**Directions Part One:**

Before you begin profiling and joining your data, you want to be sure you are following best practice for assessing data quality and aligning to your organization goals. Using the templates, document your plan for cleaning and transforming the data using quality standards.

Your Data Quality Plan should address the following:

* **Explain why high-quality data a vital business requirement is:** An organization’s data is used to make key decisions.
* **Identify organizational goals:** Identify at least three goals of organization with respect to the data.
* **Explain the characteristics that define data quality using industry standards:** It is important to ensure an organization is following best practices and data analytics industry standards to maintain quality data.
* **Describe how security policies impact data quality:** Consider the level of stakeholder’s involvement for the people who will be accessing the data from an ethical standpoint.

### Purpose Statement:

### Quality data. Especially, high quality data is best practice to be a standard for an organization or cooperation. Quality data can and will help personals on senior levels or executives make informal decisions based on the accuracy of the data presented. Thus, to help productivity for staff within the corporation or organizations, while ensuring that that data is following compliance and regulations with regards to standards within the industries. High quality data as well helps with marketing to help organizations and or corporations utilized the data and make commercials, posters, and advertisements geared towards customers with their interest at hand. High quality data is also accessible by authorized users and having quality data helps with decision making.

### Organizational Goals:

### Data assessment – This step ensures the data quality and helps to clean the data set to provide only high-quality data. It helps to create an outline as well to ensure that the data is following the Bruce Incorporation and Wayne Enterprise guidelines of compliance.

* **Data Quality Plan** – Helps to differentiate objectives for all teams that the organization has on hand. This step is an outline for the responsibility of each team.
* **Completeness** – This step we will review further to ensure the accuracy of the data and eliminate any further errors it may have to produce a high-quality dataset. Also, to help ensure that the is protected for the organization.

### Data Quality Characteristics and Procedures: Data Quality Dimensions are included below:

* **Integrity –** Information not conformed to a specific setting or following the requirements of an organization or business.
* **Uniqueness –** This step ensures that no duplicates or overlaps in data exist. Thus, measuring data sets against other datasets.
* **Validity –** Compare values of the data set to values within a defined space.
* **Completeness –** Summarized data that is compared within fields against controlled records.
* **Consistency –** Records count of values to be used in past instances for the purpose of data population.
* **Timelines –** This refers to the data that is delivered against the data that was scheduled to be delivered.

### Security and Personnel Responsibility Plan:

### People that work with data are first in line for the protection of the data. This as well goes for the organization as whole who handles data. With this manner, individuals who have access to data should handle the data in a secure manner. This well help to ensure that the data is not access by unauthorized users. At the same time. Sensitive data should be encrypted to protect the data from being exposed to vulnerabilities. As outlined previously, data policies are put in place to be implemented on all levels of access within the organization. Methods and steps are outlined as well to ensure that the data is cleansed properly, accessible by authorized users, and have the data use in a manner deemed to be of importance when combined with other datasets. This will help the company and its shareholders be at ease, knowing that the datasets are maintained and secured in high-quality standards.

**Directions Part 2:**

Profile the data and select and report on your findings. Ensuring to use best practices while assessing the data. Using the information provided, assess the quality of the data to translate and load the data.

* **Explain your plan to resolve the errors and anomalies in the data profile:** Using filters, identify errors and anomalies in the data set and determine a plan for resolution.
* **Evaluate the data types present in the data set based on the data dictionary:** Compare the data with the information within the dictionary. Examine the data and compare it to the database.
* **Evaluate how you will transform your data to import the data into the existing database:** Document the specific transformation that are necessary to join the new data sets to the existing database.
* **Effectively communicate data quality issues and the steps towards resolutions to stakeholders in a summary:** Explain you plan for cleaning and preparing the data in the organizational merge.

## Data Set Anomalies

The first row is an example.

| **Key Value** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| *Example:*  *101339052 (Employee Number field)* | *Example: Last name missing* | *Example: Delete row* |
| Employee Name | Missing Last Name | Row Delete |
| Race | Character Instead of code | Change to code |
| Sex | Character Instead of code | Change to code |
| Marital Status | Character Instead of code | Change to code |

## Data Types

You may need to add rows. The first three rows provide example answers. You will need to assess all header names from your chosen data set.

| **Header Name From File** | **Data Types Note** |
| --- | --- |
| *Example: State* | *Example: OK* |
| *Example: DOB* | *Example: Won’t Use* |
| *Example: Dates Employed* | *Example: Needs to be rounded* |
| Marital Status | Integer |
| Pay Rate | Rounded |
| Race | Code or convert |
| Employee Number | Ok |
| Zip code | Ok |
| Citizen | Ok |
| Sex | Ok |
| Hire Date | Ok |
| Termination Date | Ok |
| Status of Employee | Ok |
| Termination reason | Ok |
| Position | Ok |
| Department | Ok |
| Employee Source | Ok |
| Performance Score | Ok |
| Age | Ok |
| State | Ok |
| DOB | Change to age instead of DOB |
| Days Employed | Ok |

## Specific Transformations Needed to Join the Data

You may need to add rows. The first row is an example. Responses may have one to three Excel functions.

| **Header Name From File** | **Excel Function One** | **Excel Function Two** | **Excel Function Three** |
| --- | --- | --- | --- |
| *Example: Smoker Status* | *Example:*  *IF(A1:A52=”Yes”,1,0)* |  |  |
| Citizen | IF(J1:J106=”Citizen”,1,0) | IF(J1:J106=”Noncitizen”,0,1) | IF(J1:J106=”N/A”,2,0) |
| Sex | IF(I1:106=”Male”,1,0) | IF(I1:106=”Female”,0,1) | IF(I1:106=”N/A”,2,0) |
| Status of Marriage | IF(H1:106=”Single”,1,0) | IF(J1:J106=”Married”,0,1) | IF(J1:J106=”N/A”,2,0) |

## Executive Summary

Upon analyzing the dataset that was provided from Wayne Enterprise. There were some discrepancies that I have noticed. Fields were either missing or had to be changed to the correct information for the dataset to be free of errors and no duplicates. Meaning, the dataset within Excel must be cleansed further for it to migrate into Wayne Enterprise database. With further analysis, the fields in question have been cleansed and corrected to match Bruce Incorporation database. Thus, the dataset can be analyzed.

**Directions Part Three:**

**Clean data** using the errors and anomalies identified in a Data Profile.

**Transform data** using the transforms identified from a Data Profile.

* **Perform a validation of the transformed and cleaned data.** Perform a count reconciliation of the data rows in a clean Excel worksheet. Use the template to calculate MIN, MAX, and AVERAGE of an identified variable.
  + **Counts:** Analyze the data to identify the counts of your variables to ensure all original rows are accounted for.
  + **MIN, MAX, AVERAGE:** Determine the distribution of data using the identified variables.
* **Summarize changes made to the data set when data was cleaned and transformed.** Use the template to complete your summary.
  + **Distribution data:** Describe how the distribution data tells you if the import changes the existing distribution.
  + **Summarize why the validation steps help to ensure the final import data is clean and of high quality.**

## File Names Table

Perform the following data validation on each of the import, existing, and merge data tables.

* The COUNT of rows in each table. You should reconcile these counts to make sure the number of rows you inserted made it into the merged table.

|  | **Excel Files** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Source File** | **Anomalies** | **Import Data** | **Existing Data** | **Merge Data** |
| **Count** | 106 | 1 | 105 | 206 | 310 |

* Use the identified variable for **your chosen data set** as listed below to determine MIN, MAX and AVERAGE for each table:

**Data Set:** Twitter

**Variable:** Airline Score by Airline name

**Data Set:** Human Resources

**Variable:** Pay Rate by Department

**Data Set:** Supermarket

**Variable:** Gross Income by location

Determine the MIN, MAX, and AVERAGE in your uploaded table, the MIN, MAX, and AVERAGE in the existing table before you inserted the new rows, and the MIN, MAX, and AVERAGE in the existing table after you insert new rows.

|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Data** | **Merge Data** |
| **MIN** | 15 | 14 | 14 |
| **MAX** | 80 | 64 | 80 |
| **AVERAGE** | 35.326 | 29.1 | 31.2006 |

## Summary

Distribution data is a function of possible listings while showing possible values of the dataset and how they occur most often. The data from distribution can inform us of the changes of existing distribution and using plots of probability. It can result from imported data. A sure-fire way to ensure that your dataset is in the current distribution is by importing the data and if it follows a straight line (plot) it can be seen as corrected data. Vice versa, if the plots do not line up, then it means that the distribution data is not correct and may have some errors within the dataset.

Validating data is a mostly used method within any industry. It can be used to validate data that may fall between ranges of adequate and nonsufficient. Meaning, it falls within values of the data field. This step helps to correctly ensure that the dataset is cleansed while being of high-quality. In this step, it can help to find incomplete data and data that needs to be corrected. Lasty, it helps with authorize users entering invalid data within certain parameters.