NYC Data Analysis

Andrew Piatetsky
Period 4 Intro to Computer Science 2
TNPG - Andrew Piatetsky Software Development Company™

Page 1

Section 0 -

I want to create a program that takes a large data set of public job openings and use it to find meaningful patterns. This project at its most basic is very simple, but actually creating the graphs and organizing the data will be very challenging. I'll also have to learn more about the matplotlib library as well as more about json files.

I want to create this program because it sounds very interesting and also sounds applicable to real life, as statistics are important in creating government policies and they influence our everyday life even if we don't realize it. Furthermore, I think finding patterns in public job openings can show how well the economy is doing and also how much demand there is for different types of jobs.

My program will work very similarly to our homeworks where we converted csv files to dictionaries in python, however I will be using json files through an API instead. This part won't be the most challenging though, the most challenging aspect of the project will be representing the data in a clear and powerful way through graphs.

The program will work in the following way:

It'll import a large data set in the form of a json file, then convert that json file into a list of dictionaries. Then, I will write a few functions that will clean up the data. Cleaning up the data means that it will filter out all of the uninteresting data such as various ID numbers and other government classifications that are not useful to me. Finally, I will create functions that turn the data into specific data structures that can then be used to create graphs and diagrams. For example, to create a histogram you need to input lists of numbers or strings instead of dictionaries. Then, I will upload the graphs to the website, either manually or automatically, and have appropriate descriptions about them and their importance.

This all sounds fairly straightforward, but the project will be mostly about graphs and different patterns which I have to find and then turn into visual data. There are still many unknowns such as which type of charts and graphs I will use and what statistics I will show.

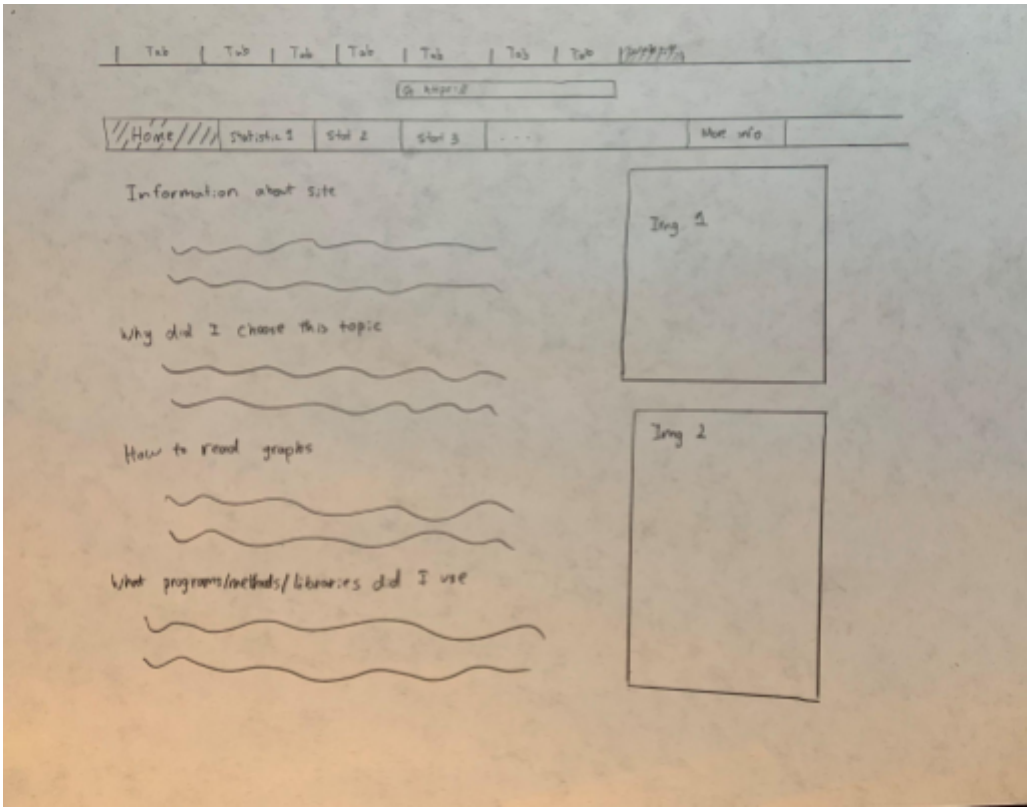
The only libraries that I will be using are matplotlib, numpy, and pprint (I'll use pprint for myself to make the data easier to see). We went over these in class but after looking at their websites for just a few hours there are still so many things that I can learn.

