Before you turn this problem in, make sure everything runs as expected. First, **restart the kernel** (in the menubar, select Kernel → Restart) and then **run all cells** (in the menubar, select Cell → Run All).

Make sure you fill in any place that says `YOUR CODE HERE` or "YOUR ANSWER HERE", as well as your name and collaborators below:

In [ ]:

```
NAME = "Paturi Jayanth Varun"
COLLABORATORS = ""
```

---

## General Instructions

1 - Start by downloading this jupyter notebook to your local machine

2 - Open a tab in your browser and type https://colab.research.google.com/

3 - This will open a small window. Choose the last option on the upper menu, "Upload". Then choose the jupyter notebook you have saved in step 1

4 - You can start working on your assignment by answering the questions in the corresponding cells.

5 - If you have any questions , please reach out to your instructor and TAs

## MAT115: Statistics (MAT115/116) - Assignment 1

## Introduction to Variables Location Based Assignment

This assignment is a location based-assignment that will require you to interact with the city around in you in a new way. Simply put, the objective is to measure a variable in a city in the Guntur district. You will identify a measurable variable in the city and then create an estimate using the Fermi estimation technique. Next, you will complete the data collection, calculate descriptive statistics on the data, and create relevant data visualizations. You will also have a chance to apply your knowledge of probability and simulation to solve a problem. This is an individual assignment. Everything you submit should be your own words and reflect your own understanding of the material.

**NOTES:**

Anything marked as optional will only be scored if it is completed correctly. You must upload two files:

- **Primary Resource**: A PDF of your entire assignment. Run all cells before converting the notebook to a PDF, and double check to make sure that the PDF is complete with all sections visible. Email attachments will not be accepted. If you're having difficulty converting your notebook to a PDF, try the tips available here
- **Secondary Resource**: A zipped folder containing the .ipynb file and your original photo files.

### PART 1: VARIABLE SELECTION [#variables]

Select a neighborhood in a city in the Guntur district. 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 at different locations in the city. You need to make at least 10 different measurements of this variable, one for each location. The locations must be at least 100 meters away from each other.
- You must be able to calculate the mean, median, mode, and standard deviation of the variable.
- Be clear about your choice of locations to make the variable measurements.
- Get creative! Try to choose an interesting and informative variable and make sure to justify why the variable you have chosen is interesting.

**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 locations where you will conduct your measurement and provide an image that clearly identifies these locations on a map. (<150 words)

The variable which I have selected is the number of shirts sold in 1 day from a brand outlet. I have selected this variable in order to know that which brand has largest sales in one day. It helps us to analyse that how many people prefer to buy a specific brand among other brands.

My variable is a Quantitative discrete variable as the number of shirts sold each day is countable and has a fixed whole number as value. I measured a proportion of my variable since i wanted to know that how many people prefer a specific brand.

Procedure: I went to Guntur District to measure the values for my variable.Then i visited 10 brand outlets as referred below.I requested them to keep a count on the number of shirts sold in thier particular brand in a day.Like this requested the same for 10 brand outlets.At the end of the day they reported the sales of shirts in one day.

The 10 locations I used to measure my variable are following: 1.Allen solly,krishna plaza,lakshmipuram road,guntur,andhra pradesh. 2.Basics,krishna plaza,lakshmipuram road,guntur,andhra pradesh. 3.U.S.Polo,krishna plaza,lakshmipuram road,guntur,andhra pradesh. 4.Flying Machine,krishna plaza,lakshmipuram road,guntur,andhra pradesh. 5.Louis Phillipe,Annapurna Complex,Lakshmipuram, CM Nagar, Guntur, Andhra Pradesh 522007. 6.Indian Terrain,Door No - 5-87-13 Ground floor, Lakshmipuram Main Rd, Below Partha Dental Clinic, Guntur, Andhra Pradesh. 7.Blackberrys,Shop No. 5, 87-100, Lakshmipuram Main Rd, Pandaripuram, Guntur, Andhra Pradesh 522101. 8.Vanhuesen,Shop No. 1-4,Lakshmipuram Main Rd, Opp. Chaitanya Techno School, Ashok Nagar, Guntur, Andhra Pradesh 522006. 9.Chennis,Chandramouli Nagar, Guntur, Andhra Pradesh 522007. 10.Woodland,Ground Floor, Ananda Nilayam, Lakshmipuram Main Rd, Guntur, Andhra Pradesh 522006.

I measured the data, directly by asking the number of shirts sold in their brand outlet. It doesn’t have any units as it has constant whole number or numerical values.

**2. Discuss variable relationships.**

- **2.1** (<150 words)
  - A. Describe a scenario in which your variable could be an independent variable.
  - B. What could be the dependent variable(s)?
  - C. What are some possible extraneous or confounding variables in this scenario?

a scenario in which my variable is independent is "the distance of the shop"and "irrespective of the quality of the shirt".the dependent variable is budget if he has alot of money he may go to the bigger store. c

- **2.2** (< 150 words)
  - A. Describe a scenario in which your variable could be a dependent variable.
  - B. What could be the independent variable(s)?
  - C. What are some possible extraneous or confounding variables in this scenario?

The scenario in which my variable is dependent variable like "cost of the shirt","Number of costumers" and "quality of the shirt".These will be some independent variable on which our variable will depend. The confounding variable that will affect the relation between these dependent and independent variables are "Brand of Shirt".

# PART 2: ESTIMATION AND MEASUREMENT [#variables]

**Important note:** *if there is any reason to believe that you did not authentically complete the location based portion of this assignment, this will be refered to the Academic Committee, and you risk receiving zeros in all your grades (as per the course policy in the syllabus). Please follow the instructions here carefully and include the original photo files in the zip folder along with the ipynb.*

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. Take some photos to document this experience. You must include:
   - A photo just of your “back of the napkin” estimate (it can and should be quite rough at this point). You will properly format the calculation later.
   - 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.

- 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 too.

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.
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.

Follow the instructions in this link to upload your pictures to the jupyter notebook:

In [1]:

```
from IPython.display import Image, display
Image(filename="j1.jpeg",height=400,width=400)
```

Out[1]:

In [2]:

```
from IPython.display import Image, display
Image(filename="j2.jpeg",height=400,width=400)
```

Out[2]:

In [3]:

```
from IPython.display import Image, display
Image(filename="j3.jpeg",height=400,width=400)
```

Out[3]:

In [4]:

```
from IPython.display import Image, display
Image(filename="j4.jpeg",height=400,width=400)
```

Out[4]:

In [5]:

```
from IPython.display import Image, display
Image(filename="j5.jpeg",height=400,width=400)
```

Out[5]:

In [6]:

```
from IPython.display import Image, display
Image(filename="j6.jpeg",height=400,width=400)
```

Out[6]:

In [7]:

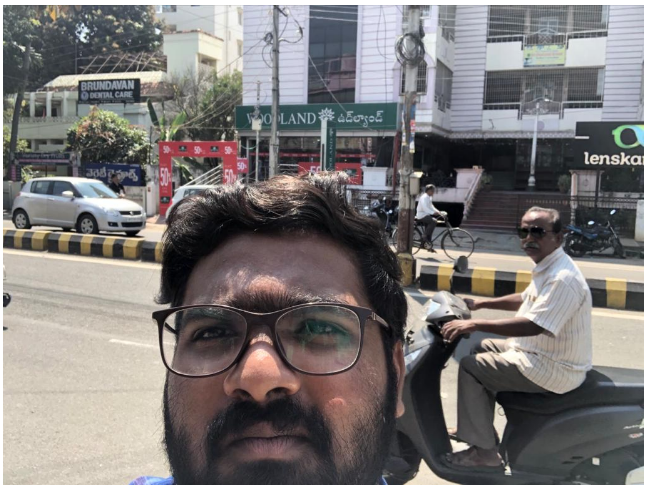
```
from IPython.display import Image, display
Image(filename="j7.jpeg",height=400,width=400)
```

Out[7]:

In [2]:

```
from IPython.display import Image, display
Image(filename="j8.jpeg",height=400,width=400)
```

Out[2]:



In [9]:

```
from IPython.display import Image, display
Image(filename="j9.jpeg",height=400,width=400)
```

Out[9]:

In [10]:

```
from IPython.display import Image, display
Image(filename="j10.jpeg",height=400,width=400)
```

Out[10]:

In [11]:

```
from IPython.display import Image, display
Image(filename="j11.jpeg",height=400,width=400)
```

Out[11]:

In [ ]:

## PART 3: ANALYSIS

1. Analyze the data in Python [#pythonprogramming]:
   - **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 .cvs file, and import into Python. Print your data in the cell below.

In [6]:

```
import pandas as pd

df1=pd.read_csv("jayanth data.csv")
df1
```

Out[6]:

| | name of brand outlet | no of shrits sold in one day |
|---|---|---|
| 0 | Allen Solly | 18 |
| 1 | Basics | 16 |
| 2 | U.S.Polo | 12 |
| 3 | Flying Machine | 16 |
| 4 | Louis Phillipe | 6 |
| 5 | Indian Terrain | 19 |
| 6 | Blackberrys | 15 |
| 7 | Vanhuesen | 17 |
| 8 | Chennis | 5 |
| 9 | Woodland | 11 |

- **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!

**Note**: Round your final answers up to 2 decimals.

In [29]:

```
a=[18,16,12,16,6,19,15,17,5,11]
n =len(a)

get_sum=sum(a)
mean = get_sum / n

print("mean / Average is: "+str(mean))
```

```
mean / Average is: 13.5
```

In [28]:

```
a = [18,16,12,16,6,19,15,17,5,11]
n = len(a)
a.sort()

if n % 2 == 0:
    median1 = a[n//2]
    median2 = a[n//2 - 1]
    median = (median1 + median2)/2
else:
    median = a[n//2]
print("median is: " + str(median))
```

```
median is: 15.5
```

In [27]:

```
import statistics
set1 =[18,16,12,16,6,19,15,17,5,11]
print("Mode of given data set is % s" % (statistics.mode(set1)))
```

```
Mode of given data set is 16
```

In [26]:

```
def data_range(list_var):
    range = max(list_var) - min(list_var)
    print("Range of the given variable is: ", range)

list_a = [18,16,12,16,6,19,15,17,5,11]
data_range(list_a)
```

```
Range of the given variable is:  14
```

In [25]:

```
import statistics
sample = [18,16,12,16,6,19,15,17,5,11]
print("standard Deviation of sample is % s "
                % (statistics.stdev(sample)))
```

```
standard Deviation of sample is 4.881939505292269
```

In [ ]:

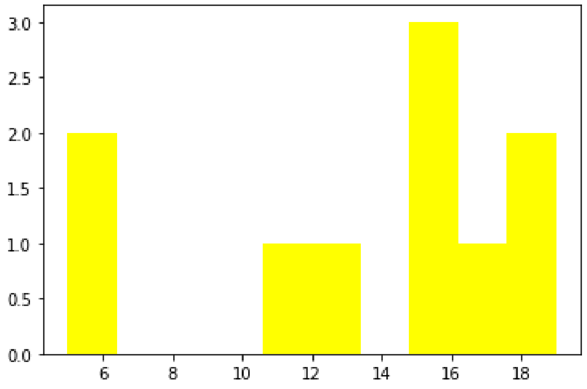
```
# Please ignore this cell. This cell is for us to implement the tests
# to see if your code works properly.
```

- **1.3** Create a histogram for your data, properly formatting your figure.

In [2]:

```
import matplotlib.pyplot as plt
data = [18,16,12,16,6,19,15,17,5,11]
plt.hist(data,bins=10,color="yellow")
plt.show
from scipy.stats import skew
print("The skewness of the data is",round(skew(data),2))
```

```
The skewness of the data is -0.72
```



**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. [#professionalism, #descriptivestats, #pythonprogramming] (<200 words)

From the above cell I calculated skewness of my lst. My skewness of the distribution is -0.72. My distribution is negatively skewed. I calculated skewness of my distribution from scipy.stats library and imported skew in that library. When the output is positive then we can say that the distribution is positively skewed whereas the output is negative then we can say that the distribution is negatively skewed. With the help of descriptive statistics and python programming I calculated skewness of my data.

## PART 4: PROBABILITY CONSIDERATIONS [#probability, #pythonprogramming, #codereadability]

**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)

yes,mean of my data is interpreted as the expected value of a random variable.

