BINA NUSANTARA

BIG DATA

for Beginner

Prof. Dr. Widodo Budiharto

November 2018

Bab 1. Data Science dan Big Data

1.1. Pengantar

Ketika seorang Direktur Keuangan bertanya ke data scientist tentang informasi yang melibatkan data yang sangat besar, misalnya *"Berapa total penjualan di Eropa dan Asia*", *"Apa tren biaya produksi di Indonesia", "Berapa breakdown revenue terhadap negara dan product line", "Kenapa penjualan tertentu berakhir dengan deal sedangkan yang lainnya gagal?", "Kenapa pelanggan meninggalkan kita?",* maka dengan mudah semua pertanyaan tersebut dapat dijawab menggunakan **Big Data** dan **Big Data Analytics**. Big Data adalah sumber inovasi yang telah menarik perhatian pengambil keputusan baik di sektor publik dan swasta. Memanfaatkan inovasi teknologi dalam data besar dapat berkontribusi pada pertumbuhan ekonomi dan pembangunan berkelanjutan dan untuk menangkap pertumbuhan eksplosif data besar. Mengolah, mengekslorasi dan memahami data selain melalui pengembangan aplikasi big data seperti R juga dapat menggunakan program yang profesional dan mudah misalnya menggunakan IBM® Watson Analytics™ atau IBM® Cognos Analytics. Sistem ini mampu memerlihatkan *insights* tentang organisasi kita, dan sharing dengan orang lain. Selain itu, menggunakan program di atas, kita juga dapat membuat dashboard professional dan infografis.

Watson Analytics / Cognos Analytics can help you understand your data better and find insights that are hidden in your data. You can get answers and new insights to make confident decisions in minutes – all on your own.

Big data sangat berperan penting di bidang kecerdasan buatan saat ini dan mulai populer di sekitar tahun 2012, sedangkan data science adalah sebuah terminologi yang mulai ngetren di tahun 2013; ketika Hadoop dan big data sudah menjadi istilah penting di berbagai organisasi. Hadoop sendiri sempat popular beberapa tahun lalu, dan hingga saat ini masih terdengar. Data scientist adalah pekerjaan yang menjanjikan, dimana memiliki keahlian memanipulasi tumpukan data digital untuk kemudian menemukan pola atau pattern yang memrediksi secara presisi tentang arah perilaku masa depan. Bagi perusahaan/ organisasi di Indonesia, sudah saatnya menyiapkan strategi pemanfaatan analisis Big data untuk mendukung strategi dan pengambilan keputusan yang tepat.

Pada aktifitas penelitian saat ini, peranan pengolah data statistik sangat penting karena membantu para peneliti untuk memeroleh insight dibalik sekumpulan informasi yang dimiliki. Perkembangan teknologi open source berdampak positif pada berkembangnya berbagai peranti lunak pengolah data. Secara umum, ada beberapa kategori software analisis data, yang bersifat komersil seperti SPSS dan SAS, serta bersifat open source seperti R.

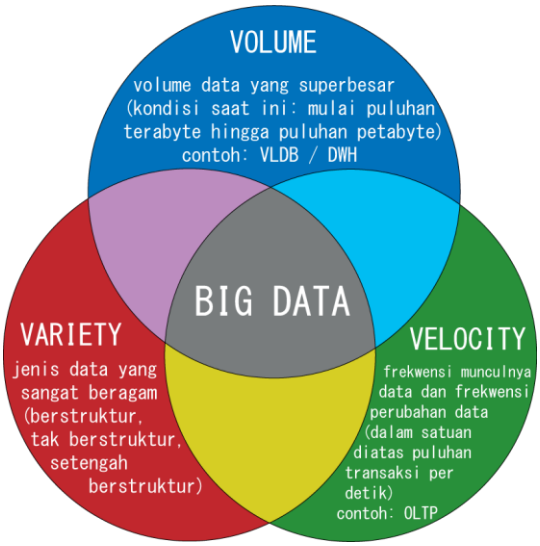
1.2. Definisi Data Science dan Big Data

Meskipun Istilah Big data terbilang baru, usaha untuk menyimpan data berukuran besar merupakan aktifitas yang sudah lama dilakukan di dunia Teknologi Informasi. Definisi umum
Big data dan Data science adalah:

"Big data is a term that describes the large volume of data – both structured and unstructured that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves"

"Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining and predictive analysis similar to Knowledge Discovery in Databases(KDD)."

Big data seperti yang ditunjukkan pada gambar 1.1 membuat kesempatan baru yang penting bagi organisasi untuk mengasilkan nilai baru dan membuat keuntungan kompetitif dari asset yang paling berarga yaitu informasi.



Gambar 1.1 Gambaran Big Data

Bagi bisnis, big data membantu efisiensi, kualitas dan personalisasi produk dan layanan, menghasilkan kepuasan pelanggan yang meningkat dan keuntungan. Big data analytics mengintegrasikan data terstruktur dan tikda terstruktur dengan query dan feed realtime, yang memungkinkan membuka jalur inovasi dan *insight*. Menurut McKinsey & Co, Big data didefinisikan sebagai:

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architecture and analytics to enable insights that unlock new sources of business value.

