KJSCE/IT/SYBTech/SEM IV/AD/2022-23



**Experiment No.8**

**Title: Execution of ETL process**

**Batch:A2 Roll No.:16010421073 Experiment No.:8**

**Aim:** Execution of ETL process

# Resources needed: Different RDBMS such as MySQL, Postgres and Excel, CSV, Rapidminer 5.3/ Latest vision

**Theory**

# Data Warehouse:

An analytics-focused type of data management system called a data warehouse is intended to assist and allow business intelligence (BI) activities. Large amounts of historical data are frequently included in data warehouses, which are only designed to be used for queries and analysis. Application log files and transaction apps are only two examples of the many different sources from which the data in a data warehouse often comes.

Big data from various sources is centralised and combined in a data warehouse. Because of its analytical skills, businesses can get more out of their data and make better decisions. It gradually compiles a historical record that data scientists and business analysts can find quite useful. Because to these features, a data warehouse can be regarded as an organization's "single source of truth."

# ETL:

Extract, Transform, Load (ETL) refers to a process in database usage and especially

in data warehousing. Data extraction is where data is extracted from homogeneous or heterogeneous data sources; data transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; data loading where the data is loaded into the final target database, more specifically, an operational data store, data mart, or data warehouse.

One may improve their chances of achieving better connection and scalability by employing a well-established ETL framework. A decent ETL tool must be able to interface with the several different relational databases and read the various file formats employed by a business. ETL solutions have started to move into Enterprise Application Integration, or even Enterprise Service Bus, systems that now encompass a lot more than simply the extraction, transformation, and loading of data. Converting CSV files into formats usable by relational databases is one frequent use case for ETL technologies. ETL solutions make it feasible for users to input csv-like data feeds/files and import it into a database with as little code as possible, facilitating a typical translation of millions of records. ESTL instruments

# RapidMiner:

RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. RapidMiner is written in the Java programming language. RapidMiner provides a GUI to design and execute analytical workflows. Those workflows are called “Processes” in RapidMiner and they consist of multiple “Operators”. Each operator performs a single task within the process, and the output of each operator forms the input of the next one.

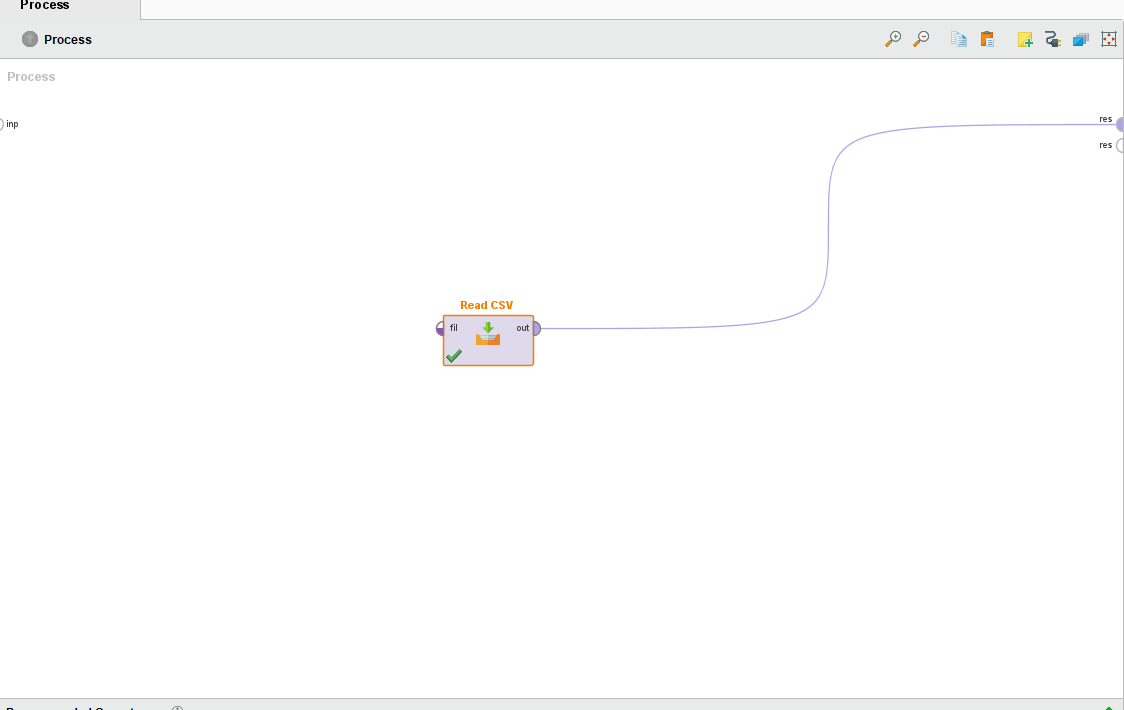
Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. RapidMiner provides learning schemes, models and algorithms and can be extended using R and Python scripts.