**2.** 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 location. On your way back to the campus, 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 location. Suppose that you tried to just guess randomly which napkin goes with which location. In other words, you randomly assign each napkin to a given location.

```
* What is the probability that you are unlucky, and sadly NONE of the napkins are matched t
o the correct location (you guessed all of them wrong)? Estimate this probability using a s
imulation. Be sure to interpret the result appropriately. See hints below.
```

In [11]:

```
import numpy as np
a=[18,16,12,16,6,19,15,17,5,11]
sum=0
for i in range(1000):
    b=np.random.choice(a,10,replace = False)
    for j in range(0,9):
        if(b[j]==a[j]):
            sum+=1
            break
c=1-(sum/1000)
print(c)
```

```
0.32099999999999995
```

**3. [Optional]:** What is the expected number of napkins that will be correctly matched to the corresponding location? Estimate this probability using a simulation and interpret the result appropriately.

In [9]:

```
# YOUR CODE HERE
raise NotImplementedError()
```

```
---------------------------------------------------------------------------
NotImplementedError                       Traceback (most recent call last)
<ipython-input-9-15b94d1fa268> in <module>
      1 # YOUR CODE HERE
----> 2 raise NotImplementedError()

NotImplementedError: 
```

**4. [Optional]:** Determine the probability distribution as a function of the number of correctly matched napkins and create a visualization.

In [ ]:

```
# YOUR CODE HERE
raise NotImplementedError()
```

**5. [Optional]:** Interpret the distribution based on your previous results.

**5. [Optional]:** Interpret the distribution based on your previous results.

YOUR ANSWER HERE

**6. [Optional]:** Compute the probability or expected value found above or both analytically (without a simulation).

YOUR ANSWER HERE

### Hints:

- To simplify the problem, you can disregard your actual variable data if you wish, and simply make a new list in Python consisting of the numbers 0 through 9: napkins = [0,1,2,3,4,5,6,7,8,9]. Pretend that this is your stack of napkins with the variable measurements in the correct order. Notice that this data satisfies napkins[i] == i, for all values of i from 0 to 9. Think of the index i as the location label.
- A random permutation of this list can be created with the following code: rand_napkins = np.random.choice(napkins,10,replace=False). You should be able to explain how this function works and why it is relevant for the problem.
- You want to check whether rand_napkins[i] == i, for each value of i from 0 to 9.
- You'll need to use a loop to create many random lists and repeat the checking procedure, keeping track of the number of matches each time.

## PART 5: REFLECTION[#probability, #variables]

Reflect on your application of the LOs in this assignment. How are the connections in the city mapped to the connections between the different LOs. Also reflect on how your prediction and estimation from parts 1 and 2 compare to the results. (<200 words)

1. Variable selection : The variable which I have selected is the number of shirts sold in 1 day from a brand outlet.
2. Python programming : By using python programming I calculated values of central tendency which is very easy to create programm for it.
3. Probability : By using python programming I calculated probability for napkin question.
4. Code readability : I included comments in the programm which is very easy to understand my code. Any person can easily understand my code.
5. descriptive statistics: I calculated values of central tendency by the help of few libraries. Without these libraries it is very difficult to write the codes. So, by the help of it I drawn histogram.
6. Professionalism : I have done my assignment without copying from anyone. I have done the assignment without any errors.

### PYTHON TIPS

Part of the purpose of this assignment is to expose you to and give you practice in using tools for working with data in Python. The following may be useful.

- Participating actively in the weekly structured study sessions will help prepare you to complete the Python portion of this assignment. The weekly session material can be found here.
- Your peer tutors and professors are here to help! Make use of office hours for assistance.
- For other resources to learn Numpy, you can read or watch any of the tutorials found online, such as https://docs.scipy.org/doc/numpy/user/quickstart.html. You do not need to learn everything about this library, just the basics of arrays and reading their entries.
- To learn to plot the necessary figures, read as much of http://matplotlib.org/users/beginner.html as is necessary to perform the required tasks. Additionally, there is an enormous amount of freely available instructional material, with examples, that can be found online.
- As a best practice, your graphics in Jupyter notebooks should be 'inline.' If your version does not do this automatically, include %matplotlib inline at the top of your script.
- Reminder: no matter what, your code needs comments. Read this resource about the importance of comments and this one for further guidance.