EDA checklist

and introduction to Bokeh

Bokeh

Bokeh is an interactive visualization library for modern web browsers

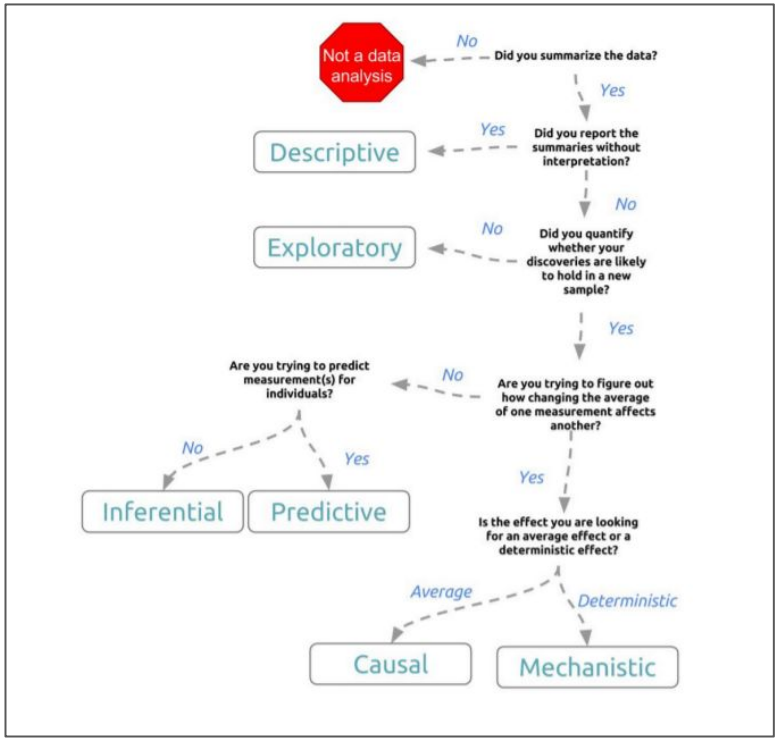
Why Bokeh?

- The more libraries you know, the better equipped you will be to use the right visualization tool for the task.
- Bokeh is slightly faster and integrates better than Plotly

Notebook Demo

EDA Journey

Asking the question



Source: The data analytic question

Types of analysis

- Descriptive(eg. Census data)
- Exploratory(eg. Discovery of Four Planet System)
- Inferential(eg. Air pollution vs Life expectancy)
- Prediction(eg. Match data in Sports)
- Causal(eg. Clinical Drug tests)
- Mechanistic(eg. Better wing design in decreased drag)

Question checklist

1. Did you specify the type of data analytic question (e.g. exploration, association causality) before touching the data?

2. Did you define the metric for success before beginning?

3. Did you understand the context for the question and the scientific or business application?

4. Did you consider whether the question could be answered with the available data?

Tidying the data

1. Is each variable one column?
2. Is each observation one row?
3. Do different data types appear in each table?
4. Did you record the recipe for moving from raw to tidy data?
5. Did you create a code book?
6. Did you record all parameters, units, and functions applied to the data?

Checking the data

1. Did you plot univariate and multivariate summaries of the data?

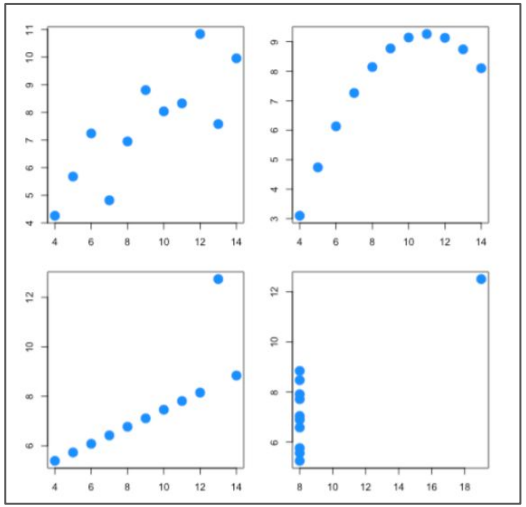
   - There is no such thing as too many plots
   - Important to check clear coding errors, label switching, units are in scale, etc

