รายงานสรุปเนื้อหาและการนำไปใช้ประโยชน์จากการเข้าอบรม สัมมนา หรือประชุมวิชาการ

ข้าพเจ้า นางสาววรรณวิมล นาดี ตำแหน่งอาจารย์ สังกัด สาขาเทคโนโลยีสารสนเทศ ขอนำเสนอรายงานสรุปเนื้อหาและการนำไปใช้ประโยชน์ จากการเข้าร่วมฝึกอบรมหลักสูตร R Programming เพื่อนำไปพัฒนาการเรียนการสอน และ งานด้านวิจัยทางด้าน Big Data, Dataming, Text mining หรือ Machine learning ระหว่างวันที่ 4–5 เดือน กุมภาพันธ์ 2560 ณ เดอะคอนเน็คชั่น เอ ดูคูซีน แขวงจอมพล กรุงเทพมหานคร ตามหนังสือขออนุญาตเดินทางไปราชการ เลขที่ ศธ 0523.4.8/37 ลงวันที่ 26 มกราคม 2560

สรุปเนื้อหาและการนำไปใช้ประโยชน์ของการเข้าร่วมฝึกอบรมดังต่อไปนี้

- แนะนำ Big Data
- การติดตั้งและการใช้งาน R โปรแกรม
- แนะนำประเภทตัวแปร ตัวดำเนินการทางคณิตศาสตร์ control flow statement และ Function ที่ใช้ในภาษา R
- การเรียกใช้ Dataset การนำข้อมูลเข้า การส่งข้อมูลออก การเลือกข้อมูล
- การสร้างกราฟในรูปแบบต่างๆ

บทความนี้จะเป็นการแนะนำเบื้องต้นเกี่ยวกับ Big Data และ สรุปเนื้อหาการใช้งานเบื้องต้นของภาษา R

- **แนะนำ Big Data**

เนื่องจากการพัฒนาเทคโนโลยีสารสนเทศ และระบบต่างๆ อย่างต่อเนื่อง และการใช้เทคโนโลยีสารสนเทศที่ในชีวิตประจำวันที่หลากหลายรูปแบบและหลากหลายช่องทาง เช่น การใช้ social network การใช้อินเทอร์เนต การเข้าเว็บเพื่ออ่านข่าว การค้นหาข้อมูล การซื้อขายของออนไลน์ การนำอุปกรณ์ต่าง ๆ เชื่อมต่อกับอินเทอร์เน็ตมากขึ้น เพื่อการทำงาน ความบันเทิง หรือแม้แต่สุขภาพ ทำให้มีเกิด Internet of Thing คือ อินเทอร์เน็ตของทุกสิ่งๆทุกอย่าง หรือทุกสิ่งทุกอย่างล้วนแล้วแต่ใช้อินเทอร์เน็ต ไม่ว่าจะเป็นการเช็คเรื่องราวข่าวสาร การสร้างธุรกิจ การศึกษาข้อมูล การลงทุน เก็บประวัติสุขภาพ หรือแม้แต่การจัดการชีวิตประจำวัน ที่อินเทอร์เน็ตก็เข้ามามีส่วนในแทบทุกกระบวนการ จากพฤติกรรมการใช้งานเทคโนโลยีสารสนเทศหลากหลายรูปแบบเหล่านี้ ทำให้การเติบโตของข้อมูลในปัจจุบันมีลักษณะหลากหลายรูปแบบ และมีปริมาณของข้อมูลที่มากขึ้นเรื่อยๆ จนกลายเป็นข้อมูลขนาดใหญ่ ซึ่งข้อมูลขนาดใหญ่นี้ในแวดวงของเทคโนโลยีมักจะใช้คำว่า Big Data

Big Data หมายถึง "อภิมหาข้อมูล หรือ ข้อมูลที่มากมายมหาศาล” ซึ่งข้อมูลเหล่านี้เกิดจากการพัฒนาเทคโนโลยี และระบบต่างๆ ทำให้แต่ละองค์กรมีการเก็บข้อมูลต่างๆ ไว้อย่างมากมาย

มหาศาล และเป็นรูปแบบไม่มีโครงสร้าง (Unstructured) ทั้งรูปแบบข้อความต่าง ๆ ยังเปลี่ยนไปจากเดิม จากรูปแบบข้อความ (Text) เป็นรูปแบบไฟล์ Media มากขึ้น ซึ่งเมื่อนำข้อมูลปริมาณมากๆ เหล่านั้น มาผ่านกระบวนการวิเคราะห์ การประมวลผล และแสดงผล สกัดเอาคุณค่าออกมาจากข้อมูลขนาดใหญ่ด้วย เทคนิค หรือเทคโนโลยีในการกลั่นกรองซึ่งเกินขอบเขตหรือขีดจำกัดของการจัดการข้อมูลแบบเดิม ๆ จะมีประโยชน์อย่างมากต่อการตัดสินใจของผู้บริหารองค์กร ดังนั้น Big Data จึงเป็นแนวคิดที่จะช่วยให้เกิดการบริหารจัดการข้อมูลให้ได้ประโยชน์สูงสุดอย่างมีประสิทธิภาพ

โดย Big Data จะมีคุณลักษณะที่สำคัญ (ตัวย่อ 5V) คือ Volume, Velocity ,Variety, Veracity และ Value แต่ส่วนใหญ่จะใช้คุณลักษณะแค่ 3V คือ Volume, Velocity ,Variety

- **Volume:** คือ ขนาดของข้อมูล เป็นข้อมูลมหาศาล ขนาดใหญ่ (data storage) จะต้องอยู่ในระดับ Terabytes (เท่ากับ ๑,๐๒๔ Gigabyte) ขึ้นไป ถัดขึ้นไปก็เป็น Petabyte (เท่ากับ ๑,๐๒๔ Terabyte) และ Exabyte (เท่ากับ ๑,๐๒๔ Petabyte)
- **Velocity:** คือ ข้อมูลจะมีการเปลี่ยนแปลงตลอดเวลาและรวดเร็วโดยมักเป็นข้อมูลแบบ Real – time เช่น ข้อมูลจาก Social Media ข้อมูลการทำธุรกรรมการทาง เงิน จะทำให้เราได้รับรู้ข้อมูลในมิติต่างๆ และทันต่อเหตุการณ์ปัจจุบัน
- **Variety:** คือ ข้อมูลมีรูปแบบที่หลากหลายทั้งที่เป็นรูปแบบมีโครงสร้าง (จัดเก็บในระบบฐานข้อมูล) ไม่มีโครงสร้าง (เช่น ไฟล์รูปภาพ วิดีโอ เป็นต้น) หรือกึ่งโครงสร้าง
- **Veracity:** คือ ความถูกต้อง ครบถ้วนของข้อมูล
- **Value :** คือ ข้อมูลที่สามารถนำมาใช้งานได้จริง

