***, so that voilà dashboards can be served alongside the Jupyter notebook application.

Verbose options avaialable: --strip-sources=False, --theme=dark

Custom templates

PREFIX/share/jupyter/voila/templates/template_name/

```
├── conf.json                # Template configuration file
├── nbconvert_templates/     # Custom nbconvert templates
├── static/                  # Static directory
└── templates/               # Custom tornado templates
```

Deploy notebooks on cloud

– – –

Check this out for deploying your notebooks on heroku

https://github.com/voila-dashboards/voila-heroku

Streamlit

– – –

It is the easiest and quickest way to build web apps that can present your ML models and data attractively using it’s awesome UI elements and Markdown.

https://www.streamlit.io/