รายงาน

เรื่องวิเคราะห์การเกิดแผ่นดินไหวจากสถิติที่ผ่านมา

โดย

นางสาวอภิญญา ประดิษฐผลเลิศ 5410501497

นายคณินท์ วรางคณากูล 5410503082

นางสาวฐิติมา ตโมหรณวงศ์ 5410503341

นางสาวปาณิสรา อินโต 5410504208

เป็นส่วนหนึ่งในรายวิชา 01204465

ปีการศึกษา 2557

ภาควิชาวิศวกรรมคอมพิวเตอร์คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเกษตรศาสตร์

**Data Mining Project Report**

**Problem**

แผ่นดินไหวเป็นภัยธรรมชาติที่สามารถก่อให้เกิดความเสียหายและภัยพิบัติต่อบ้านเมือง ที่อยู่อาศัย สิ่งมีชีวิต จึงมีการรวบรวมข้อมูลการเกิดแผ่นดินไหวในแต่ละครั้งเพื่อให้คนทั่วไปสามารถนำข้อมูลไปใช้ประโยชน์ได้ แต่การที่มีข้อมูลจำนวนมหาศาล ทำให้ยากต่อการเลือกดู จึงเป็นสาเหตุที่ทำให้เกิด โปรเจคนี้ขึ้นมา

**Mining Objectives**

เนื่องจากในปัจจุบันนี้หลายองค์กรได้ทำการเก็บรวบรวมข้อมูลเกี่ยวกับการเกิดแผ่นดินไหว โดยจะทำการเก็บข้อมูลในส่วนของปัจจัยต่างๆที่ส่งผลต่อการเกิดแผ่นดินไหว รวมถึงข้อมูลสภาพแวดล้อม ณ ตำแหน่งที่เกิดแผ่นดินไหว เช่น ค่าละติจูด ลองจิจูด ซึ่งโปรเจคนี้จะนำข้อมูลมาวิเคราะห์โอกาสในการเกิดสึนามิ ระดับความรุนแรงของแผ่นดินไหวที่จะเกิดขึ้น บริเวณที่ควรมีการจัดตั้งสถานีเตือนภัยแผ่นดินไหว รวมถึงการวิเคราะห์ช่วงเวลาที่จะมีการเกิดแผ่นดินไหวบ่อยครั้งที่สุด

**Why it necessitates Data Mining**

เนื่องจากการเกิดแผ่นดินไหวนั้นเป็นปรากฏการณ์ที่ค่อนข้างมีผลกระทบที่รุนแรงและอันตรายต่อผู้คนที่อยู่ในบริเวณใกล้กับที่เกิดแผ่นดินไหว อีกทั้งยังสร้างความเสียหายให้กับสภาพแวดล้อมอีกด้วย จึงเห็นความจำเป็นว่าถ้ามีการเตือนภัยก่อนที่แผ่นดินไหวจะเกิดรวมถึงรู้ข้อมูล ช่วงเวลาที่มักจะเกิด ระดับความรุนแรงก็จะยิ่งทำให้เรารับมือกับสถานการณ์ที่จะเกิดได้ดีขึ้น แต่ปัจจุบันข้อมูลที่มีเป็นข้อมูลที่คนทั่วไปไม่สามารถทำความเข้าใจได้โดยง่าย โปรเจคนี้จึงนำข้อมูลเหล่านี้มาวิเคราะห์แล้วแสดงผลในรูปแบบที่คนทั่วไปสามารถทำความเข้าใจได้ง่าย

**Presentation of the data to be mined**

* ที่มาของ Dataset

<https://www.firebase.com/docs/open-data/earthquakes.html>

* Data set ประกอบด้วย
  + africa 14 records
  + europe 373 records
  + asia 1054 records
  + northamerica 41119 records
  + southamerica 292 records
  + antarcrica 52 records
  + ocenic 1183 records
* รายละเอียดของ Dataset

มีการแยกไฟล์ของ dataset ออกเป็นทวีปต่างๆทั้ง 7 ทวีป ดังนี้

* + africa.csv
  + europe.csv
  + asia.csv
  + northamerica.csv
  + southamerica.csv
  + antarcrica.csv
  + ocenic.csv

โดยในแต่ละไฟล์นั้นจะประกอบด้วย attribute ที่เหมือนกัน ดังนี้

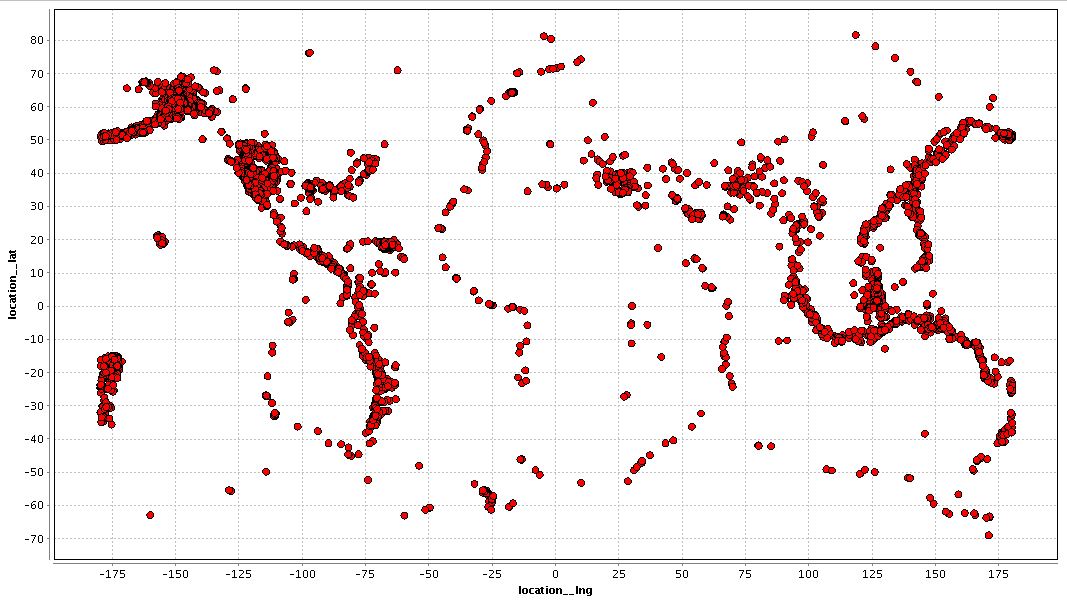
* alert : สถานะของการเตือน (red, green, orange)
* cdi
* code : โค้ดพิเศษสำหรับเหตุการณ์นี้
* dmin : ระยะตามแนวนอน (latitude) ที่ใกล้ที่สุดกับสถานีที่ใช้วัด
* gap : มุมที่ทำกับสถานี
* la : latitude
* long : longitude
* mag : ระดับความรุนแรง(ริกเตอร์)
* magtype : ระดับความรุนแรง แบ่งประเภท Ml (local), Mb (body-wave), MS (surface-wave), Mw (moment), Md (duration), Mwp, Mwb, Mww, Mi, Me, Mlg
* net : เครือข่ายที่ตรวจพบ (แปลตามรูปใน dropbox)
* place
* time : เวลาเริ่ม โดยเก็บอยู่ในรูปของ timestamp
* type : earthquake, quarry, quarry\_blast, sonicboom, explosion, anthropogenic\_event
* update : เวลาสุดท้ายของการเกิดแผ่นดินไหว โดยเก็บอยู่ในรูปของ timestamp
* tsunami : 1 คือเกิด , 0 คือไม่เกิด

**Knowledge discovery process**

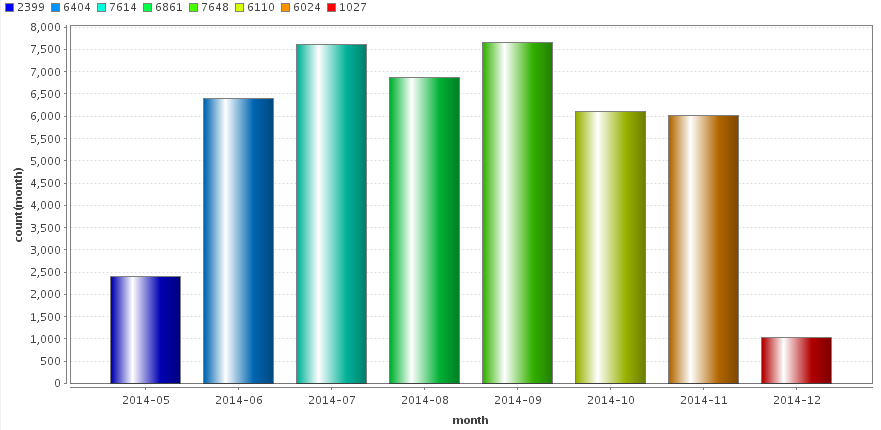
* Data exploration

จะทำการสำรวจ คุณภาพ ความสัมพันธ์ และลักษณะทั่วๆไปของข้อมูล โดยแสดงข้อมูลที่สำรวจด้วยกราฟแบบต่างๆ ดังนี้

1. กราฟแสดงการกระจายตัวของการเกิดแผ่นดินไหวในสถานที่ต่างๆจากทั่วโลก
   * แกน X แสดง longitude
   * แกน Y แสดง latitude

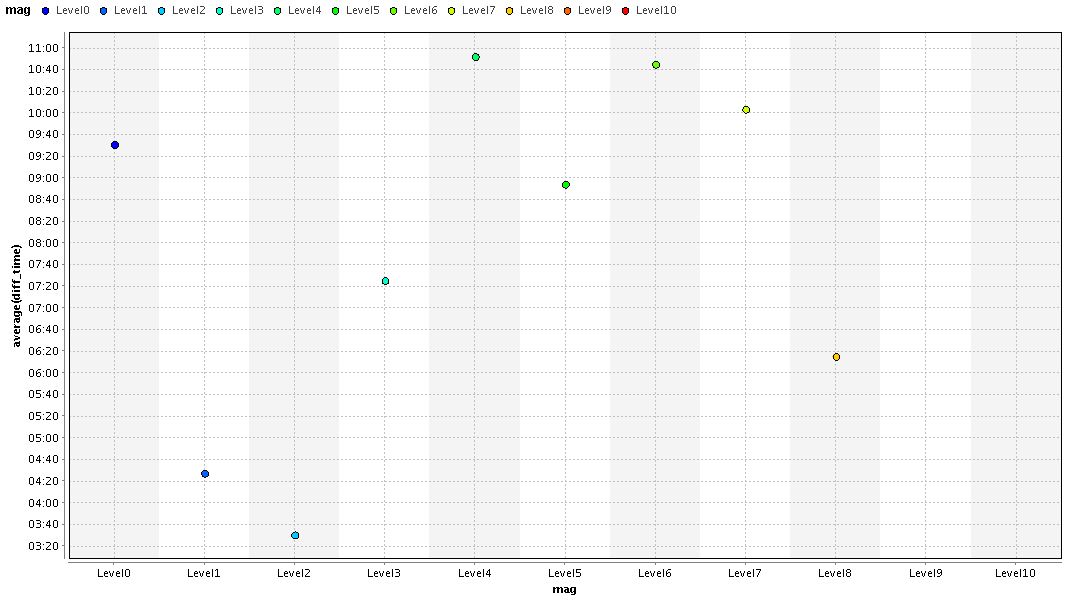


1. กราฟแสดงจำนวนการเกิดแผ่นดินไหวจากทั่วโลกในแต่ละเดือน
   * แกน X แสดง เดือน
   * แกน Y แสดง ความถี่ในการเกิดแผ่นดินไหวทั่วโลก



จากกราฟ พบว่าในช่วงตั้งแต่เดือนมิถุนายนเป็นต้นไป แผ่นดินไหวเกิดค่อนข้างเยอะ เมื่อเทียบเดือนจำนวนการเกิดแผ่นดินไหวของเดือนพฤษภาคม จากการสังเกตดูแล้ว คาดว่าเป็นเพราะตั้งแต่เดือนมิถุนายนเป็นต้นไป จะเป็นช่วงมรสุมจึงทำให้เกิดแผ่นดินไหวบ่อยกว่าช่วงที่ไม่ใช่มรสุม

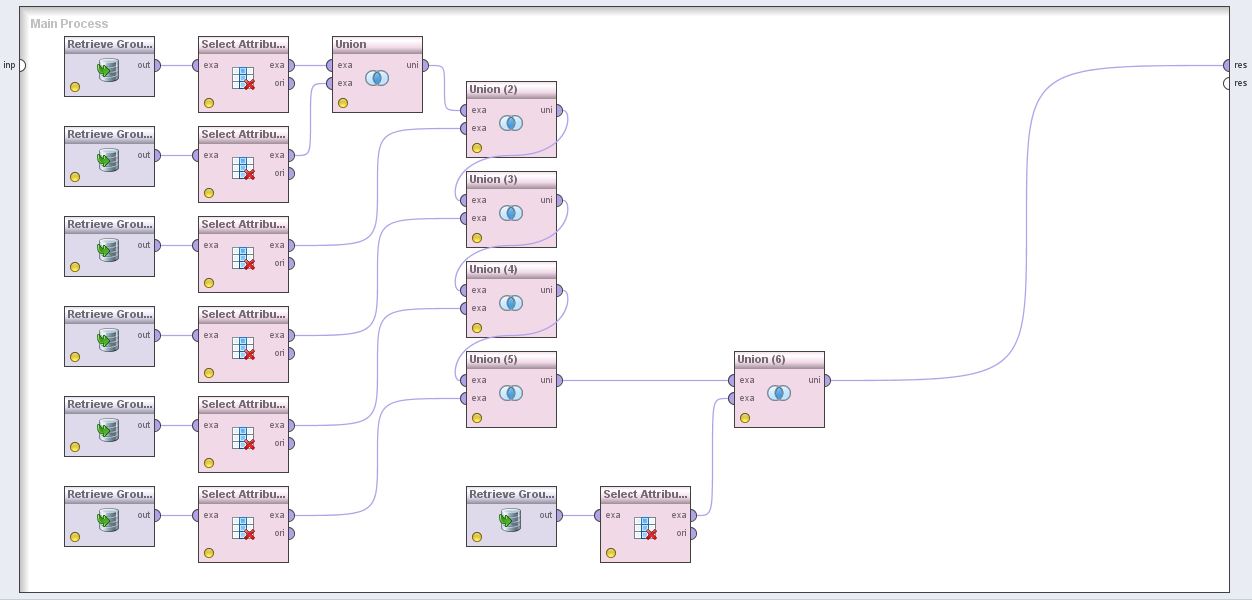
1. กราฟแสดงค่าเฉลี่ยของจำนวนเวลาของการเกิดแผ่นดินไหวจากทั่วโลกเทียบกับระดับความรุนแรงของการเกิดแผ่นดินไหว
   * แกน X แสดง จำนวนเวลา (ชม:นาที:วินาที)
   * แกน Y แสดง ระดับความรุนแรงของการเกิดแผ่นดินไหว



* Data preprocessing

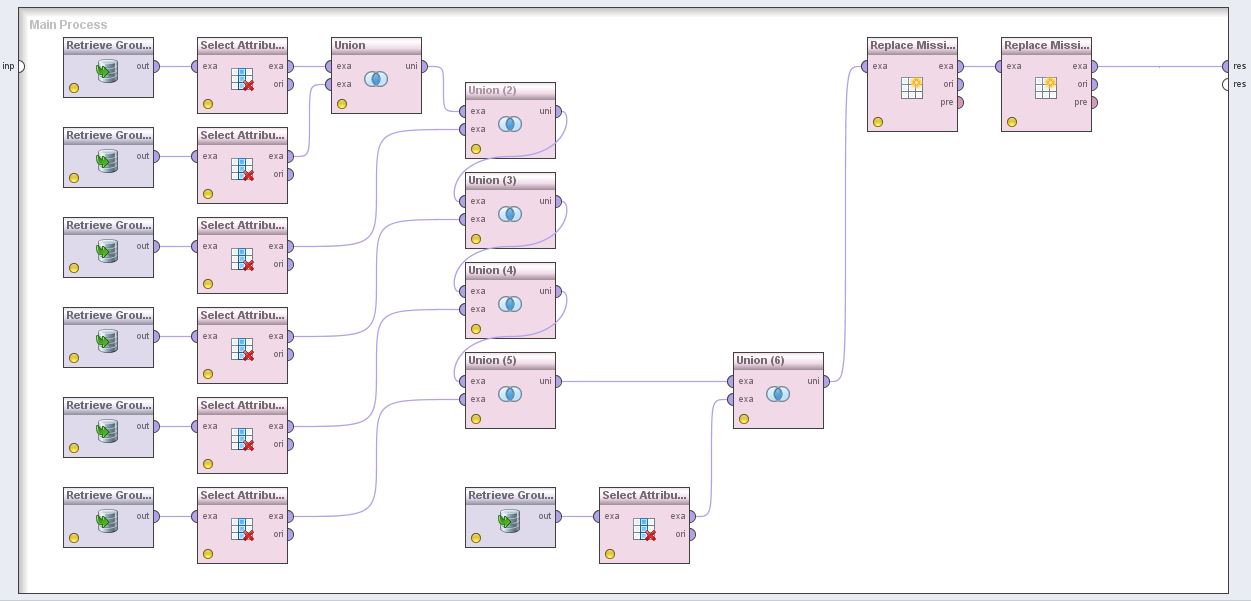
1. Select Attribute and Union

เนื่องจากข้อมูลที่ต้องการใช้มาจาก dataset หลายชุด และแต่ละชุดมี attribute ที่ไม่เหมือนกัน จึงต้องมีการ select attribute ที่จำเป็นต่อการพิจารณาและมีอยู่ในทุกdataset แล้วนำมา union กันทั้งหมด เพื่อรวมข้อมูลทั้งหมดมาวิเคราะห์ต่อไป



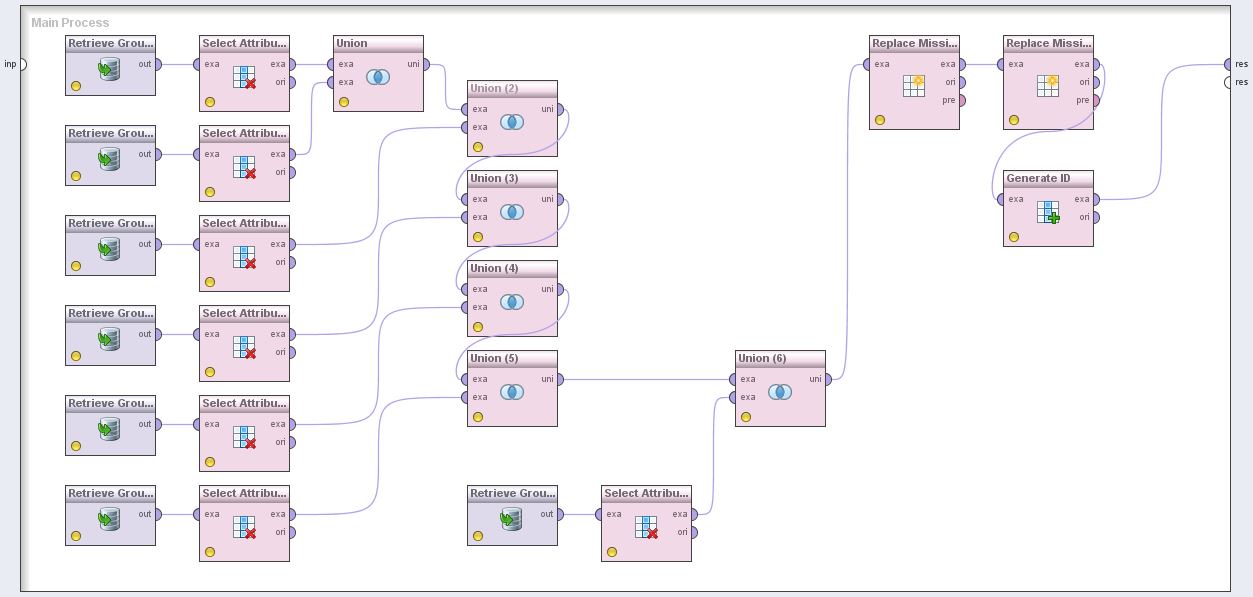
1. Replace Missing Values

เนื่องจากข้อมูลที่ดึงมาใช้นั้น มีหลาย record ที่มีข้อมูลไม่ครบถ้วน จึงต้องมีการจัดการกับค่าที่หายไปเหล่านั้น โดยวิธีการที่เราเลือกใช้คือการแทนที่ด้วยค่าเฉลี่ยของข้อมูลทั้งหมดใน attribute นั้นๆ



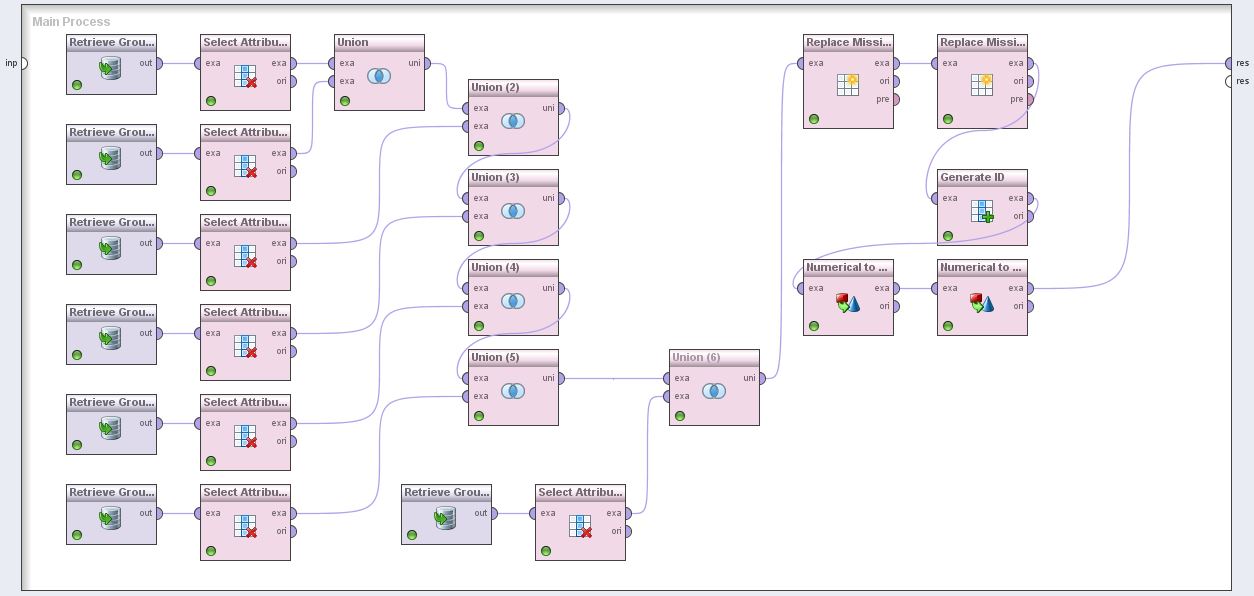
1. Generate ID

เนื่องจากแหล่งข้อมูลมีการเก็บรวบรวมข้อมูลในรูปแบบของ jsonทำให้ไม่มี id ที่ใช้สำหรับอ้างอิงข้อมูล เพื่อเรียกใช้งานข้อมูลที่ต้องการมาวิเคราะห์และเปรียบเทียบ จึงต้องมีการใช้ generate id



