Lecture 7:
Dealing with Missing Data

COSC 526: Introduction to Data Mining
Spring 2020

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

Today we will explore …

Discussions:

- Discussion of your solutions for Assignment 6
- Overview of new Assignment 7

New concepts:

- Reasons for missing data
- Strategies for dealing with datasets with missing data

Project:

- Introduce your dataset(s)

Assignment 6
Discussion of Solutions

Each problem builds on the previous one

Lectures 3-4

- We learned about the **MapReduce** paradigm

Assignment 5

- We defined three sequential methods:
  - mapSequential
  - reduceSequential
  - reduceByKeySequential

Assignment 6

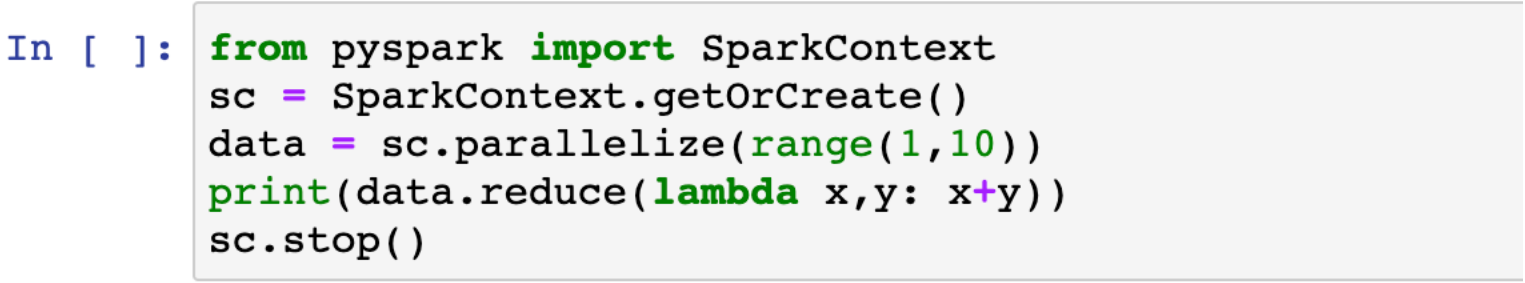
- We learned about the **Apache Spark** implementation of the **MapReduce** programming model
- We used **PySpark** (the Spark Python API) to solve text parsing problems *with the power of parallel processing*

Assignment 7

- We use PySpark's parallel version to perform clustering operations

Assignment 6: Test PySpark

- Run the cell below to verify that your Java, Spark, and PySpark installations are successful



```python
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
data = sc.parallelize(range(1,10))
print(data.reduce(lambda x,y: x+y))
sc.stop()
```

Do we have any problem running Java, Spark, and PySpark?

Assignment 6: Problem 0

- Now that we are in Jetstream, clone Assignment05 in Jetstream, open your completed Assignment05, and rerun it
    - Note that you are running your job on a remote 6-core node and not on your laptop
- Executing the same code on different machines is a valuable test of the portability of your code

Do we have any problem running JetStream?
Can we execute our notebooks within Jetstream?
What is our feedback on Jetstream?

Assignment 6: Problem 1

- **We redo the problems from Assignment05 using Apache Spark**
  - We move **from *sequential* text processing** in Python (i.e., Assignments 4 and 5) **to the *parallel* implementation in *Apache Spark***
- *Note that the code you wrote for the sequential version should work with the parallel version*
- *You will only need to adapt the code to use **Spark's parallelized data structure, the RDD***

What were the challenges in completing this task?

Assignment 6: Problem 2

- We analyze the text for letter frequency
  - If you’ve taken a crypto course and/or have seen substitution ciphers then you are probably aware that ’e’ is the most common letter used in the English language
- **Use the pre-processed text words to count the frequency of each letter in the text using the parallel MapReduce methods of Spark**

What was the frequency of the top 5 letters you found?

Assignment 6: Problem 3

- If we really wanted to crack a substitution cipher (or win on "Wheel of Fortune") then we should be aware that, although 'e' is the most common letter used in English, it may not be the most common first letter in a word
- **Count the positional frequencies of each letter using the parallel MapReduce methods of Spark**
  - Count the number of times each letter appears as the first letter in a word, as the last letter in a word, and as an interior letter in a word (i.e. a letter that is neither first nor last)

What are the number of times the top 5 most frequently found letters appear:

As the first letter in a word?
As a last letter?
As an interior letter?

Assignment 6: Problem 4

- As you did the previous assignments, use matplotlib to create histograms for Problems 1-3

Did you further improve your graph representations?
Can you reuse the same code with other datasets?