ตัวอย่างข้อมูล Big Data

- ข้อมูลเครือข่ายสังคม (Social Networks)
- ข้อมูลการบริการทางเว็บ (Web Server Log)
- ข้อมูลจากอุปกรณ์ตรวจตราการจราจร (Traffic Flow Sensors)
- ข้อมูลภาพถ่ายดาวเทียม (Satellite Imagery)
- ข้อมูลด้านการกระจายเสียง (Broadcast Audio Streams)
- ข้อมูลธุรกรรมทางธนาคาร(Banking Transaction)
- ข้อมูลด้านการตลาดการเงิน (Financial Market Data)
- ข้อมูลการสื่อสารจากโทรศัพท์เคลื่อนที่ (Telemetry from Automobiles)

Big Data จะเป็นประโยชน์ต่อการใช้งานหลายประการ เช่น การใช้งานข้อมูลเกี่ยวกับการค้นคว้า วิจัย เอกสาร เครือข่ายทางสังคม หรือข้อมูลเฉพาะต่างๆ เช่น โรงพยาบาล คลังต่างๆ เป็นต้น

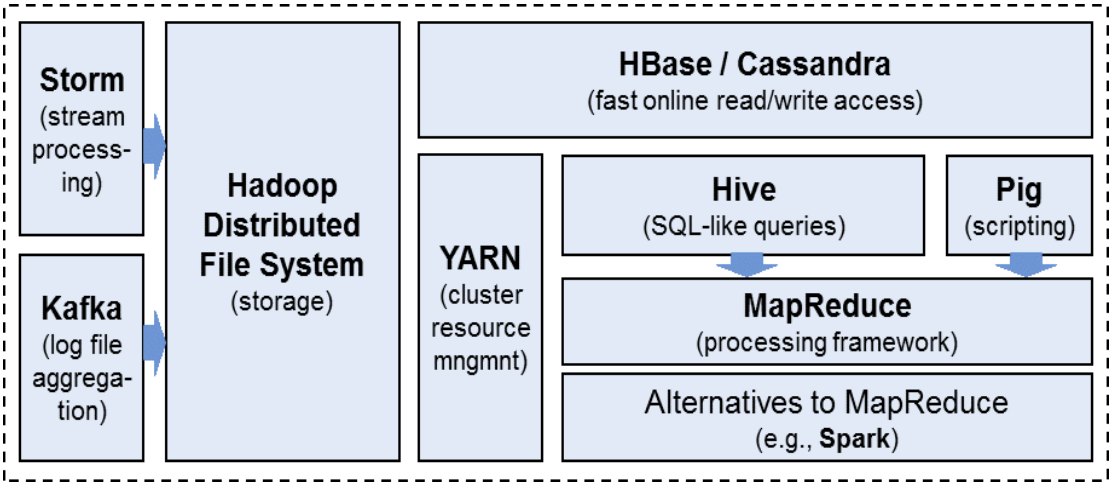
ซึ่ง Big Data นี้เหมาะสำหรับการนำมาวิเคราะห์ข้อมูลดิบ หรือข้อมูลกึ่งโครงสร้างต่างๆ นำไปใช้ในการวิเคราะห์พฤติกรรมลูกค้า หรือธุรกิจที่เกี่ยวข้อง เพื่อหาการแก้ไขหรือหาวิธีการจัดการให้ธุรกิจเป็นไปตามที่คาดหวัง ไม่ว่าจะเป็นด้านธุรกิจ ที่จะเพิ่มโอกาสทางธุรกิจทำให้เกิดนวัตกรรมด้านเทคนิคที่สามารถรวบรวมและจัดเก็บข้อมูลได้ง่ายยิ่งขึ้น และทางด้านการเงินที่สามารถคิดเป็นเปอร์เซ็นต์ค่าใช้จ่ายไอทีได้ด้วย จะเห็นได้ว่าข้อมูลด้านต่าง ๆ ที่กระจัดกระจายซึ่งมีอยู่มากมายมหาศาล เมื่อนำเอาแนวคิดBig Data มาวิเคราะห์ประมวลผล จะก่อเกิดประโยชน์อย่างมากต่อองค์กรและผู้รับบริการ โดย Big Data มีประโยชน์ที่เห็นได้ชัดเจนมีอยู่ 2 ประการใหญ่ คือ

1. การวิเคราะห์ข้อมูลที่ทำ ให้เห็นความรู้ที่ซ่อนอยู่ เช่น ข้อมูลสภาพอากาศจากเครื่องมือตรวจวัดจำนวนมาก ทั้งดาวเทียม เรดาร์ทุ่นในมหาสมุทร ทำให้สามารถพยากรณ์อากาศได้อย่างแม่นยำ
2. สามารถทราบพฤติกรรมและความต้องการที่แท้จริง ก่อให้เกิดผลิตภัณฑ์หรือบริการใหม่ ๆ ที่เหมาะสมตามความต้องการของผู้ใช้ ทำให้เกิดความพึงพอใจและประทับใจในบริการ

