Final Production Deliverables and Submission Requirements

Heading:

 <b>

Body:

<B> TAKE A CLOSER LOOK WITH THE TOP 10 DISHES

- ☑ ~~Your project needs an engaging and clear Title. (Unit 1 Assignment *does not count*; it should reflect your chosen data story.)~~

In your code using markdown or HTML (as relevant):

- ☑ ~~Write a full introduction to your problem. This needs to be written to engage a general audience (think about tone and clarity—do not assume reader familiarity with your work or the assignment). Work from the answers to your questions in Step 4. A, but also be sure to speak from experience (you've finished *building* the visualizations now and you've gotten feedback from your instructors and user evaluations). Be sure to include the following elements (*the story outside your story*):~~

As food critic Jonathan Gold put it,

"Food is a pretty good prism through which to view humanity."

And perhaps nowhere else is a better microcosm of the human spectrum than New York City. I agree with Gold. I love to navigate the world through what we eat. You can learn history, anthropology, economics, language, and connection through food. Generally, I am a proponent of learning about these huge phenomena through specific and interesting means. When looking for a data set, I really struggled to find anything that I thought was both significant and casually intriguing. It's hard for me to be invested in a project if I don't care about what project I choose to pursue. However, I stumbled across the New York Public Library (NYPL)'s "What's On the Menu" project, which is a collection of 45,000 menus from New York City dating from the 1840s to the present. They bring together volunteers who transcribe this data so the public can access it.

I wondered, what has been hidden in these archives and what can be distilled using visualization and other analysis skills we have been using. Originally, I wanted to answer questions such as,

- What foods or ingredients became more popular over time?"

- Are there menu trends that correlate with historical events or cultural shifts?
- Can you identify regional differences in menu offerings?
- Are there specific dishes that are more popular in certain neighborhoods?
- Have menu offerings become more diverse over time?
- What are the most popular dishes over different decades?
- Are there any dishes that have consistently been popular?

When I got the data, I saw that I was going to have to choose the cleanest dataset due to the sometimes chaotic results of sourcing all your entries from volunteers. This was "dish.csv", which was quite clean but was more limited in its columns, forcing me to adapt to a less specific set of questions. In terms of potential harm, at the front of my mind is the ethical question of using data that other people have put a lot of time into. I feel somewhat exploitative, but also, on the other side using this data might be inaccurate due to the lax environment it was curated in.

But why could this be important? Well, I believe artifacts like menus are severely underrated sources; it might not be what you think of when you think of "data." However, there is much to be gained from menus. With price you can measure inflation; with names of dishes, we can discover diversity in language and trends in naming conventions. Trends can be revealed by using the first and last appeared years. Most importantly, menus allow us to think about history as a living thing with living people who dined, ordered cocktails, and socialized in the same buildings and with the same ingredients we do.

- ☑ ~~Introduce your audience to the larger problem by employing narrative elements to engage your reader, defining terms, and exploring the problem's significance.~~
- ☑ ~~Describe the process of finding your data set, not just the happy ending, but share some of the pitfalls of the process.~~
- ☑ ~~Outline how and why your story/question changed from the start of this process.~~
- ☑ ~~Explain the benefits of visualizing this data.~~
- ☑ ~~Explore the potential harm of this same visualization.~~

- ☑ ~~In the next section, briefly describe any transformations of the dataset you needed to perform to get the data into the format needed for your visualization.~~

Alright, as I mentioned, this dataset was fairly clean. The most obviously messy thing to me was the names of the dishes. Unlike a previous dataset I was using, there was no pattern to the name string so it was not easy to use it in a way that would be consistent and make sense. As you will see later, after consulting with both Tyrica Terry Kapral and Dominic Bordelon in Hillman, I decided to use NLTK to sort through that column, which I hadn't been planning on. Besides "description," which was completely empty, there were no empty columns technically. The ones that were problematic were simply populated with zeros. Luckily, I think I avoided circumstances where this would significantly impact my data (for instance I didn't make a zero decade when determining the top three dishes per decade and when I determined the least popular, I made it >= 1 because having a meal that appears zero times simply does not make sense. This made me question the entries, but we had to move on.

