Data Science Career Track
EDA Cheat Sheet

Why We Use EDA

Sometimes the consumer of your analysis won't understand why you need the time for EDA and will want results right away Here are some of the reasons you can give to convince them it's a good use of time for everyone involved.

Reasons for the analyst

- Identify patterns and develop hypotheses.
- Test technical assumptions. Inform model selection and feature engineering.
- Build an intuition for the data.

Reasons for the consumer of the analysis

- Ensures delivery of technically-sound results.
- Ensures the right question is being asked. Tests business assumptions.
- Provides context necessary for maximum applicability and value of results.
- Leads to insights that would otherwise not be found.

Things to keep in mind:

- You're never done with EDA. With every analytical result, you want to return to EDA, make sure the result makes sense, test other questions that come up because of it.
- Stay open-minded. You're supposed to be challenging your assumptions and those of the stakeholder who you're performing the analysis for.
- Repeat EDA for every new problem. Just because you've done EDA on a dataset before doesn't mean you shouldn't do it again for the next problem. You need to look at the data through the lens of the problem at hand and you will likely have different areas of

investigation.

EDA Major Tasks

Exploratory data analysis consists of the following major tasks, which we present linearly here because each task doesn't make much sense to do without the ones prior to it. However, in reality, you are going to constantly jump around from step to step. You may want to do all the steps for a subset of the variables first or you might jump back because you learned something and need to have another look.

1. Form hypotheses/develop investigation themes to explore
2. Wrangle data
3. Assess the quality of data
4. Profile data
5. Explore each individual variable in the dataset
6. Assess the relationship between each variable and the target
7. Assess interactions between variables
8. Explore data across many dimensions

Throughout the entire analysis you want to:

- Capture a list of hypotheses and questions that come up for further exploration.
- Record things to watch out for/ be aware of in future analyses.
  Show intermediate results to colleagues to get a fresh perspective, feedback, domain knowledge. Don't do EDA in a bubble! Get feedback throughout especially from people removed from the problem and/or with relevant domain knowledge.
- Position visuals and results together. EDA relies on your natural pattern recognition abilities so maximize what you'll find by putting visualizations and results in close proximity.

Wrangling

Basic steps to follow:

Make your data tidy.

i. Each variable forms a column
ii. Each observation forms a row
iii. Each type of observational unit forms a table

Transform data: sometimes you will need to transform your data to be able to extract information from it. This step will usually occur after some of the other steps of EDA unless domain knowledge can inform these choices beforehand.

Log: when data is highly skewed (versus normally distributed like a bell curve), sometimes it has a log-normal distribution and taking the log of each data point will normalize it.
Binning of continuous variables: Binning continuous variables and then analyzing the groups of observations created can allow for easier pattern identification. Especially with non-linear relationships.

Simplifying of categories: you really don't want more than 8-10 categories within a single data field. Try to aggregate to higher-level categories when it makes sense.

Helpful packages
pandas

Data quality assessment and profiling
Before trying to understand what information is in the data, make sure you understand what the data represents and what's missing.

Basic steps to follow:
Categorical: count, count distinct, assess unique values
Numerical: count, min, max
Spot-check random samples and samples that you are familiar with Slice and dice

Questions to Consider
What data isn't there?
Are there systematic reasons for missing data? Are there fields that are always missing at the same time? Is there information about what data is missing? Is the data that is there right?
Are there frequent values that are default values? Are there fields that represent the same information? What timestamp should you use? Are there numerical values reported as strings?
Are there special values? Is the data being generated the way you think?
Are there variables that are numerical but really should be categorical? Is data consistent across different operating systems, device types, platforms, countries? Are there any direct relationships between fields (e.g. a value of x always implies a specific value of y)? What are the units of measurement? Are they consistent? Is data consistent across the population and time? (time series) Are there obvious changes in reported data around the time of important events that affect data generation (e.g. version release)? (panel data)