Assignment 7

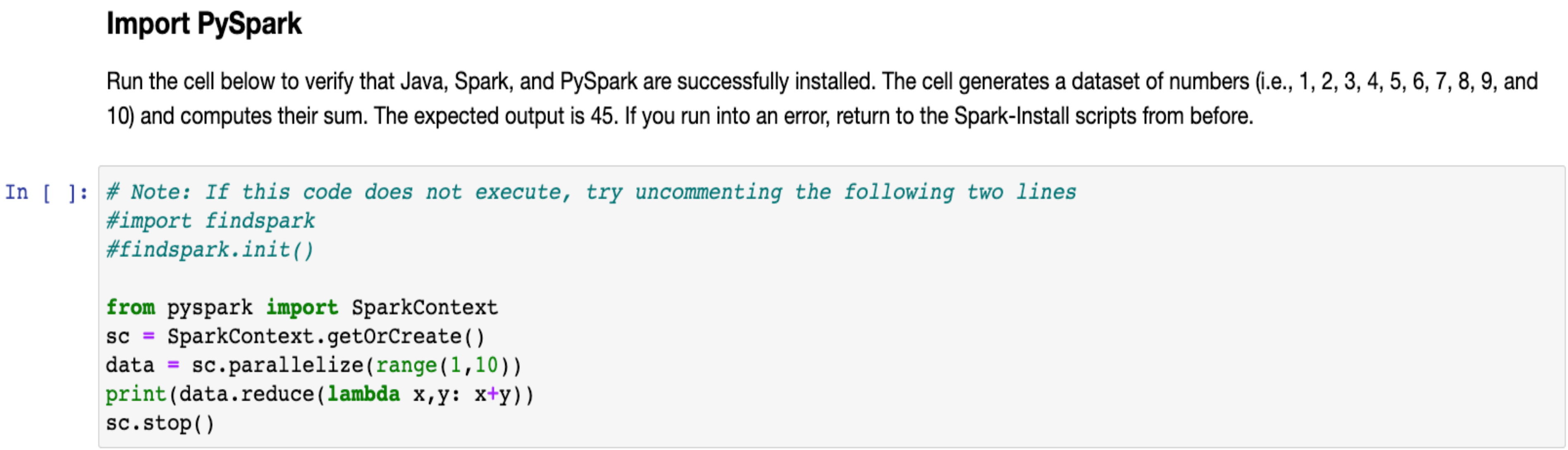
Assignment 7

- During the past lecture, we learned about k-means clustering.
- In Assignment 6, we learned to use PySpark's parallel versions of the map and reduce functions.
- In this assignment, we will be implementing the k-means algorithm in parallel with PySpark.

Open your Assignment 7 and refer to the notebook content for the discussion of the next few slides.

Assignment 7

- Test your environment



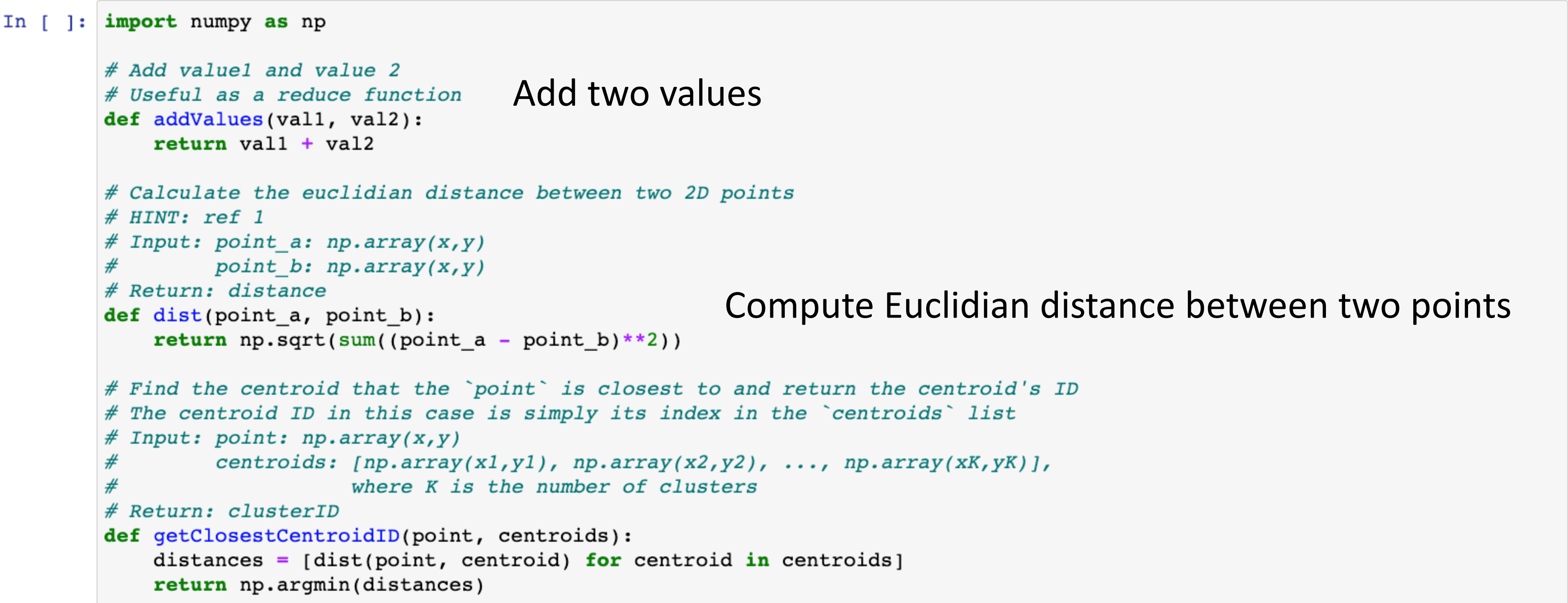
Import PySpark

Run the cell below to verify that Java, Spark, and PySpark are successfully installed. The cell generates a dataset of numbers (i.e., 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10) and computes their sum. The expected output is 45. If you run into an error, return to the Spark-Install scripts from before.