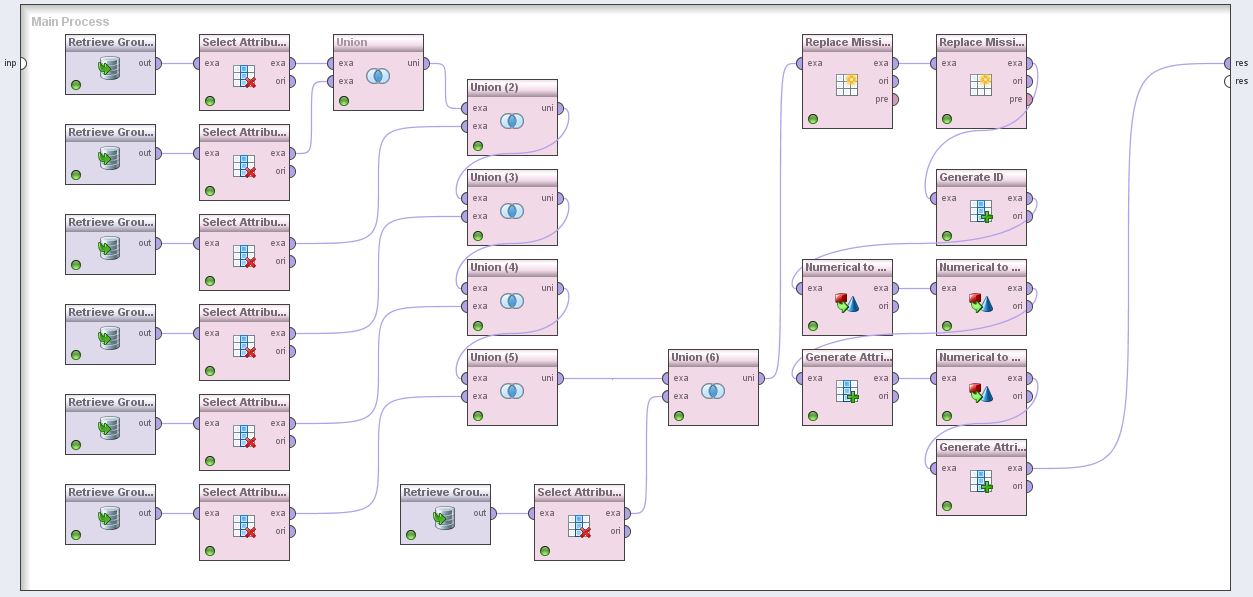
1. Numerical to Date

เนื่องจากข้อมูลบาง attribute เป็นข้อมูลชนิด Timestamp แต่โปรแกรมอ่านค่าของข้อมูลเข้ามาเป็นข้อมูลชนิด Numerical จึงต้องมีการแปลงค่าให้ของในรูปของวัน เดือน ปี และเวลาตาม time zone



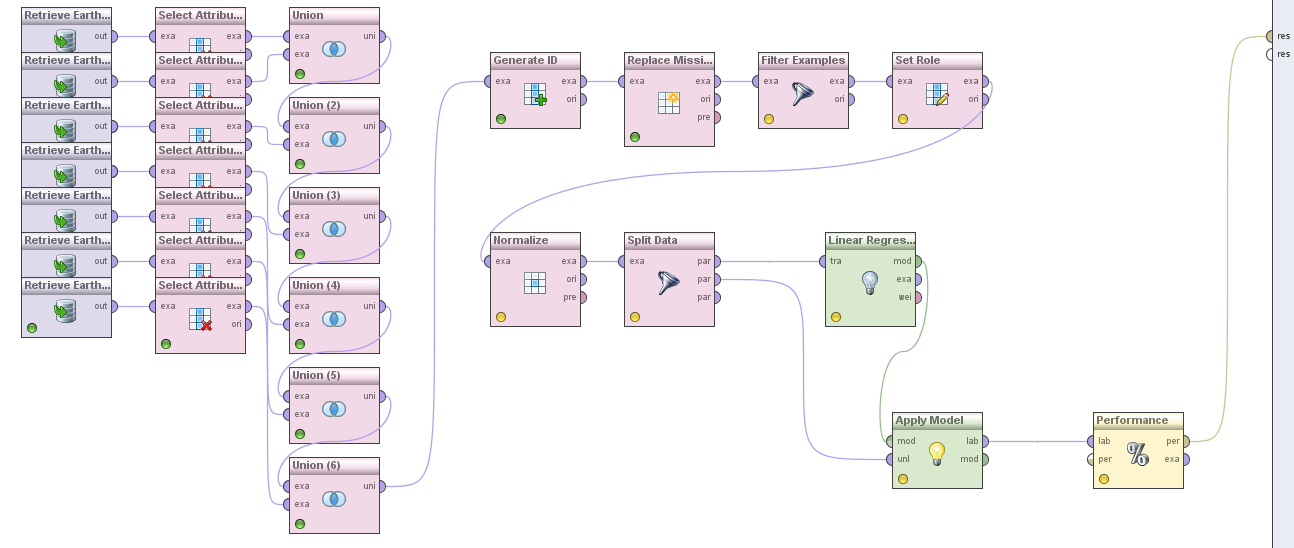
1. Generate Attribute

จากข้อมูลข้างต้น มีข้อมูลของเวลาเริ่มต้นของการเกิดแผ่นดินไหว (time) และข้อมูลของเวลาสิ้นสุดของการเกิดแผ่นดินไหว (update) ซึ่งสามารถหาช่วงเวลาของเกิดแผ่นดินไหวได้ จึงมีการ generate attribute เพื่อหาความต่างของเวลาจาก function: date\_diff() ซึ่งผลลัพธ์ที่ได้ออกมาเป็นข้อมูลชนิด Numerical จึงต้องแปลงชนิดของข้อมูลให้เป็น Timestamp แล้วแสดงผลออกมาในรูป “HH:mm:ss” จากการ generate attribute ด้วย function: date\_str\_custom()



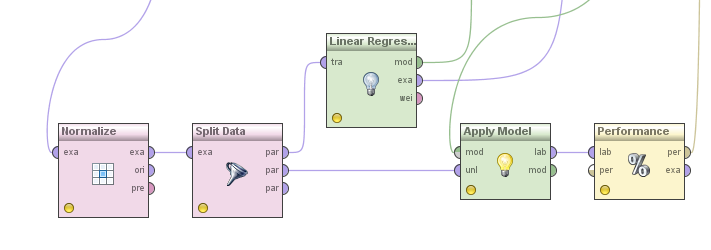
1. Filter Examples

เนื่องจากข้อมูลบางอย่างไม่สามารถทำการแทนที่ด้วยค่าเฉลี่ยของ attribute นั้นๆได้ จึงต้องทำการกรองเอาข้อมูลที่ไม่ครบถ้วนนั้นออกไปก่อนที่จะนำข้อมูลทั้งหมดมาทำการวิเคราะห์ เช่น ข้อมูลของละติจูด ลองจิจูด



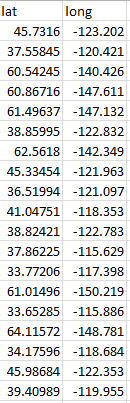
1. Normalize

เนื่องจากข้อมูลในแต่ละ attribute มีความแตกต่างกันค่อนข้างมาก เราจึงทำการ normalize ข้อมูลทั้งหมดก่อนที่จะนำข้อมูลเหล่านั้นไปวิเคราะห์



* Techniques(s) to be used
* Clustering (K-means)

เป็นวิธีที่ใช้เพื่อจัดกลุ่มข้อมูลที่เป็นจุดที่ควรจัดตั้งสถานีเตือนภัยแผ่นดินไหว โดยทำการวิเคราะห์จากอัตราการเกิดแผ่นดินไหวและความรุนแรงที่เกิดขึ้น จากนั้นเมื่อได้ตำแหน่งมาแล้วจึงค่อยนำตำแหน่งนี้ไปแมพกับแผนที่

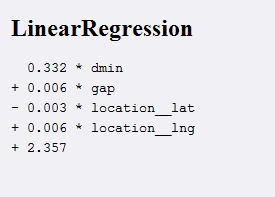


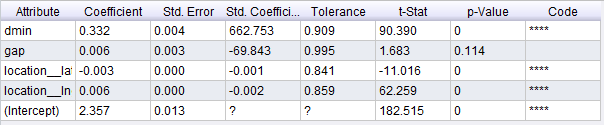
รูปภาพ แสดงผลลัพธ์ที่เป็นตำแหน่งที่ควรจัดตั้งสถานี

ซึ่งข้อมูลดังกล่าวนี้ได้ถูกแปลงเพื่อไว้ใช้สำหรับการ visualize เรียบร้อยแล้ว

* Regression

เป็นการทำนายค่าออกมาในรูปแบบของเลขจำนวนจริง โดยเราใช้เทคนิคนี้ในการทำนายข้อมูล 2 อย่างด้วยกัน (คนละโมเดล) คือ

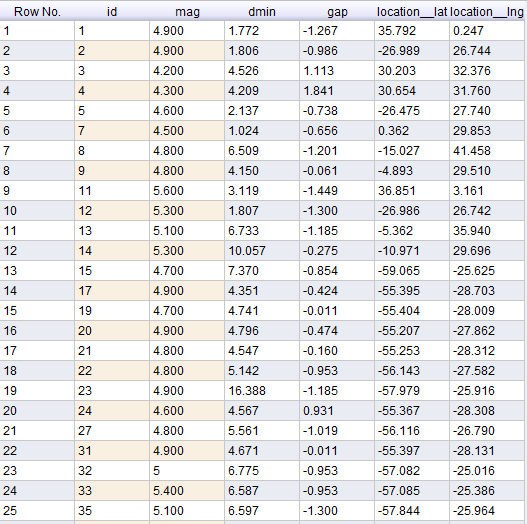
1. ทำนายระดับความรุนแรงของการเกิดแผ่นดินไหว ขนาดเป็นริกเตอร์ โดยในการทำนายนั้น เราใช้ regression มาเพื่อสร้างโมเดลหรือสมการเพื่อหาค่าระดับความรุนแรงนั่นเอง ซึ่งโมเดลที่ได้นั้น อยู่ในรูปของสมการความสัมพันธ์ระหว่าง ค่าที่ต้องการทำนาย (ระดับความรุนแรง) กับตัวแปรอื่นที่ได้ทำการวิเคราะห์และผ่านขั้นตอนของการเตรียมข้อมูลมาแล้วว่ามีความเกี่ยวข้องกับการเกิดแผ่นดินไหว ได้แก่ dmin , ค่าละติจูด และค่าลองจิจูด ตามสมการดังรูป



รูปภาพ แสดงผลลัพธ์จากการทำ linear regression



รูปภาพ แสดงค่า root mean squared ของการทำ linear regression



รูปภาพ ตัวอย่างข้อมูลที่ใช้วิเคราะห์ โดยวิธี linear regression

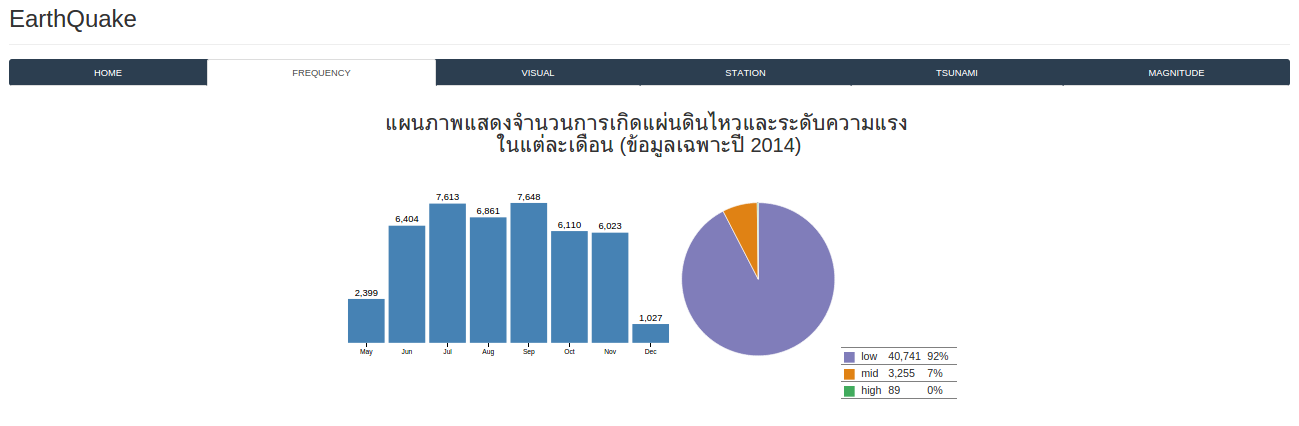
1. ทำนายการเกิดสึนามิจากเหตุการณ์การเกิดแผ่นดินไหว และนำผลลัพธ์ไปคำนวณเป็นเปอร์เซ็นต์ว่าจะมีโอกาสที่จะเกิดสีนามิเป็นกี่เปอร์เซ็นต์ และมีโอกาสที่จะไม่เกิดสึนามิกี่เปอร์เซ็นต์

**Application**

จากข้อมูลที่ได้นั้นสามารถทำเป็น web application สำหรับให้ข้อมูลเกี่ยวกับการเกิดแผ่นดินไหว ทั้งในส่วนของการทำนายระดับความรุนแรงของการเกิดแผ่นดินไหวในแต่ละพื้นที่ ทำนายผลการเกิดสึนามิจากเหตุการณ์การเกิดแผ่นดินไหว รวมถึงการสรุปข้อมูลจำนวนการเกิดแผ่นดินไหวในแต่ละเดือนออกมาเป็นแผนภาพเพื่อง่ายต่อการทำความเข้าใจ

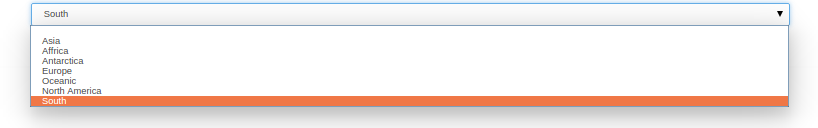
* หน้า Frequency

แสดงจำนวนการเกิดแผ่นดินไหวและระดับความรุนแรง (น้อย,ปานกลาง,มาก) โดยข้อมูลที่ใช้ในการแสดงจะเป็นข้อมูลที่เกิดขึ้นตั้งแต่เดือน พฤษภาคม 2014 จนถึงเดือนธันวาคม 2014 โดยแต่ละเดือนก็จะบอกระดับความรุนแรงของแผ่นดินไหวที่เกิดว่าแต่ละระดับเกิดจำนวนกี่ครั้งใน 1 เดือน

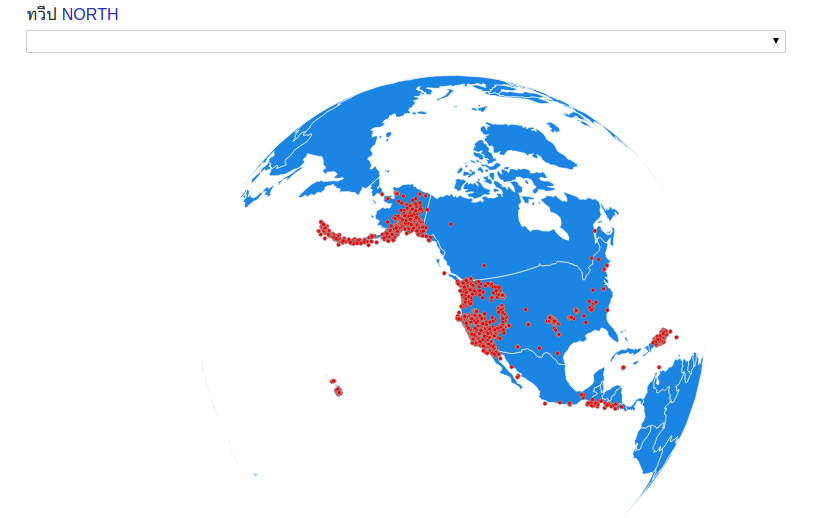


* หน้า Visual

ในหน้านี้จะทำการแสดงผลข้อมูลตำแหน่งของสถานีที่ได้ทำการ visualize เรียบร้อยแล้ว แต่จะทำการแสดงผลเฉพาะทวีปที่เลือกเท่านั้น



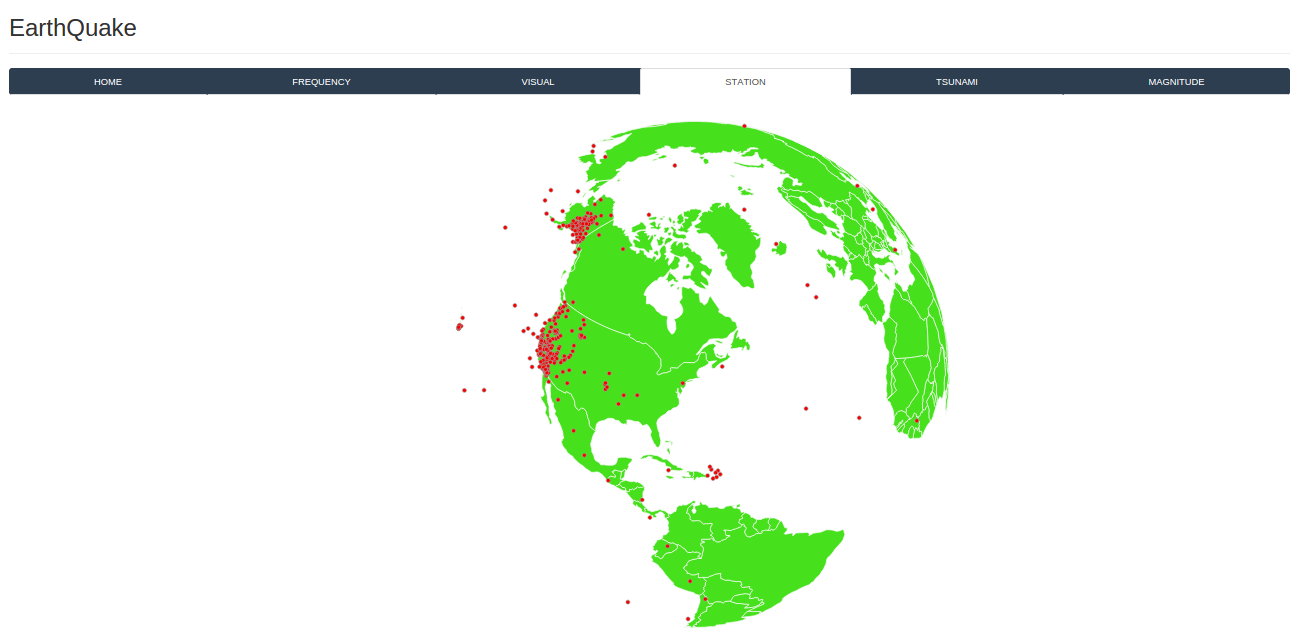
หน้า Visual ในส่วนของการเลือกทวีปที่ต้องการให้แสดงผล