อย่างไรก็ตามการเลือกใช้เทคโนโลยี เทคนิค และ เครื่องมือ ที่ประสิทธิภาพ เหมาะสม มาวิเคราะห์ข้อมูล ก็จะสามารถช่วยตอบโจทย์ธุรกิจทิศทางขององค์กรได้ดี ซึ่งปัจจุบันนี้มีเครื่องมือที่ได้รับความนิยมเข้ามามีส่วนช่วยในการจัดการก็คือ Hadoop ที่ถูกพัฒนามาจาก Open Source Technology สามารถเก็บข้อมูลขนาดใหญ่และนำไปประมวลผลได้ แต่การวิเคราะห์ Big Data นั้นเป็นเพียงแค่การวิเคราะห์ข้อมูลดิบแบบย่อยเท่านั้น หากต้องการข้อมูลที่เจาะลึกมากขึ้นไปอีกก็ต้องเพิ่มขั้นตอนการวิเคราะห์แบบ Analytics (Big Data + Analytics = Values) ที่จะทำให้ได้ข้อมูลในเชิงลึกมากขึ้นไปอีกด้วย ซึ่ง Big Data Analytics สามารถแบ่งการวิเคราะห์ได้ 4 ประเภท Descriptive Analytics Diagnostic Analytics Predictive Analytics and Prescriptive Analytics

เทคโนโลยีที่เกี่ยวกับ Big Data สามารถแบ่งออกเป็น 3 กลุ่มใหญ่ๆ คือ

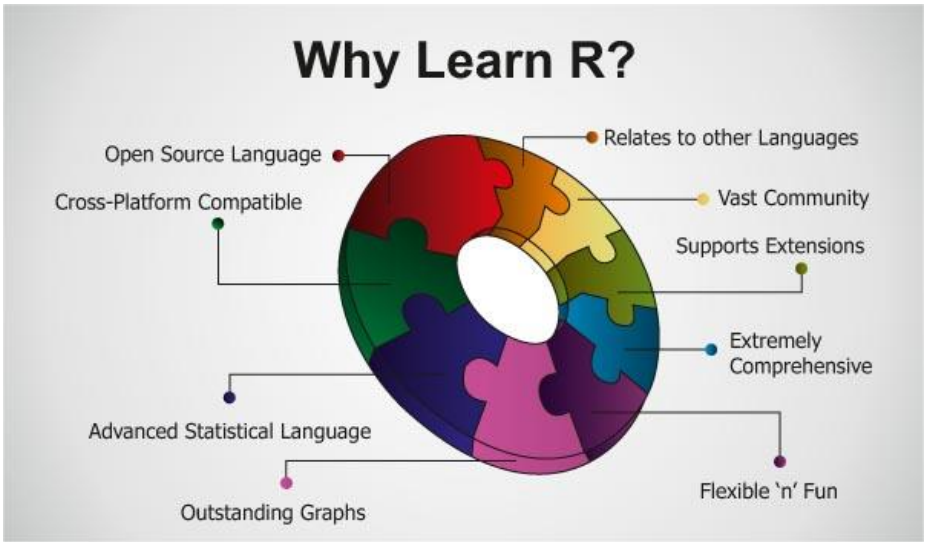
1. Storage คือ การจัดเก็บข้อมูล จะเกี่ยวข้องกับ Volume และ Variety เทคโนโลยีที่นิยมได้แก่ Hadoop
2. Processing คือ การประมวลผล จะเกี่ยวข้องกับ Volume และ Velocity จะมีการประมวลผล 2 แบบ คือ แบบ Batch Processing และ แบบ Streaming Processing เทคโนโลยีที่นิยม คือ Apache Spark
3. Analytic คือ การวิเคราะห์ข้อมูลเชิงลึก โดยเทคนิคในการวิเคราะห์ประกอบไปด้วย Data mining, Predictive analytic, Text analytic, Video analytic, Social media analytic, Sentiment analytic, Location analytic, Machine Learning.



ภาพรวมของ Big Data Ecosystem

- **แนะนำ โปรแกรม R กับ Data mining**

R เป็นภาษาคอมพิวเตอร์ภาษาหนึ่งที่นิยมใช้สำหรับการวิเคราะห์ข้อมูลเชิงสถิติ, คณิตศาสตร์การประมวลผลต่างๆ และ การแสดงผลทางด้านกราฟฟิก ภาษา R ถูกพัฒนาในปี 1993 โดย Ross Ihaka and Robert Gentleman ภาษา R ถูกสร้างอยู่บนพื้นฐานของภาษา S ซึ่งพัฒนาขึ้นมาเพื่อใช้ในงานสถิติ เป็นโปรแกรมประเภท Opensource และ Free software โปรแกรมถูกออกแบบมาให้มีความยืดหยุ่นต่อการใช้สูตรต่างๆ ทางคณิตศาสตร์ เช่น การคำนวณอาร์เรย์ เมทริกซ์ และ การประมวลผลทางสถิติเช่น linear, non-linear modeling, classification, time-serie analysis, clustering เป็นต้น นอกจากนี้ ยังสามารถที่จะนำไปใช้ร่วมกับภาษาอื่นเช่น ภาษา Python เพราะ R มีฟังก์ชั่นทางสถิติที่ใช้ช่วยในการประมวลผลข้อมลมหึมา (big data) ได้โดยง่าย นอกจากนี้ยังสามารถทำงานกับแหล่งข้อมูล (Data sources) ที่หลากหลาย ด้วยคุณสมบัติดังกล่าวปัจจุบันภาษา R จึงถูกนำมาใช้เป็นเครื่องมือในการวิเคราะห์ข้อมูลทางด้าน Data mining และ Machine learning

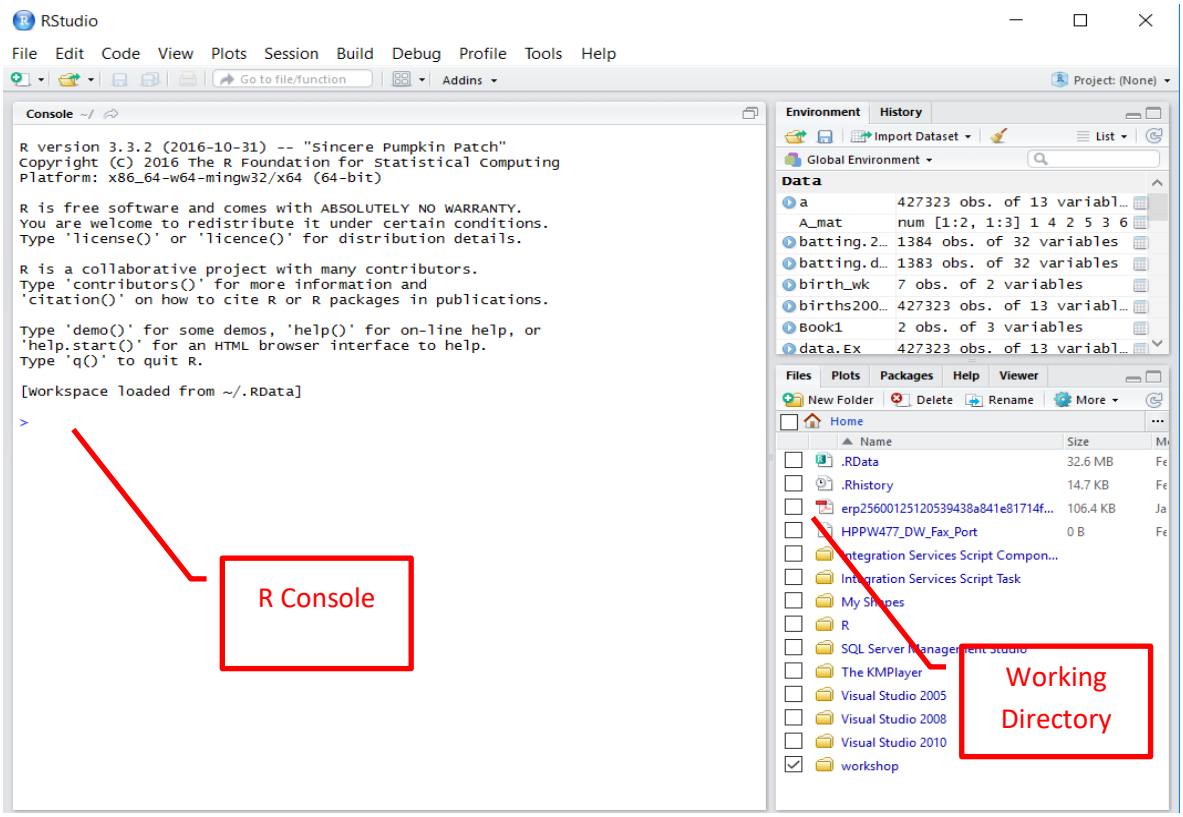


R ประกอบด้วยชุดเครื่องมือการทำงานต่างๆ ต่อไปนี้

- การจัดการข้อมูล และ จัดเก็บข้อมูล ในรูปแบบต่างๆ
- มี operator สำหรับการคำนวณข้อมูลในรูปแบบของ array และ matric
- สามารถทำงานร่วมกับเครื่องมือวิเคราะห์ข้อมูลอื่นๆ ได้ง่าย
- ส่วนการแสดงข้อมูล เพื่อให้ง่ายต่อการวิเคราะห์
- สามารถพัฒนาด้วยภาษาโปรแกรมที่มีประสิทธิภาพ ทำให้มีความยืดหยุ่นสูง
- สามารถสร้างส่วนขยายด้วยการสร้าง package
- สามารถพัฒนาด้วยภาษา C ได้

- **การติดตั้งโปรแกรม R ผู้ใช้สามารถเข้าไปดาวน์โหลด 2 โปรแกรมเว็บไซต์ดังนี้**
    1. ติดตั้งโปรแกรม R เวอร์ชั่น 3.3.2 หรือใหม่กว่า สามารถ download ได้ที่ https://cran.r-project.org/bin/windows/base/
    2. ติดตั้งโปรแกรม R studio 0.99.902 สามารถ download ได้ที่ https://www.rstudio.com/products/rstudio/download/

ส่วนใหญ่แล้วจะนิยมใช้ R studio มากกว่าใช้งาน R ผ่าน R console เนื่องจากมีเครื่องมือที่สะดวก ยืดหยุ่น และใช้งานง่าย เราสามารถนำชุดคำสั่งเดิมๆที่เก็บไว้มาเรียกใช้งานได้ใหม่อีกหลายๆครั้ง โดยไม่ต้องเสียเวลาพิมพ์หรือเรียกคำสั่งใหม่ทีละคำสั่ง เมื่อทำการติดตั้ง R studio เสร็จจะได้หน้าจอดังนี้



จากประกอบไปด้วยส่วนประกอบหลักดังนี้

– **R Console:** ใช้ในการเขียนคำสั่งต่างๆของภาษา R โดยเรียกผ่าน command panel หรือเรียก R command เช่น พิมพ์ 3+2 ผลลัพธ์ที่ได้ [1]5 โดยตัวแปรในภาษา R ตัวพิมพ์ใหญ่เล็กถือว่ามีความแตกต่างกัน เราสามารถตรวจสอบ version ของ R ที่ใช้งานอยู่ด้วยคำสั่ง version จะได้ผลลัพธ์ดังภาพด้านล่าง