In [ ]:

```
# Note: If this code does not execute, try uncommenting the following two lines
#import findspark
#findspark.init()

from pyspark import SparkContext
sc = SparkContext.getOrCreate()
data = sc.parallelize(range(1,10))
print(data.reduce(lambda x,y: x+y))
sc.stop()
```

Assignment 7: Support tools



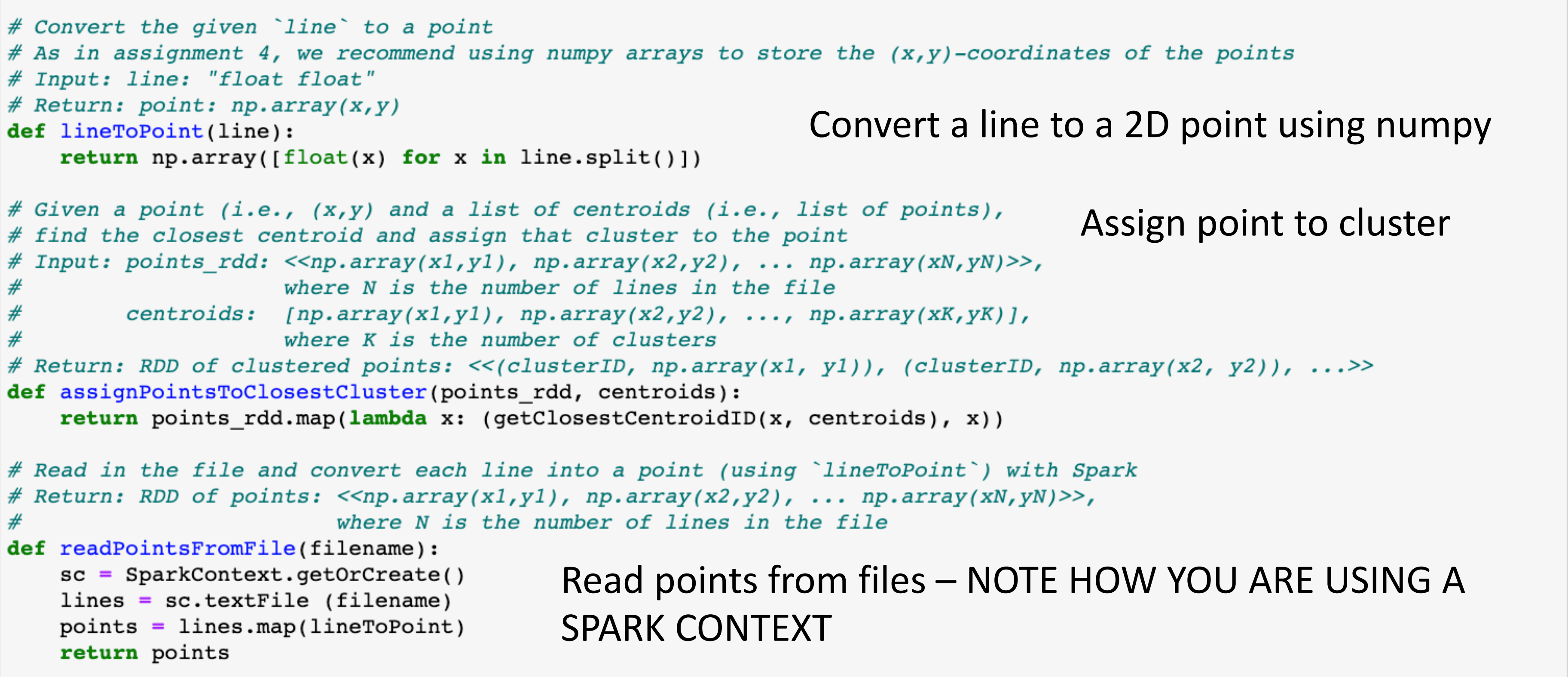
```
In [ ]: import numpy as np

        # Add value1 and value 2
        # Useful as a reduce function
        def addValues(val1, val2):
            return val1 + val2

        # Calculate the euclidian distance between two 2D points
        # HINT: ref 1
        # Input: point_a: np.array(x,y)
        #        point_b: np.array(x,y)
        # Return: distance
        def dist(point_a, point_b):
            return np.sqrt(sum((point_a - point_b)**2))

        # Find the centroid that the `point` is closest to and return the centroid's ID
        # The centroid ID in this case is simply its index in the `centroids` list
        # Input: point: np.array(x,y)
        #        centroids: [np.array(x1,y1), np.array(x2,y2), ..., np.array(xK,yK)],
        #                   where K is the number of clusters
        # Return: clusterID
        def getClosestCentroidID(point, centroids):
            distances = [dist(point, centroid) for centroid in centroids]
            return np.argmin(distances)
```

Add two values

Compute Euclidian distance between two points

Get the id of the closer centroid to a given point

Assignment 7



```
# Convert the given `line` to a point
# As in assignment 4, we recommend using numpy arrays to store the (x,y)-coordinates of the points
# Input: line: "float float"
# Return: point: np.array(x,y)
def lineToPoint(line):
    return np.array([float(x) for x in line.split()])
```

Convert a line to a 2D point using numpy

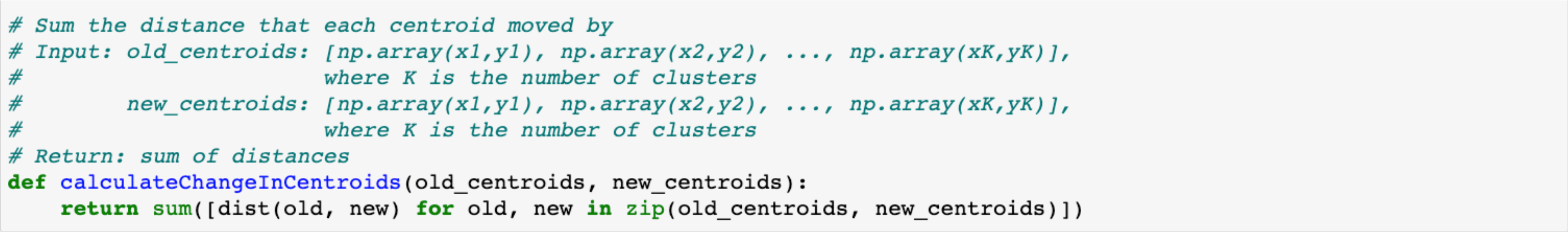
```
# Given a point (i.e., (x,y) and a list of centroids (i.e., list of points),
# find the closest centroid and assign that cluster to the point
# Input: points_rdd: <<np.array(x1,y1), np.array(x2,y2), ... np.array(xN,yN)>>,
#                    where N is the number of lines in the file
#        centroids:  [np.array(x1,y1), np.array(x2,y2), ..., np.array(xK,yK)],
#                    where K is the number of clusters
# Return: RDD of clustered points: <<(clusterID, np.array(x1, y1)), (clusterID, np.array(x2, y2)), ...>>
def assignPointsToClosestCluster(points_rdd, centroids):
    return points_rdd.map(lambda x: (getClosestCentroidID(x, centroids), x))
```

Assign point to cluster

```
# Read in the file and convert each line into a point (using `lineToPoint`) with Spark
# Return: RDD of points: <<np.array(x1,y1), np.array(x2,y2), ... np.array(xN,yN)>>,
#                        where N is the number of lines in the file
def readPointsFromFile(filename):
    sc = SparkContext.getOrCreate()
    lines = sc.textFile (filename)
    points = lines.map(lineToPoint)
    return points
```

Read points from files – NOTE HOW YOU ARE USING A SPARK CONTEXT

Assignment 7



```
# Sum the distance that each centroid moved by
# Input: old_centroids: [np.array(x1,y1), np.array(x2,y2), ..., np.array(xK,yK)],
#                       where K is the number of clusters
#        new_centroids: [np.array(x1,y1), np.array(x2,y2), ..., np.array(xK,yK)],
#                       where K is the number of clusters
# Return: sum of distances
def calculateChangeInCentroids(old_centroids, new_centroids):
    return sum([dist(old, new) for old, new in zip(old_centroids, new_centroids)])
```

Calculate change in centroid

Assignment 7

- Problem 1: **implement the calculateClusterMeans function and then test it against a provided test case.**
- Problem 2: **read in the points from the provided file, cluster the points into** K **clusters, and then return the clustering results (cluster centroids and clustered points).**
- Problem 3: **devise a way to visualize the clusters you have created.**

Assignment 7: things to consider

- How do you choose KK for your dataset?
- Do you always get the same results from k-means? Is it non-deterministic? Is this an error in your code or a feature of the algorithm?
- How would you optimize the code to work for larger datasets (e.g., 100GBs of points)?
- How would you generalize the code to work for larger-dimensionality datasets?

Breakout sessions

Reading

Reading

Therese D. Pigott. A Review of Methods for Missing Data. (2001)

Missing Data: what strategy?

Data Collection

Missing data = incomplete observations

- Critical data issues:
  - Reasons for missing data
  - Scale and distribution of the values in the data

Can you make an example of incomplete observations?

- **In social sciences?**
- **In astronomy?**
- **In earth sciences?**
- **In medical studies?**

Case Study

- Study of an asthma education intervention in eight schools
- Randomly chosen set of students aged 8 to 14 with asthma
- Observations over two weeks period post-treatment
- Students complete:
  - Scale to measure self-efficacy beliefs with regard to their asthma
  - Questionnaire rating severity of their symptoms

(Velsor-Friedrich, see attached paper in GitHub)

Case Study

- Students simply forgot to visit the school clinic to fill out the form → Missing completely at random (MCAR)
  - Complete cases are representative of the originally sample
- Students missed school because of severity of their asthma symptoms and failed to complete the symptom severity rating
  - Missing variable is directly related to study
  - Example of non-ignorable missing data!!!
- Younger children missed ratings of symptom severity because they had a harder time interpreting the rating form → missing at random (MAR)
  - Values are missing for reasons related to another observable variable

Mechanisms to Deal with Missing Data

- Data are MCAR or MAR
  - Ignore the reasons for missing data in the analysis of the data
  - Simplify the model-based methods used for missing data analysis
- Use more than one method for collecting important information
  - E.g., income + years of education or type of employment