Helpful packages
missingno
pivottablejs
pandas_profiling

Example backlog

- Assess the prevalence of missing data across all data fields, assess whether its missing is random or systematic, and identify patterns when such data is missing Identify any default values that signify missing data for a given field.
- Determine sampling strategy for quality assessment and initial EDA
- For DateTime data types, ensure consistent formatting and granularity of data, and perform sanity checks on all dates present in the data
- In cases where multiple fields capture the same or similar information, understand the relationships between them and assess the most effective field to use
- Assess the data type of each field For discrete value types, ensure data formats are consistent
- For discrete value types, assess the number of distinct values and percent unique and do a sanity check on types of answers
- For continuous data types, assess descriptive statistics and perform a sanity check on values Understand relationships between timestamps and assess which to use in analysis
- Slice data by device type, operating system, software version and ensure consistency in

data across slices

- For device or app data, identify version release dates and assess data for any changes in format or value around those dates

Exploration
After quality assessment and profiling, exploratory data analysis can be divided into 4 main types of tasks:
1. Exploration of each individual variable
2. Assessment of the relationship between each variable and the target variable
3. Assessment of the interaction between variables
4. Exploration of data across many dimensions

Step 1: Exploring each individual variable

Basics steps to complete during Step 1
Quantify:
Location: mean, median, mode, interquartile mean *Spread*: standard deviation, variance, range, interquartile range
Shape: skewness, kurtosis
For time series, plot summary statistics over time.
For panel data:
Plot cross-sectional summary statistics over time Plot time-series statistics across the population

Questions to consider when working on Step 1
What does each field in the data look like?
Is the distribution skewed? Bimodal? Are there outliers? Are they feasible? Are there discontinuities? Are the typical assumptions seen in modeling valid?
Gaussian
Identically and independently distributed
Have one mode
Can be negative
Generating processes are stationary and isotropic (time series)
Independence between subjects (panel data)

Step 2: Exploring the relationship between each variable and the target
How does each field interact with the target?
Assess each relationship's:
Linearity Direction
Rough size
Strength
Methods:
Bivariate visualizations Calculate correlation

Step 3: Assessing interactions between variables
How do the variables interact with each other?

Basics steps to complete during Step 3:

Bivariate visualizations for all combinations
Correlation matrices
Compare summary statistics of variable x for different categories of y

Step 4: Exploring data across many dimensions
Are there patterns across many of the variables?

Basics steps to complete during Step 4:
Categorical:
Parallel coordinates
Continuous Principal component analysis
Clustering

Helpful packages
ipywidgets: making function variables interactive for visualizations and calculations
mpld3: interactive visualizations

Example backlog

- Generate list of questions and hypotheses to be considered during EDA
- Create univariate plots for all fields
- Create bivariate plots for each combination of fields to assess correlation and other relationships
- Plot summary statistics over time for time series data
- Plot distribution of x for different categories y
- Plot mean/median/min/max/count/distinct count of x over time for different categories of y
- Capture list of hypotheses and questions that come up during EDA
- Record things to watch out for/ be aware of in future analyses Distill and present findings

Visualization guide
Here are the types of visualizations and the python packages we find most useful for data exploration.

Univariate
Categorical: Bar plot Continuous:
Histograms Kernel density estimation plot Box plots

Bivariate
Categorical x categorical
Heat map of contingency table Multiple bar plots
Categorical x continuous
Box plots of continuous for each category Violin plots of continuous distribution for each category
Overlaid histograms (if 3 or fewer categories)
Continuous x continuous
Scatter plots Hexibin plots Joint kernel density estimation plots Correlation matrix heatmap

Multivariate
Pairwise bivariate figures/ scatter matrix

Timeseries
Line plots
Any bivariate plot with time or time period as the x-axis.

Panel data
Heat map with rows denoting observations and columns denoting time periods.
Multiple line plots
Strip plot where time is on the x-axis, each entity has a constant y-value and a point is plotted every time an event is observed for that entity.

Geospatial
Choropleths: regions colored according to their data value.

Helpful packages
matplotlib: basic plotting
seaborn: prettier versions of some matplotlib figures
mpld3: interactive plotting
folium: geospatial plotting