# Milestone 1: Business Problem solving, SQL, Data Wrangling, Power Query & Excel Data Modeling

## Case Study

In this milestone, we will expand on the scenario from Lesson 1 Assignment.

**Background:** An eCommerce firm (let’s call it Cables.com) sells 2 types of USB Cables and 3 types of Laptop Power Cables in US, Canada and Mexico. They manufacture cables locally in each country that they sell in. Each country has only 1 manufacturing plant that produces all 5 products. Each plant has a Plant Manager, each product has a Product Manager. The Executive team is responsible for the overall health of the business.

**The problem:**  The VP of Engineering, Dr.Who, based in US, has recently noticed an increased level of Returns for USB Cables sold to customers in the US. After a focused investigation of the returned cables, it was determined that the instructions for manufacturing the cables were recently updated incorrectly. This caused the cables to short out after sustained use (over 12 hours of continuous use). This defect has resulted in a loss of approximately $5 million in the 6 months it took to notice the issue and deploy a fix in place. The timeline of the fix was:

1. May 1st 2017 - Dr.Who suspected there could be an issue at play.
2. Jul 1st 2017 - The instructions for manufacturing were fixed but due to high business season, it was decided not to implement the changes until Sept 1st. A strategic decision was taken to at least start logging the manufacturing defects using this new instruction set. (Note - Manufacturing defects are recorded when an inspector manually compares a sample product with the manufacturing instructions to determine deviation from expected specs) From Jul 1st 2017, the particular defect in focus stared getting recorded because the product now did not comply with the expected specs.
3. Sep 1st 2017 - New manufacturing instructions were fully implemented for both manufacturing and inspection. The defect rate falls starting Sep 1st 2017 because the product being manufactured meets the expected specs.

**The initiative:** The Executive team, on advice of the Dr.Who, has now made it a strategic initiative to build mechanism to track defect rates for products and manufacturing plants. The goal is to detect any adverse trends in defect rates much earlier. Note: To keep things simple let’s assume that there can be only 1 type of defect in an item i.e. a manufacturing defect. But, in reality you will encounter many different types of defects ranging from manufacturing defect, packaging defect, defects in instructions that go with the product, packaging defect, shipping defect etc.

**The Implementation Plan:** Dr.Who has developed a plan of action which involves building visualizations from the defect data and making it available to Plant Managers (in each country), Product Managers (in US), and Executives (in US). Each month, Plant Managers, Product Managers and Executives will be required to meet and discuss the “health” of their operations. Dr.Who has been able to find out that the following information about the databases:

1. **Manufacturing Defects Database:** It records the manufacturing defects (for years 2016-2017) by sampling approximately 1 of 100 items manufactured. An important note to consider here is that the sampled items are inspected for defects against the expected specs (based on manufacturing instructions). So, in this case, if the manufacturing instructions are incorrect, then a true defect will not get noticed/recorded.
2. **Returns Database:** Database that records all returns agains the date the item was manufactured and the plant where it was manufactured.
3. **Customer Complaints database:** This database is currently being developed and is expected to be available in 30 days. Once available, it will also contain data from 6 months ago.  
   Links to the databases and access permissions need to be requested via self-service IT Request tool.

**Your Responsibility:** reporting to Dr.Who , you are the Analyst tasked with developing the Visualizations for the project. This implies that you will be expected to:

1. Connect to data source(s)
2. Query relevant data
3. Determining relevant metrics and charts for each level of management (Plant Managers, Product Managers, Executives)
4. Get feedback and improve your dashboard

Consider the information that is available to you and follow the template below to guide you through this assignment:

## Milestone Tasks

### Datasources

For this milestone, assume that your datasources are the following (for just 1 plant in US):

1. Manufacturing Defects Database : CSV file named MFG\_DEFECT\_DATA with daily tally of num of ITEMS\_MANUFACTURED, ITEMS\_SAMPLED and ITEMS\_DEFECTIVE
2. Returns Database : CSV file named RETURNS\_DATA with a tally of MFG\_DATE and the num of returns that were received for items manufactured on that date (i.e. in that batch) as RETURNS\_RECEIVED

(All data files are available on the course folder)

### Questions

Perform the following using R, SQL and/or Excel as prompted. You will be required to submit your code and you Excel workbook for grading.

1. Load the two CSV files in R using fread function from data.table package. They can have any name in R (example, dt1, dt2) but for simplicity sake, call them MFG\_DEFECT\_DATA and RETURNS\_DATA.
2. Use the data.table syntax to add a new column to MFG\_DEFECT\_DATA called MFG\_DEFECT\_RATE where
   * MFG\_DEFECT\_RATE = ITEMS\_DEFECTIVE/ITEMS\_SAMPLED
3. In both datasets, the MFG\_DATE column is currently in character format. Convert it to a Date format in both datasets.
   * Hint: as.Date(unformatted\_date\_field, pattern\_of\_the\_unformatted\_date). The patter here is "%m/%d/%y" (small letter y denotes 2 digit year. capital lettet Y denores four digit year)
4. Now using write.csv(dataset\_name, file="complete\_file\_path with .csv suffix", row.names = FALSE) syntax, write the modified datasets to csv file. Keep the names of the files the same (i.e MFG\_DEFECT\_DATA.csv and RETURNS\_DATA.csv). If you are writing to the same location where you read the data from, then this operation will overwrite the two files.
5. Using Power Query in Excel read and combine (i.e JOIN) the MFG\_DEFECT\_DATA.csv and RETURNS\_DATA.csv files. Keep the following in mind as you proceed.
   1. You will need to create two "queries" in Power Query of which MFG\_DEFECT\_DATA will be your base query
   2. You will then need to use the Merge queries option
   3. For a successful join, you will need to determine:
      * the common KEY between the two tables
      * the appropriate type of JOIN
   4. The resulting table should have all the columns from MFG\_DEFECT\_DATA and the RETURNS\_RECEIVED data table from RETURNS\_DATA table
   5. Its a good practice to give descriptive names to the auto-populated steps in Power Query
   6. Save the table resulting from the JOIN in a tab called JOIN Result in the workbook -you will need to submit it.
6. In the same Excel workbook, build a data model using the joined dataset (from the step above) and another dataset called PRODUCT\_NAMES (contains names corresponding to the PRODUCT\_ID values)
7. Use the resulting dataset to create a pivot table showing the following:
   * X-axis = Dates [filtered to show 1/1/2017 to 12/31/2017]
   * Y-axis = Integer
   * Two Series = Two series RETURNS\_RECEIVED, ITEMS\_DEFECIVE
   * Filtered for product name USB C Cable
   * If you have successfully created a chart, you will be able to distinctly identify the first MFG\_DATE for which the number of RETURNS\_RECEIVED were unusually high.

Please save this Excel workbook and submit it along with your R script file.