- [ ] Now we're getting into the visualizations (*the story within the story.)* Provide a short caption/paragraph "moment" to provide context as appropriate. Consider how your captions help clarify what you are demonstrating through your visualizations–and how you can more effectively connect the information shown (clearly) in your visualization to your audience.
- [x] ~~Perhaps this is a good moment for a quote. Perhaps not. It depends.~~
    - [x] ~~(Worried about style and tone? Work with your instructors, review sample stories, and consider your workshop feedback. Quigley will say "*Write with strong verbs and cut dead wood.*" For the "technical" side of this caption moment, you can do a few Google Scholar searches to see how figures are included in research papers; and reference both the Purdue OWL's section on Data Visualization and on using figures in APA or IEEE citation style. Using APA will increase the likelihood of publishing your document in a digital composition journal. Using IEEE will position you for journals related to computer science.)~~
- [ ] Coda: Write a few paragraphs at the end of your Jupyter Notebook or HTML file further explaining your method and your particular visualization approach. Be sure to provide examples / specific details to support any claims you make. Answer the following:
    - [ ] How did your design decisions work for your test audience/classmate user feedback, and how did they fall short?
    - [ ] If you didn't do so above, explain some of your "failed attempts" at visualization. How did you change tactics to improve outcomes?
    - [ ] Where might you revise in another round of iteration to better evidence your data's story?
    - [ ] Are there any stories your current work occludes, marginalizes, or minimizes? How might an alternative direction of this work include or highlight those stories?
    - [x] ~~What did you like best about learning how to create data narratives?~~

Thank you for reading my project!

This was my first time ever working on something like this, and my first time really working in Python. I was so excited to do so, because I love data journalism and data in general. I think I would go into data science if I was slightly more mathematically motivated :) It took a long time to find a dataset that spoke to me, but it was so fun once I got started. Perusing other people's examples of cool visualizations was so fun, as was talking to experts at the

University to see what is possible. I also enjoyed using HTML to make it look more appealing and slightly more similar to an actual data story on a reputable platform.

That being said, I wish I was able to delve deeper into the field and experiment more with different datasets, especially ones that have geospatial elements or more precise dates. Not to mention more consistent entry of the data itself. Right now, I think my story is interesting but decently basic, with the majority of visualizations being bar charts. I've also never taken a statistics class, so the learning curve of asking the right questions about the data in the first place was a bit of a fun challenge. Also, I wasn't anticipating using natural language processing tools, and I was amazed at the capacity at which I could work, even with such a limited skillset.

Given the scope of the data, I am proud at what I was able to achieve. In another iteration, I would have loved to work directly with the librarians and archivists who have made this possible.

- ☑ ~~Provide a YouTube link. Record a short video walking your audience through your data story. (This is taking the place of an in-class presentation; should be roughly 2-3 minutes long, and present an overview of your work (perhaps some background information not self-evident in the draft itself such as "Here's why I choose visualization types X and Y", "I was going to do Z but dropped it because..") Include at least one specific question you have for your workshop partners about your current data story draft.~~

https://www.timeout.com/newyork/blog/10-awesome-vintage-menus-from-the-new-york-public-librarys-digital-collection-021816

For each visualization:

1. Most Common Words

What's in a name? Well, a lot. The "name" column of this dataset intrigued me first. It was eclectic, from complex names such as "Bouchees aux Macaroni" to "celery." Given that this dataset includes decades up until the 60s, I was interested in how the naming of dishes has changed. This is every word in the name column and how many times they appear, listed from most to least.

- Find a question
- NLTK English dictionary-- what's not in there?
    - How to filter out

- What are the most common or least/unique *key/in title* ingredients
- Hundreds of counts
- Problem: the tokenizer takes words, and sometimes they need to be groups of words
- How to do ngrams
    - number of words or tokens
- Key ingredients...
    - Pre-processing tools: open refine
- NLTK - tokenize all words and make into new data set
    - or do with pandas
- commonly mentioned countries with nltk or some other library?
    - countries dictionary?
    - country stems?
- ANALYSE STOPWORDS
- take first 3 numbers and add zero after that

Elizabeth Liang Hudak
Unit I - Data Storytelling

Questions and Stuff - 1 February

Step 4. A Plan / conduct your project.
In the planning / early stages of your project, be sure to work through the following. You will need to work from the answers to tell the story of your research journey.

