

CS50 Assignment 2: Variables with LBA

Part 1: Variable Selection [#variables]

Select a neighborhood in San Francisco. Visit this neighborhood and spend at least 30 minutes exploring the neighborhood to find your variable.

Important notes:

- The variable must be something that can be measured on each block in a specific 10+ block area. This means you need at least 10 different measurements (one for each block).
- You must be able to calculate the mean, median, mode, and standard deviation of the variable.
- Be clear about how your blocks are defined.
- Get creative! Try to choose an interesting and informative variable.

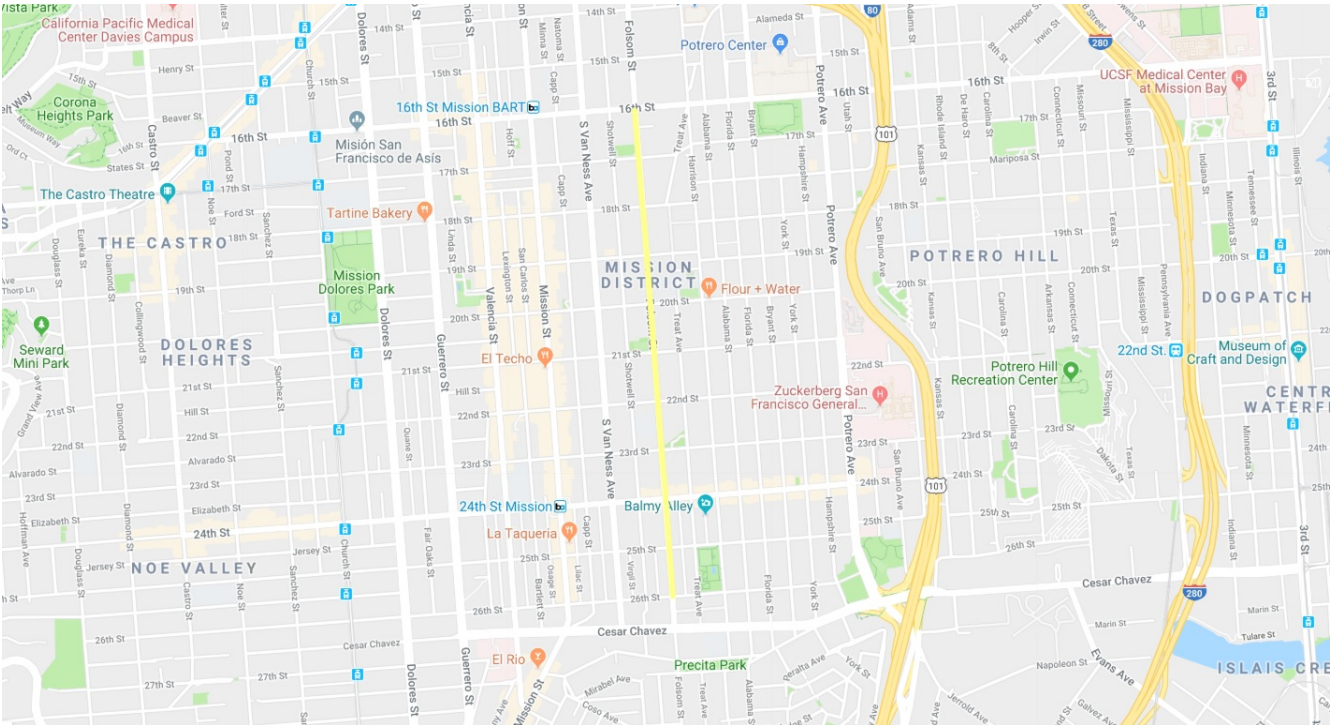
1.1 Define and operationalize your variable here. Describe how you selected your variable. Specifically identify the type of variable, and whether you will be measuring a total, proportion, or average. Also identify the units it will be measured in and explain in detail how you will measure it. Make sure that your explanation is clear enough that another student would understand how to make the same measurement. Give the address of the 10 or more blocks where you will conduct your measurement and provide an image that clearly identifies these blocks on a map. (<150 words)

1.1: I had been to Mission District several times before, specifically Folsom Street. I noticed that there were a lot of houses that had outside stairs leading to the entrance. At first, I decided to choose this as my variable. However, this variable was uncommon in the blocks where there were no houses. So, I walked around a couple more times and found another variable. I saw that there were driveways both in the businesses and houses in the region. Hence, I chose driveways as my variable. In this case, the number of driveways is a discrete and quantitative variable. I will measure the total number of driveways. This variable is unitless because I will count it. The way I will measure this is by walking through the block and counting the number of driveways. The address of the ten blocks is Folsom Street from 16th Street to 26th Street.

In [20]:

from IPython.display import Image
Image("Folsom Neighborhood.jpeg")

Out[20]:



1.2.1 Describe a scenario in which your variable could be an independent variable. What could be the dependent variable(s)? What are some possible extraneous or confounding variables in this scenario? (<150 words)

1.2.1: A scenario in which my variable could be an independent variable is if we want to know how the number of driveways affects the number of "No Parking" signs in the neighborhood. The dependent variable could be the signs because it would depend on the independent variable which is the number of driveways. It is a common practice to have these signs in front of the driveway. In this scenario, a confounding variable could be the number of driveways that are no longer in use. Since they are no longer in use, the number of signs would be affected because there would probably be less than expected. In this scenario, an extraneous variable could be No parking signs that are in other places where there are no driveways. For example, there could be a no parking sign in front of a house even if they don't have a driveway.

1.2.2 Describe a scenario in which your variable could be a dependent variable. What could be the independent variable(s)? What are some possible extraneous or confounding variables in this scenario? (< 150 words)

1.2.2: A scenario in which my variable could be a dependent variable is if we want to know how the number of garages affects the number of driveways. The independent variable could be the garages in the neighborhood and the number of driveways would depend on them. This is under the assumption that every garage has a driveway that leads up to it. In this scenario, a confounding variable could be garages that share a driveway. For example, if we have two garages, we might have only one driveway. This would affect the number of driveways because even if there are several garages, there could be many shared driveways. In this scenario, an extraneous variable could be a new business put in place where there used to be a garage and a driveway. There would be still be a driveway but there would not be a garage anymore.

Part 2: Estimation and Measurement [#estimation]

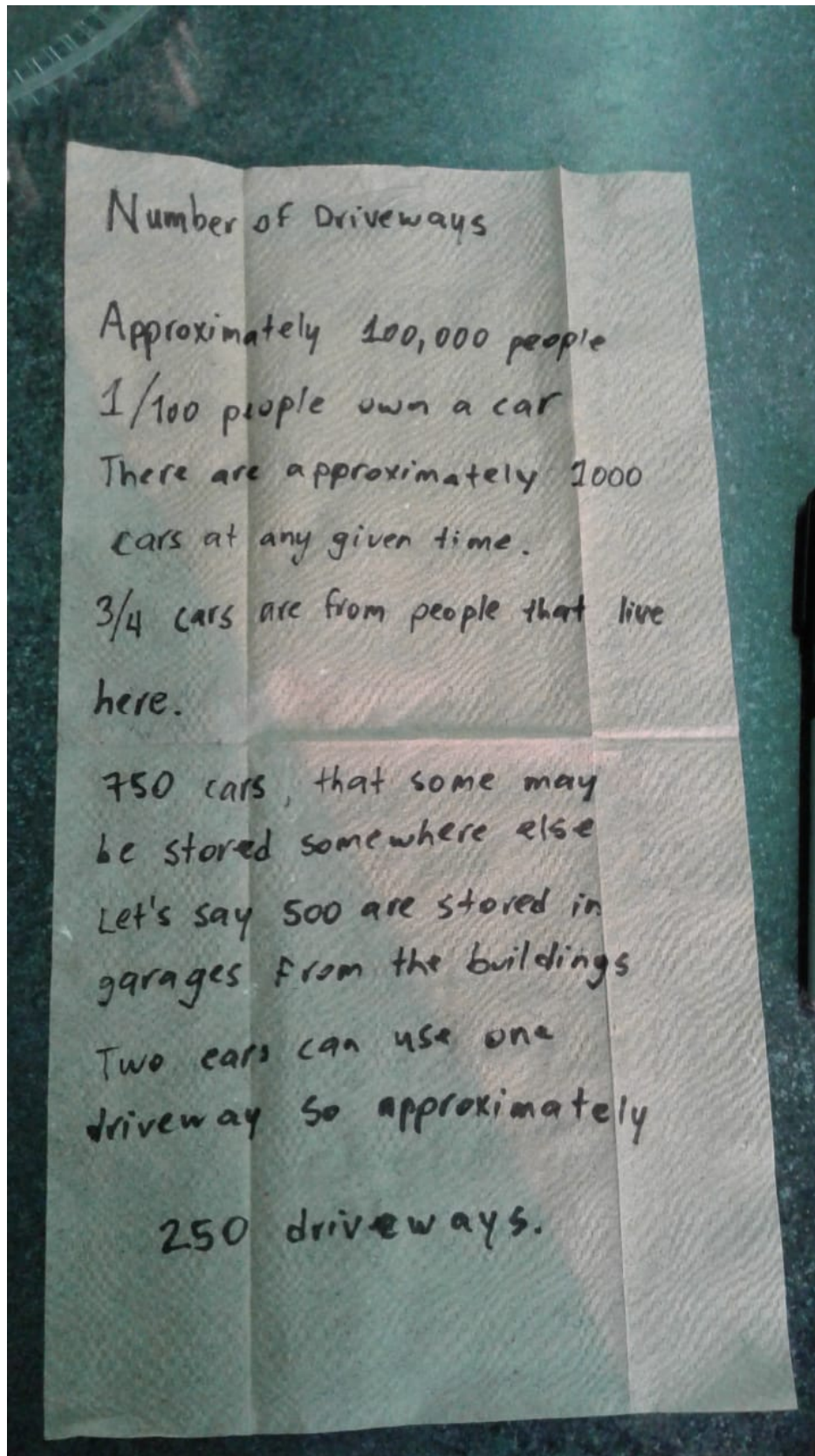
2.1 Go to a Cafe in the neighborhood of your choice to produce a Fermi estimate of your variable. Use a napkin at a cafe to begin your Fermi estimate. You may not (yet) make any measurements. Your estimate should aim to involve at least 5 steps where you compute intermediate values. You will have to describe each step clearly, show your work, state any assumptions you're making, and discuss whether your answer seems plausible (but it's not necessary to do so on the napkin; see step 4 below).

2.2 Take some photos to document this experience. You must include:

1. A photo of your "back of the napkin" estimate (it can and should be quite rough at this point). You will properly format the calculation later.
2. A selfie in the cafe in which you constructed your Fermi estimate. Clearly show your face, your Fermi estimate, and some of the interior of the cafe.
3. A selfie outside of the cafe showing your face and the exterior of the cafe, including the name. Bonus points if you are also holding your completed Fermi estimate in the photo.