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

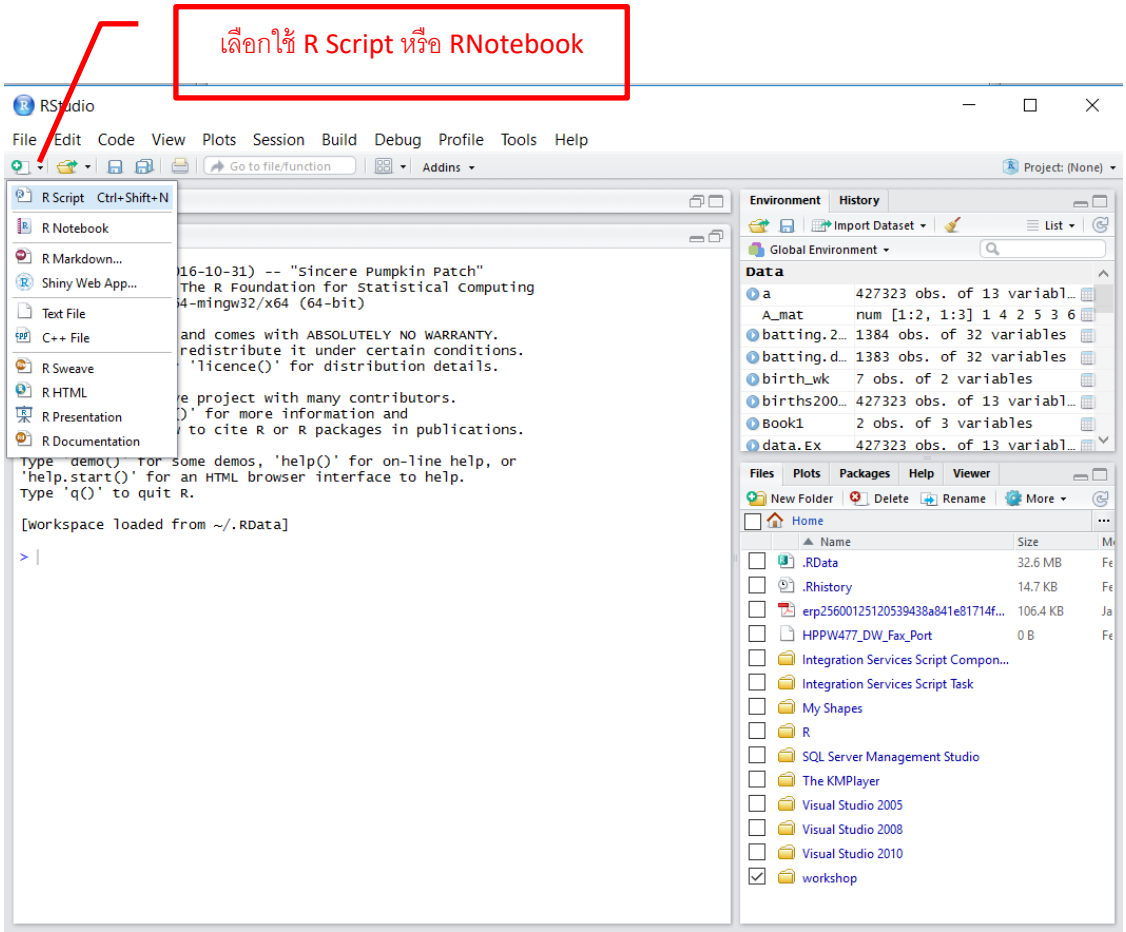
> version
               _
platform       x86_64-w64-mingw32
arch           x86_64
os             mingw32
system         x86_64, mingw32
status
major          3
minor          3.2
year           2016
month          10
day            31
svn rev        71607
language       R
version.string R version 3.3.2 (2016-10-31)
nickname       Sincere Pumpkin Patch
>
```

– **Working Directory:** เป็นส่วนที่ใช้ในการกำหนดโฟล์เดอร์เพื่อ saveค่า หรือ คำสั่ง ต่างๆ โดยสามารถตรวจสอบว่าขณะนี้กำลังทำงานอยู่บนโฟลเดอร์ใดได้โดยใช้คำสั่ง getwd() ซึ่งจะแสดงผลลัพธ์โฟล์เดร์ที่ถูกใช้งานอยู่ปัจจุบัน

– **Environment:** จะเป็นส่วนในการ import dataset จากไฟล์นามสกุลชนิดต่างๆ เช่น CSV, Excel, SPSS, SAS นอกจากนี้ยังทำการ save สภาวะแวดล้อมของคำสั่งที่เขียนไว้ก่อนหน้านี้ เมื่อทำการเรียกไฟล์เดิมขึ้นมาใช้ ก็จะทำการดึงสภาวะแวดล้อมเดิมขึ้นมาใช้ได้อย่างต่อเนื่อง

การเขียนคำสั่งต่างๆเพื่อให้ R ทำงานนั้นจะสามารถเรียกผ่าน command panel หรือจะเขียนในรูปแบบ Script ก็ได้ โดยการเขียนคำสั่งบน command panel นั้นจะไม่สามารถบันทึกค่าไว้ได้ เมื่อปิดโปรแกรม ดังนั้นเมื่อมีการใช้คำสั่งจำนวนมากๆ ในการทำงานจึงต้องมีการเขียนในรูปแบบ script ซึ่งจะทำให้เรา save ไฟล์ไว้ในรูปแบบ file.r ที่สามารถนำกลับมาแก้ไข หรือ run ใหม่ได้ เราสามารถใช้เครื่องหมาย # ในการ comment หรือเขียนอธิบายโค๊ด

การเขียน R script สามารถที่จะเลือกเขียนโดยเลือกใช้เครื่องมือ R Script หรือ R Notebook โดยทำการคลิกเลือกที่สัญลักษณ์  ในบทความนี้จะแนะนำการใช้ R Notebook ในการสร้างคำสั่งต่างของภาษา R ก่อนอื่นมารู้จักประเภทข้อมูลของภาษา R ก่อน

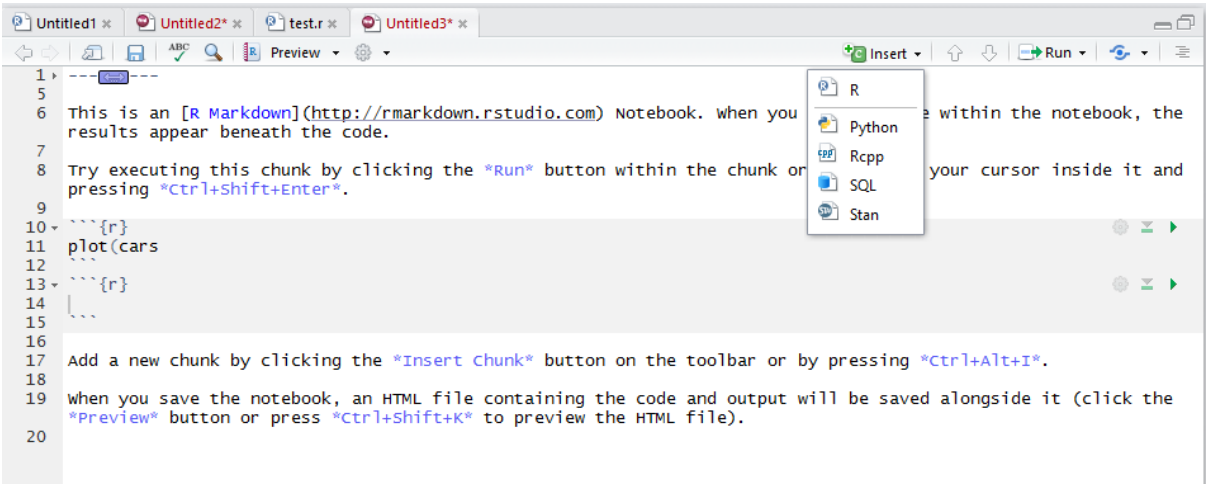


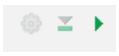
- **การเขียนคำสั่งบน R Notbook**

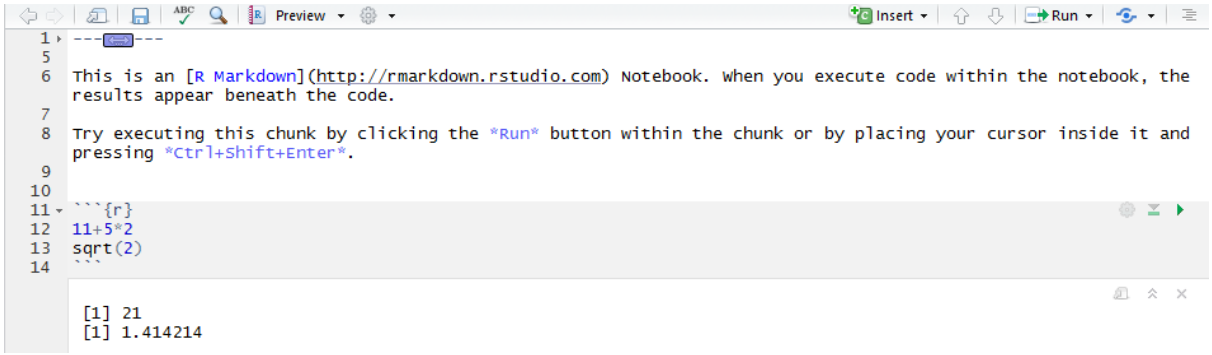
1. ในการเขียนคำสั่งภาษา R บน R Notepad จะอยู่ภายใต้รูปแบบคำสั่ง R Chunk ดังนี้

