Github-- sampe code from other users

Previous Work (eg neuefische notebooks)

Kaggle

Stack Overflow-- sample code (esp for diagrams)

Stackoverflow, Documentation for libraries (Sklearn etc)

Kahn Academy (to explain stats, other concepts)

zindi webpage

sklearn documentation

Dataset

feature understanding

Blog posts (Towards Data Science, Medium)

Books

YouTube tutorials

Wikipedia

google

Colleagues

my colleagues at neue fische

Friends with statistical knowledge

cheat sheets

Documentation

Publications / magazines/ blogs about the sector/field the project is part of

Resources

Kickstarter and crowdfunding websites

Understanding of ML models

numpy

vs code

git

Sololearn

Power BI

Resources: Limited
Amount of people, time, IT
Sources: Less limited
things like Books, ...

budget

laptop (computational resources)

books

youtube

Academic Papers

stakeholder

Previous professional experience

Wallstreet/ Business Magazines

expenditure reports

field analysis reports

experience

collecting info

Deep diving into questions of stakeholders and translating them

Use Cases

wired.com

Discussion with stakeholders, understanding goals

Business Understanding

Wikipedia, for understanding background

Social Media

self-description on the the business website

Stakeholder analysis

their product and cycle

Online Blogs

usual customers

make a solid case for stakeholder

Competition comparatives

Putting yourself in the shoe of your client

Industry reports

deadlines

ask people that work in this business

Trend research on topic and in a general sense

social media

failed companies/ projects

choose metric

Identify the appropriate delivery for the customer.