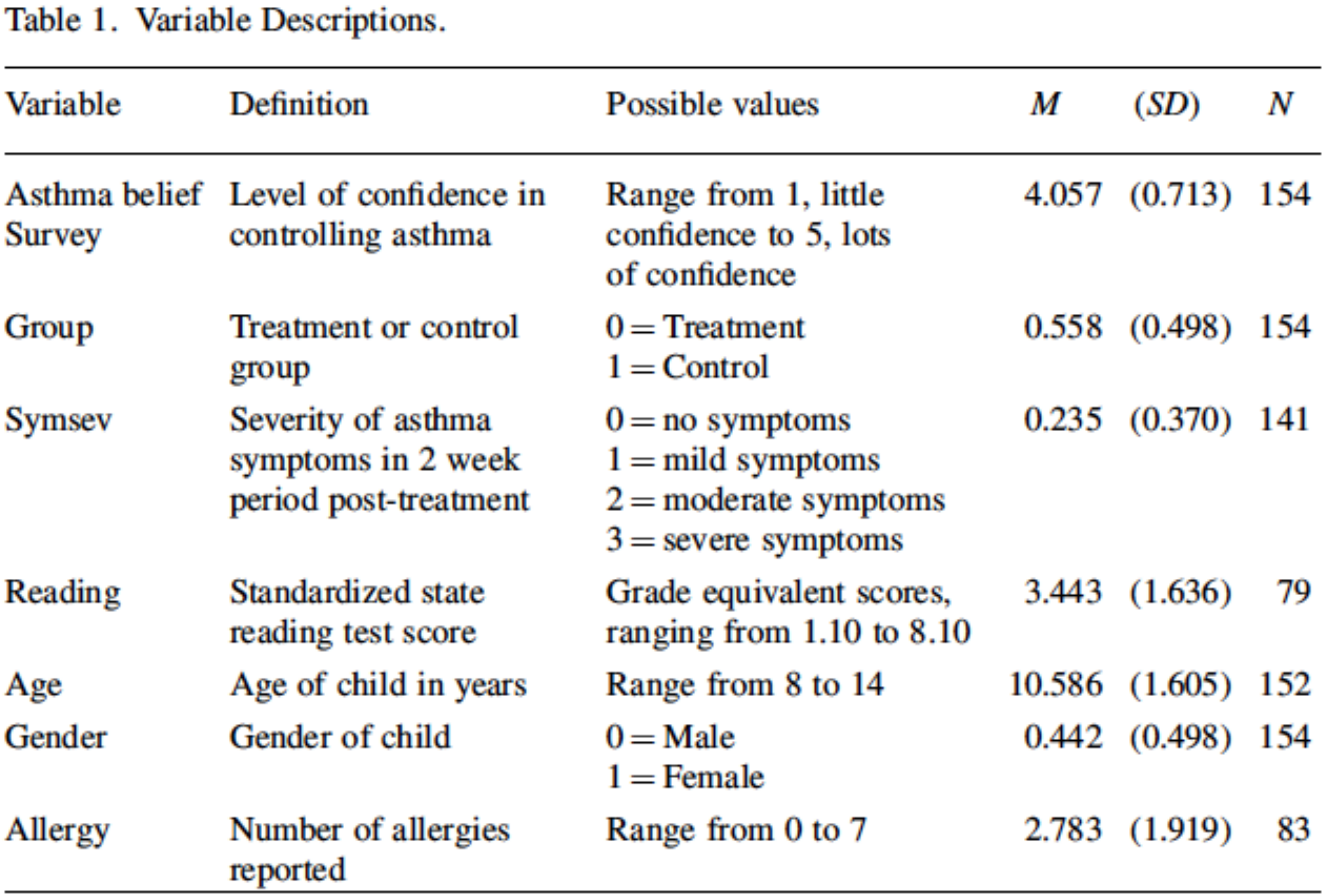


Table 1. Variable Descriptions.

Variable	Definition	Possible values	*M*	(*SD*)	*N*
Asthma belief Survey	Level of confidence in controlling asthma	Range from 1, little confidence to 5, lots of confidence	4.057	(0.713)	154
Group	Treatment or control group	0 = Treatment 1 = Control	0.558	(0.498)	154
Symsev	Severity of asthma symptoms in 2 week period post-treatment	0 = no symptoms 1 = mild symptoms 2 = moderate symptoms 3 = severe symptoms	0.235	(0.370)	141
Reading	Standardized state reading test score	Grade equivalent scores, ranging from 1.10 to 8.10	3.443	(1.636)	79
Age	Age of child in years	Range from 8 to 14	10.586	(1.605)	152
Gender	Gender of child	0 = Male 1 = Female	0.442	(0.498)	154
Allergy	Number of allergies reported	Range from 0 to 7	2.783	(1.919)	83

