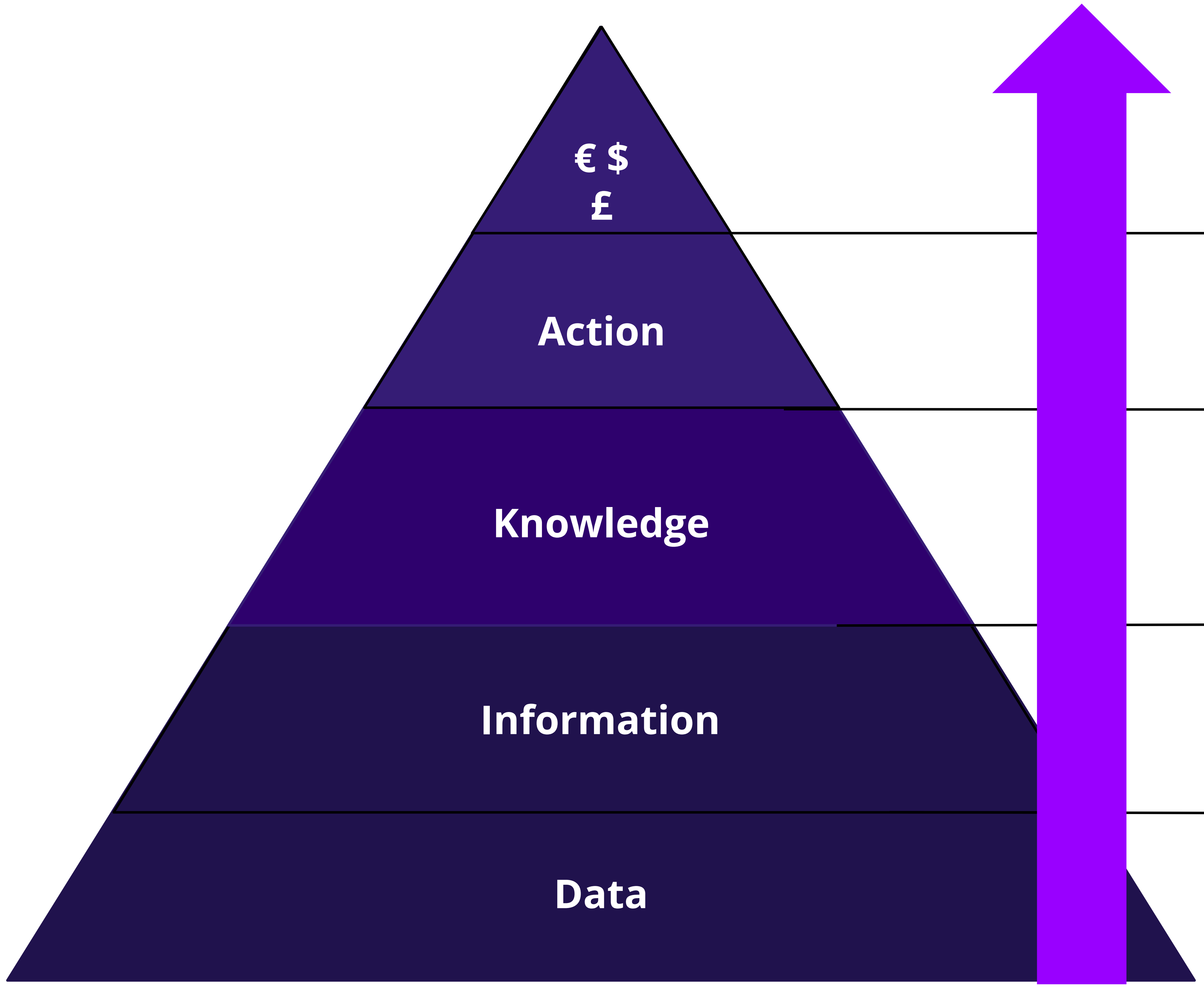
Exploratory Data Analysis (EDA)