# Activities:

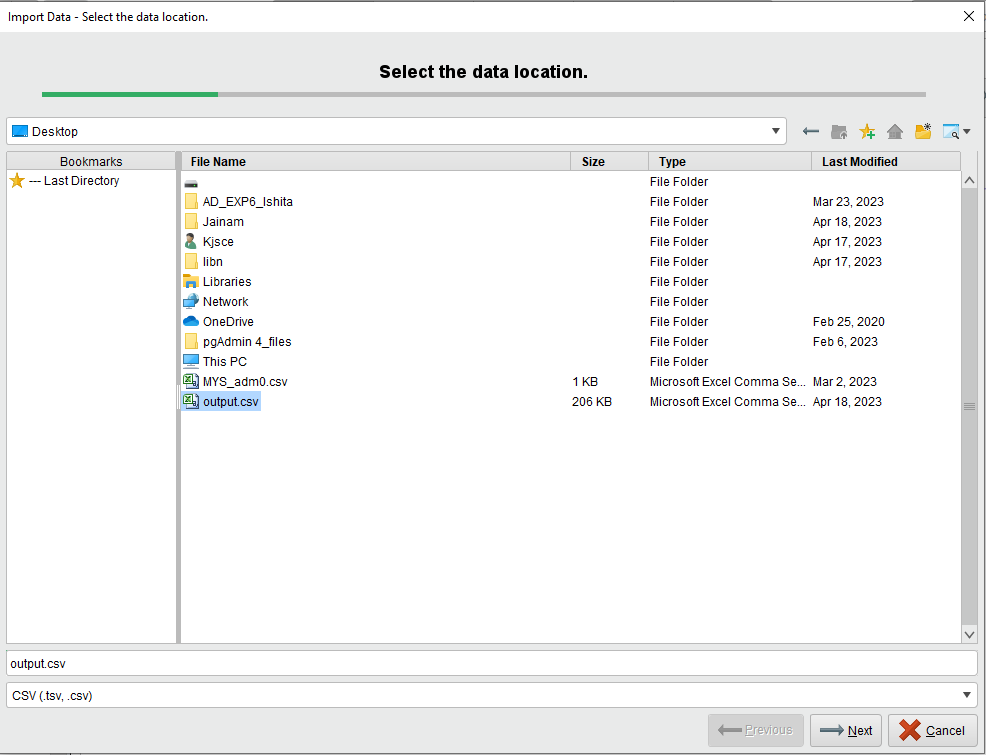
**For ETL:**

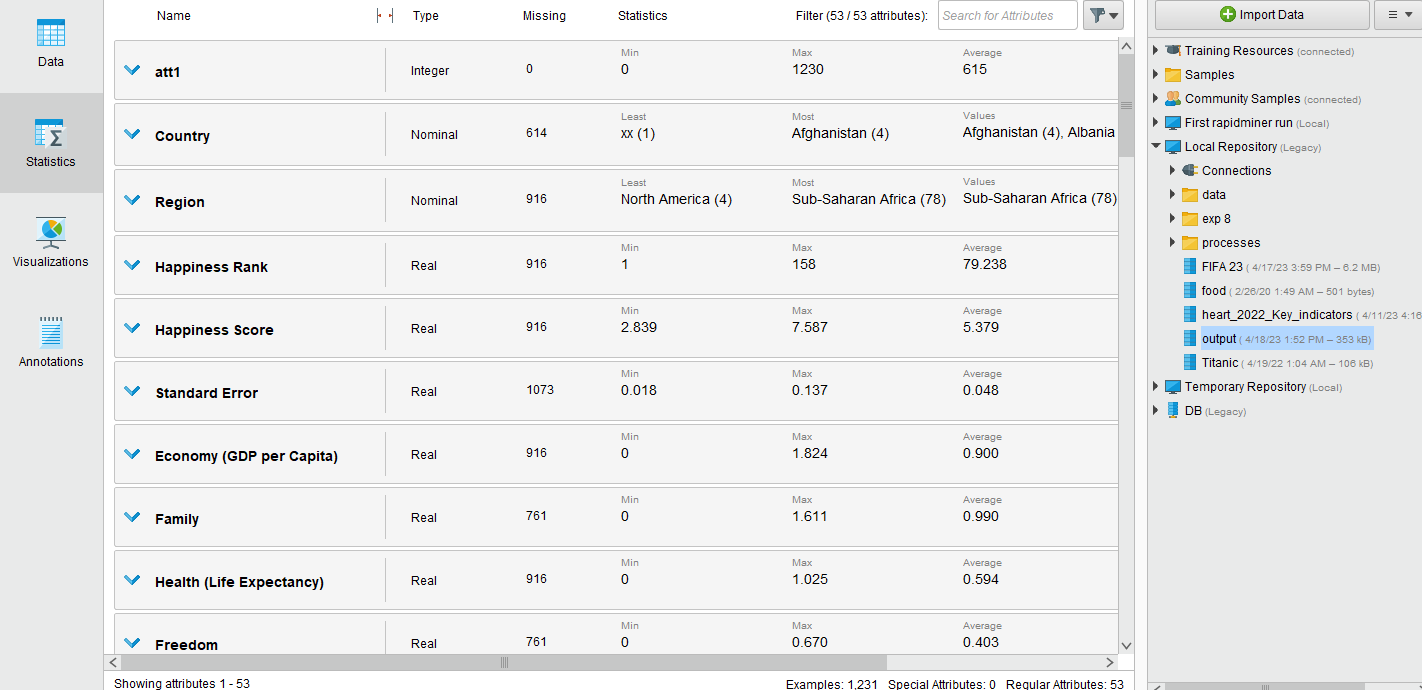
1. Go through the tutorial provided by RapidMiner
2. Install https://rapidminer.software.informer.com/download/#downloading
3. Extract data from 2 to 3 heterogeneous sources such as excel, MYSQL, Postgres etc.
4. Download any data set from *htt*[*ps://w*](http://www.kaggle.com/datasets)*ww.k*[*aggle.com/datas*](http://www.kaggle.com/datasets)*e*[*ts*](http://www.kaggle.com/datasets) or similar website
5. Apply five different transformations and filters to the data with specific requirement
6. Prepare a report for the activities 2 and 4 (ETL part) with steps and visualisations applied.