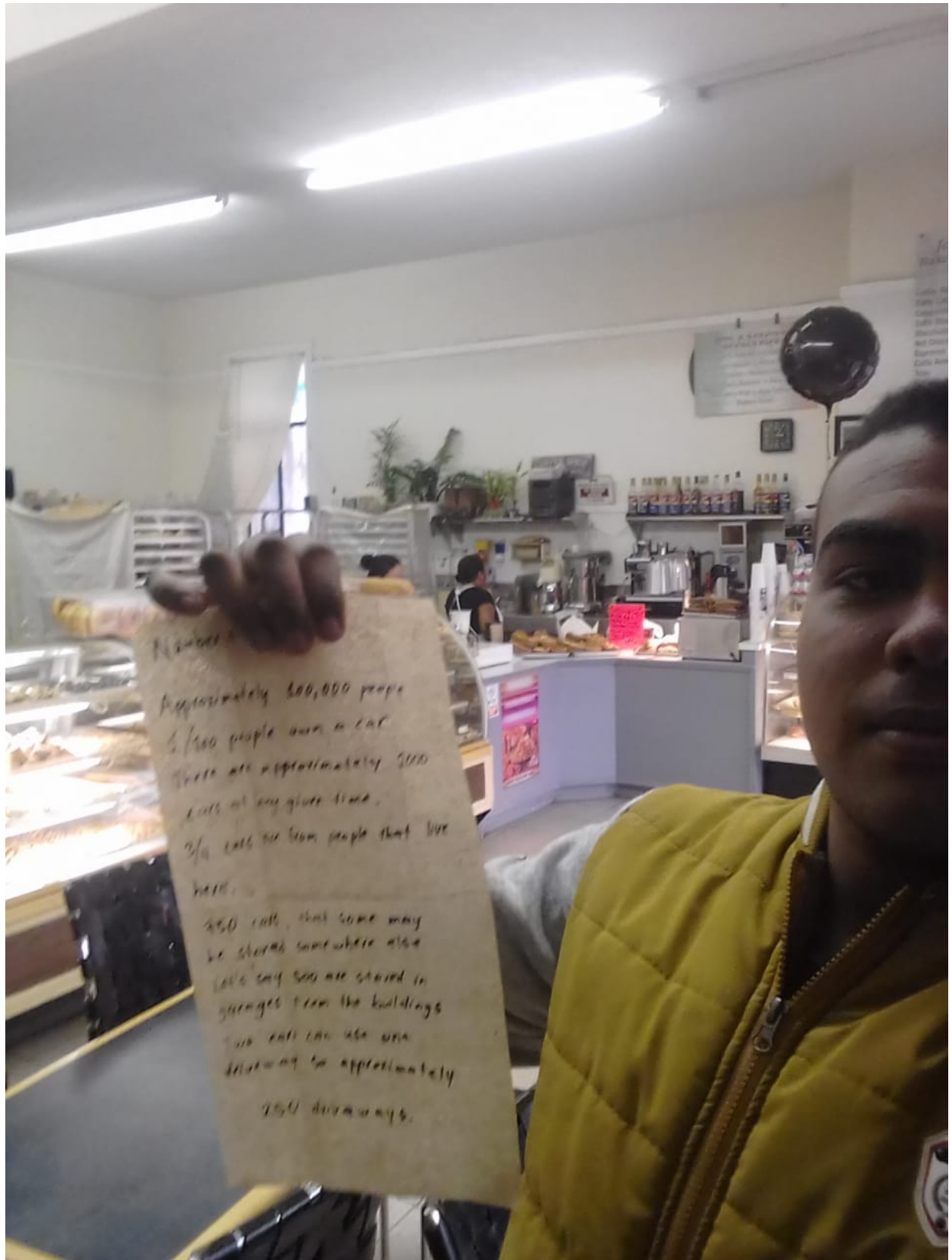
In [21]: Image("Back of the Napkin.jpeg")

Out[21]:



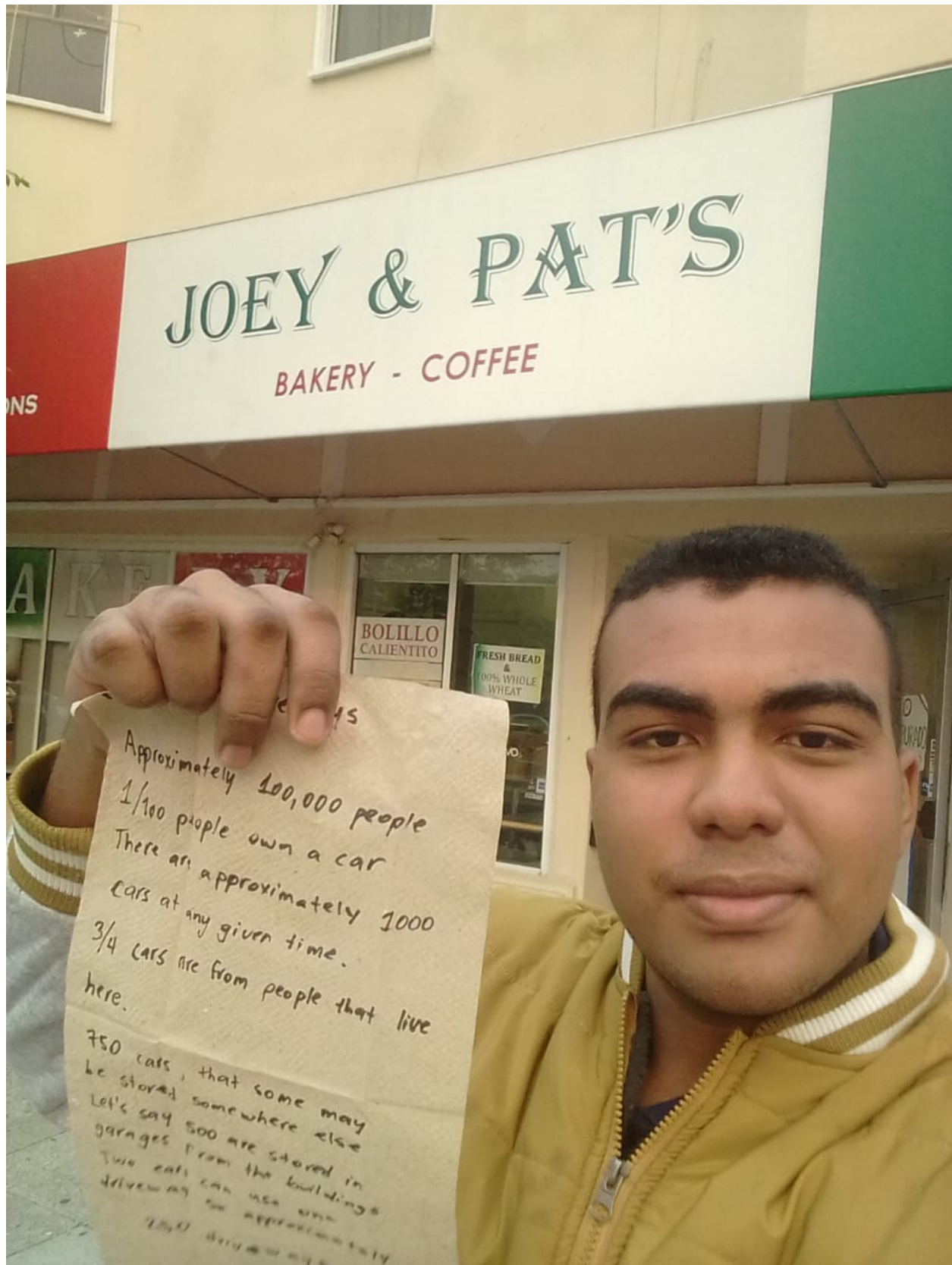
In [10]: Image("Selfie Inside.jpeg")

Out[10]:



In [11]: Image("Selfie Outside.jpeg")

Out[11]:



2.3 Typeset your full estimation in the Python notebook. Here, be sure to clearly explain all steps, justify all assumptions, and comment on whether the answer seems plausible.

Number of Driveways

There are approximately 100,000 people that live, work, or transit this part of the neighborhood. Out of those people, 1 out of 100 or 1% own a car. This means that there are about 1,000 cars. Since some cars may be from visitors or workers from the area, we don't need those for our estimate. We will assume that driveways are mainly used by people who live here to store their cars in garages. We can say that 3 out of 4 or 75% of the cars are from people that live there. This means 750 cars that may be stored somewhere else or parked in front of the house. We can say that 500 cars are stored in the garages. We can also assume that one driveway can be used by two cars. So, we end up with approximately 250 driveways.

2.4 It's time to collect your data! Once again, take some photos to document your experience. Include at least two photos of your variable collection process. At least one photo should include your face and the variable you are counting.

In [13]: `Image("Variable.jpeg")`

Out[13]:




```
In [14]: Image("Selfie with Variable.jpeg")
```

```
Out[14]:
```



Part 3: Analysis

3.1 Analyze the data in Python [#algorithms]:

3.1.1 Use any method to import your collected data into Python. You can simply type the data directly into a Python list or numpy array. Or, you can put the data in a Google sheet, export to a .csv file, and import into Python. Print your data here.

```
In [132]: import pandas as pd
#This line of code imports the pandas library as pd for better code writing.

collected_data = pd.read_csv('Driveways Folsom.csv')
#The variable is assigned to a pandas function that reads the .csv file.

Drwys = collected_data['Number of Driveways']
# The variable is assigned to get the List of the values below "Number of Driveways".

collected_data
#Calling the variable makes it appear.
```

Out[132]:

	Block	Number of Driveways
0	First Block (16th Street to 17th Street)	7
1	Second Block (17th Street to 18th Street)	5
2	Third Block (18th Street to 19th Street)	1
3	Fourth Block (19th Street to 20th Street)	4
4	Fifth Block (20th Street to 21st Street)	15
5	Sixth Block (21st Street to 22nd Street)	14
6	Seventh Block (22nd Street to 23rd Street)	8
7	Eight Block (23rd Street to 24th Street)	8
8	Ninth Block (24th Street to 25th Street)	6
9	Tenth Block (25th Street to 26th Street)	8

3.1.2 Using Python, calculate the mean, median, mode, range, and standard deviation of your variable. Print these values. If you use a library function, you need to explain how it works with detailed comments. Do not blindly use library functions!

```
In [185]: import numpy as np
#This line of code imports the numpy library as np for better code writing.
from scipy import stats
#Imports stats from scipy library.

def myrange(A):
    return (max(A) - min(A))
#This function takes the maximum value and subtracts it from the minimum value.

mode = stats.mode(Drwys)
#I assign a variable to the library function to make it easier to print.
#This library function returns an array of the most common value.

stdn = round(np.std(Drwys),1)
#I assign a variable to the library function to make it easier to print.
#I use the method round() to return the library function rounded one digit after decimal point.
#This library function computes the standard deviation of the given data.

print("The mean is" ,np.mean(Drwys),)
#The mean is calculated using the np.mean library function that takes the data we have in Drwys and finds the average.
#It takes in the number of driveways, adds up the data and divides the summation with the number of blocks.
print("The median is" ,np.median(Drwys),)
#The library function takes in the data and calculates the median value of it.
#It arranges the data in ascending order and finds out the mid value.
#Since, there are 10 elements within the list, it takes the mid two values and finds their mean.
print("The mode is" ,mode[0][0],)
#The mode function finds out the total number of times the same element is repeated and prints the element repeated the most.
#The mode variable is printed, but since it is an array of the most common we use [0][0] to just give us the value.
print("The range is" ,myrange(Drwys),)
#The range function we had written is printed.
print("The standard deviation is" ,stdn, "driveways.")
#This shows how much our data deviates from the mean.
#It takes each data from the list and subtracts it with the mean.
#Then, it squares all those values and adds them up. Finally, it square roots the data to give the final answer.

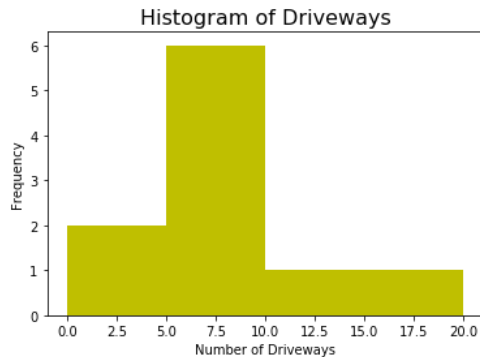
The mean is 7.6
The median is 7.5
The mode is 8
The range is 14
The standard deviation is 4.0 driveways.
```

3.1.3 Create a histogram for your data, properly formatting your figure.


```
In [166]: import matplotlib.pyplot as plt
#This line of code import the library needed to plot a histogram.

plt.hist(Drwys,bins = [0,5,10,15,20],facecolor = 'y')
#Given the data in Drwys, this makes a histogram with the bins I set between
#0 and 20, and the color of the histogram is given as well.

plt.xlabel("Number of Driveways")
#This is the Label of the x axis.
plt.ylabel("Frequency")
#This is the Label of the y axis.
plt.title("Histogram of Driveways", fontsize = 16, color = 'black')
#This is the title of the Histogram.
plt.show()
#This shows the histogram.
```



3.2 Interpret the descriptive stats: What can you say about the neighborhood based on these values? Is the distribution skewed? Is your visualization in agreement with the descriptive statistics? Explain. [#dataviz, #descriptivestats] (<200 words)

3.2 Based on these values I can say that there are generally several driveways in the neighborhood, this could be because there are houses and businesses that use them. The distribution is skewed based on the mean-median which is positive, hence the histogram should be right-skewed. The use of the descriptive stats can help us understand the shape of our data visualization. The values that our variable took were between 1 and 15. The right-skewed histogram implies that there is a higher chance of finding a block where the number of driveways is between 5 and 10. The visualization is in agreement with the descriptive stats and further helps to give information about the neighborhood and the variable.