neue fische

SPICED

Data Analytics transform Data to Dollars

1 - Specialised skills are needed to progress up each step



Action → **Outcome** : Collect Feedback Data

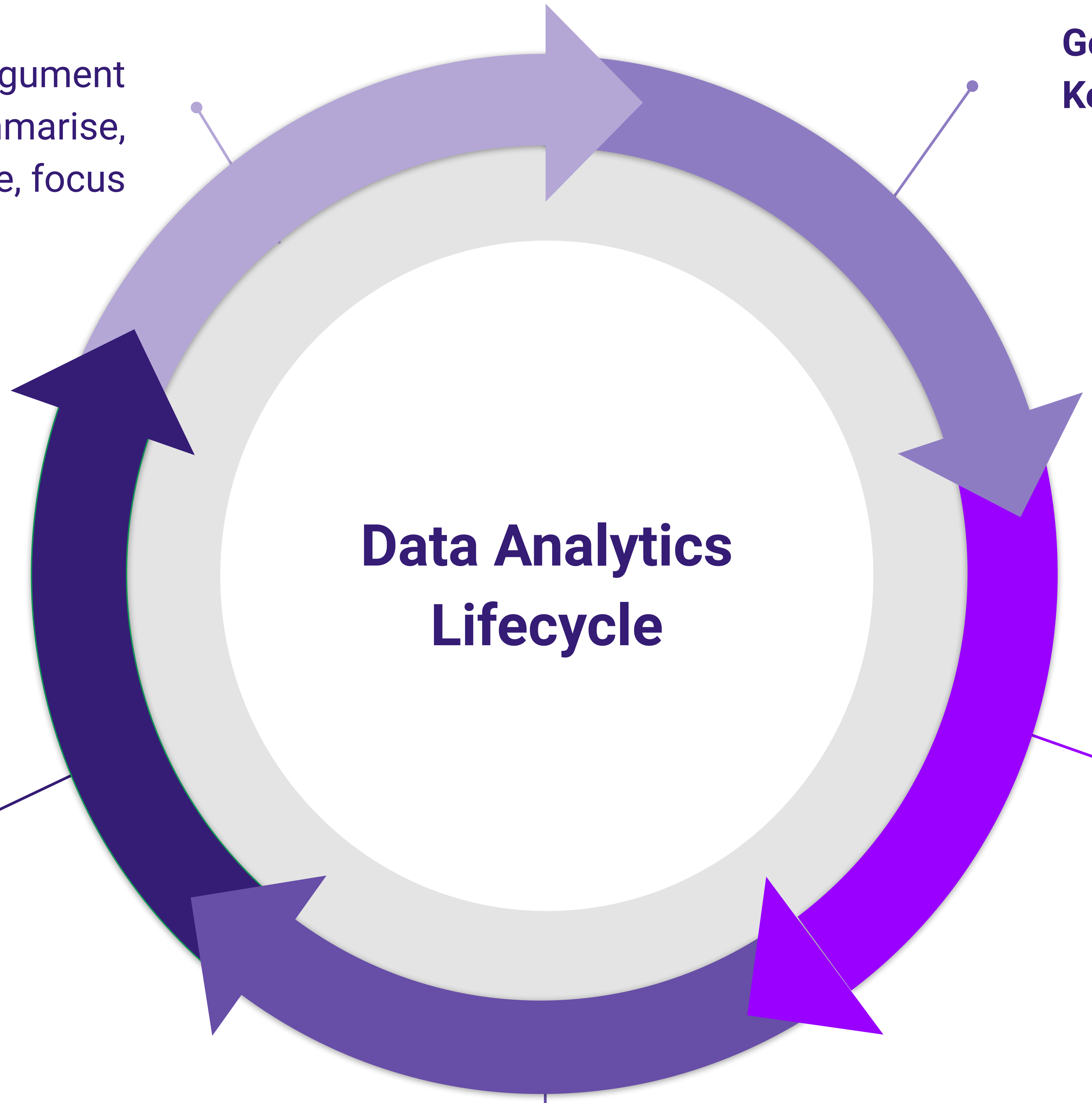
Knowledge → **Action** : Judgement, Communication

Information → **Knowledge** : Math and Statistics

Data → **Information** : Programming, SQL, Subject matter understanding

The World → **Data** : Observation, Measurement, Collection

The Data Analytics Workflow



Data Analytics Lifecycle

1. Business Understanding

Goal: Understand the current process and problems
Key Activity: Speak with Stakeholders