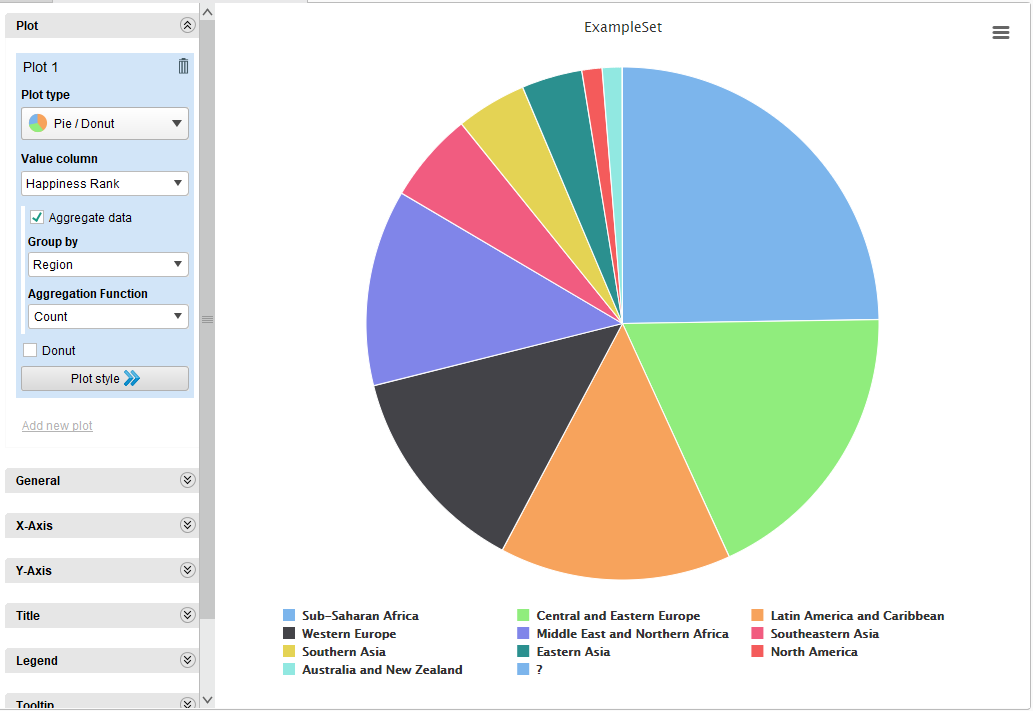
# Results:

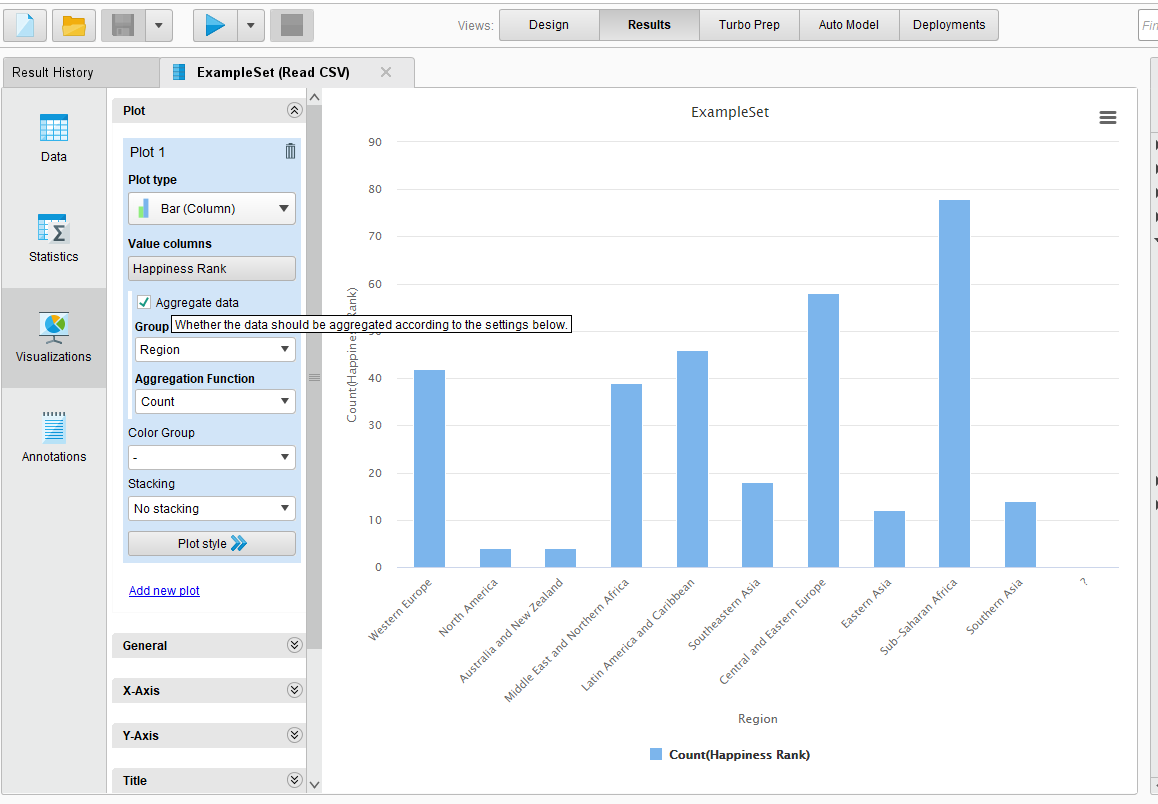


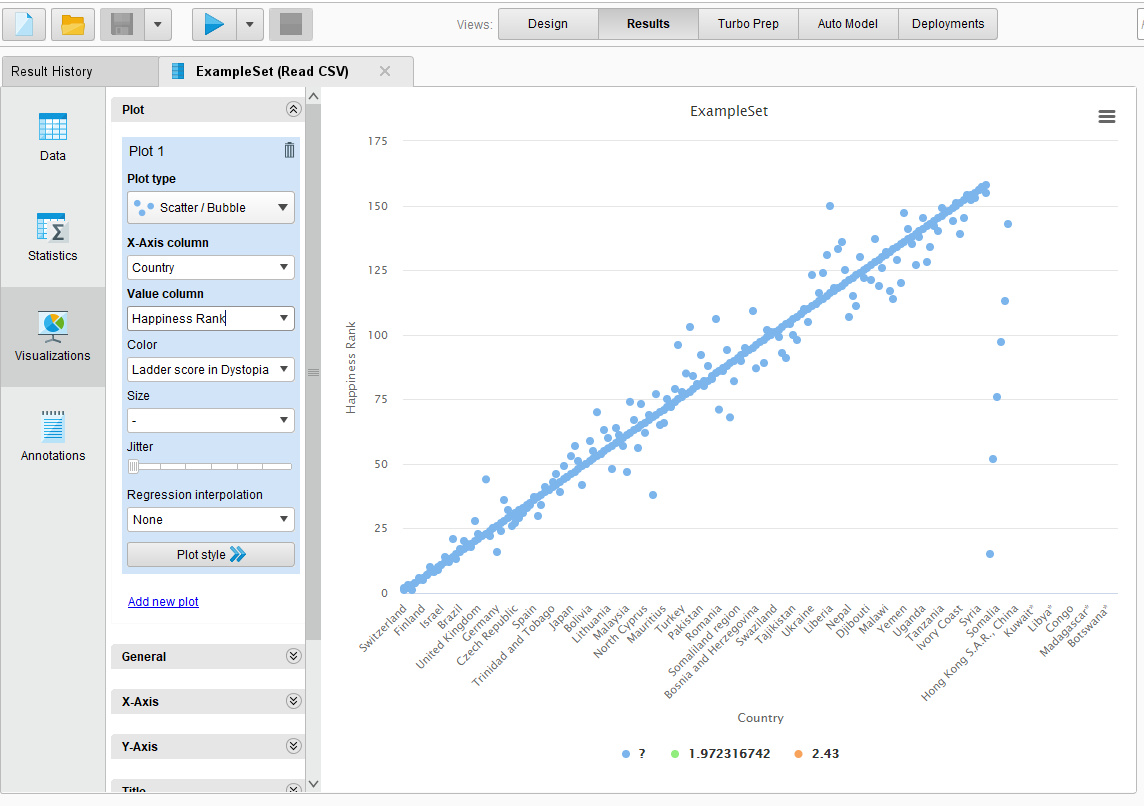
On double clicking, read CSV-



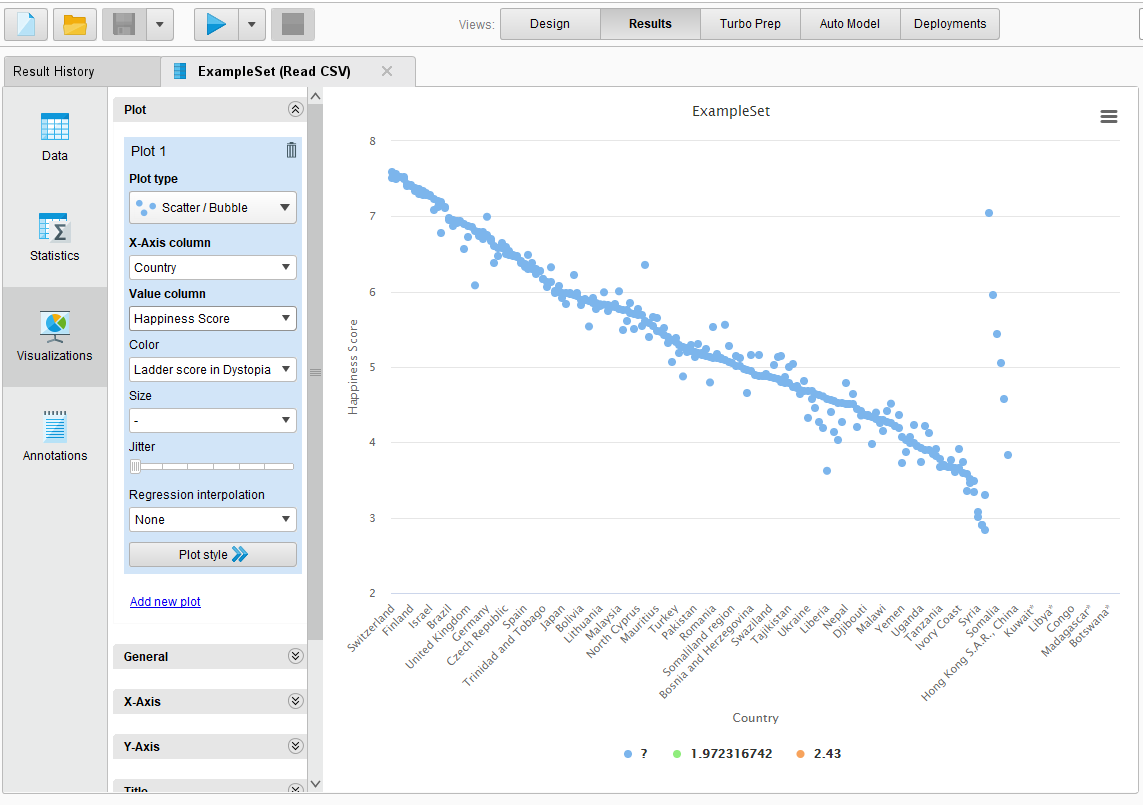