```
```{r}
คำสั่งต่างๆ
```
```

โดยสามารถสร้าง R Chunk ขึ้นมาใหม่โดยเรียกใช้จากปุ่ม Insert a new R Chunk 



2. เมื่อต้องการดูผลลัพธ์จากการการเขียนคำสั่ง สามารถคลิกได้ที่ปุ่ม play  หรือ crtl+Enter โดยโปรแกรมจะทำการรันผลลัพธ์ของทุกคำสั่งที่อยู่ภายใต้ R Chunk นั้น ดังรูป



3. นอกจากนี้เราสามารถที่จะใช้เมนู Preview แสดงคำสั่ง และ ผลลัพธ์ที่เขียนขึ้นและสามารถ save file เป็น pdf เก็บไว้ได้
4. ในการจัดเก็บ file ให้ save เป็นนามสกุล R รวมถึงให้ save file environment เก็บคุณสมบัติต่างๆไว้ด้วยก่อนออกจากโปรแกรม เมื่อเรียกไฟล์เดิมกลับมาใช้งานใหม่อีกครั้ง

- **ประเภทข้อมูลใน R**

ข้อมูลพื้นฐานที่ใช้ใน R คือ เว็กเตอร์(Vector) ซึ่งเป็นข้อมูลมิติเดียว มีข้อมูลตัวเดียวหรือหลายตัวก็ได้ ได้แก่ character, numeric, integer, complex number และ logical (True/False) โดยข้อมูลประเภท character และ string จะอยู่ในเครื่องหมาย ' ' หรือ " " เช่น เราพิมพ์ >y=c(3,7,9,11) เป็นการสร้าง vector ที่มีค่า 4 ตามลำดับคือ 3,7,9,11 ไปเก็บไว้ที่ตัวแปร y เราสามารถอ้างถึงค่าแต่ละตัวใน vector โดยอ้างถึงลำดับที่ เช่น y[2] จะมีค่าเท่ากับ 7 เป็นต้น

ข้อมูล Vector

```
y= c(3,7,9,11)
y
```

```
[1]  3  7  9 11
```

```
x= c("Monday","Tuesday","Wednesday")
x
```

```
[1] "Monday"    "Tuesday"   "Wednesday"
```

- **Basic Commands and Operations**

ตัวดำเนินการทางคณิตศาสตร์ของ R ก็ใช้สัญลักษณ์พื้นฐานเช่นเดียวกับภาษาคอมพิวเตอร์ทั่วไป คือ + – * / ^ นอกจากนี้จุดเด่นของ R คือการมีฟังก์ชันให้ใช้ที่หลากหลาย โดยเฉพาะอย่างยิ่งฟังก์ชันที่เกี่ยวกับการวิเคราะห์ข้อมูล เช่น mean(), max(), min(), average()}, sqrt(), var(), summary(), plot() เป็นต้น ดังตัวอย่าง

```
40 ```{r}
41 3+5.56
42 76-45
43 3*6
44 85/4
45 sqrt(2)
46 log(10)
47 exp(1)
48 3+4-5/2
49 ```
```

```
[1] 8.56
[1] 31
[1] 18
[1] 21.25
[1] 1.414214
[1] 2.302585
[1] 2.718282
[1] 4.5
```

นอกจากนี้เรายังสามารถสร้างฟังก์ชันเพื่อใช้งานเฉพาะของเราเองได้ ซึ่งมีลักษณะการทำงานเหมือนกับการเรียกใช้ฟังก์ชันในภาษา C เช่น function การนับจำนวน 1 ถึง n

Function for counting from 1 to n

```
countToN = function(n){
 for (i in 1:n) print(i)
}
countToN(8)
```

คำสั่ง

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
```

ผลลัพธ์

- **Control Flow Statements**

คำสั่งควบคุมการทำงานของโปรแกรมเมื่อมีการใช้คำสั่งหรือเงื่อนไขจำนวนมากจะทำได้โดยอาศัย control flow ดังนี้

- การใช้คำสั่ง IF-Else ในการตรวจสอบเงื่อนไขก่อนการดำเนินการ

If else

```
z=NA
if(!is.na(z)){
if (z>0) {
  print("z is positive")
} else if (z <0){
  print("z is negative")
}else{
  print("z is zero")
}
}else print ("z is NA")
```

```
[1] "z is NA"
```