2. Did you check for outliers?

3. Did you identify the missing data code?

Exploratory analysis

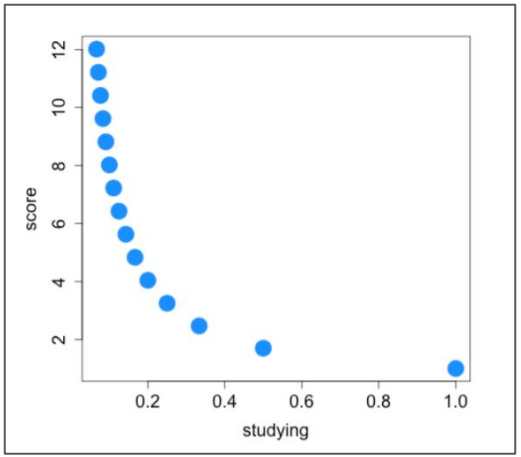
- Plot as much of the actual data as you can
- Plots are better than summaries

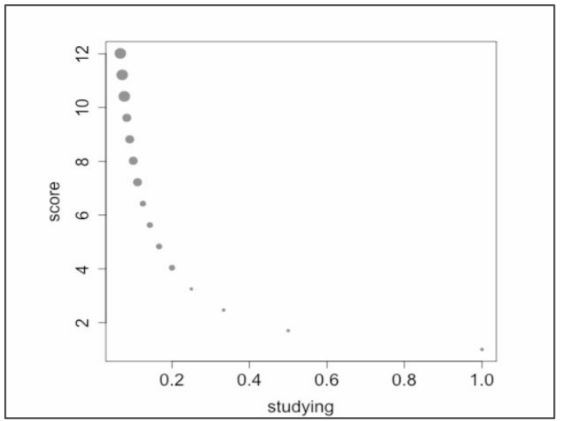


Source: The data analytic question

Exploratory analysis

- For large data sets, subsample before plotting
- Use log transforms
- Use color and size to check for confounding





Source: The data analytic question

Exploratory analysis

1. Did you identify missing values?

2. Did you make univariate plots (histograms, density plots, boxplots)?

3. Did you consider correlations between variables (scatterplots)?

4. Did you check the units of all data points to make sure they are in the right range?

5. Did you try to identify any errors or miscoding of variables?

6. Did you consider plotting on a log scale?

7. Would a scatterplot be more informative?

Inference

1.Did you identify what large population you are trying to describe?

2. Did you clearly identify the quantities of interest in your model? Do you care about it?

3. Did you consider potential confounders?

4. Did you calculate a measure of uncertainty for each estimate on the scientific scale?

TIP: When possible, perform exploratory and confirmatory analysis on separate data sets

Prediction

1. Did you identify in advance your error measure?

2. Did you immediately split your data into training and validation?

3. Did you use cross validation, resampling, or bootstrapping only on the training data?

4. Did you create features using only the training data?

5. Did you fix all features, parameters, and models before applying to the validation data?

6. Did you apply only one final model to the validation data and report the error rate?

Prediction is about tradeoffs

- Interpretability versus accuracy
- Speed versus accuracy
- Simplicity versus accuracy
- Scalability versus accuracy

Written analyses

1. Did you describe the question of interest?

2. Did you describe the data set, experimental design, and question you are answering?

3. Did you specify the type of data analytic question you are answering?

4. Did you specify in clear notation the exact model you are fitting?

5. Did you explain on the scale of interest what each estimate and measure of uncertainty means?

6. Did you report a measure of uncertainty for each estimate on the scientific scale?

Figures

1 .Does each figure communicate an important piece of information or address a question of interest?

2. Do all your figures include plain language axis labels?

3. Does every figure have a detailed caption that explains all axes, legends, and trends in the figure?

Bonus: Top 10 worst graphs

Reproducibility

1.Did you avoid doing calculations manually?

2. Did you create a script that reproduces all your analyses?

3. Did you save the raw and processed versions of your data?

4. Did you record all versions of the software you used to process the data?

5. Did you try to have someone else run your analysis code to confirm they got the same answers?

Study Jam Activity

You have been hired as Data Scientist by the “EFA(Europe Football Association) They have shared with you all the data about their matches.

As a data scientist, you need to:

- Figure out the right question

- Perform EDA

Helpful Resources

- The elements of data analytic style

- European soccer database notebooks