1. **a. What’s the difference between structured and unstructured data?**

Also referred to as “small data”, referred to the data models we’ve been using throughout this course by using Excel, Python, SQL, and Tableau. Though large in size, it is organized in structured columns that have a commonality to them. While, unstructured data is randomly scattered, unorganized information that needs to be organized for the analyst to understand how the information is useful.

**b.Can you give examples that you’ve encountered for both types?**

**Structured Data** - In our last section using Python and information from InstaCart customers, information about what type of food in each aisle were organized according to a number and a commonality. For example - Frozen Food would be characterized by anything that was found in the store in aisle 7.

**Unstructured Data -** Random order #s of Amazon customers that do not share a common region, age group, ordered products, and timestamps in an Amazon RMS system.

1. **Given that much of big data is produced by machines and sensors, how trustworthy do you think that big data is?** Because there is so much information collected and these programs coded by human data are never going to be 100% accurate since there can always be malfunctions or errors in the systems that are collecting this data. That’s why it’s important to break down the big organization data into smaller segments so that pieces of the data can be further analyzed and useful to cross-reference different data sources to come to conclusions.

**What characteristic of big data relates to the question of trustworthiness?** “Veracity” is the characteristic that describes how accurate and trustworthy big data is. If the veracity of the data is “high” then it means that it’s trustworthy and “low” veracity means that it’s not as trustworthy.

1. **Assume that you receive a table containing customer data. You notice that some values are missing or incomplete, and the formatting is inconsistent in some columns. Based on what you’ve learned so far, how would you go about cleaning this table? Think about what you would do first, second, third, etc.**

As we’ve learned previously, it’s important to start by cleaning your data to make it organized and easier to understand and get rid of information in the data frame that is not valuable. By removing inconsistencies, missing, and duplicate data in the set you are then able to organize the remaining data into organized columns and make sure that consistency remains throughout your data set and finally cross-reference your data to make sure it’s as accurate as possible before analyzing.

1. **Can you describe tools such as Hadoop and Apache Spark and their role in big data?**

Hadoop and Spark are powerful systems designed to help companies manage and process massive amounts of data, overcoming bottlenecks that a single application, like Python, cannot handle alone. Hadoop operates by distributing data across multiple computer systems, using a disk-based approach to process large datasets. While, Spark uses an in-memory processing model, which allows it to work with APIs and perform computations much faster than Hadoop, as it minimizes the need for disk storage. Spark's speed and efficiency make it particularly well-suited for real-time data processing and iterative tasks.

1. **How has the application of analytics to big data led to new discoveries and innovation? Can you give some examples?**

* **Weather Predictions -** Data Scientists use past data from weather patterns such as air temperature, water temperature, humidity, and pressure to forecast where devastating weather patterns will go by modeling the moves hurricanes As the hurricane moves, real-time data from sensors continuously updates these models, improving accuracy.
* **Sports -** Big Data has helped sports teams analyze the way and how well their players perform. Sports Scouts when recruiting for teams while evaluating a players value and statics to determine how well a player might perform that season rather than relying on traditional biases like physic and build.
* **GPS Systems and Apps -** Real time data can help GPS systems coordinate the best route for drives with real time information like traffic, accidents, and closures to predict the best and fastest route for the driver.