**Communication with Stakeholders:**

1. What questions do you have about the data?

* What is the primary purpose of analyzing the data? Does it to understand the most frequent users and where they live? Is it finding the top brands from the scanned receipts?
* Where did we get this data? What are the primary sources?
* What are the data quality issues within our data, and how to resolve them? Like missing values, wrong data types, duplicate values, etc.
* Understanding the distribution of each numerical and categorical column by plotting histogram and table summary.
* If we are working on a machine learning problem, understanding the dependent and independent variables through correlations.

1. How did you discover the data quality issues?

Ways to discover data quality issues by using tools like Python, SQL, and excel:

* Finding duplicates
* Flattening the dictionary values to string
* Identifying missing values
* Identifying wrong data formats
* Identifying the misspellings and non-uniformity in the data
* Identifying the gaps in the timeline if it’s related to the date

1. What do you need to know to resolve the data quality issues?

Ensuring and sustaining data quality is significant as it directly relates to how the organization is doing its business and the entire life cycle of the data. To resolve data quality issues, we should understand several things like

* **Accuracy** - Data should be accurate like the way it is described
* **Relevancy** - It should be relevant and need to meet the requirements of intended use.
* **Completeness** - Data should not have missing values or missing records.
* **Timeliness** - Data should be up to date and should not have gaps in the timeline.
* **Consistency** - Data should be in the expected data format and should be cross-referenceable.

1. What other information would you need to help you optimize the data assets you're trying to create?

Better Data Governance framework provides our organization with complete approach to collecting, storing, managing, and securing data. To optimize the data assets, I would need more information about,

* **Data Architecture** – Information about the overall structure of data and data related resources.
* **Data Modeling and Design** - Information about database building, design, testing, and maintenance.
* **Data Security** – Information about data security to ensure privacy, confidentiality, and appropriate access.
* **Data Quality** – Information about the process of defining, monitoring, maintaining data integrity, and improving data quality
* **Reference and master data** – Information of managing shared data to reduce data redundancy and ensuring data quality through standardized definition and data values.

1. What performance and scaling concerns do you anticipate in production and how do you plan to address them?

**Issues that affect scaling and performance in production:**

* Inefficient Database Engine
* Limited resources in Memory, CPU’s and GPU’s
* Complicated database schema and bad Indexing
* App server limitations and wrong server configuration
* Poorly performed database queries and inefficient code
* Lack of monitoring tools

**Solutions for the scaling and performance issues:**

* Knowing the product’s qualitative and quantitative growth scale
* Efficient uses of hardware and software resources
* Understanding the product resources and requirements
* Capacity planning on hardware and dependency scaling
* Traffic Management and scalable data storage
* Better task handling and processing.