8/21/23

Data science Lifecycle

6 steps:

1. Question
2. Collect data
3. Wrangle data
4. Analyze data
5. Visualize information
6. Communicate information

9/5/23

Data sciences fundamentals

Collection types are very useful in data science because data sets are the collection of data.

Pythons collection types are:

- List
  Ordered, changeable,duplicates allowed
- Set
  Ordered, changeable, duplicates not allowed
- Dictionary
- Tuple

You can represent data sets with code using one of the following:

Column oriented- grouping by features

Row oriented- grouping by a single observation

Eg.

Name	City	Age
Matt	Los Angeles	27
Dave	San Francisco	30
Tim	Oakland	33

Row oriented would be

Matt, Los Angeles, 27   Dave ,San Francisco   Tim, Oakland, 33

Column oriented would be

Matt, Dave, Tim   Los Angeles, San Francisco, Oakland   27,30,33

Indexing

Inorder to access values in a collection type, we need to index into it

List – name[index(whole number only]
Dictionary – name[key]

Iterations

You can repeat processes with loops or recursion in python.

There are: for loops and while loops.

Useful methods

Dictionary; values(), items(), keys()

Lists : len(), append(), sort()

Other: range(), split(), type(), int(), str()

Central Tendency

Measures of central tendency

Mean, Median and Mode are statistical measures that help us describe the behavior of a collection of data points.

Pandas fundamentals.

Pandas:

A python library that can make analyzing data easier.

Dataframes:

- A pandas object that is used to store a dataset.
- Information is organized in rows and columns
- Dataframes simplify common operations, like sorting data.

Series:

- A pandas object used to create dataframes
- Seen as a one dimensional list of data.

  Its as if

Indexing into Dataframes

Main techniques

1. df.loc [ ]
2. df.iloc[ ]

19/9/23

Selection : the process of accessing a subset of a dataframe. You can select subsets using loc and iloc.

Filtering: Selecting values of a dataset where certain conditions are true.

Popular pattern;
Df[condition]

Combining data frames

Three techniques:

Concatenate: Naively combines along an axis

Merge: combine through shared column

Join: Combine using shared indices.

Inner join, outer, left, right,

Distributions:

Distributions are graphs that tell us about some characteristic of a population. Mean and median are important parts of these graphs!

Data visualizations.

A graph or picture that helps humans understand important patterns in a dataset.

Seaborn Fundamentals

A python library that can make visualizing data easier.

Bar Charts

A bar graph is a graph type that uses bars to depict a value associated with a category.

Histograms

A graph that shows the frequency distribution of a variable in a dataset
Sns.distplot(), sns.distplot(), sns.ecdfplot(), sns.kdeplot(), sns.rugplot(), sns.jointplot().

Scatter plots

A graph that uses points to show the relationship between 2 quantitative variables in a dataset.

Scatterplot(), lineplot(), stripplot(), swarmplot()

Techniquesn to collect data
Observe a sample-
Survey a sample- ask people directly. Also with questionnaires.

Experiment on a sample
Usedata someone else has responsibly collected

Sourcing Digital Datasets: API Requests:
The act of using HTTP requests in order to access datasets collected and maintained by other people.
Common API requests:
Get - see
Post - modify data and create resources
,Put,- modify data
Delete- remove existing resources

Sourcing Digital Datasets: Web Scraping
The act of extracting data from websites using the structure of its HTML.

HTML

Hypertext Markup Language is used to display content on a web page. Look for angle brackets <>

General page structures:

Two major sections:
Head- contains important metadata

Body- all content that is seen on a page.

Tag structures:

- HTML is made up of tags
- Each tag does something different
- Most have an opening and closing tag.
  Eg- <h1>Content</h1>

Tag attributes

Some tags need more information in order to work. In order to do this, you need to use attributes.

Eg, <img src = “url”>

**the image tag doesn’t have a closing tag.*

Important metadata tags and attributes
Tags;

- <title>?</title

Accessibility:
We want to make sure our websites are accessible to as many people as possible

Accessibility considerations:

- Low- bandwidth users(users with unstable internet connections)
- Visually impaired users(use screen readers to surf the web)
- Low English proficiency users(content displayed must be understandable)

General structures

Two major sections

- selector
  Targeted Html tag
  - □ General
  - □ Class
  - □ Id

Class selectors

- Used to style a subset of the html tags used.
- Has more priority than the generic HTML tag selector
- Start selector

ID selector

- Used to style a single html tag used
- Has the most priority of all selectors.
- Start selector with a hashtag(#) in order to use.

The box model
Every HTML tag makes a box.

JavaScript- is the programming language of the web. It is used to give your website behavior.

D3:
Is a javascript library that is used to create beautiful and interactive data visualizations.

Purpose of an introduction

- Introduces audience to question
- Provides context to the topic or the subject you're talking about
- Hooks the reader and sets the tone
- Outlines how you're going to answer the question
- Provides important background information that is tailored to your specific audience.(i.e define important terms.

Characteristics of a great/strong introduction

- Includes at least one piece of background information
- It includes a hook
- Simple structure. concise/shot.
- Includes your question. Provides a "mystery" context – we don't know the answer but we'll discover together
- Stays on topic and builds credibility.

What to avoid when writing an introduction?

- Doesn't stay on topic! It's off topic or super vague. Using confusing statements.
- Making it too long. Overwhelming the reader with too much information.
    - ☐ Avoid showing data too early. Stick to anecdotes instead or tease with preview(make sure it is related to your question).
- Boring. Use emotion, think about the tone of your writing!
-

Data Stories;

Central Insights:

- Gives story a direction. It helps the writing process because you know where you end goal is.
    - ☐ Maintains focus for the reader
- It helps the reader know what to do, which is the point of our data story.
- Without a central insight, there is no point to writing a data story.
- With a weak central insight, your reader might conclude something that you don't want them to.

Characteristics of a great/strong central insight

- It is a clear and comprehensible sentence or two.
    - ☐ Think about grammar and uses engagement techniques(tone of voice).
- Actionable! Clearly outlines a specific next step
    - ☐ It's important to make this relevant to an audience member. This is something that builds in the story so that it helps make the central insight relevant
    - ☐ Mayble outline the consequences of not taking action.

What to avoid when doing central insight

- Not clear. Your grammar is off, enough that readers can't understand what you wrote.

- Don't include the raw analysis. Instead, simplify the analysis in a way that makes that next step/significance easy to understand.
- Too long, not engaging. Too many central insights. How do we pick which one to tackle!?
- Does Not relate to your question or analysis. It's off topic.
- It doesn’t answer the question Fully. Leaves conclusion up for interpretation.
- Don’t show bias or lie.
- Too general. If it's not specific, people won’t take action.

List the types of narrative structure you might consider using in your final project

- Freytag’s pyramid
    - □ Great for stories with clear conflict and resolution
- Aristotle’s tragedy
    - □ Great for simple data stories
- Campbell’s hero journey
    - □ Complex stories that have an obvious “hero” perspective.

Types of story points

- Change over time: shows how a variable changes over some period of time
- Relationship: shows how variables impact each other.
- Intersection: shows when variables surpass or fall below other variables.
- Project forward - predicts what will happen in the future.
- Compare and contrast: shows how two variables are different.
- Drill down: start high-level, then talk about specific subset of data.
- zoom out - start specific, then talk about how it applies in a general context.
- Cluster: describes what data have in common.
- Outlier: describe what is not common or unusual