1. First, Assess your chosen data
    a. What did you expect to find, what wasn't there?
        i. I really wished there was a date that the AirBnB started listing, or some variable that let me know when the property popped up.
    b. How did what you DID find change your thinking about the data?
        i. The latitude and longitude made me wonder what made properties more expensive: did it have anything to do with location? were those neighborhoods wealthy to begin with or was it a different pattern?
    c. How will the realities of the data be foundational to your visualization?
        i. To be honest, I am not certain that I will use this dataset. I am really indecisive and have put a lot of pressure on myself to make it good (I'm a huge fan of the Pudding, too)
        ii. I would like to introduce other, simpler visualizations of different datasets (like census, tourism) in order to offer a comparison

2. Working from part 1, try and compose a draft statement of your research question and visualization approach (use this statement to share in class workshops / with instructors during office hours/discussion boards, and as you iterate).
    a. Research Question: What factors influence the popularity and pricing of Airbnb listings in Asheville, and how have these dynamics evolved over time?

3. Consider: What are the parameters of the story you're going to tell?
    a. What is the scale? Scale is quite small, which I appreciate because I think it can be more personal and easier to get invested in
    b. What is the focus? Focus on tourism but also I guess on property values/desirable areas in Asheville?
    c. What are the elements of this story? This narrative explores different facets of the Asheville Airbnb dataset such as property and host details, temporal trends, and spatial patterns. Visualizations highlight factors such as room types, geographic distributions, and host characteristics.
    d. Why does this story matter? (What is the "point" of your story?) Growing up in the South, I have seen firsthand how tourism has increased and even changed in nature in the Carolinas, especially as the towns get more trendy. I really want to show this and Airbnbs seemed like a reasonable way to visualize it in a targeted way.
4. Have answers to the following questions:

a. Who is your target audience? People who can relate to this phenomenon in any way.
b. Who will benefit from your visualization story? I think that, since the pandemic era boom in Airbnbs and people fleeing from big cities, this kind of story hasn't been talked about and yet is still quite relevant to these smaller communities that experienced such a big change in the last few years. I hope my story will keep these damaging trends in people's minds.
c. What is the context of your data/story? (Think about *Design Justice*, here)
d. Who are you trying to convince with your message?
e. What narrative strategy should you employ? The best visual stories I know are backed up with equally strong written contexts, so I am looking forward to creating a tone of voice that I can frame the data viz with.
f. What are some potential unintended consequences of your visualization (i.e., can it be misinterpreted, might it cause harm, and if so, what are the implications?)
    i. Consider both those you want to benefit from and who might reap benefits, even ones you may not agree with, beyond your direct audience appeals.

Pitch Day - 30 January

Pitch 1: NYPL's Menu Dataset

- Questions to be Asked:
    - What foods or ingredients became more popular over time?
    - Are there menu trends that correlate with historical events or cultural shifts?
    - How have major events (e.g., wars, economic changes) influenced menus?
    - Can you identify regional differences in menu offerings?
    - Are there specific dishes that are more popular in certain neighborhoods?
        - Have menu offerings become more diverse over time?
    - What are the most popular dishes over different decades?
    - Are there any dishes that have consistently been popular?

Pitch 3: NYC Popular Baby Names

- Questions to be Asked:
    - What is the distribution of counts for the given baby names in the dataset?
    - Are there a few names that dominate, or is the distribution more even?
    - Not sure if I can answer this successfully, but somehow can I see if overlap between non-POC and POC names is growing? Which "direction" is it growing (are POC using more names associated with white people or vice versa?)
    - How is the popularity of names distributed among genders?
        - Are names shifting gender? Is there more similarity than there used to be?

- Explore the relationship between the count and rank of names.
- Are there names that are consistently high in rank but have lower counts, or vice versa?
- Could I do a letter-by-letter analysis somehow and see common trends linguistically?

Pitch 3: Tuition Assistance Program (TAP) Recipients & Dollars by Income, Age Group and Program Information: Beginning 2000

- Questions to be Asked:
  - How does the distribution of TAP recipients vary across different sector types (e.g., public vs. private)?
  - Is there a difference in TAP dollars awarded between different sector types?
  - Analyze the distribution of TAP recipients based on their financial status (e.g., financial independent vs. dependent).
  - How does financial status relate to the amount of TAP dollars awarded?
  -

Pitch 4: Legally Operating Businesses

- Questions to be Asked: