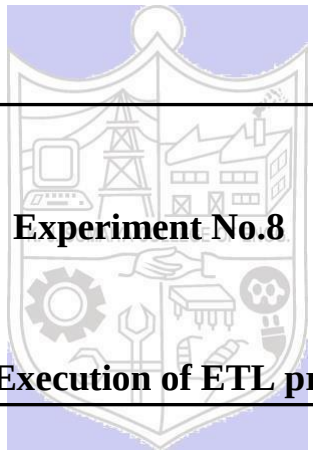


Experiment No.8

Title: Execution of ETL process

Name: Chinmay Mhatre **Batch: A2** **Roll No.: 16010421059**

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 employinga 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

Rapid Miner:

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. Rapid Miner 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. Rapid Miner provides learning schemes, models and algorithms and can be extended using R and Python scripts.

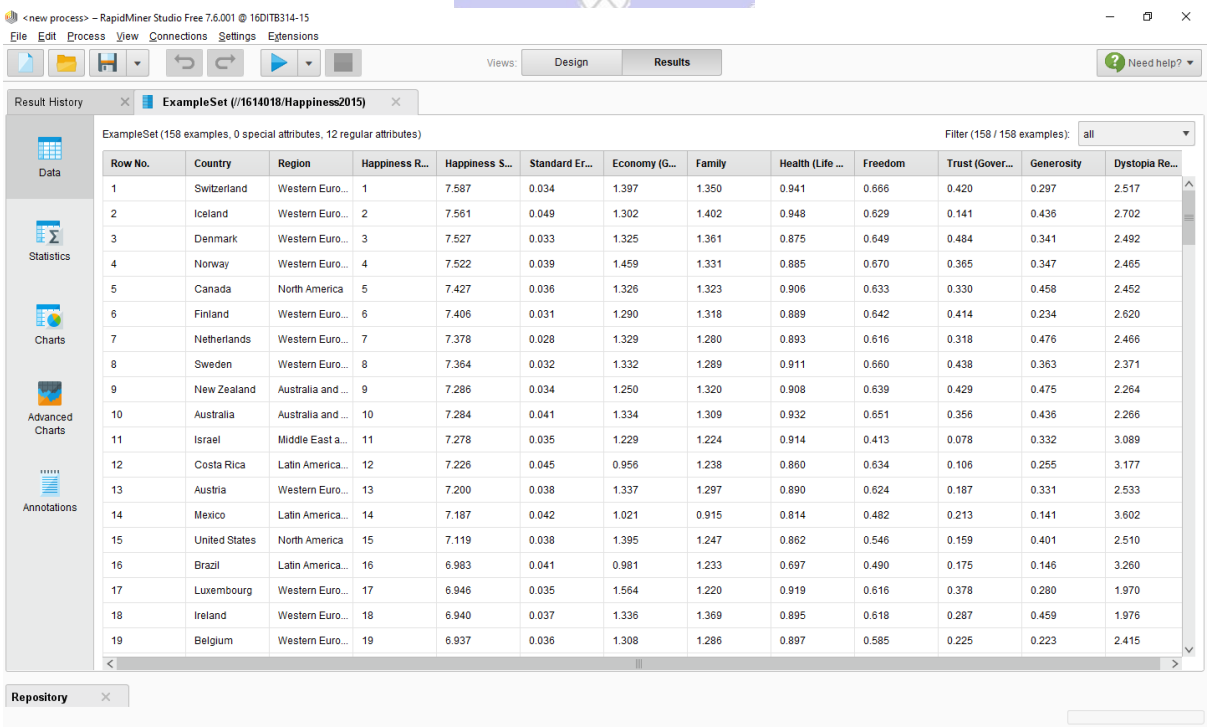
Activities:
For ETL:

1. Go through the tutorial provided by Rapid Miner
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 *https://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:

Report for ETL

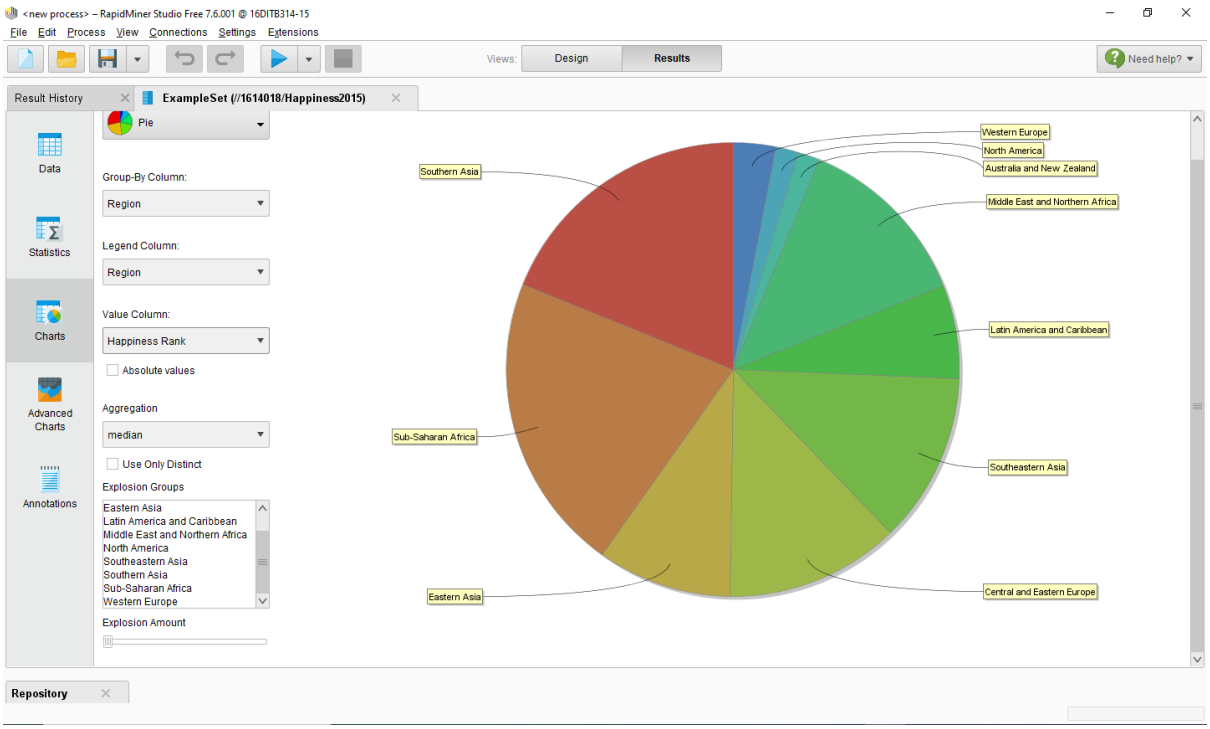
1. Check Basic Statistics for imported data



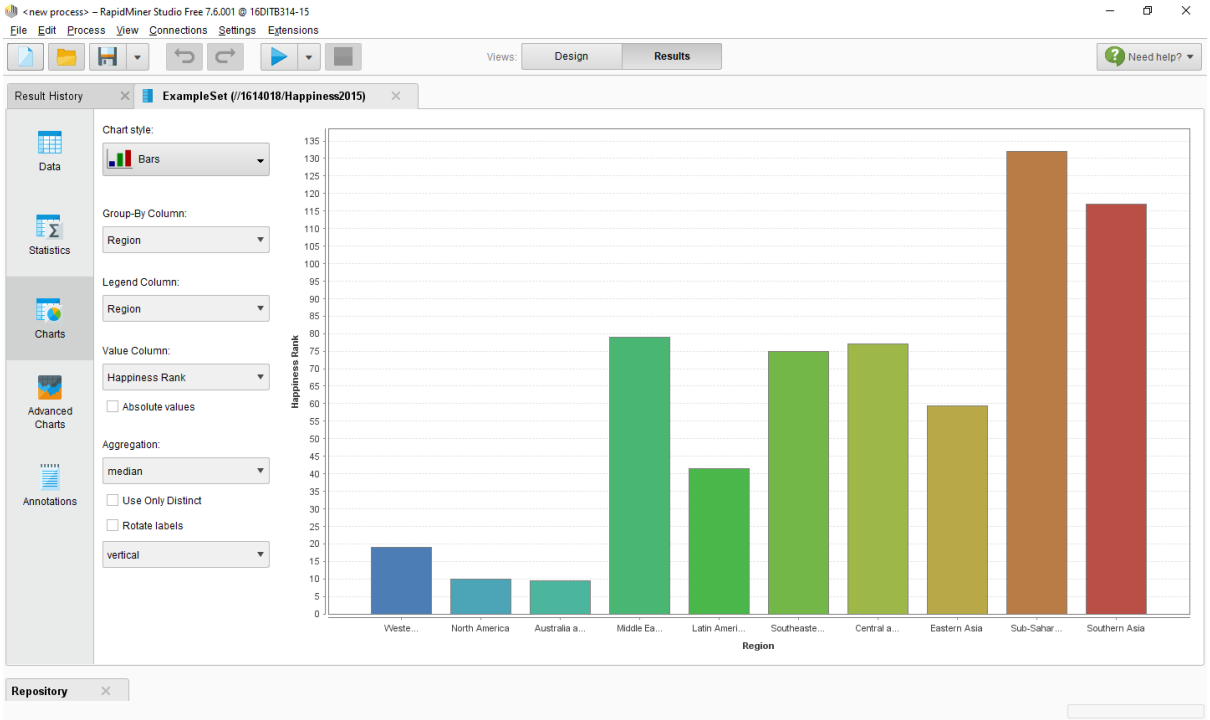
Row No.	Country	Region	Happiness R...	Happiness S...	Standard Er...	Economy (G...	Family	Health (Life ...	Freedom	Trust (Gover...	Generosity	Dystopia Re...
1	Switzerland	Western Euro...	1	7.587	0.034	1.397	1.350	0.941	0.666	0.420	0.297	2.517
2	Iceland	Western Euro...	2	7.561	0.049	1.302	1.402	0.948	0.629	0.141	0.436	2.702
3	Denmark	Western Euro...	3	7.527	0.033	1.325	1.361	0.875	0.649	0.484	0.341	2.492
4	Norway	Western Euro...	4	7.522	0.039	1.459	1.331	0.885	0.670	0.365	0.347	2.465
5	Canada	North America	5	7.427	0.036	1.326	1.323	0.906	0.633	0.330	0.458	2.452
6	Finland	Western Euro...	6	7.406	0.031	1.290	1.318	0.889	0.642	0.414	0.234	2.620
7	Netherlands	Western Euro...	7	7.378	0.028	1.329	1.280	0.893	0.616	0.318	0.476	2.466
8	Sweden	Western Euro...	8	7.364	0.032	1.332	1.289	0.911	0.660	0.438	0.363	2.371
9	New Zealand	Australia and ...	9	7.286	0.034	1.250	1.320	0.908	0.639	0.429	0.475	2.264
10	Australia	Australia and ...	10	7.284	0.041	1.334	1.309	0.932	0.651	0.356	0.436	2.266
11	Israel	Middle East a...	11	7.278	0.035	1.229	1.224	0.914	0.413	0.078	0.332	3.089
12	Costa Rica	Latin America...	12	7.226	0.045	0.956	1.238	0.860	0.634	0.106	0.255	3.177
13	Austria	Western Euro...	13	7.200	0.038	1.337	1.297	0.890	0.624	0.187	0.331	2.533
14	Mexico	Latin America...	14	7.187	0.042	1.021	0.915	0.814	0.482	0.213	0.141	3.602
15	United States	North America	15	7.119	0.038	1.395	1.247	0.862	0.546	0.159	0.401	2.510
16	Brazil	Latin America...	16	6.983	0.041	0.981	1.233	0.697	0.490	0.175	0.146	3.260
17	Luxembourg	Western Euro...	17	6.946	0.035	1.564	1.220	0.919	0.616	0.378	0.280	1.970
18	Ireland	Western Euro...	18	6.940	0.037	1.336	1.369	0.895	0.618	0.287	0.459	1.976
19	Belgium	Western Euro...	19	6.937	0.036	1.308	1.286	0.897	0.585	0.225	0.223	2.415

2. Check Happiness Rank by Region (Data Visualisation using Median Aggregation)

Pie Chart

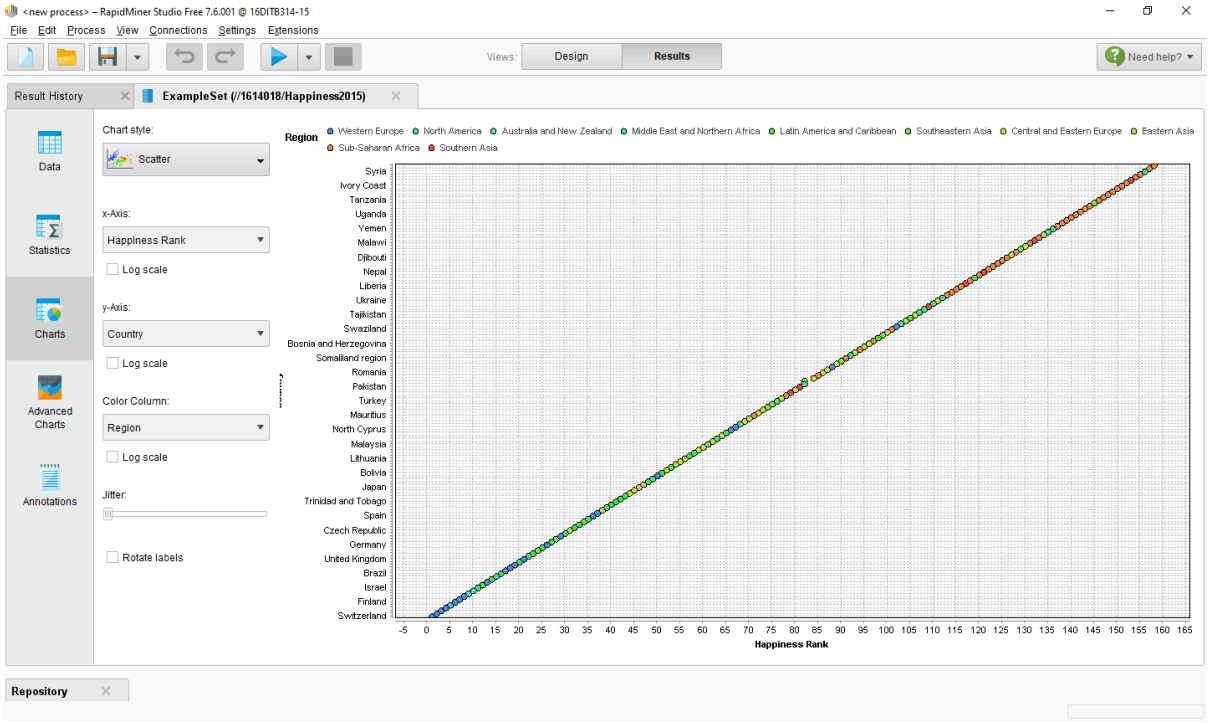


Bar Chart

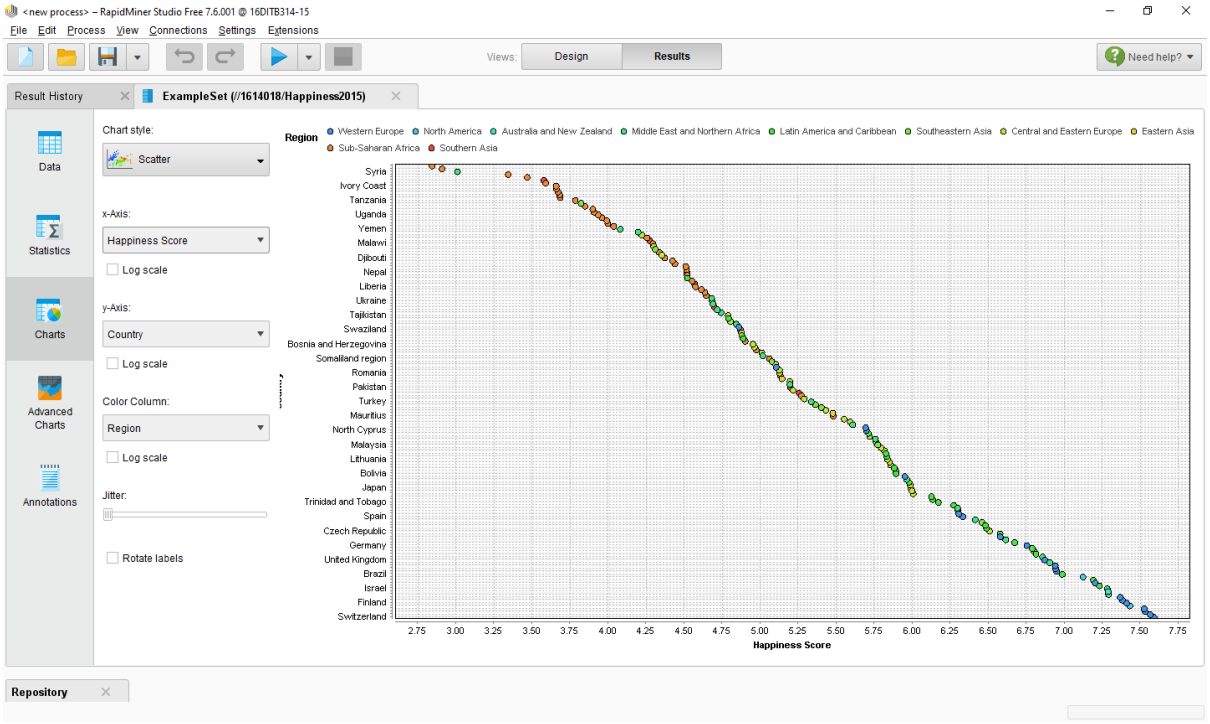


3. Apply Various chart styles for data analysis

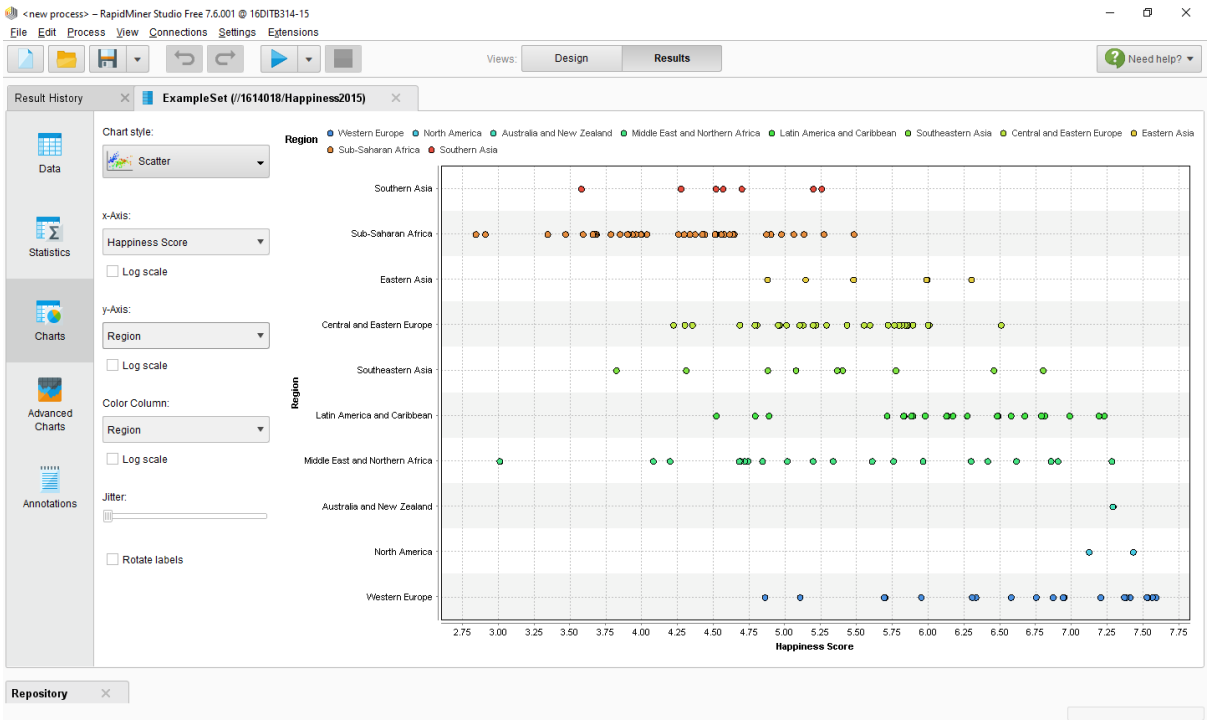
i) Happiness Rank vs Country



ii) Happiness Score vs Country

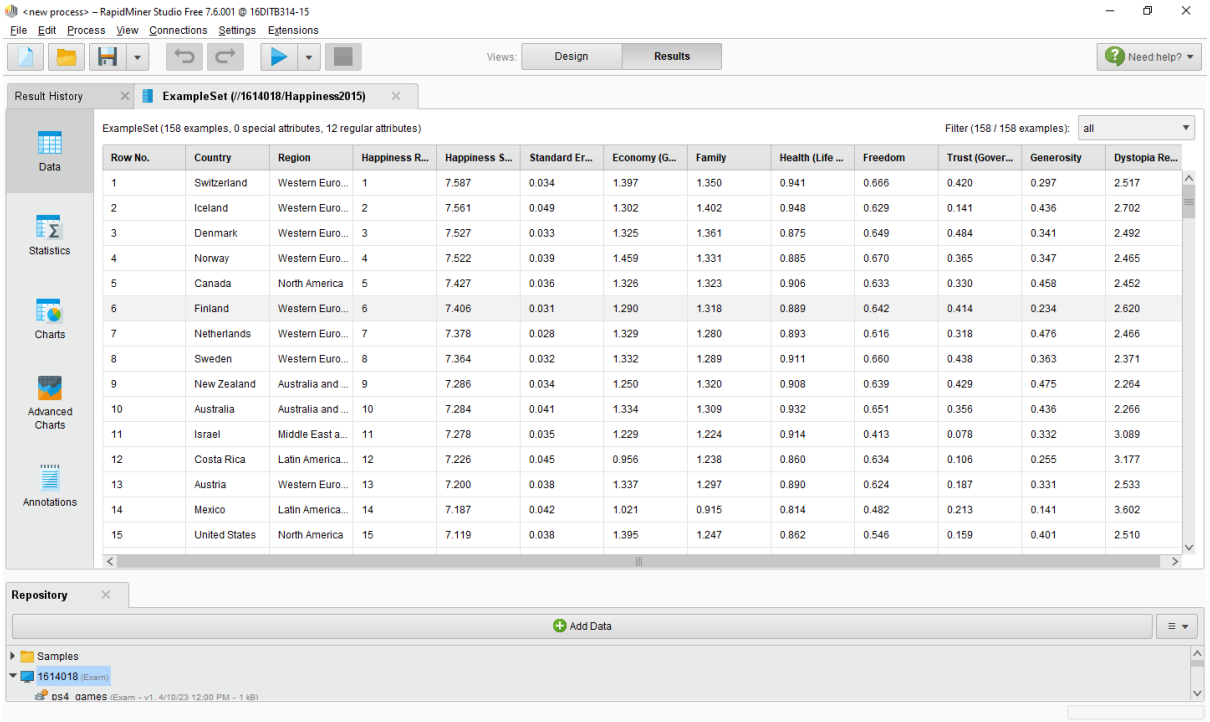


iii) Happiness Score vs Region

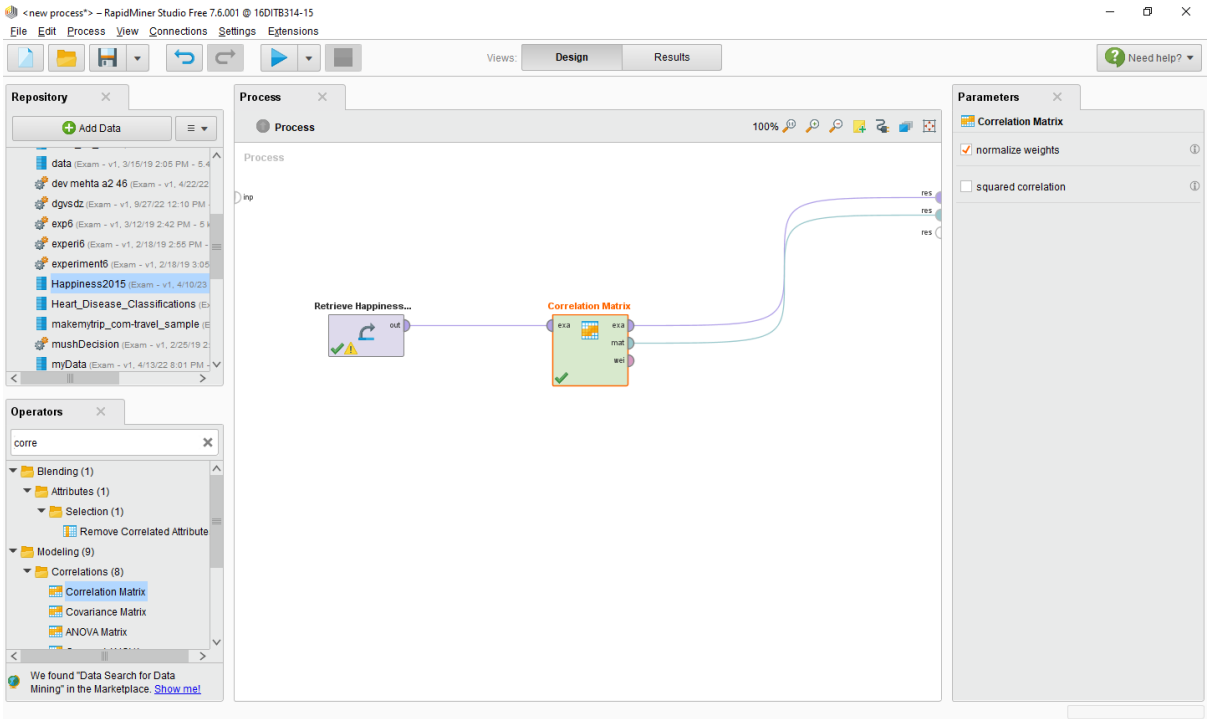


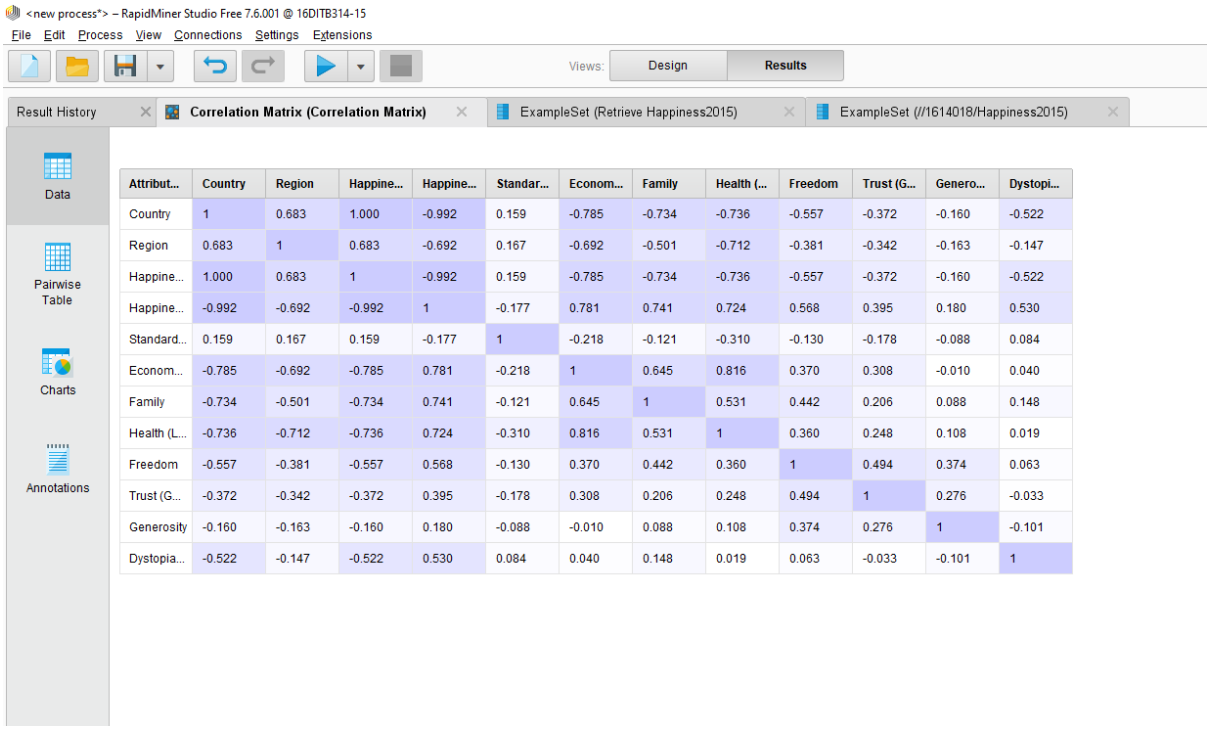
iv) Happiness Score vs Economy

4. Basic Display using Rapid Miner



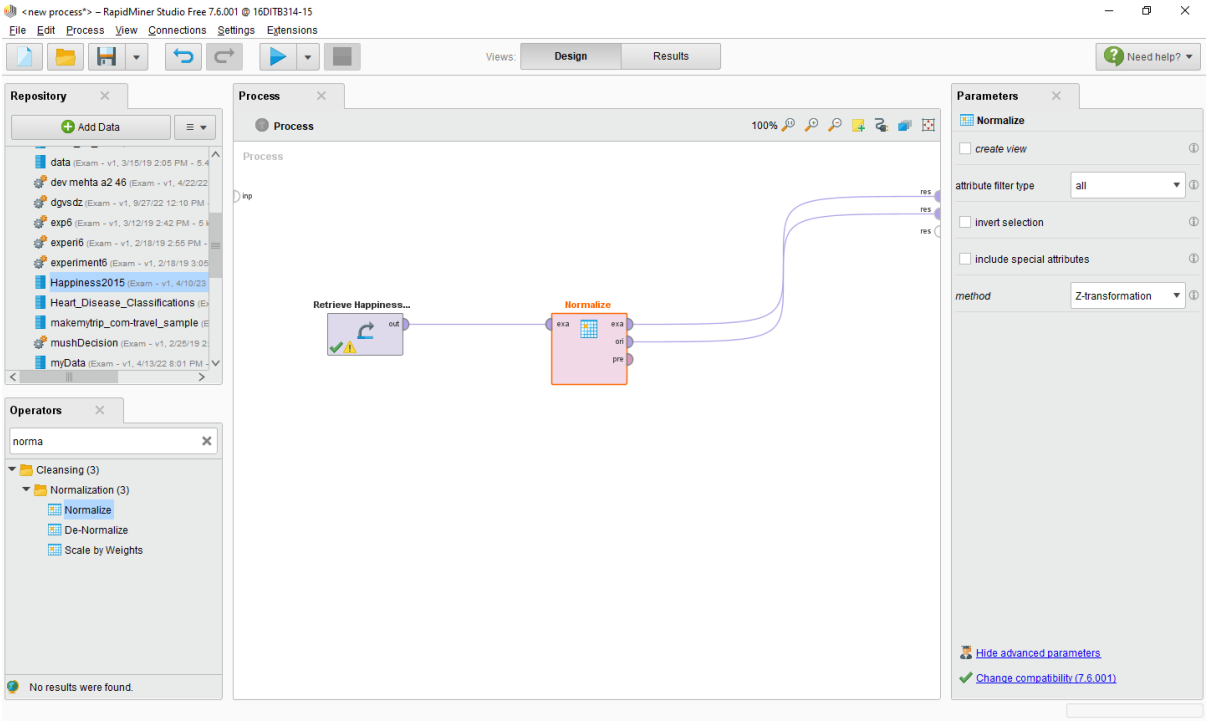
5. Correlation Matrix

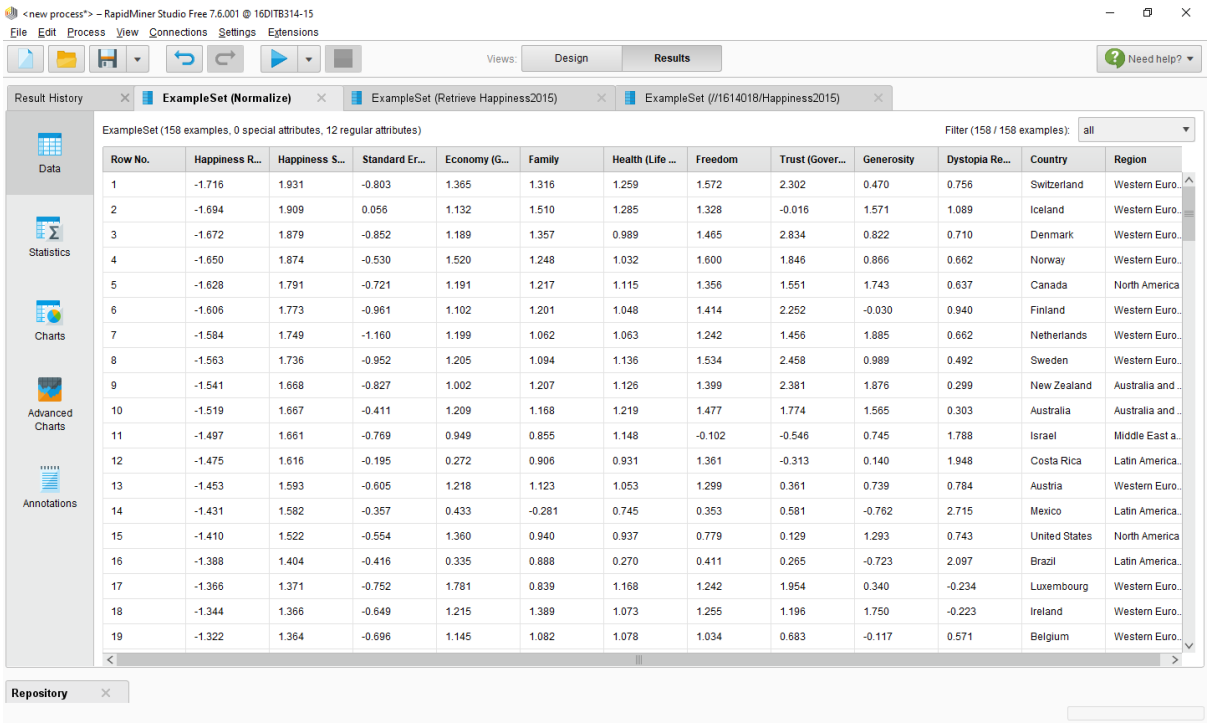




Attribut...	Country	Region	Happine...	Happine...	Standar...	Econom...	Family	Health (...	Freedom	Trust (G...	Genero...	Dystopi...
Country	1	0.683	1.000	-0.992	0.159	-0.785	-0.734	-0.736	-0.557	-0.372	-0.160	-0.522
Region	0.683	1	0.683	-0.692	0.167	-0.692	-0.501	-0.712	-0.381	-0.342	-0.163	-0.147
Happine...	1.000	0.683	1	-0.992	0.159	-0.785	-0.734	-0.736	-0.557	-0.372	-0.160	-0.522
Happine...	-0.992	-0.692	-0.992	1	-0.177	0.781	0.741	0.724	0.568	0.395	0.180	0.530
Standard...	0.159	0.167	0.159	-0.177	1	-0.218	-0.121	-0.310	-0.130	-0.178	-0.088	0.084
Econom...	-0.785	-0.692	-0.785	0.781	-0.218	1	0.645	0.816	0.370	0.308	-0.010	0.040
Family	-0.734	-0.501	-0.734	0.741	-0.121	0.645	1	0.531	0.442	0.206	0.088	0.148
Health (L...	-0.736	-0.712	-0.736	0.724	-0.310	0.816	0.531	1	0.360	0.248	0.108	0.019
Freedom	-0.557	-0.381	-0.557	0.568	-0.130	0.370	0.442	0.360	1	0.494	0.374	0.063
Trust (G...	-0.372	-0.342	-0.372	0.395	-0.178	0.308	0.206	0.248	0.494	1	0.276	-0.033
Generosity	-0.160	-0.163	-0.160	0.180	-0.088	-0.010	0.088	0.108	0.374	0.276	1	-0.101
Dystopia...	-0.522	-0.147	-0.522	0.530	0.084	0.040	0.148	0.019	0.063	-0.033	-0.101	1

6. Data Normalization





Row No.	Happiness R...	Happiness S...	Standard Er...	Economy (G...	Family	Health (Life ...	Freedom	Trust (Gover...	Generosity	Dystopia Re...	Country	Region
1	-1.716	1.931	-0.803	1.365	1.316	1.259	1.572	2.302	0.470	0.756	Switzerland	Western Euro...
2	-1.694	1.909	0.056	1.132	1.510	1.285	1.328	-0.016	1.571	1.089	Iceland	Western Euro...
3	-1.672	1.879	-0.852	1.189	1.357	0.989	1.465	2.834	0.822	0.710	Denmark	Western Euro...
4	-1.650	1.874	-0.530	1.520	1.248	1.032	1.600	1.846	0.866	0.662	Norway	Western Euro...
5	-1.628	1.791	-0.721	1.191	1.217	1.115	1.356	1.551	1.743	0.637	Canada	North America
6	-1.606	1.773	-0.961	1.102	1.201	1.048	1.414	2.252	-0.030	0.940	Finland	Western Euro...
7	-1.584	1.749	-1.160	1.199	1.062	1.063	1.242	1.456	1.885	0.662	Netherlands	Western Euro...
8	-1.563	1.736	-0.952	1.205	1.094	1.136	1.534	2.458	0.989	0.492	Sweden	Western Euro...
9	-1.541	1.668	-0.827	1.002	1.207	1.126	1.399	2.381	1.876	0.299	New Zealand	Australia and ...
10	-1.519	1.667	-0.411	1.209	1.168	1.219	1.477	1.774	1.565	0.303	Australia	Australia and ...
11	-1.497	1.661	-0.769	0.949	0.855	1.148	-0.102	-0.546	0.745	1.788	Israel	Middle East a...
12	-1.475	1.616	-0.195	0.272	0.906	0.931	1.361	-0.313	0.140	1.948	Costa Rica	Latin America...
13	-1.453	1.593	-0.605	1.218	1.123	1.053	1.299	0.361	0.739	0.784	Austria	Western Euro...
14	-1.431	1.582	-0.357	0.433	-0.281	0.745	0.353	0.581	-0.762	2.715	Mexico	Latin America...
15	-1.410	1.522	-0.554	1.360	0.940	0.937	0.779	0.129	1.293	0.743	United States	North America
16	-1.388	1.404	-0.416	0.335	0.888	0.270	0.411	0.265	-0.723	2.097	Brazil	Latin America...
17	-1.366	1.371	-0.752	1.781	0.839	1.168	1.242	1.954	0.340	-0.234	Luxembourg	Western Euro...
18	-1.344	1.366	-0.649	1.215	1.389	1.073	1.255	1.196	1.750	-0.223	Ireland	Western Euro...
19	-1.322	1.364	-0.696	1.145	1.082	1.078	1.034	0.683	-0.117	0.571	Belgium	Western Euro...

Outcomes:

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

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

From the above experiment we downloaded a database from kaggle and imported data from the dataset into RapidMiner. Using RapidMiner we performed Data Visulaization techniques using different chart styles such as Bar Charts, Pie Charts, Scatter Chart. Then we applied Correlation Matrix and Data Normalization using Rapid Miner process tools.

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

Signature of faculty in-charge with date

References:

- **https://www.oracle.com/in/database/what-is-a-data-warehouse**
- **Paulraj Ponniah, "Data Warehousing: Fundamentals for IT Professionals", Wiley India**