POPULATION: 154

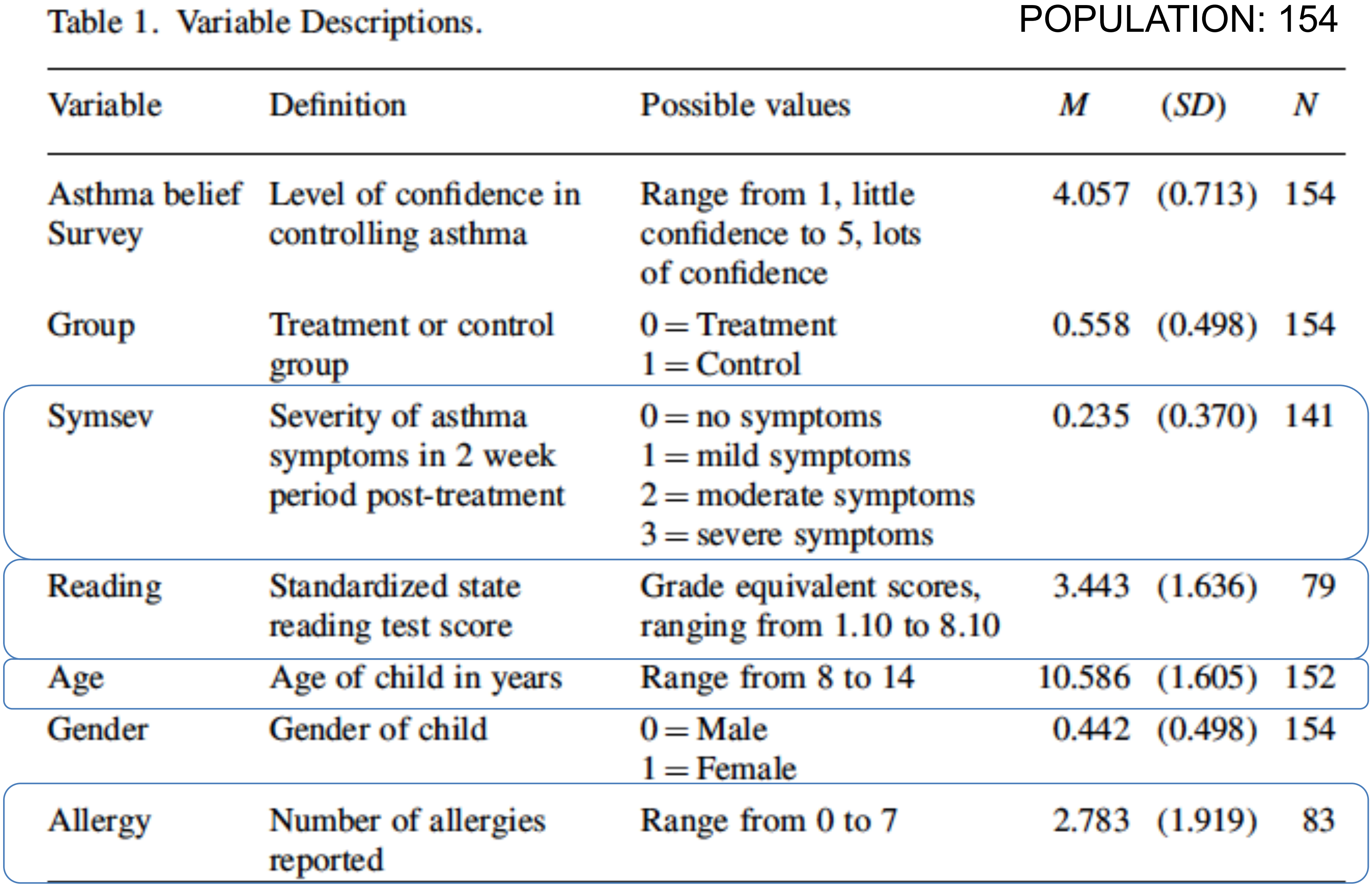
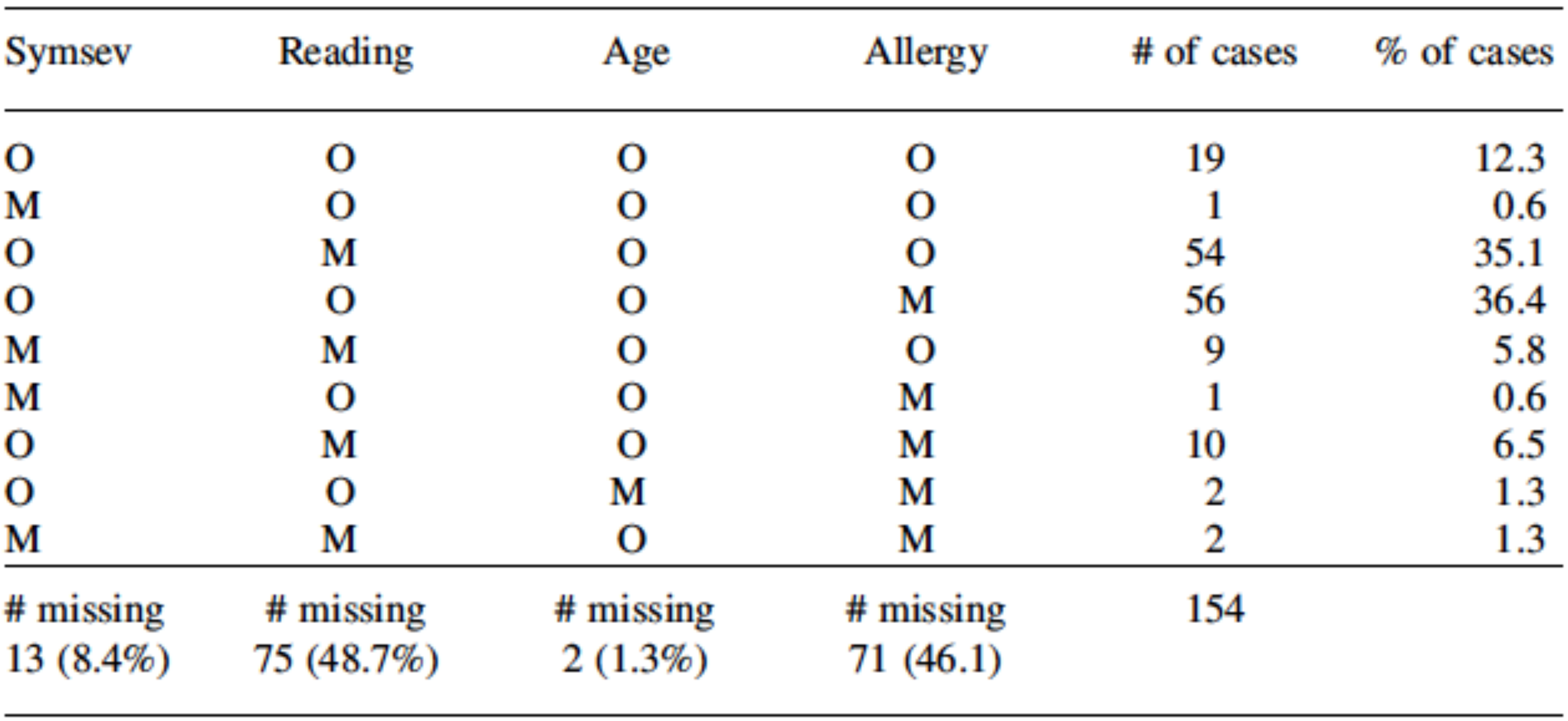