2. Define Solution

Goal: Set Requirements
Key Activity: Build solution with hypothesis

3. Get Data
Goal: Identify and validate data sources
Key Activity: Data Mining and EDA

4. Challenge hypothesis

Goal: Confirm or reject hypothesis
Key Activity: Apply statistical and logical analysis

5. Result communication and Decision Making

Goal: Create convincing argument
Key Activity: Data Visualisation, Summarise, reduce, focus

Today's Objective

Exploratory Data Analysis

Why?

- Initial investigations on your data are key in order to understand them - which again is necessary for further data analysis and future predictions

What we aim for today:

- Understanding the concept of EDA
- Knowledge about the steps within an EDA

EDA = Detective Work

Performing initial investigations

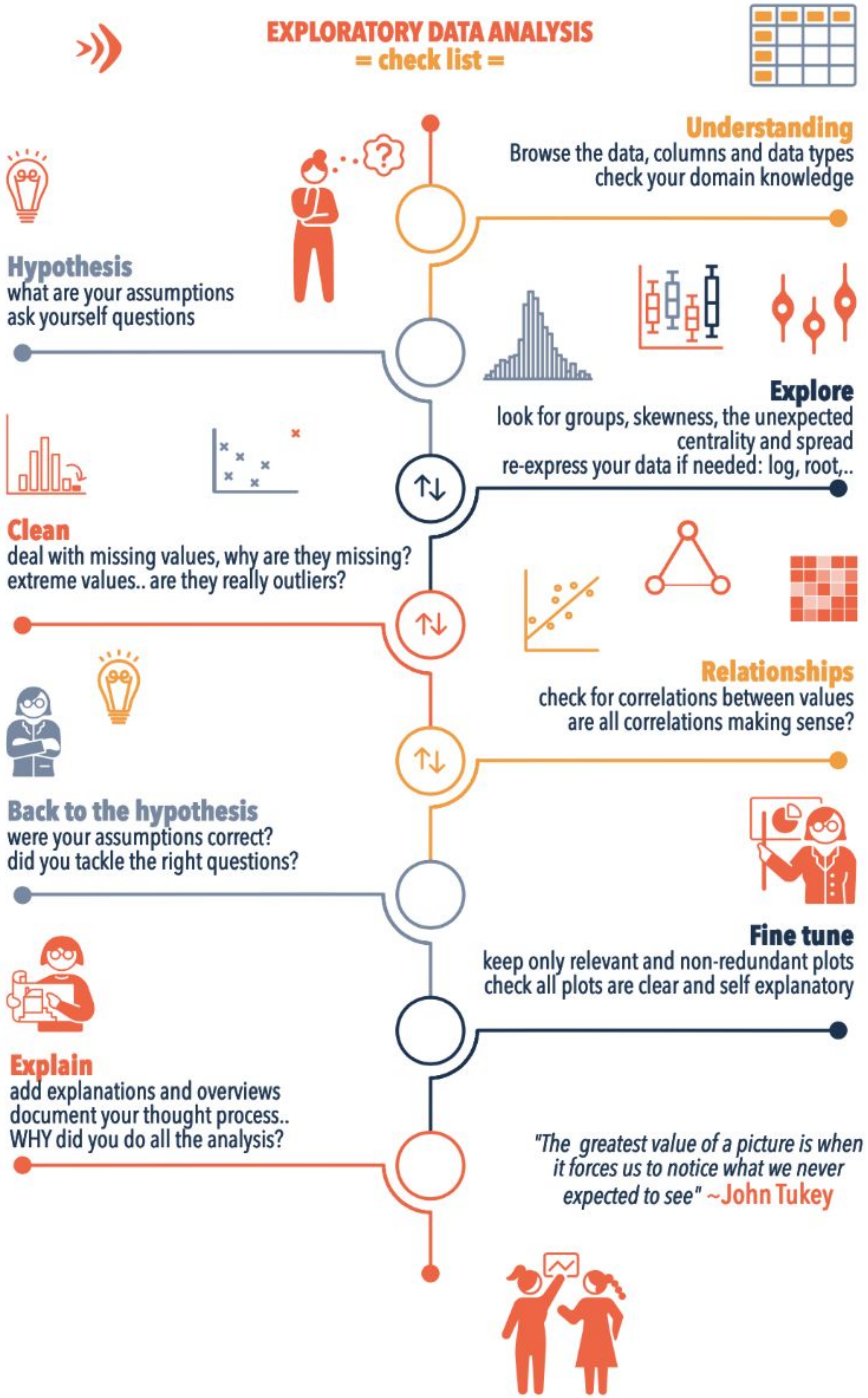
- Get to know your data set
- discover patterns
- spot anomalies
- test hypothesis
- check assumptions