McKinsey & Co; Big data: The next frontier for innovation, competition and productivity [1][2].

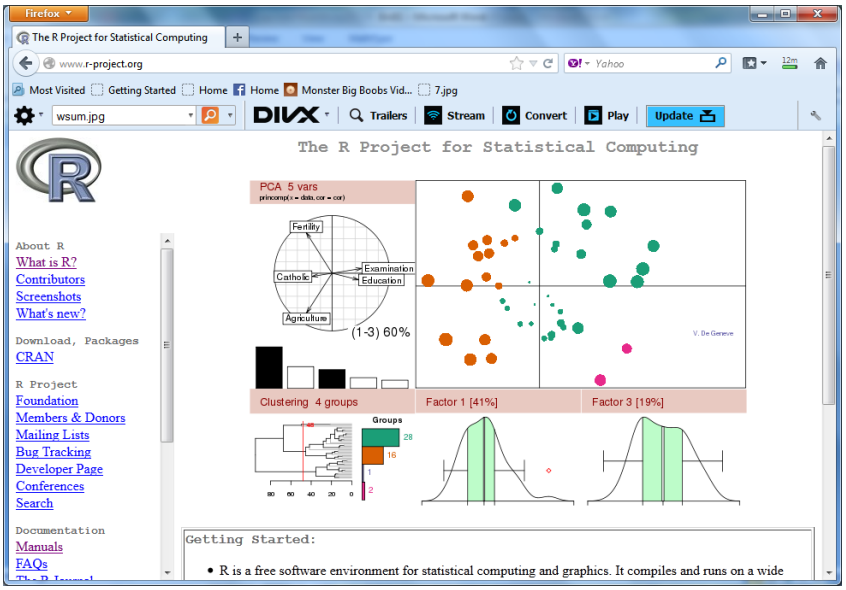
Sedangkan pada awal 2000an, analis industri bernama Doug Laney mendefinisikan big data dalam 3 V sebagai berikut:
Volume. Organisasi mengumpulkan data dari berbagai sumber, termasuk transaksi bisnis, social media dan informasi dari sensor atau mesin-mesin data.
Velocity. Aliran Data dengan kecepatan tak terbatas dan secara realtime terus menerus seperti RFID tags, sensor dan alat ukur pintar.
Variety. Data hadir dalam berbagai format, dari terstruktur hingga tak terstruktur, dokumen teks, email video, data karyawan, audio dan transaksi finansial.

Beberapa perusahaan besar dan popular yang memiliki Big Data dengan gambaran:
• Facebook: memiliki lebih dari 40 PB dan capture 100 TB/hari.
• Yahoo!: memiliki 60 PB data.
• Twitter: menerima 8 TB/hari.
• EBay: memiliki 40 PB data dan capture 50 TB/hari.

1.3 Mengenal Bahasa R

R adalah bahasa pemrograman open source yang berhubungan dengan komputasi dan pengolahan data untuk statistika dan yang berhubungan dengan penampilan grafik menggunakan tools yang disediakan oleh paket-paketnya yang sangat berguna di dalam penelitian dan industri. Saat ini, riset di Data Science atau Big Data menggunakan R sebagai tool untuk Statistical computing yang handal.

Paket statistik R bersifat multiplatforms, dengan file instalasi binary/file tar tersedia untuk sistem operasi Windows, Mac OS, Mac OS X, Linux, Free BSD, NetBSD, iSolaris, AIX, dan HPUX. Secara umum, sintaks dari bahasa R adalah ekuivalen dengan paket statistik Splus, sehingga sebagian besar keperluan analisis statistika, dan pemrograman dengan R adalah hampir identik dengan perintah yang dikenal di Splus. R menyediakan berbagai aplikasi statistika (linear dan nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) dan teknik penampilan simulasi gafik. Kelebihan R lainnya adalah kemudahan yang dirancang dengan baik untuk plot grafik berkualitas, termasuk simbol matematika dan rumus jika diperlukan, termasuk berbagai perbaikan di sisi tampilan aplikasi. Gambar 1.2 berikut menampilkan situs resmi R untuk tutorial dan pengunduhan di www.r-project.org dengan versi R 3.3. Paket ini diluncurkan pada tahun 2016 serta tersedia pada fasilitas network komprehensif R sejak 2006 (CRAN, http://CRAN.R-project.org/).