หน้า Visual ในส่วนของการแสดงผล

* หน้า Station

จะแสดงผลในรูปแบบของแผนที่โลกซึ่งสามารถหมุนลูกโลกจำลองดูตำแหน่งของสถานีได้ทั่วทั้งโลก โดยจะแสดงตำแหน่งที่มีการจัดตั้งสถานีด้วยจุดสีแดง ดังรูป



หน้า station แสดงตำแหน่งของสถานี

* หน้า Tsunami

จะทำนายโอกาสที่จะเกิดสึนามิจากเหตุการณ์การเกิดแผ่นดินไหว โดยสิ่งที่ผู้ใช้จะต้องป้อนข้อมูล ได้แก่ Magnitude Latitude Longtitude โดย Latitude Longtitude สามารถจิ้มเอาบนแผนที่โลกได้

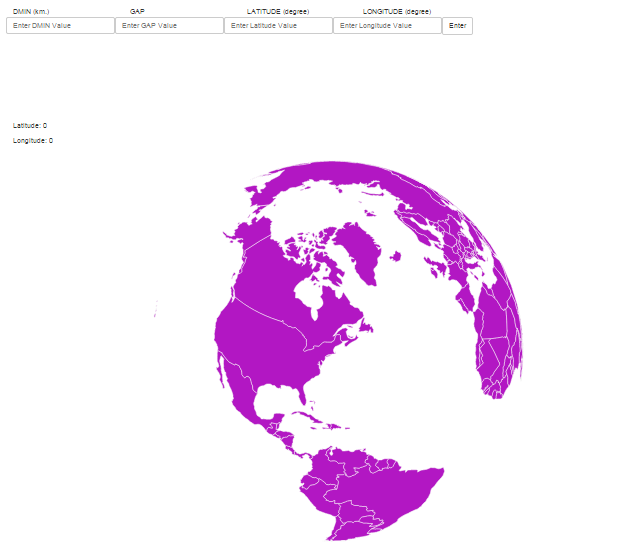


หน้า Tsunami สำหรับรับ Input

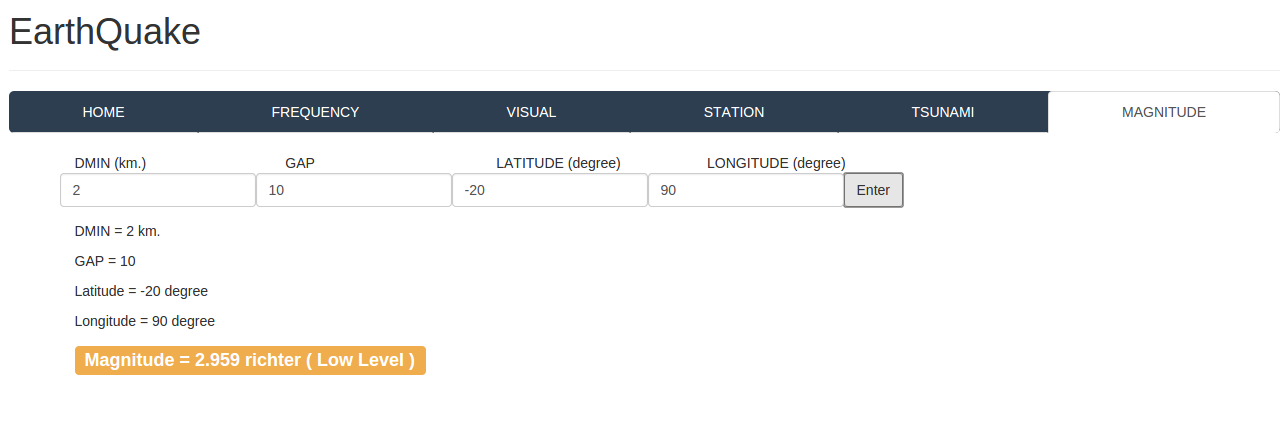
หน้าแสดงผลลัพธ์จากการทำนาย

* หน้า Magnitude

จะเป็นในส่วนของการทำนายระดับความรุนแรงของการเกิดแผ่นไหว ออกมาในหน่วยของริกเตอร์ โดยสิ่งที่ผู้ใช้จะต้องป้อนข้อมูล ได้แก่ ค่าของ dmin, gap, latitude และ longitude



หน้า Magnitude สำหรับรับ Input



หน้าแสดงผลลัพธ์จากการทำนาย

**Data Source**

* <https://www.firebase.com/docs/open-data/earthquakes.html>

**Toolsที่ใช้**

1. Rapid Miner 5.3 เป็นเครื่องมือที่ใช้ในการศึกษาและวิเคราะห์ข้อมูล
2. <https://json-csv.com/?gclid=CjwKEAiAkpCkBRCtstKQo5ia5nESJACsCikRD8iR-cKJeWBO_nsXK9uvEopXhusAxb_-01Y4VynYVBoCuNDw_wcB>ใช้สำหรับแปลง dataset จากแหล่งข้อมูลจากไฟล์ jsonเป็นไฟล์ csv
3. d3.js ใช้เป็นเครื่องมือในการเขียนส่วนของ web application