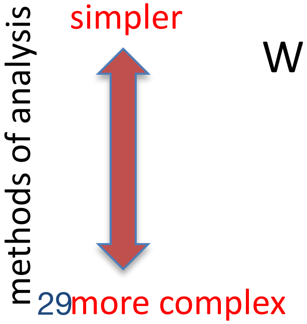
Table 1. Variable Descriptions.

Variable	Definition	Possible values	*M*	(*SD*)	*N*
Asthma belief Survey	Level of confidence in controlling asthma	Range from 1, little confidence to 5, lots of confidence	4.057	(0.713)	154
Group	Treatment or control group	0 = Treatment 1 = Control	0.558	(0.498)	154
Symsev	Severity of asthma symptoms in 2 week period post-treatment	0 = no symptoms 1 = mild symptoms 2 = moderate symptoms 3 = severe symptoms	0.235	(0.370)	141
Reading	Standardized state reading test score	Grade equivalent scores, ranging from 1.10 to 8.10	3.443	(1.636)	79
Age	Age of child in years	Range from 8 to 14	10.586	(1.605)	152
Gender	Gender of child	0 = Male 1 = Female	0.442	(0.498)	154
Allergy	Number of allergies reported	Range from 0 to 7	2.783	(1.919)	83

Table 2. Missing Data Patterns.



Symsev	Reading	Age	Allergy	# of cases	% of cases
O	O	O	O	19	12.3
M	O	O	O	1	0.6
O	M	O	O	54	35.1
O	O	O	M	56	36.4
M	M	O	O	9	5.8
M	O	O	M	1	0.6
O	M	O	M	10	6.5
O	O	M	M	2	1.3
M	M	O	M	2	1.3
# missing 13 (8.4%)	# missing 75 (48.7%)	# missing 2 (1.3%)	# missing 71 (46.1)	154	



What is the reasons for the missing data?

Can we accept the MCAR assumption?

Can we accept the MAR assumption?

Does missing data result from a non-ignorable response mechanism?

Commonly-Used Missing Data Methods

- Complete-Case Analysis: Cases that are missing variables in the proposed model are dropped from the analysis, leaving only complete cases
  - Assume that missing data are MCAR
  - Adequate amount of data remains for the analysis?

Commonly-Used Missing Data Methods

- Available Case Analysis: with X1 complete and X2 partially complete, all cases are used to estimate the mean of X1, but only the complete cases contribute to an estimate of X2, and the correlation between X1 and X2.
  - Different sets of cases are used to estimate parameters of interest in the data

Strategy 1	X1 =	x11	x12	x13	x14	x15	→ work on a population of 5 individuals
	X2 =		x22		x34	x25	→ work on a population of 3 individuals
Strategy 2	X1 =		x12		x14	x15	→ work on a population of 3 individuals
	X2 =		x22		x34	x25	→ work on a population of 3 individuals

Commonly-Used Missing Data Methods

- Single-Value Imputation: Fill in the missing value with a plausible one, e.g., mean for cases that observe the variable
  - Analyst continues with the statistical method as if the data are completely observed
  - Single value changes the distribution of that variable by decreasing the variance that is likely present
  - Bias in the estimation of variances and standard errors are compounded