Gambar 1.2 Situs resmi dan pengunduhan software R[4]

R menyediakan tools statistik dan machine learning (*linear dan nonlinear modeling, classic statistical tests, time-series analysis, classification dan clustering*) dan grafik yang cocok untuk big data dan visualisasi seperti[5]:

• Data extraction

• Data cleaning

• Data loading

• Data transformation

• Statistical analysis

• Predictive modeling

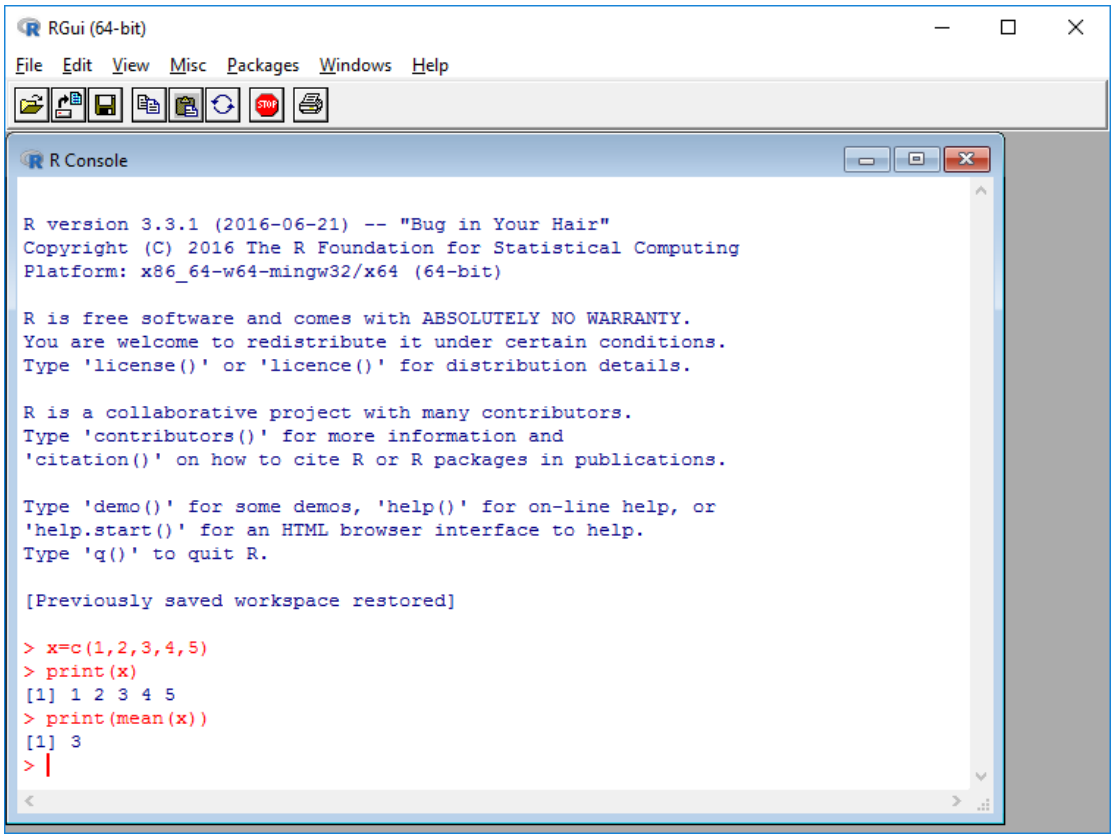
• Data visualization

1.4 Memulai Bahasa R

Setelah proses instalasi, maka Anda dapat mencoba memelajari bahasa R. Sebagai contoh, variabel dasar yang menyimpan data pada R adalah vektor.

>x=c(1,2,3,4,5)

Untuk mencetak nilai, gunakan perintah print(x), sedangkan untuk mengetahui nilai rata-rata (mean) dari x gunakan perintah **print(mean(x))** seperti gambar berikut:



Gambar 1.3. Penanganan dan pencetakan data dalam bentuk vektor

Pada pemrograman R, segalanya dianggap sebagai obyek. Sebuah vektor, matrix, data frame, bahkan variabel adalah obyek. Obyek di R antara lain:

1. Karakter
2. Numerik (Real Number)
3. Integer (Whole Number)
4. Bilangan komplek
5. Logika (True / False)

Sebagai contoh:

```
> a <- c(1.8, 4.5)   #numerik
> b <- c(1 + 2i, 3 - 6i) #komplek
> d <- c(23, 44)   #integer
> e <- vector("logical", length = 5)
```

Data Frame adalah tipe data untuk menyimpan data yang berukuran lebih dari satu. Contohnya:

```
> df <- data.frame(name = c("ash","jane","paul","Edy"), score = c(67,56,87,91))
> df
name score
1 ash 67
2 jane 56
3 paul 87
4 Edy 91

> dim(df)
[1] 4 2 # ukuran dimensi
```

Untuk membuat matrik, berikut contohnya:

Matrik.R:

```
M <- matrix(1:12, nrow=3)        # membuat matrik 3 x4
dim(M)                           # menampilkan dimensi
M[2, 3]                          # cetak elemen di  baris 2 dan kolom 3 2 * matrix(rep(1, 12),
nrow=3)                          # multiply every element by a constant
A <- array(1:12, dim=c(2, 2, 3))  # membuat array 3 dimensi
A[1, 2, 1]                       # print single element
A[1, , ]                         # print a matrix subset
```

Hasilnya:

```
> A[1, 2, 1]    # print single element
[1] 3
> A[1, , ]      # print a matrix subset
    [,1] [,2] [,3]
[1,]    1    5    9
[2,]    3    7   11
```

Latihan:

1. Installah Program R dan beberapa paket pendukung yang penting, misal untuk paket Psikologi:

- Unduh paket psych serta paket penting lainnya seperti ggpolt2, scales, dan lainnya menggunakan perintah:

>install.packages()

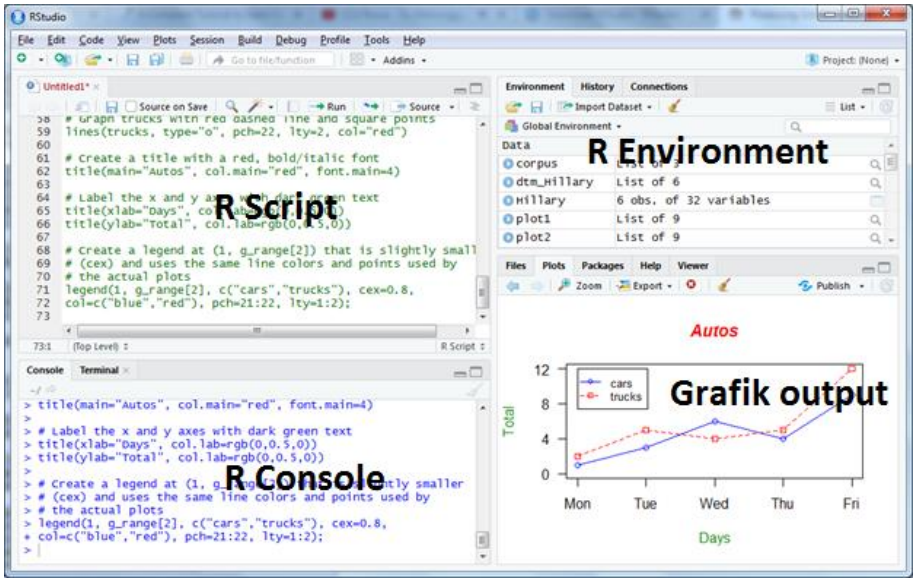
- Serta mengaktifkan paket yang dibutuhkan, misalnya menggunakan perintah **library(psych).** Masukkan data atau load data dari file teks atau data contoh dari web.

- Gunakan berbagai metode statistik deskriptif dan grafik deskriptif yang simple dari contoh-contoh yang diberikan di internet[4].

2. Buat program grafik untuk menampilkan fungsi Sinus.

   Sinus.R:
   t=seq(0,10,0.1)
   y=sin(t)
   plot(t,y,type="l", xlab="time", ylab="Sine wave")

3. RStudio merupakan salah satu software yang bagus untuk membuat dan mengedit program R. Install program Rstudio yang dapat diunduh pada https://www.rstudio.com/products/rstudio/download/#download dengan hasil sebagai berikut:



Gambar 1.4 Program RStudio.

Bab 2. Pemrograman Dasar Big Data menggunakan R

2.1 Mengakses Data

Dengan hadirnya teknologi Big Data seperti Hadoop, maka R dapat digunakan untuk melakukan analysis data yang berukuran besar. Big data biasanya disimpan dalam file atau yang dapat diakses melalui web. Sebagai contoh, untuk mengkses data yang disimpan di dalam file .CSV, berikut contohnya:

Bacacsv.R:

```
MyData <- read.csv(file="d:/BigDATA/DataPenjualan.csv", header=TRUE, sep=",")
MyData
```

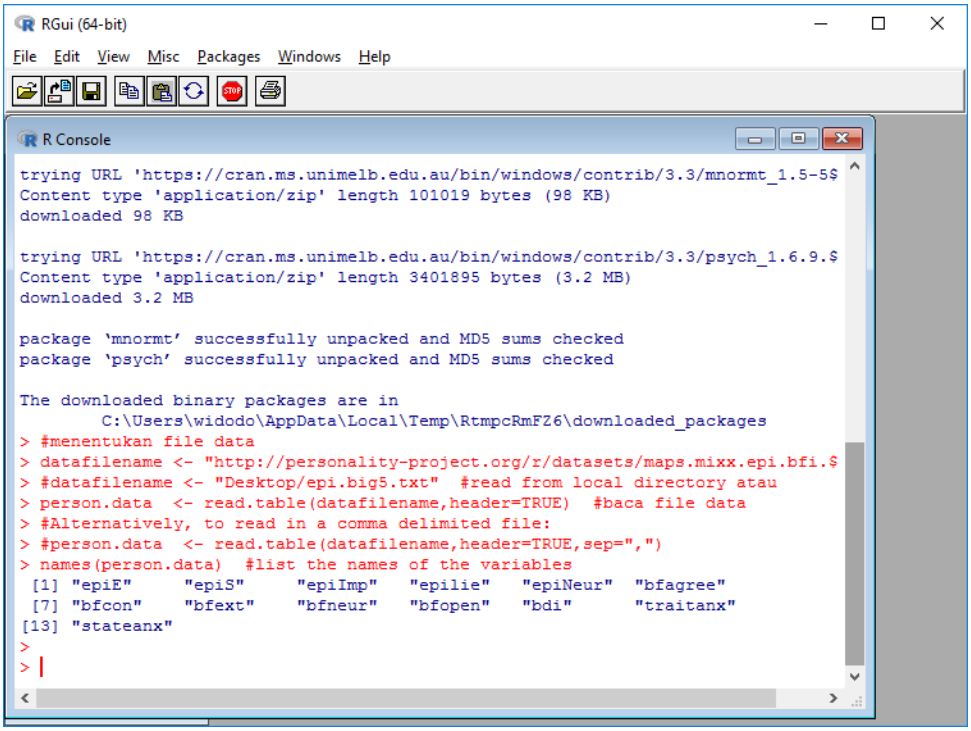
Hasilnya:

```
> MyData
  Camry New Rush Veloz
1    22      45     67
```