- การใช้คำสั่ง For looping ในการดำเนินการแบบวนลูปหรือทำซ้ำเป็นรอบๆจนกว่าจะครบตามเงื่อนไข

For loop

```
x=1
for (i in 1:5){
x = x*i
print(paste("x=",x,"i=",i))

}
```

```
[1] "x= 1 i= 1"
[1] "x= 2 i= 2"
[1] "x= 6 i= 3"
[1] "x= 24 i= 4"
[1] "x= 120 i= 5"
```

- การใช้ While looping ในการดำเนินการวนลูป

While loop

```
j=1
while(j<10)
{ j=j+1}
x=10
i=j=1
while(i<10| j<30)
{
  j=j*3
  i=i*2
  x=x+x
  print(paste("i=",i,"j=",j,"x=",x))
}
```

```
[1] "i= 2 j= 3 x= 20"
[1] "i= 4 j= 9 x= 40"
[1] "i= 8 j= 27 x= 80"
[1] "i= 16 j= 81 x= 160"
```

- **Basic Manipulation**

ในกรณีที่เรามี Data Set แล้วเราต้องการนำมาวิเคราะห์ข้อมูลในโปรแกรม R เราสามารถทำได้ดังนี้ ในที่นี้จะขอใช้ตัวอย่าง Dataset ที่ติดมากับ package เช่น package nutshell เป็นชุดข้อมูลประกอบหนังสือ R in Nutshell dataset สามารถทำได้ขั้นตอนต่อไปนี้

1. ให้ทำการเปิด R notebook ขึ้นมาใหม่ ให้ทำการพิมพ์คำสั่ง data() เพื่อเรียกดูข้อมูลที่ load อยู่ใน working space ได้

```
data()
iris
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
 [ reached getOption("max.print") -- omitted 142 rows ]
```

2. ในการเรียกใช้ package เราจะต้องทำการติดตั้ง package และใช้เรียก library ขึ้นมาก่อน ในตัวอย่างจะเรียกใช้ package ชื่อว่า nutshell และ ใช้ dataset ชื่อ births2006.smpl สามารถเขียนตามคำสั่งดังนี้ โดยคำสั่ง str() เป็นการเรียกดูโครงสร้างของข้อมูล และข้อมูลใน dataset

```
if(!require("nutshell"))
{install.packages("nutshell")
 library("nutshell")
  }
data(package ="nutshell")
data(births2006.smpl)
str(births2006.smpl)
```

```
'data.frame':   427323 obs. of  13 variables:
 $ DOB_MM   : int  9 2 2 10 7 3 5 4 10 4 ...
 $ DOB_WK   : int  1 6 2 5 7 3 2 7 3 4 ...
 $ MAGER    : int  25 28 18 21 25 28 33 31 18 24 ...
 $ TBO_REC  : int  2 2 2 2 1 3 2 3 1 2 ...
 $ WTGAIN   : int  NA 26 25 6 36 35 26 25 46 43 ...
 $ SEX      : Factor w/ 2 levels "F","M": 1 2 1 2 2 2 2 1 1 2 ...
 $ APGAR5   : int  NA 9 9 9 10 8 9 9 9 9 ...
 $ DMEDUC   : Factor w/ 18 levels "1 year of college",..: 18 4 18 18 6 18 18 4 18 6 ...
 $ UPREVIS  : int  10 10 14 22 15 18 10 19 15 13 ...
 $ ESTGEST  : int  99 37 38 38 40 39 38 38 40 40 ...
 $ DMETH_REC: Factor w/ 3 levels "C-section","Unknown",..: 3 3 3 3 3 3 1 1 1 3 ...
 $ DPLURAL  : Factor w/ 5 levels "1 Single","2 Twin",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ DBWT     : int  3800 3625 3650 3045 3827 3090 3430 3204 3227 3459 ...
```

- **การ Save/การ Exporting Data**

สามารถการเขียนข้อมูลลงบนไฟล์ประเภท .txt และ .csv ได้ด้วยคำสั่งดังนี้

```
### write.table/write.csv
write.csv(data.Ex, file = "birth2006.csv", row.names = F)
write.table(data.Ex, file = "birth2006.txt", row.names = F, sep = "\t", quote = F)
```

- **การ Load/การ Importing Data**

การนำข้อมูลเข้ามาคำนวณใน working space นั้นทำได้สำหรับไฟล์หลายชนิด ในที่นี้ยกตัวอย่างการนำเข้าข้อมูลแบบ text file ทั้ง .csv และ .txt ได้ด้วยคำสั่ง read.table/read.csv ดังนี้

–Import .csv file สามารถเรียกใช้คำสั่งดังนี้

```
data.read = read.csv("birth2006.csv", header = TRUE)
str(data.read)
```

```
'data.frame':   427323 obs. of  13 variables:
 $ DOB_MM   : int  9 2 2 10 7 3 5 4 10 4 ...
 $ DOB_WK   : int  1 6 2 5 7 3 2 7 3 4 ...
 $ MAGER    : int  25 28 18 21 25 28 33 31 18 24 ...
 $ TBO_REC  : int  2 2 2 2 1 3 2 3 1 2 ...
 $ WTGAIN   : int  NA 26 25 6 36 35 26 25 46 43 ...
 $ SEX      : Factor w/ 2 levels "F","M": 1 2 1 2 2 2 2 1 1 2 ...
 $ APGAR5   : int  NA 9 9 9 10 8 9 9 9 9 ...
 $ DMEDUC   : Factor w/ 18 levels "1 year of college",..: 18 4 18 18 6 18 18 4 18 6 ...
 $ UPREVIS  : int  10 10 14 22 15 18 10 19 15 13 ...
 $ ESTGEST  : int  99 37 38 38 40 39 38 38 40 40 ...
 $ DMETH_REC: Factor w/ 3 levels "C-section","Unknown",..: 3 3 3 3 3 3 1 1 1 3 ...
 $ DPLURAL  : Factor w/ 5 levels "1 Single","2 Twin",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ DBWT     : int  3800 3625 3650 3045 3827 3090 3430 3204 3227 3459 ...
```

