Beginning Data Analysis Routine

This notebook contains a checklist of many of the different things you can do during exploratory data analysis.

Develop your own Data Analysis Routine

- Use this notebook as a starting point for developing your own data analysis routine. Ideally, you will create and continually modify this document so that it contains all the ideas and steps you usually take during the beginning of an analysis.

Have the Mindset of a Detective

- As an analyst, it is your job to extract information from data
- Go beyond the keyboard to investigate as if you are detective
- Have courage to ask the important questions
- You can also think of yourself as making a documentary about the data. You will ultimately tell some story about it. Make it accurate and interesting.

Project Genesis

These are some things to do at the genesis of your data analysis project

- Create a folder in your file system to hold all your files for the analysis
- Create a documents/spreadsheets to store the names, titles, contact information and notes of all the people connected to your project
- Find and introduce yourself to all the people connected to your project
- Connections to others is key to making your projects work. The more you are visible to others the more information will freely pass your way
- Be aware of all the people that are directly and indirectly connected to your project. Meet all of them
- Stakeholders
- Domain Experts
- Other data scientists
- Database admins - data engineers
- Solutions architects
- Project managers
- Web developers

Before Looking at Data

Once you have been given access to data, in a text document or Jupyter notebook answer the following questions:

- What process generates this data?
- Is it generated from industrial equipment, a website, internal software?
- When was it created?
- How often is it updated?
- What database(if any) is it stored in?
- Who are the admins of the database?
- Can you view the schema?
- What is the process that the raw data has gone through before it reached your hands? Has it already been pre-processed before it reaches you?
- Is there a data dictionary describing every column?
- What systems use the data?
- Have their been previous data scientists working with this dataset?
- How has data changed over time? Which columns have been added/subtracted?
- Is data for some columns not being collected?

Subject Matter Research

- Read articles, watch videos, talk to local subject matter experts
- Read articles/papers by academics who have already studied the field using statistical analysis
- Could be beneficial to do some analysis first as to not bias your results

First Look at Data

- Find data dictionary
- Even if one exists, create a column to keep track of notes for each variable
- Make sure your data dictionary has the column name, data type, range of values and notes on each column
- If the data comes from a relational database, ask to see the schema
- Number of rows and columns
- Find number of missing values per column

Is the Data Tidy?

- Data must be tidy before analysis starts.
- Most data from relational databases will be tidy
- Data from spreadsheets or scraped from the web/pdfs might not be
- Find data type of each column - continuous, categorical (ordinal or nominal), or date
- Rearrange column order in a sensible manner - categorical first, continuous last. Group common variables together.

Univariate vs Bivariate and Graphical vs Non-Graphical

Univariate	Graphical	Non-Graphical
Categorical	Bar char of frequencies (count/percent)	Contingency table (count/percent)
Continuous	Histogram/rugplot/KDE, box/violin/swarm, qqplot, fat tails	central tendency -mean/median/mode, spread - variance, std, skew, kurt, IQR

Bivariate/multivariate	Graphical	Non-Graphical

Bivariate/multivariate	Graphical	Non-Graphical
Categorical vs Categorical	heat map, mosaic plot	Two-way Contingency table (count/percent)
Continuous vs Continuous	all pairwise scatterplots, kde, heatmaps	all pairwise correlation/regression
Categorical vs Continuous	bar, violin, swarm, point, strip seaborn plots	Summary statistics for each level

Univariate Analysis

- Look at one variable at a time.

Categorical variables

- There is less available options with categorical variables
- Count the frequency of each variable
- Low frequency strings might be outliers
- You might want to relabel low frequency strings 'other'
- Find the number of unique labels for each column
- In pandas, change the data type to categorical (better when there aren't too many unique values)
- Bar plots of counts
- String columns allow for feature engineering by splitting the string, counting certain letters, finding the length of, etc... Feature engineering can be done later when modeling

Continuous variables

- There are a lot more options for continuous variables
- Use the five number summary - with `.describe`
- Boxplots are great ways to find outliers
- Use histograms and kernel density estimators to visualize the distribution.
- Know the shape of the distribution
- Think about making categorical variables out of continuous variables by cutting them into bins.

Use bootstrapping to get more 'samples'

- Bootstrapping is done by resampling your data with replacement and gives you a 'new' random dataset
- This helps you get multiple looks at the data
- You can get estimates for the mean and variance of continuous columns this way.

Outliers in one dimension

- Use your natural human ability to look at boxplots to find thresholds for what an outlier might be
- Generate a new column of data that is 0/1 for outlier or not. This will quickly help you find them later.

Duplicated data

- Lots of data gets accidentally duplicated. Check for duplicates or near duplicates of rows and columns
- If any columns are calculated entirely by that of another column or columns (like with depth from the diamonds data), ensure the calculation holds.

Making new binary columns to label some finding

- Just like it was described above to make a 0/1 column for outliers, you can do the same for any other finding
- You can drop the duplicated rows or you can make a binary column labeling them.
- Same for rows that do not have a correct calculation.

Bivariate and Multivariate EDA

Categorical vs Categorical

- Create two way contingency table of frequency counts
- Create a heat map
- Find expected counts and possibly do a chi-squared test

Categorical vs Continuous

- Use the seaborn categorical plots

Continuous vs Continuous

- Plot all combinations of scatterplots
- Use a hierarchical clustering plot