Sebagai contoh nyata, perhatikan kasus proyek personality berbasis R. Misal kita ingin membaca data dari file remote untuk beberapa ratus subyek pada 13 skala personality (5 dari Eysenck Personality Inventory (EPI), 5 dari 5 Big Five Inventory (BFI), Beck Depression Inventory dan 2 anxiety scales). Pada program di bawah, data yang sudah dibaca disimpan di tabel person.data dan hasilnya ditampilkan seperti gambar 2.1:

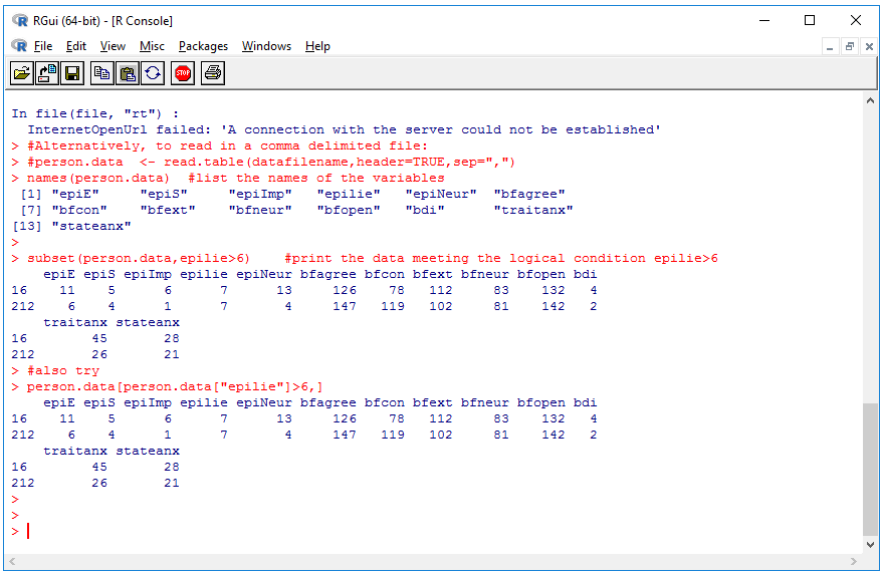
Personality1.R:

```
#menentukan file data
datafilename <- "http://personality-project.org/r/datasets/maps.mixx.epi.bfi.data"
#datafilename <- "Desktop/epi.big5.txt"  #read from local directory atau
person.data  <- read.table(datafilename,header=TRUE)  #baca file data
#Alternatively, to read in a comma delimited file:
#person.data  <- read.table(datafilename,header=TRUE,sep=",")
names(person.data)  #list the names of the variables
```



Gambar 2.1 Hasil penampilan person

Untuk memilih subset data, gunakan fungsi subset. Contoh berikut menampilkan subset untuk menampilkan *lie scales* yang tinggi.



Gambar 2.2. Hasil penampilan person.data[person.data["epilie"]>6,]

Untuk membaca file teks dari lokal atau remote server dan disimpan pada program statistik lainnya seperti SPSS dan SAS serta Minitab, dapat digunakan contoh program berikut:

```
datafilename <- "http://personality-project.org/r/datasets/finkel.sav"   #remote file
library(foreign)   #make it active
eli.data <- read.spss(datafilename, use.value.labels=TRUE, to.data.frame=TRUE)
```