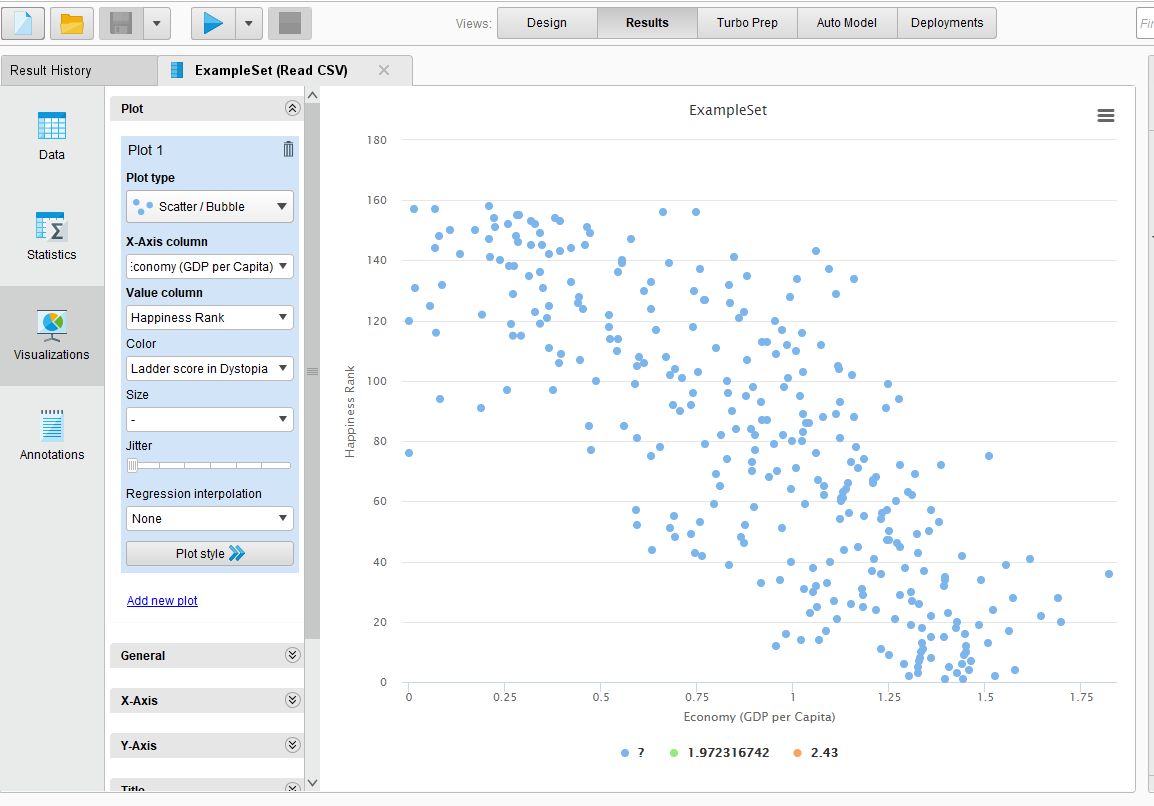
1. Switzerland has best happiness rank.



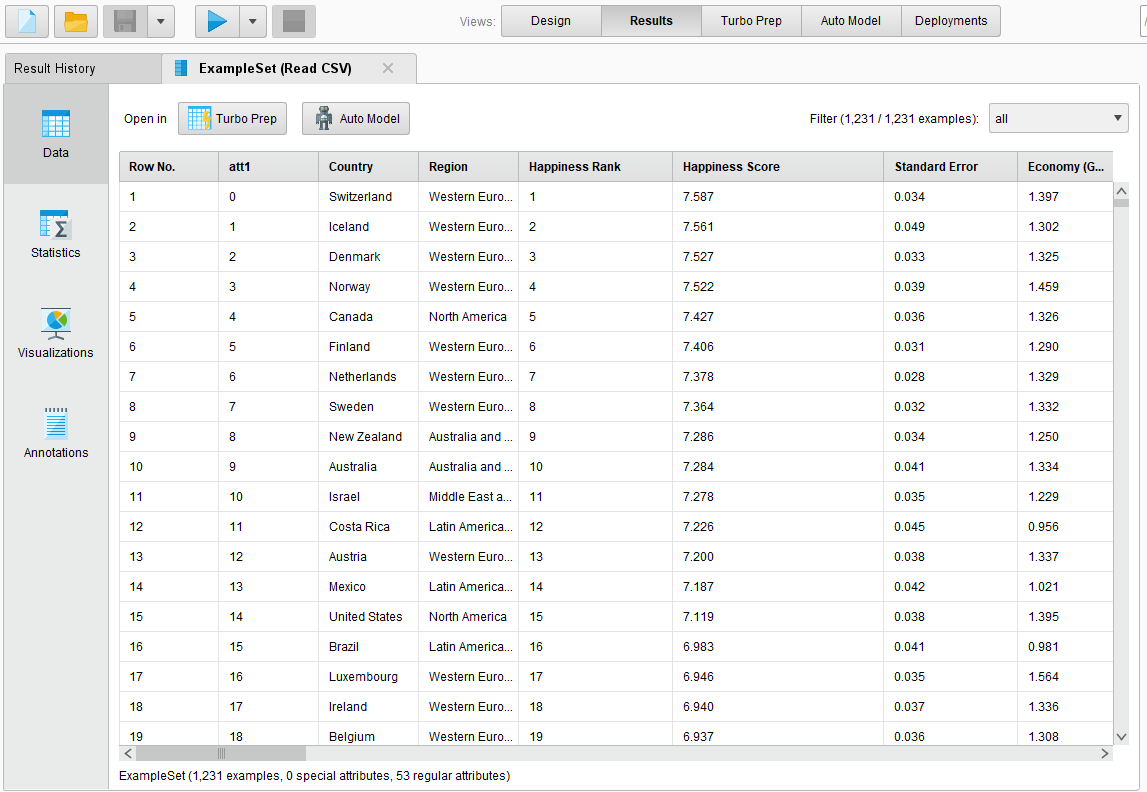
1. Switzerland has highest happiness score



