I want to work on the Nobel Prize datasets for the final project. The main question I want to answer is: which category has won the most award prize? I plan to answer this question by dividing the Nobel Prize winners into their suitable category groups, then adding the total number of prizes in each category and compare with other categories. Other questions that might come along during my data analysis are how many prizes were given out in each category, which country had won the most, whether some people declined a Nobel Prize, whom the prize went to, etc.

The code components need to be able to handle the data types that I'm planning to use. I'm now thinking about TEXT or CSV files and HTML text to use. I might need some functions that handle calculations. Here is the pseudocode of functions that might be on my project:

```
# function to calculate how many male and female laureates there are.
def genderNum(sex):
    # keep track
    countMale = 0
    countFemale = 0

    for gender in sex:
        loop through and sum up the total number of genders in each sex
        return the total number of each sex

    # then compare and find out which gender has more laureates
    if(totalFemale > totalMale):
    print(“There are more female Nobel laureates, there are { } of
them”.format(totalFemale))

    else:
    print(“There are more male Nobel laureates, there are { } of
them”.format(totalMale))

# function to find the person who had won the highest prize
import requests
from bs4 import BeautifulSoup

# let’s grab some data
nobelPrizeHTML = requests.get(“http://nobelprizeinfo.com/”).text
# parse it to BeautifulSoup
nobelPrizeSoup = BeautifulSoup(nobelPrizeHTML,”html”)

nobelPrizeTag = findall(“award prize”)
```

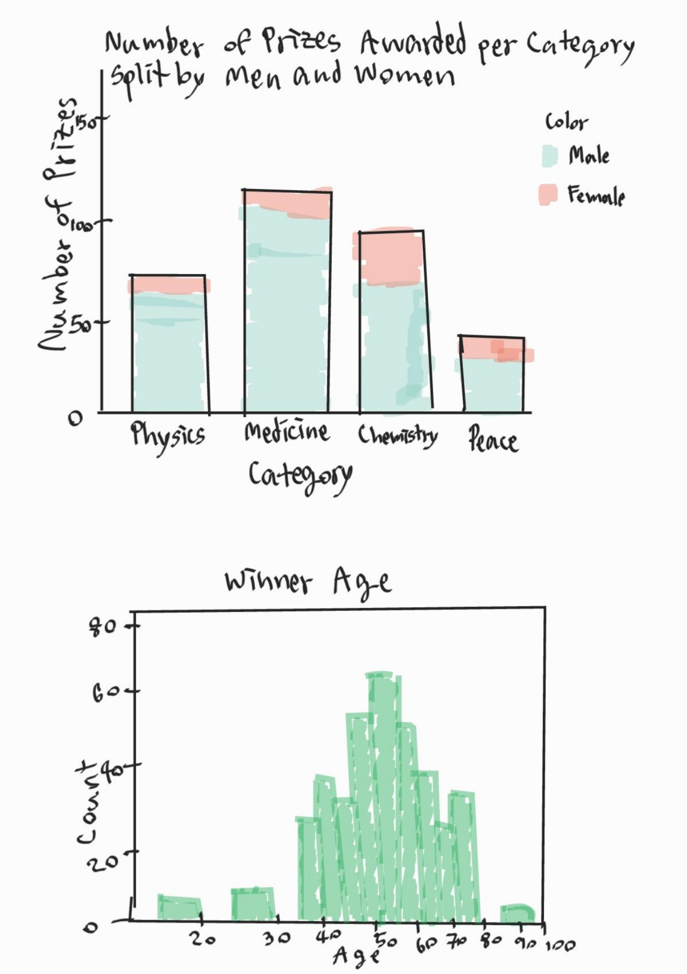
```
def findHighestPrizeWinner(nobelPrizeTag)
        highestPrize = “ ”

        for prize in nobelPrizeTag:
        # convert string to int
        if len(int(prize)) > len(int(highestPrize)):
                highestPrize = prize

        print(“The person who had won the highest prize is { }. The award prize is $
        { } dollars”).format(highestPrize,len(highestPrize))
```

The obstacles I might encounter while working on this project are the data type I choose to use, calculation, and visualization. Depending on which HTML web I choose, I know from the past homework that we have done that it can be time-consuming to get the right tags with the correct data you want. Especially if there are many nested tags, secondly, my questions involve math, so some functions might be tricky to do the calculations. Finally, my visualization needs to make sense of the big picture I am trying to convey to the audience. I must eliminate unnecessary visuals that do not contribute to the project and ensure that it is manageable.

Here are the sketches of how I think my visualizations might look:


Number of Prizes Awarded per Category
Split by Men and Women
Number of Prizes
Color
🟩 Male
🟥 Female
Physics
Medicine
Chemistry
Peace
Category
Winner Age
Count
Age