Using

- summary statistics ✓
- Python Pandas
- statistics/plots showing relationship
- visualizations

Exploratory Data Analysis

EDA Checklist



EXPLORATORY DATA ANALYSIS
= check list =

Understanding
Browse the data, columns and data types
check your domain knowledge

Hypothesis
what are your assumptions
ask yourself questions

Explore
look for groups, skewness, the unexpected
centrality and spread
re-express your data if needed: log, root,..

Clean
deal with missing values, why are they missing?
extreme values.. are they really outliers?

Relationships
check for correlations between values
are all correlations making sense?

Back to the hypothesis
were your assumptions correct?
did you tackle the right questions?

Fine tune
keep only relevant and non-redundant plots
check all plots are clear and self explanatory

Explain
add explanations and overviews
document your thought process..
WHY did you do all the analysis?

"The greatest value of a picture is when it forces us to notice what we never expected to see" ~John Tukey

Step 1: Understand your data

To Do's:

- “Browse” the data, columns and data types
- Apply your domain knowledge

Helpful functions

df.head()
df.shape
df.info()
df.columns

Step 2: Build hypotheses

To Do's:

What are your assumptions: Ask yourself questions

Pick your hypotheses
***before* looking at your data too deeply!!!**

What is the conclusion or what does it mean if all your hypothesis are confirmed?

Exploratory Data Analysis

Step 3: Explore your data

To Do's:

Have a look at

- Distribution of your data, eg.:
    - Skewness
    - Centrality and spread
- Unexpected values (e.g. outliers, “?”, …)
- Missing values
- Make list of issues, and needed changes

Helpful functions/tools

- df.describe() - Descriptive Statistics
- df.isnull() - find missing values
- Distribution plots - detecting skewness and deviation from normal distribution
- Boxplots - detecting (possible) outliers
- … and a lot more

Step 4: Clean your data

To Do's:

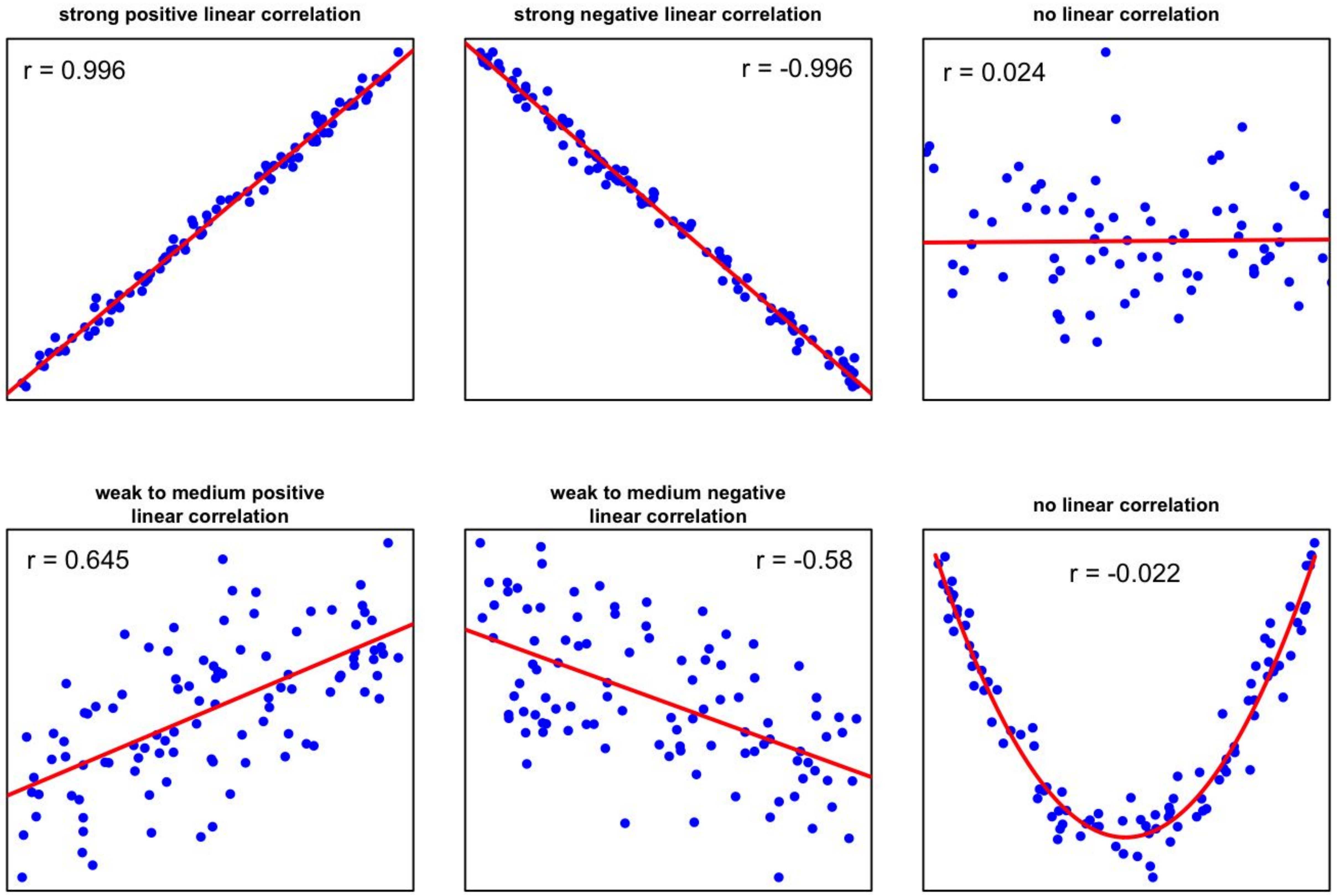
- Adapt data types if needed
- Deal with missing values, why are they missing, are they really missing?
- Extreme values - are they really outliers?
- Special characters? Special formatting?
- Imputation/augmentation
- Re-express your data if needed

Step 5: Check for relationships

To Do's:

- Make scatter plots
- Check for correlations between values
- Are all correlations making sense?

Correlations: covered later



Step 6: Back to your hypothesis

To Do's:

- Confirm or reject each of your hypothesis?
- For each rejected hypothesis, reformulate and repeat the process of validating
- What conclusion can you make from the new knowledge gained from confirming (rejecting) the hypothesis?

Step 7: Fine tuning

To Do's:

Make yourself ready for presenting your insights!