- Import .txt file สามารถเรียกใช้คำสั่งดังนี้

```
data.readtxt = read.table("birth2006.txt", header = T, sep = "\t")
str(data.readtxt)
```

- **การสร้าง Data Frame**

Data Frame ไว้สำหรับจัดเก็บข้อมูลในรูปแบบของตาราง table ซึ่งจะมีการจัดเก็บข้อมูลในรูปแบบ vector โดยประกอบด้วย row และ column โดยที่แต่ละcoulmn จะเป็นตัวแปรตัวต่างๆซึ่งมีชนิดเดียวกัน และทุกๆ column จะมีจำนวนแถวหรือ observation เท่ากัน ตัวอย่างเช่น ต้องการสร้างข้อมูลดังในตัวอย่าง

sex <fctr>	W.Hnd <fctr>	Pulse <dbl>	Smoke <fctr>	Height <dbl>
female	Right	NA	Never	172
female	Right	64	Never	NA
male	Right	NA	Never	180
male	Right	80	NA	NA
female	Right	64	Never	170
male	Right	NA	Heavy	176

6 rows

สร้าง data frame ชื่อ ex_df สำหรับเก็บข้อมูล 5 columns 6 rows ดังนี้

```{r}
ex_df = data.frame(sex=c("female","female","male","male","female","male"),
                   W.Hnd=c("Right","Right","Right","Right","Right","Right"),
                   Pulse=c(NA,64,NA,80,64,NA), Smoke=c("Never","Never","Never",NA,"Never","Heavy"),
                   Height =c(172,NA,180,NA,170,176))
ex_df
```

– กรณีที่เราเรียกใช้ Dataset ที่มีอยู่ จะต้องทำการแปลงข้อมูลให้อยู่ในรูปตาราง tbl_df ซึ่งเป็นโครงสร้างข้อมูลแบบ data.frame ก่อน ในที่นี้จะยกตัวอย่างการใช้ข้อมูล ใน package dplyr

1. เริ่มต้นให้ทำการติดตั้ง และ เรียกใช้ library ของ dplyr ดังนี้

```{r}
if(!require("dplyr"))
{install.packages("dplyr")
library("dplyr")
}
###require(dplyr data)
data(package ="dplyr")
```

3. ทำการแปลง ข้อมูลให้อยู่ในรูป tbl_df ซึ่งเป็นโครงสร้างข้อมูลแบบ data.frame ที่ใช้กับ dplyr ด้วยคำสั่ง

```r
library(dplyr)
class(mydata)
```

```
[1] "data.frame"
```

```r
## create tbl_df object
mydata.tbl = tbl_df(mydata)
class(mydata.tbl)
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

4. เมื่อสร้างเสร็จสามารถเรียนดูข้อมูลด้วยคำสั่ง mydata หรือ mydata.tbl

```
mydata.tbl
```

DOB... <int>	DOB... <int>	MA... <int>	TBO_R... <int>	WTG... <int>	... <fctr>	APGA... <int>	DMEDUC <fctr>	UPREVIS <int>	ESTGEST <int>
9	1	25	2	*NA*	F	*NA*	NULL	10	99
2	6	28	2	26	M	9	2 years of college	10	37
2	2	18	2	25	F	9	NULL	14	38
10	5	21	2	6	M	9	NULL	22	38
7	7	25	1	36	M	10	2 years of high school	15	40
3	3	28	3	35	M	8	NULL	18	39
5	2	33	2	26	M	9	NULL	10	38
4	7	31	3	25	F	9	2 years of college	19	38
10	3	18	1	46	F	9	NULL	15	40
4	4	24	2	43	M	9	2 years of high school	13	40

1-10 of 427,323 rows | 1-10 of 13 columns Previous 1 2 3 4 5 6 ... 100 Next

- **การเลือกดูข้อมูล และการกรองข้อมูล**

การเลือกดูข้อมูลบางส่วนที่สนใจในภาษา R สามารถใช้คำสั่ง filter() นอกจากนี้ยังสามารถใช้คำสั่ง select, distinct คล้ายคลึงกับภาษา sql ในการเลือกดูข้อมูลได้เช่นเดียวกัน ตัวอย่างเช่น

ต้องการเลือกดูข้อมูลเฉพาะเพศชาย เราสามารถใช้คำสั่ง filter ในการเลือกดูข้อมูลได้ดังนี้

```r
filter(mydata.tbl, SEX == "M")
```

```
filter(mydata.tbl, SEX == "M")
```

DOB... <int>	DOB... <int>	MA... <int>	TBO_R... <int>	WTG... <int>	... <fctr>	APGA... <int>	DMEDUC <fctr>	UPREVIS <int>	ESTGEST <int>
2	6	28	2	26	M	9	2 years of college	10	37
10	5	21	2	6	M	9	NULL	22	38
7	7	25	1	36	M	10	2 years of high school	15	40
3	3	28	3	35	M	8	NULL	18	39
5	2	33	2	26	M	9	NULL	10	38
4	4	24	2	43	M	9	2 years of high school	13	40
3	3	31	2	30	M	9	2 years of college	9	38
4	3	28	1	57	M	9	NULL	10	38
11	2	33	6	10	M	9	2 years of high school	5	35
1	4	19	1	22	M	9	2 years of high school	15	38

1-10 of 218,670 rows | 1-10 of 13 columns Previous **1** 2 3 4 5 6 ... 100 Next

- ตัวอย่างการใช้คำสั่ง select() ในการเลือกดูค่าข้อมูลของ 2 คอลัมน์ DOB_MM และ DBWT จากตาราง mydata

````
```{r}
select(mydata.tbl, DOB_MM, DBWT)
```
````

DOB_MM <int>	DBWT <int>
9	3800
2	3625
2	3650
10	3045
7	3827
3	3090
5	3430
4	3204
10	3227
4	3459

จะเห็นได้ว่า ประสิทธิภาพของโปรแกรม R สามารถนำมาใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่โดยมีความยืดหยุ่นต่อฟังก์ชันทางคณิตศาสตร์ และ สถิติ และ การโปรแกรมไปในตัว เหมาะสำหรับผู้ที่ทำงานวิจัยทางด้าน Datamining, Textmining และ Machine learning เป็นอย่างดี