Data Collection

private datacenter

Own measurements

collection automation

tool

mindfulness

http://www.google.com/publicdata/directory

Open data

Zindi

kaggle

University …ebsites

https://wiki.kiwix.org/wiki/Content_in_all_languages

city web portals

social media comparatives

Field research

government datasets

google and amazon api's

Survey data

https://www.data.gov/opendata/

archive.org

survey

Private or stakeholder-supplied data

website api's

http://aws.amazon.com/public-data-sets/

Published Papers

correspondent Author for scientific papers

http://www.kdnuggets.com/datasets/index.html

Questionair

Previous projects

observational data collection

experimental design

Webscraping

https://github.com/viisar/awesome-datasets

academictorrents.com

Online recorded videos

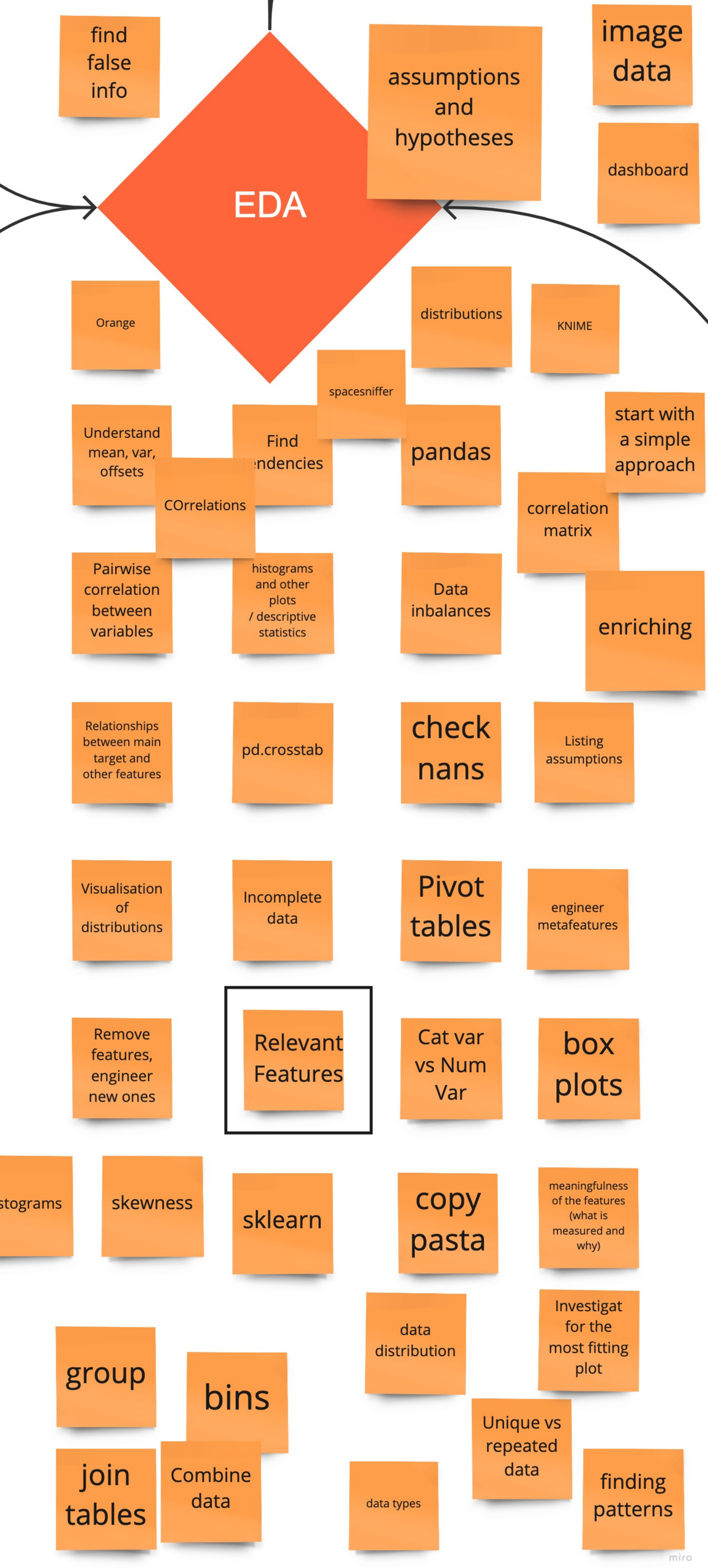
Previous data in companies

data.gov

criminal offender data bases

a lot of data funded by national research agencies has to be made public

https://github.com/awesomedata/awesome-public-datasets


find false info
assumptions and hypotheses
image data
dashboard
EDA
Orange
distributions
KNIME
spacesniffer
Understand mean, var, offsets
Find ndencies
pandas
start with a simple approach
COrrelations
correlation matrix
Pairwise correlation between variables
histograms and other plots / descriptive statistics
Data inbalances
enriching
Relationships between main target and other features
pd.crosstab
check nans
Listing assumptions
Visualisation of distributions
Incomplete data
Pivot tables
engineer metafeatures
Remove features, engineer new ones
Relevant Features
Cat var vs Num Var
box plots
stograms
skewness
sklearn
copy pasta
meaningfulness of the features (what is measured and why)
data distribution
Investigat for the most fitting plot
group
bins
Unique vs repeated data
join tables
Combine data
data types
finding patterns

ratio of one feature to another

conversion

Conversion of strings to numerical data

bins,

Inputting

fillna

Standardization

Normalization

Create new variables

droping

Feature Engineering

balancing

evaluation of importance

format data

dummies

dataVisuals

time data

Remove unnecessary variables

Evaluation metrics for model selection

confusion matrix

tuning

scatter looking at residuals

cause and effect

Hyper Param

start with a simple approach

scikitlearn

dealling with empty values/ Nan

VS correlation

Base line vs more complex models

GridSearch

Ensemble methods

Predictive Modeling

picking appropriate model for the data (e.g. how many features? observations?)

Classification models

Feature importance

supervised ML models

model weighting

neural net

Regression models

important features

REGRESSION!!!

Base line

choose the right model (under/overfit)

Correlation

Color scaling

Heat Maps

heat maps

bar charts

SHAP

Interpretation

Relate findings back to hypothses and objectives

GIS

clear colour coding

Take a step back if results do not seem to be reasonable

decision trees

assume only what you can base on the actual data

Reduce merge variables to make it easier to interpretative

Speak from data perspective, not yours

Use of multiple metrics to identify best match

Geomaps

main point of each graph

actual vs predicted values

interactive graphs

expected vs unexpected trends and predictions

scatterplot

tell a story

Remember the target

Make them feel interested

Storytelling, consistency

Explain what's on the slide

Consistent colour schematics

answer relevant business questions

know your audience

conclusion

use nice theme

easy to understand visuals

Bullet points

feedback

Use non technicsl terms

time box

Intr. to agenda

dress the part

Audience participation

show profit / loss

Online Resources

google slides

Make some jokes!

financial implications

Adapt to audience

Graphics and pictures

outlook what else could be done

present after lunch, not before!

make it simple

don't exceed the time

humanun derstand able plots

further work

digestable amounts of data / information for each slide

stakeholder perspective

demonstration implications of your model/work

corporate presentation style?

Stakeholder presentation

explain target

stick to the storyline

have a storyline

guide viewer through the stages (signposting)

Keep textand numbers to a minumum, keep plots clean/simple

Technical Presentation

Notebooks

Don't present something you can't debate about

Plots are still your friends. Not everybody will understand even when supposed to.

Animation or Videos

tell a technic al story

model behaviour

document your failings

Give your Github details, potential collars

Show the time to rune the code

explain model detailes

go into the maths

deep dive where necessary

be honest about what cannot be known / limitations

explain why you choose models

Explain usage of models / reasoning behind

simplify

Code

psedocode

Error handling documetation

Mention potential pitfalls and how they were overcome (others learn from you)

No details that are irrelevant for your team

backup slides

explain the process in detail

Don't hide data

Present the technical details