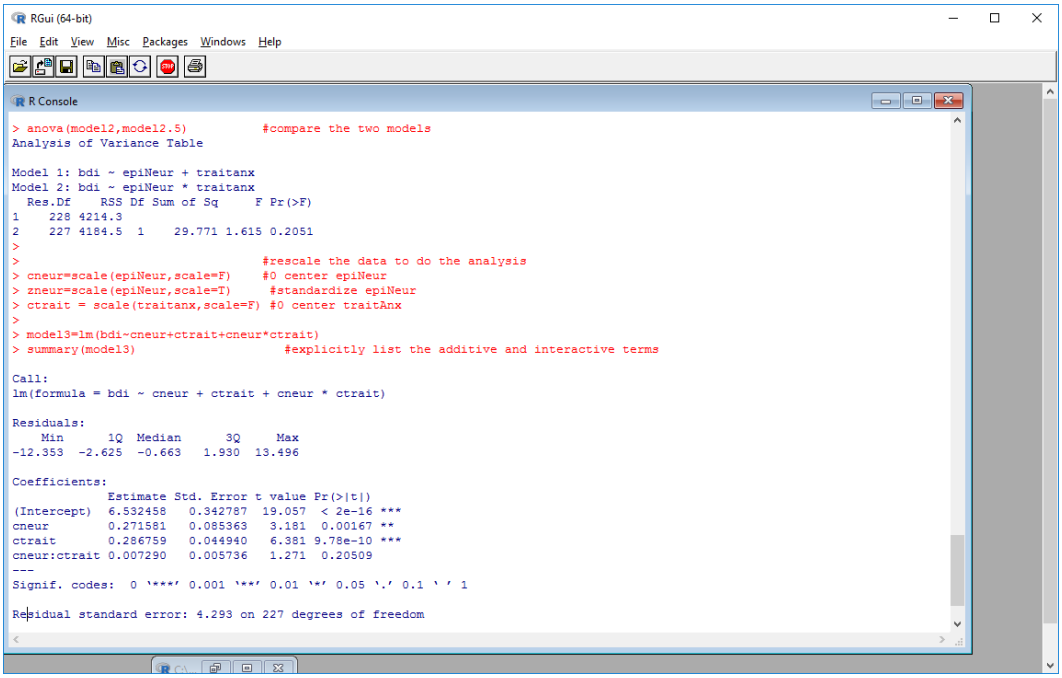
Lalu dataset dapat disimpan untuk analisa berikutnya menggunakan perintah **save**:

```
save(object,file="local name") #save an object (e.g., a correlation matrix) for later analysis
load(file)                     #gets the object (e.g., the correlation matrix back)
load(url("http://personality-project.org/r/datasets/big5r.txt"))  #get the correlation matrix
```

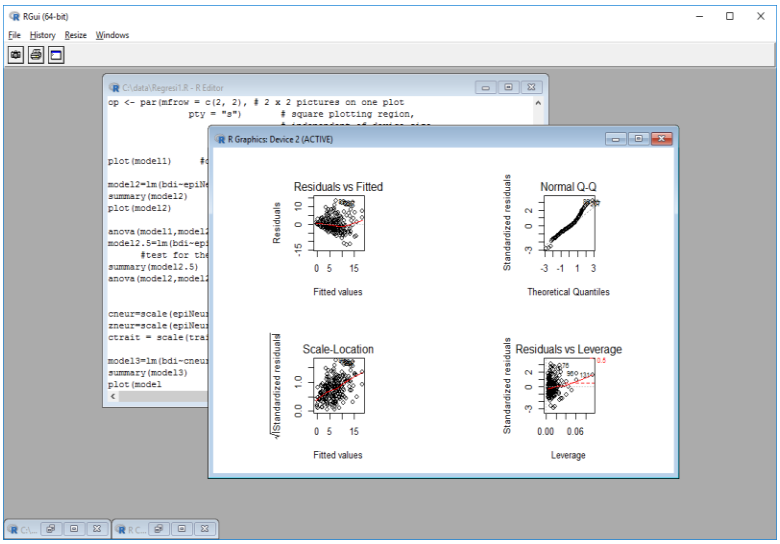
Penelitian Big Data menggunakan R sangat mudah, misalnya menggunakan regresi linear, sebagai berikut contoh program regresi linear:

Regresi.R:

```
datafilename="http://personality-project.org/r/datasets/maps.mixx.epi.bfi.data"
personality.data =read.table(datafilename,header=TRUE)  #baca data
names(personality.data)     #what variables are in the data set?
attach(personality.data)    #make the variables easier to use
model1 = lm(bdi~epiNeur)  #simple regression of beck depression on Neuroticism
summary(model1)   # basic statistical summary
op <- par(mfrow = c(2, 2), # 2 x 2 pictures on one plot
          pty = "s")      # square plotting region,
plot(model1)    #diagnostic plots in the graphics window
model2=lm(bdi~epiNeur+traitanx)    #add in trait anxiety
summary(model2)    #basic output
plot(model2)
anova(model1,model2)            #compare the difference between the two models
model2.5=lm(bdi~epiNeur*traitanx)
summary(model2.5)               #because we need to 0 center the data
anova(model2,model2.5)           #compare the two models
                                #rescale the data to do the analysis
cneur=scale(epiNeur,scale=F)    #0 center epiNeur
zneur=scale(epiNeur,scale=T)    #standardize epiNeur
ctrait = scale(traitanx,scale=F) #0 center traitAnx
model3=lm(bdi~cneur+ctrait+cneur*ctrait)
summary(model3)                 #explicitly list the additive and interactive terms
plot(model)
```



Gambar 2.3. Hasil perhitungan statistik regresi linear



Gambar 2.4. Hasil perhitungan statistik dan plot

2.2 Pemrograman Grafik

R memiliki grafis yang sangat baik dan kemampuan merencanakan, yang sebagian besar dapat ditemukan di 3 sumber utama: grafik dasar, lattice package dan paket ggplot2 yang sangat berguna untuk visualisi data di big data analytics. Secara umum untuk menampilkan plot grafik