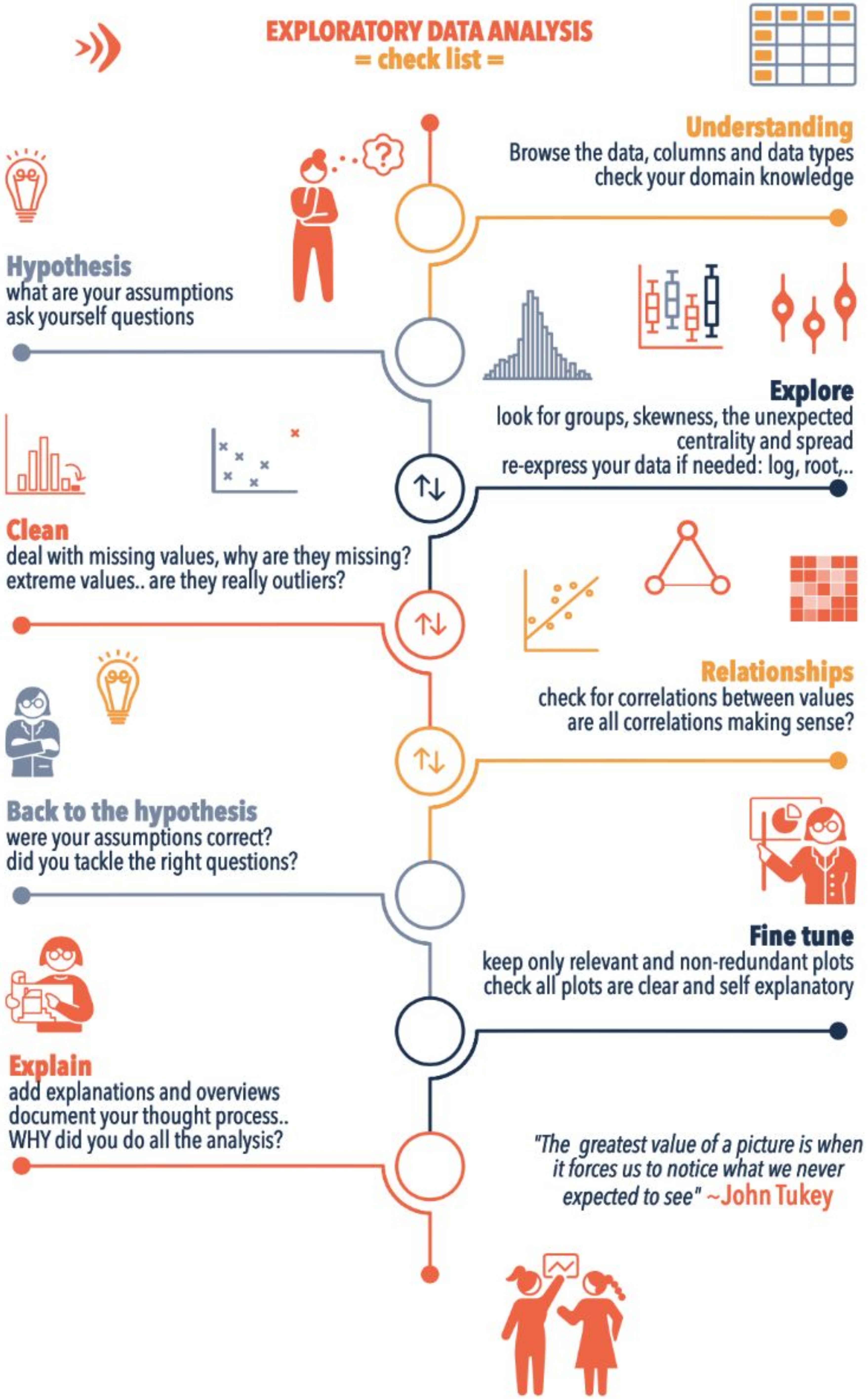
- Keep only relevant and non-redundant plots
- Check all plots are clear and self explanatory

Covered in Data Viz and Tableau

Step 8: Explain your reasoning

To Do's:

- Add explanations and overviews
- Document your thought process
- WHY did you do all the analysis?



EXPLORATORY DATA ANALYSIS
= check list =

Understanding
Browse the data, columns and data types
check your domain knowledge

Hypothesis
what are your assumptions
ask yourself questions

Explore
look for groups, skewness, the unexpected
centrality and spread
re-express your data if needed: log, root,..

Clean
deal with missing values, why are they missing?
extreme values.. are they really outliers?

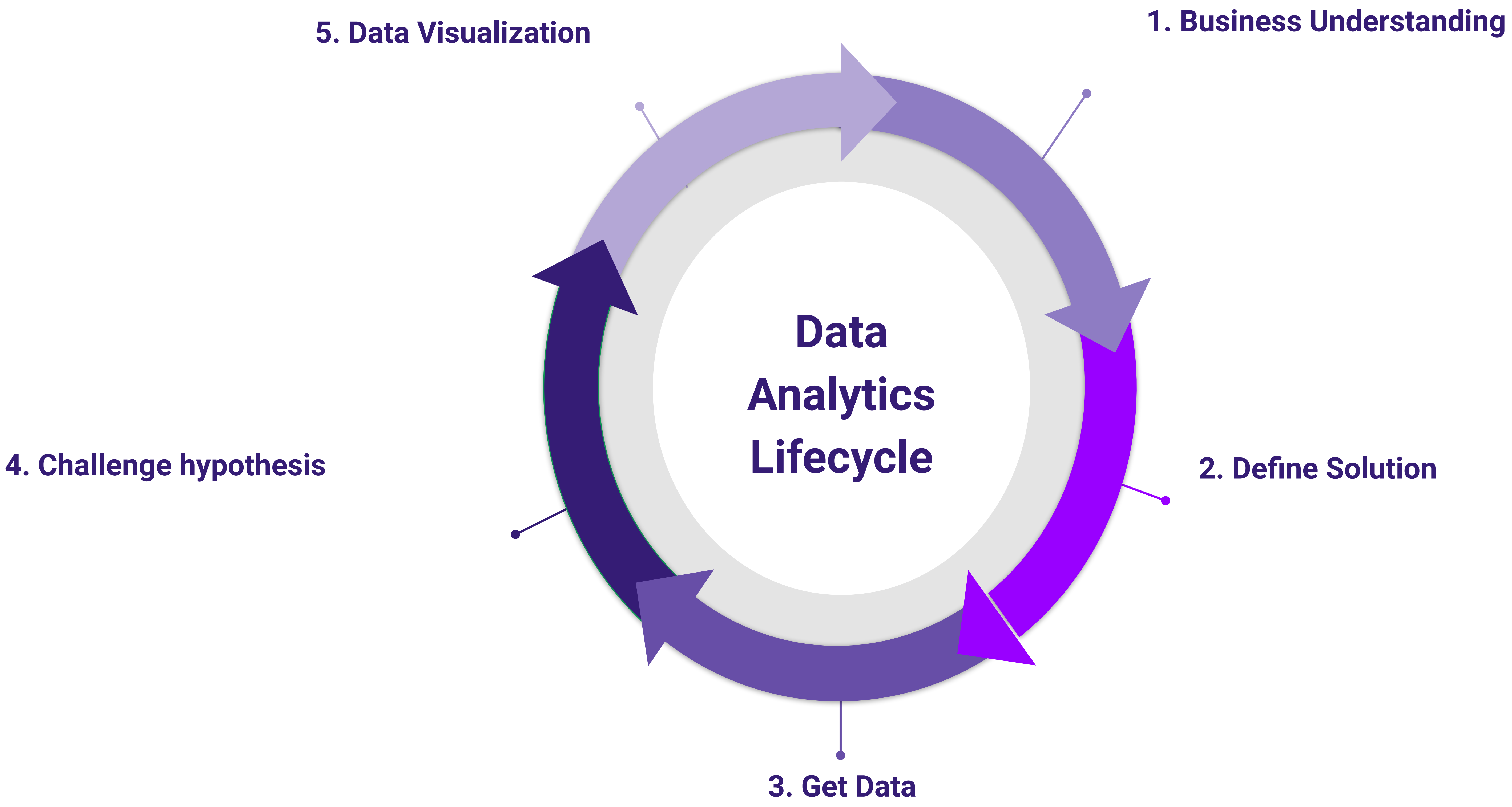
Relationships
check for correlations between values
are all correlations making sense?

Back to the hypothesis
were your assumptions correct?
did you tackle the right questions?

Fine tune
keep only relevant and non-redundant plots
check all plots are clear and self explanatory

Explain
add explanations and overviews
document your thought process..
WHY did you do all the analysis?

"The greatest value of a picture is when it forces us to notice what we never expected to see" ~John Tukey



Data Analytics Lifecycle

1. Business Understanding
2. Define Solution
3. Get Data
4. Challenge hypothesis
5. Data Visualization

... and now: 🐼, 🐼, 🐼!

References

Eda:

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Field, A. P. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)*. Los Angeles [i.e. Thousand Oaks, Calif.: SAGE Publications.

Missing values:

https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b

https://www.kaggle.com/alexisbcook/handling-missing-values

Outlier:

https://pub.towardsai.net/outlier-detection-and-treatment-a-beginners-guide-c44af0699754