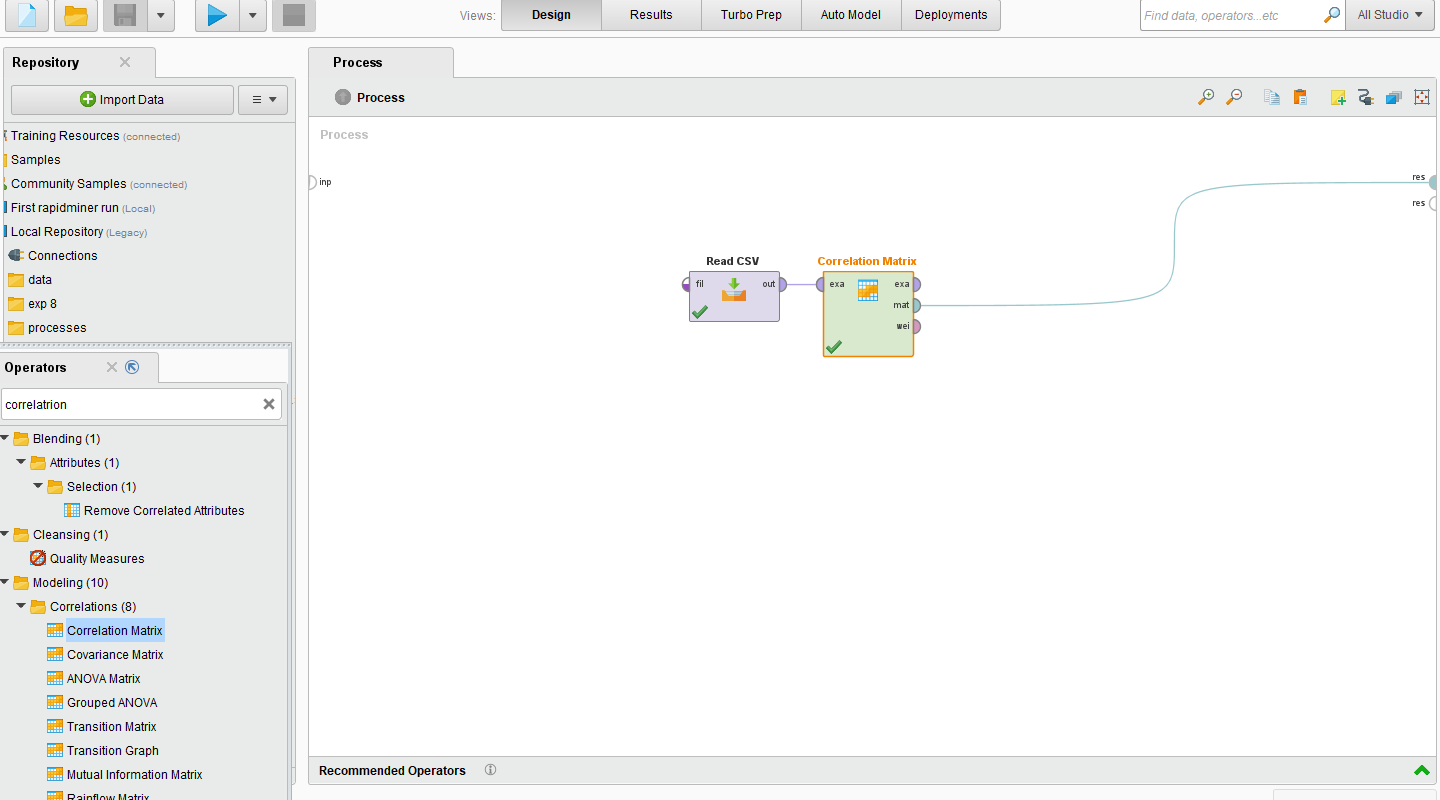
1. Australia and New Zealand has consistently high happiness score

