**ETL Report Guide**

Abbey Guillat, Regina Huber, Annie Kittendorf, Lindsey Oh, Scott Partacz, Raymond Sepulveda  
 ETL Process Date: 01/13/2022

**Introduction**

Set the stage. Introduce the problem that you are trying to solve. Identify sources of data. Describe why the data needs to be transformed.

By considering the proportion of seats held by women in national parliaments from years 1997-2020, we hope to explore the following questions:

* How do trends in the US compare over time? Ray
* How do trends in the world change over time? Annie
* How do income groups affect parliament trends over time? Abbey
* Taking income group into consideration, how do parliament trends differ between countries categorized as higher vs. lower income groups? Scott
* What does the data look like for 2020 for all countries? Regina
* What does the data look like for 2001 for all countries? Lindsey

**Data Sources**

The data was obtained from The World Bank Group (Proportion of seats held by women in national parliaments (%)): <https://data.worldbank.org/indicator/SG.GEN.PARL.ZS?view=chart>.

The two CSV files downloaded for this project are:

* API\_SG.GEN.PARL.ZS\_DS2\_en\_csv\_v2\_3469491.csv
* Metadata\_Country\_API\_SG.GEN.PARL.ZS\_DS2\_en\_csv\_v2\_3469491.csv

**Extraction**

1. Import pandas
2. Load data files using the pandas CSV reader

**Transformation**

Did you use all of the data you extracted as-is? Did you remove columns? Did you change columns’ names? Did you change your column formats? What steps were taken to get the data in a form that you could use it? Be sure to number steps when the order matters.

Since your end goal will be to load your data in SQL Server, include table mappings that identify the source data and its destination.

Cleaning API\_SG.GEN.PARL.ZS\_DS2\_en\_csv\_v2\_3469491.csv:

1. Columns 1960-1996 contain only null values for our dataset so we began by creating a list to hold these year values
2. Set column headers equal to row values with index=3
3. Drop rows [0:4]
4. Drop columns that contain all NaN values using year list
5. Drop rows which contain only NaN values

Cleaning Metadata\_Country\_API\_SG.GEN.PARL.ZS\_DS2\_en\_csv\_v2\_3469491.csv:

1. Drop all columns except for Country Code and IncomeGroup
2. Drop rows which contain NaN values for IncomeGroup

Working with both tables:

1. Perform an inner join on the two tables (on the Country Code)

**Load**

This data is not being loaded onto a database.

**Conclusion**

The data was successfully extracted and transformed for project use.