Strategy 3

X1 = x11 x12 x13 x14 x15 → work on a population of 5 individuals

X2 = x_{avg} x22 x_{avg} x34 x25 → work on a population of 5 individuals

Model-Based Methods

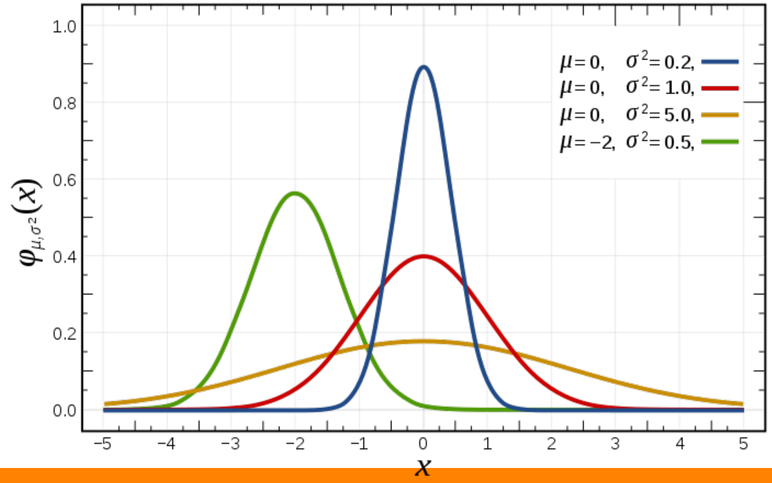
- Add assumptions about the distribution of the data and the nature of the missing data mechanism
  - Multiple imputation

Strategy 4

X1 = x11 x12 x13 x14 x15 → work on a population of 5 individuals

X2 = x_{21} x22 x_{23} x34 x25 → work on a population of 5 individuals

Assumption:



Discussion

Let’s consider the examples of incomplete observations that we listed for different sciences when we introduced the critical issues of data collections:

- **In social sciences**
- **In astronomy**
- **In earth sciences**
- **In medical studies**

For those example, when can we assume that missing data are MCAR?

NHANES Dataset

NHANES Dataset

- National Health and Nutrition Examination Survey (NHANES) is a cross-sectional survey that is conducted every two years in the United States
- Individuals are asked to complete
  - demographics questionnaire
  - 2-day, 24 hour dietary recall data for all individuals
- Sampled population in NHANES for the years 1999 thru 2016.
  - In the early years (1999 and 2001) only 1 day of dietary information was collected
  - For years 2003 thru 2016, 2 days of dietary data were collected

NHANES Dataset

- For each year, individuals have a unique identifier called a sequence number
- All files for that year can be linked by the sequence number (variable name: SEQN)
- Every year, different participants are sampled, so you cannot track the same person over time
  - You can track the average population intakes over time.

NHANES Questions

- What do the nutrient intake profiles of individuals look like? For example, do people who consume more saturated fats consume less fiber?
- Do nutrient intake profile patterns differ by gender, age, race, education, or poverty level?
- What do time trends in nutrient patterns look like?
- How consistent are individuals in their consumption over two days? For example f they are consuming high amounts of fiber on day one, are they also consuming high amounts on day 2?

Medicaid-Vital Statistics Data

Medicaid-Vital Statistics Data

- The Medicaid and Medicare Administration in the State of Delaware (DMMA) examines medical usage and health outcomes of their clients on a regular basis
- A report came out in 2014 that stated that individuals with mental illness were twice as likely to die and that they die at a much earlier age (almost 20 years earlier) compared to those without mental illness in the United States
- DMMA wanted to know if this was true of their Medicaid population
- The DSAMH_Medicaid dataset is a subset of about 6000 individuals who have Medicaid as their primary insurance and have at least one instance of mental illness.
- These data were then linked to vital statistics to determine mortality
  - There were approximately 200 deaths

Medicaid-Vital Statistics Data

- What cause of death codes are most commonly reported in the DSAMH_Medicaid dataset?
- Do the cause of death codes cluster into groups by gender? Or by disability status?
- Are there differences in the patterns of medical care (i.e., time spent in Medicaid, number of medical claims, number of hospital claims, number of emergency department claims, total billed amounts) reported by those that die versus those that didn't die?

Project Steps

Project (I)

- Step 0: search for datasets
  - Discussion in class of datasets identified

Assignment 6 – Problem 5

Find an interesting dataset.

- Is the dataset available? Do you have the dataset source (with url)?
- It your dataset large enough to allow for interesting analysis and non-trivial results?
- How were the data were collected?
- What is the significance of the data?
- Describe the number of rows, objects, or data points
- What information is contained in each rows, objects, or data points?
- What is the data types (int, str, char, float, etc.) and numerical ranges where appropriate
- What file(s) do you need to parse (if multiple files are available)?

Project (II)

- Step 1: Address (and answer) these questions
  - Do you work alone or in team? If in team, indicate who is your collaborator and what are the skills / expertise he/she is bringing to the project
  - What is the dataset you are considering?
  - What are the possible key question(s) you want to answer? Are the questions too general? Are the questions too narrow?

Project (III)

- What are the scientific questions you are answering with your tool/framework? Be as specific as possible.
- What is the tentative title of your project?
- What are the milestones you want to meet from now until the end of the semester when you will present your poster at the poster showcase? Be as specific as possible. Write the date of the deadline and the task(s) you want to achieve for that deadline.

The cost of missing something

Live chat

Tricia Wand. The cost of Missing Something

https://youtu.be/WaOUJa9fjXU

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

BIG ORANGE
BIG IDEAS