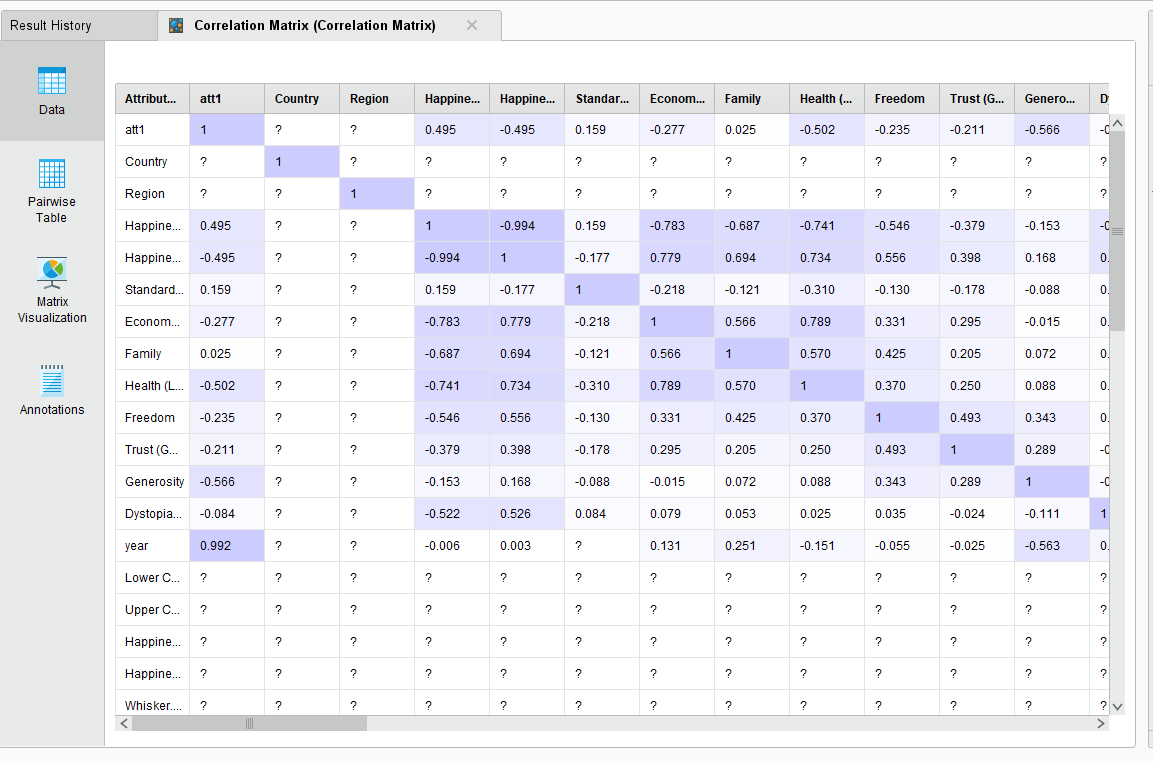


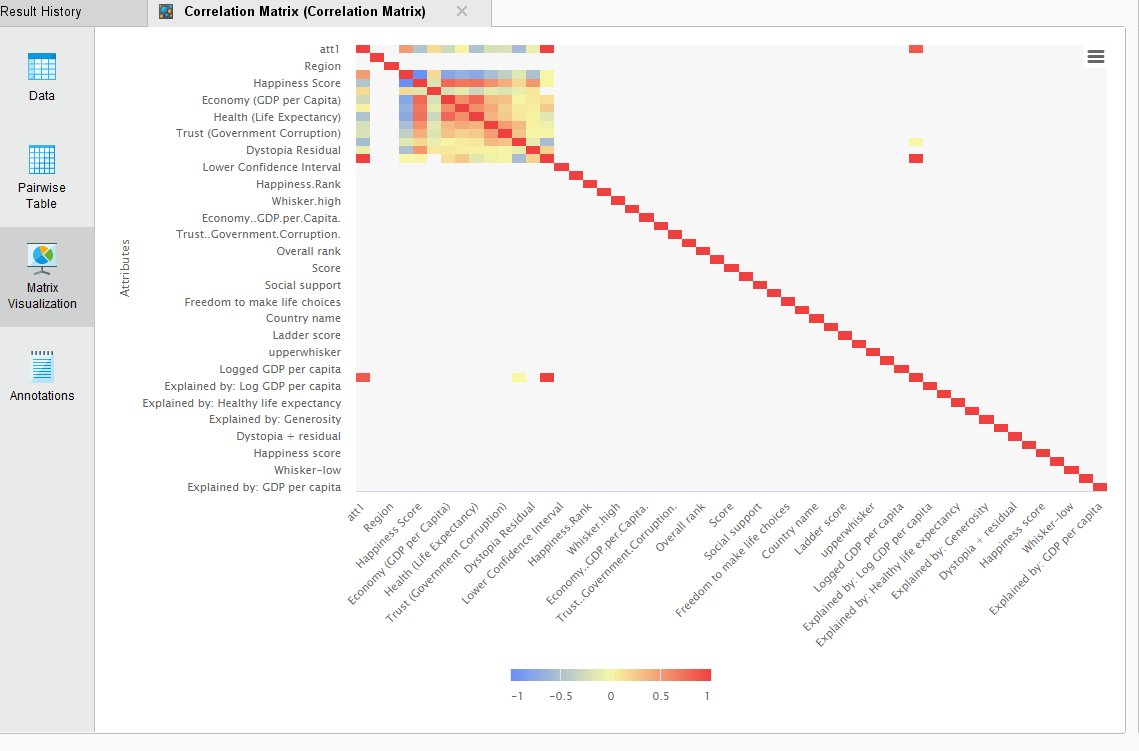
1. Countries having higher economy rank higher at happiness rank (meaning they are more happy).