Part 4: Probability Considerations [#probability, #algorithms, #simulation]

4.1 Can the mean of your data be interpreted as the expected value of a random variable? Explain why or why not in detail. (~50 words)

4.1 No, the mean of my data cannot be interpreted as the expected value of a random variable because the variable being observed is discrete. There is no way that we choose one of the variables at random and have the mean be the number of driveways found because we will have a whole number of driveways.

4.2.1 Suppose something unfortunate happened: you stole too many napkins for your Fermi estimate, so you decided to write all of your variable measurements on separate napkins, one napkin for each block. On your way back to res, the wind picked up and blew them all away! Luckily, you managed to collect all of the napkins, but now the data is totally randomly reordered, meaning that you have no idea which napkin corresponds to which block. Suppose that you tried to just guess randomly which napkin goes with which block. In other words, you randomly assign each napkin to a given block.

What is the probability that you are unlucky, and sadly NONE of the napkins are matched to the correct block (you guessed all of them wrong)? Estimate this probability using a simulation. Be sure to interpret the result appropriately.

```
In [174]: import random

napkins = [0,1,2,3,4,5,6,7,8,9]
unlucky = 0
N = 100000
prblty = []

for number in range(1,N):
    match = 0
    rand_napkins = np.random.choice(napkins,10,replace=False)

    for i in range(10):
        if(rand_napkins[i] == i):
            match = 1
            break

    if match == 0:
        unlucky += 1

    prblty.append(unlucky/number)

probability = unlucky/N
pblt = round(probability,2)

print("The probability that none of the napkins are matched is:", pblt)
```

The probability that none of the napkins are matched is: 0.37

4.2.1 After running the simulation, we find that the probability of none of the napkins being matched to the correct block is of 37%. The code presents us with the list of the napkins because we will use that to determine if the random number assigned to each napkin corresponds to it. The simulation runs the amount of times we input and begins to add up the unlucky events with the for loop. When the simulation has ended and all the unlucky outcomes have been calculated, the code divides the unlucky outcomes by the number of total simulations to obtain the probability.

4.2.2 What is the expected number of napkins that will be correctly matched to the corresponding block? Estimate this probability using a simulation and interpret the result appropriately.

```
In [186]: napkins = [0,1,2,3,4,5,6,7,8,9]

N = 100000
match_napkins = [0,0,0,0,0,0,0,0,0,0]
napkin_prblty = []
exp_number = 0

for number in range(N):
    match = 0
    rand_napkins = np.random.choice(napkins,10,replace=False)

    for i in range(10):
        if(rand_napkins[i] == i):
            match += 1

    match_napkins[match] += 1

for i in range(10):
    napkin_prblty.append(match_napkins[i] / N)
    exp_number += i * napkin_prblty[i]

xpct = round(exp_number,2)

print("The expected number of napkins that will be correctly matched to the corresponding block is: ", xpct)
```

The expected number of napkins that will be correctly matched to the corresponding block is: 1.0

4.2.2 In this code we want to make a simulation of every time that we manage to collect all the napkins and end up reordering them. We have a list that will take up the values of matched napkins every time the simulation runs. This gets increasingly better and will give a value that the simulation tends to be. So, at the end we get that the expect number of matched napkins is 1 because that is the number it tends to.

Part 5: Reflection

Reflect on what you learned about the HCs in this assignment, focussing on the connections between the HCs, and their connections to the city. Also reflect on how your prediction and estimation from parts 1 and 2 compare to the results. (<200 words)

5 I learned a lot about how methods of estimation and actual calculations can help us out to understand aspects of our everyday life. Not only do we learn new things when joining all these concepts together but also we are able to infer on things that are not apparent at first sight. This was a truly remarkable experience because I used what I learned to look at my surroundings in a different way. The HCs like #estimation and #variables help us get creative with what we observe and try to dig deeper. After gathering factual data, we are able to use #probability, #dataviz, and #descriptivstats to further find implications of the information we gathered. Using all these tools we are able to make data vizualizations that are relevant to what is being studied. Because of this, we become better problem-solvers. It is important to highlight that the tools that require little to no actual calculations are on par with the more formal ones. I was able to predict and estimate the number of driveways in my chosen neighborhood within one order of magnitude.