Section 1 -

My website will have a similar look and feel to my site.html project at the beginning of the semester because I really liked the colors and navigation bar that I used.

My website will have a few pages (the number of pages is subject to change).

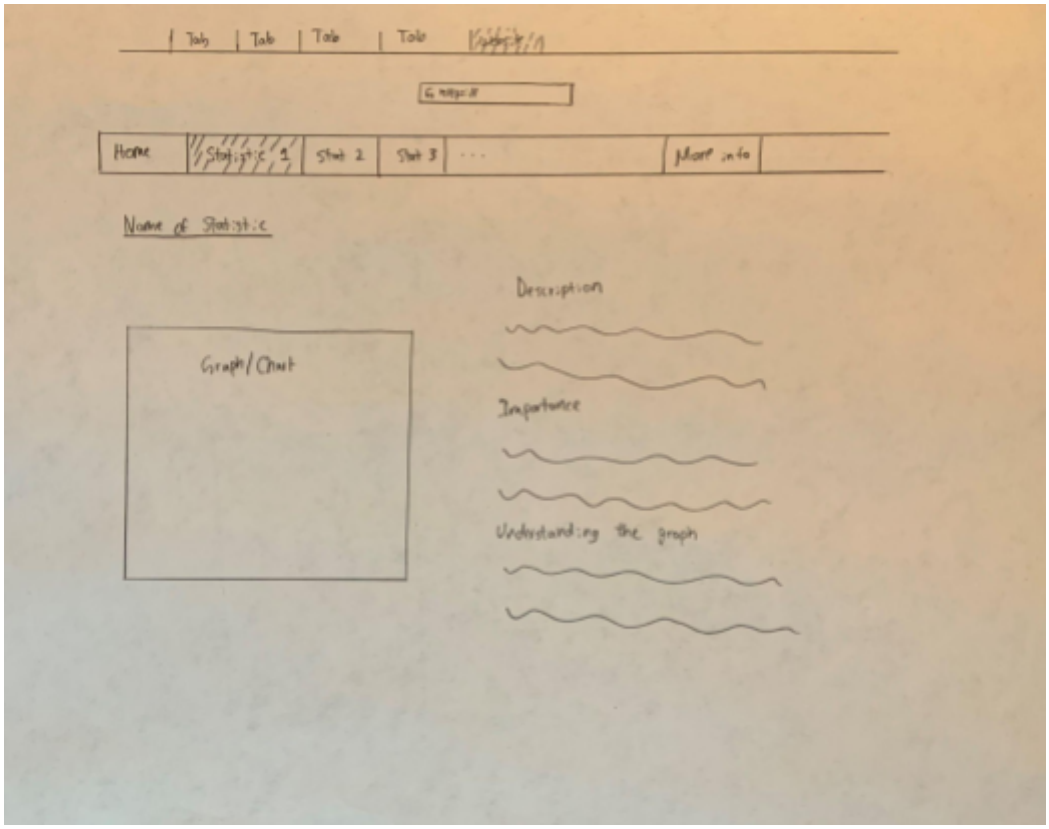
Image 1 - Home/main page



The first page will be the home page/main page where I will have some basic information about the site, libraries that I used, why this topic interested me, etc.

Then I will most likely put some images of maybe the data or something else relating to the assignment. I'm not sure yet, but the page won't just be words.

Image 2 - Statistic Pages (i.e pages with graphs)



The next few pages will include their own unique statistics and graphs.

I already know that I will have at least one pie chart, but all of the other charts and statistics are unknown at the moment.