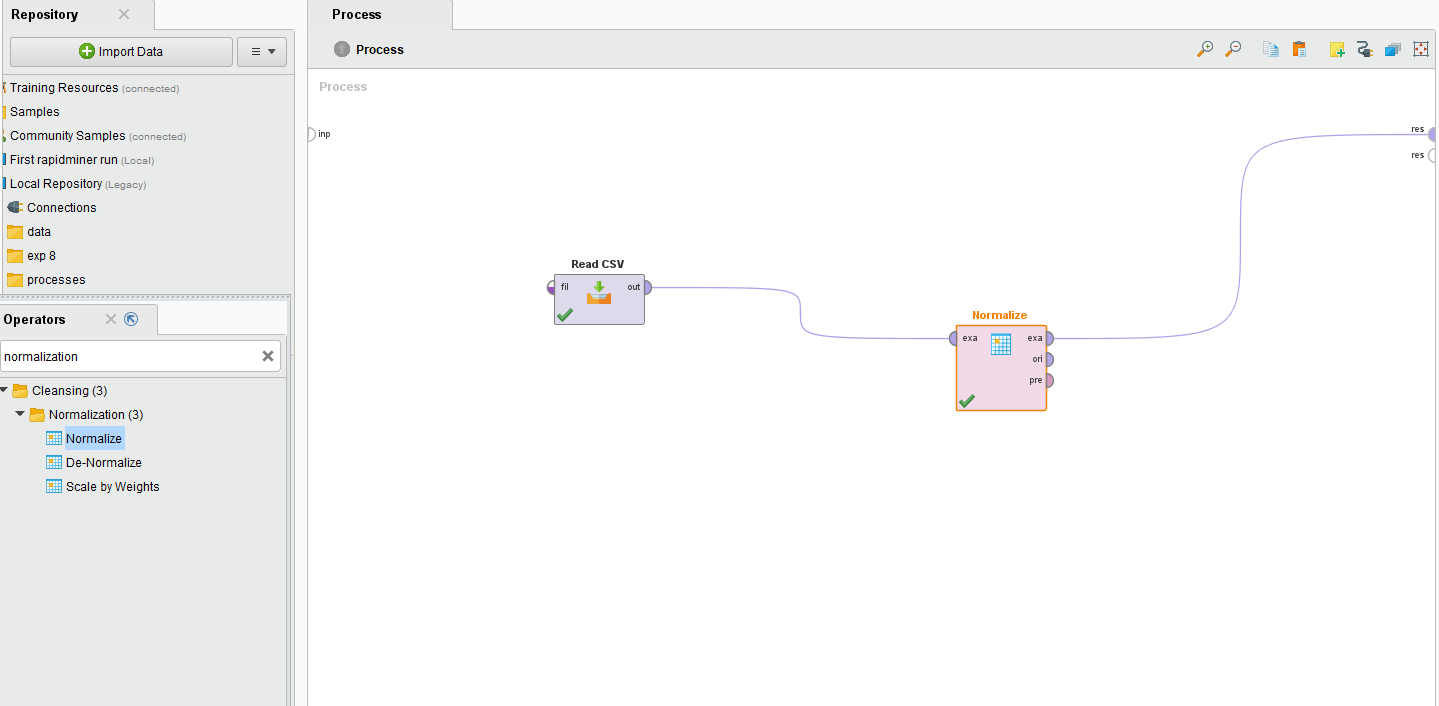
Correlation Matrix

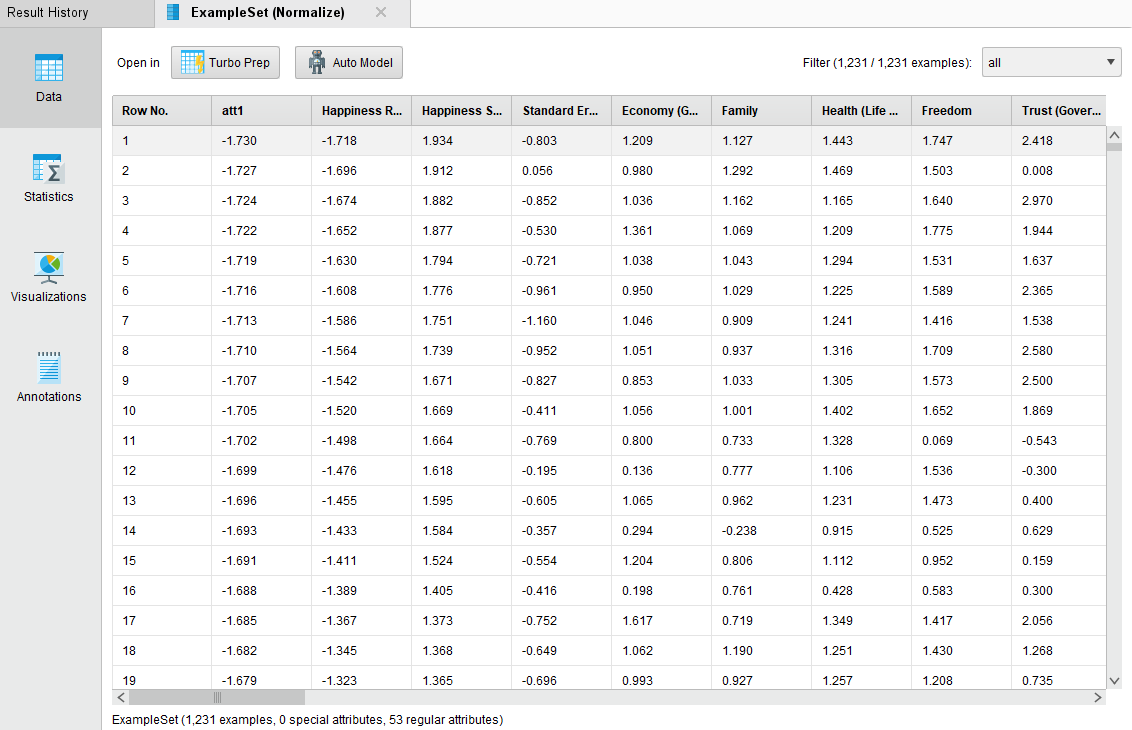






Normalization





**Report for ETL**

**Step-1) First we downloaded dataset from Kaggle.**

**Step-2) Then we**

# Outcomes:

**CO4 Apply ETL processing and Online Analytical Processing on the warehouse data.**

**Conclusion: (Conclusion to be based on the outcomes achieved)**

**Thus we successfully implemented execution of ETL process by applying different transformations on dataset and flters to a specific requirement.**

# Grade: AA / AB / BB / BC / CC / CD /DD

**Signature of faculty in-charge with date**

# References:

* https:/[/www.ora](http://www.oracle.com/in/database/what-is-a-data-warehouse)c[le.com/in/database/what-is-a-data-warehouse](http://www.oracle.com/in/database/what-is-a-data-warehouse)
* Paulraj Ponniah, “Data Warehousing: Fundamentals for IT Professionals”, Wiley India