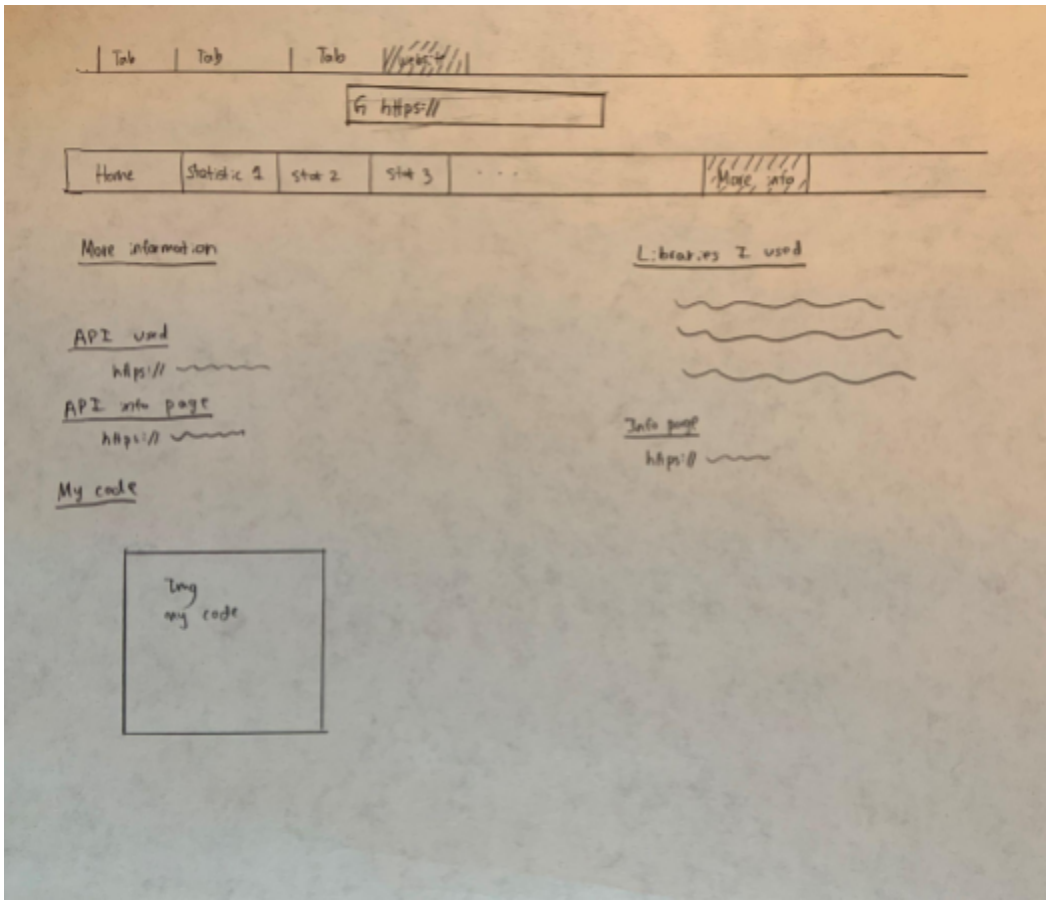
The graphs will all have clear labels and will be color coded as matplotlibs dictates. I don't know what the graphs will look like because I still have to play around with the library.

I will also describe the importance of the statistic, how to read the graph and a general description of the topic and chart.

As mentioned, there will be a few of these "statistic" pages. The exact number is unknown for now, but I estimate around 3 or 4.

This image is just an outline for one of the statistic pages, but they will all follow a similar format and will look very similar as well.

Image 3 - More info



This page will be the final page and I imagine it's navigation bar link will be on the complete right side of the bar.

This page will just have some links to different resources used such as the data set, API, and also various libraries that I used in my code.

I will also have an image of my python code on this page so that people can see just by viewing my website how I did my project.

This will be the most simple page, but I still think that it is important to disclose everything that I worked with.

Obviously there will be lots of code that'll be creating all of the graphs, but this is just an outline of what my website will look like visually.

Page 2

For this project I'll be using an API to get the data. It is a SODA API meaning that it is a json file and that it is free to access and use as I please.

Link to the API info page: https://data.cityofnewyork.us/City-Government/NYC-Jobs/kpav-sd4t

I will have to learn how to access the json data, but I don't think it will be that difficult.

Task Breakdown - I'm working by myself for this project so I will have to do everything.

Timeline -

June 1 - June 5, start figuring out how to use the API and how to convert it to useful data (I can't do too much because I have AP tests)

June 5 - June 7, create one graph.

June 7 - June 11, continue working to create the rest of the graphs and start building the website (I can't do too much either because I have AP tests)

June 11 on, finish the website and finish all of the graphs.

Priority list -

1. Converting json into a usable form - make sure I can work with json
2. Interpreting the json file - turning it into lists, dictionaries, etc
3. Finding statistics and patterns to graph - looking for interesting patterns
4. Creating the basic website - Nav bar, css file link, some basic information
5. Creating the graphs using matplotlibs - using matplotlibs as a resource to create graphs of interesting patterns
6. Finish the website and add graphs and information.
7. Adding more to the website - automatic graph uploading from API
8. Finding more patterns and making more complex graphs - have to look around matplotlibs more to achieve this

***I'm working by myself so if I find that something isn't feasible, I won't include it in the final build. The beginning is definitely feasible and so are the other steps up to 6. The other ideas (7 and 8) are the ones I have not thought about, but seem interesting and fun to tackle, if I have time.