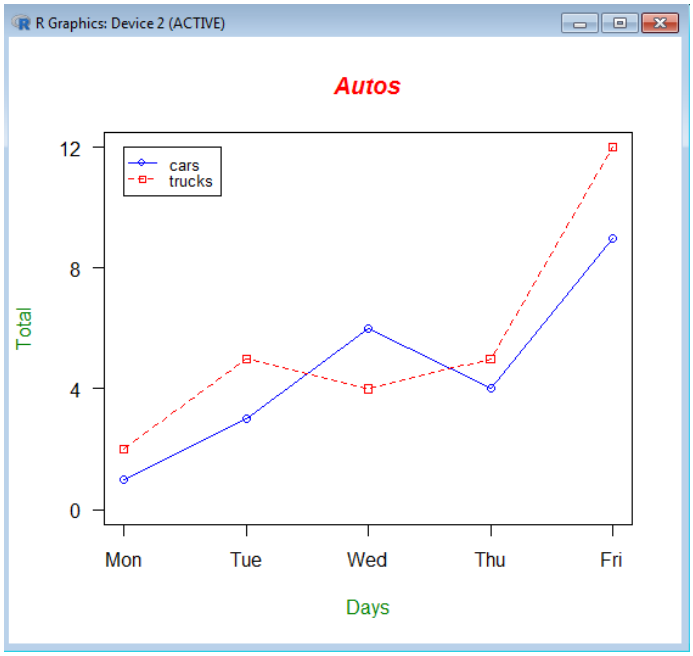
menggunakan fungsi **plot** dan fungsi **text** berparameter untuk menampilkan teks pada posisi tertentu:

```
> plot(pressure)
> text(150, 600, "Pressure (mm Hg)\nversus\nTemperature (Celsius)")
```

Berikut contohnya:

Graphics1.R:

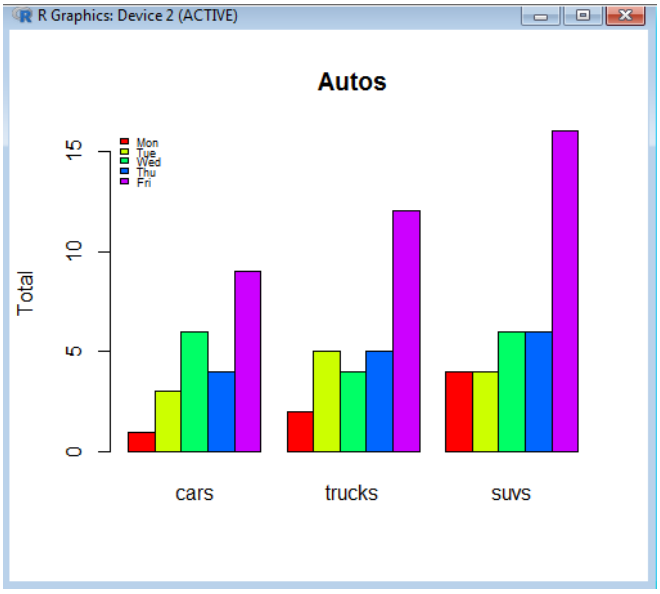
```
# Program membaca data mobil dan truk dari vektor
# Definisikan 2 vector
cars <- c(1, 3, 6, 4, 9)
trucks <- c(2, 5, 4, 5, 12)
# Calculate range from 0 to max value of cars and trucks
g_range <- range(0, cars, trucks)

# Graph autos using y axis that ranges from 0 to max
# value in cars or trucks vector.  Turn off axes and
# annotations (axis labels) so we can specify them ourself
plot(cars, type="o", col="blue", ylim=g_range, axes=FALSE, ann=FALSE)
# Make x axis using Mon-Fri labels
axis(1, at=1:5, lab=c("Mon","Tue","Wed","Thu","Fri"))

# Make y axis with horizontal labels that display ticks at
# every 4 marks. 4*0:g_range[2] is equivalent to c(0,4,8,12).
axis(2, las=1, at=4*0:g_range[2])
# Create box around plot
box()
# Graph trucks with red dashed line and square points
lines(trucks, type="o", pch=22, lty=2, col="red")
# Create a title with a red, bold/italic font
title(main="Autos", col.main="red", font.main=4)

# Label the x and y axes with dark green text
title(xlab="Days", col.lab=rgb(0,0.5,0))
title(ylab="Total", col.lab=rgb(0,0.5,0))

# Create a legend at (1, g_range[2]) that is slightly smaller
# (cex) and uses the same line colors and points used by
# the actual plots
legend(1, g_range[2], c("cars","trucks"), cex=0.8, col=c("blue","red"), pch=21:22, lty=1:2);
```



Gambar 2.5 Hasil grafik

Grafik2.r:

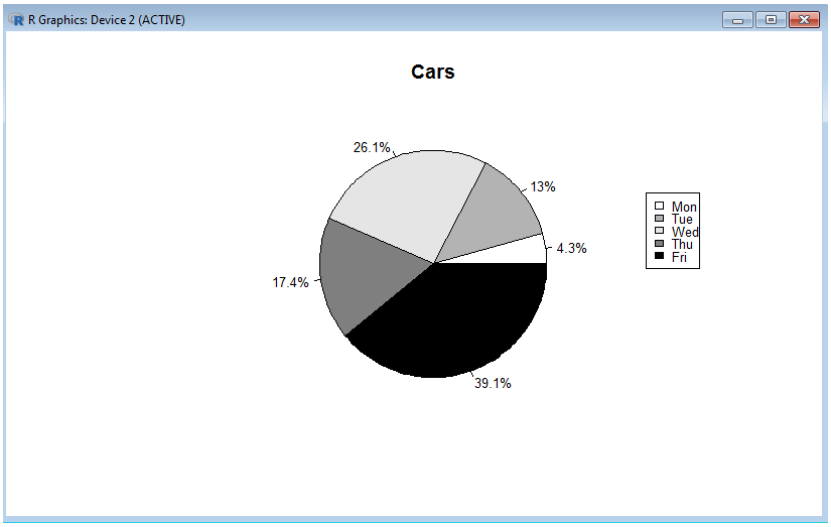
```
# Membaca data dari autos.dat
autos_data <- read.table("D:/BigDATA/autos.dat", header=T, sep="\t")

# Graph autos with adjacent bars using rainbow colors
barplot(as.matrix(autos_data), main="Autos", ylab= "Total",
      beside=TRUE, col=rainbow(5))

# Place the legend at the top-left corner with no frame
# using rainbow colors
legend("topleft", c("Mon","Tue","Wed","Thu","Fri"), cex=0.6, bty="n", fill=rainbow(5));
```



Gambar 2.6 Hasil Grafik

Grafik3.R:

```
# Define cars vector with 5 values
cars <- c(1, 3, 6, 4, 9)

# Define some colors ideal for black & white print
colors <- c("white","grey70","grey90","grey50","black")

# Calculate the percentage for each day, rounded to one
# decimal place
car_labels <- round(cars/sum(cars) * 100, 1)

# Concatenate a '%' char after each value
car_labels <- paste(car_labels, "%", sep="")

# Create a pie chart with defined heading and custom colors
# and labels
pie(cars, main="Cars", col=colors, labels=car_labels, cex=0.8)

# membuat legend di posisi kanan
legend(1.5, 0.5, c("Mon","Tue","Wed","Thu","Fri"), cex=0.8,  fill=colors)
```

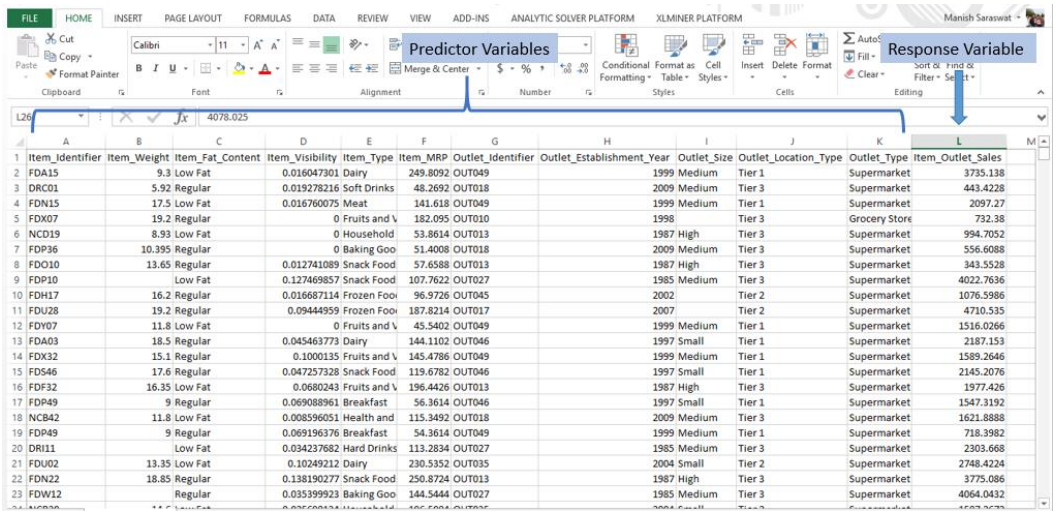


Gambar 2.7 Hasil grafik.

2.3 Exploratory Data Analysis

Data Exploration merupakan tingkatan penting pada model prediktif di big data. Terdapat beberapa istilah penting di dalam **Exploratory Data Analysis**, yaitu:

- **Response *Variable (Dependent Variable)***: pada data set, response variable (y) ialah dimana kita membuat prediksi, dalam hal ini kita memrediksi kolom 'Item_Outlet_Sales'.
- **Predictor Variable *(Independent Variable)***: pada dataset, predictor variables (Xi) ialah segala yang mana prediksi dibuat menggunakan response variable.
- **Train Data**: model prediktif dibangun dari melatih data set.
- **Test Data**: setelah model dibangun, akurasi diters menggunakan test data.

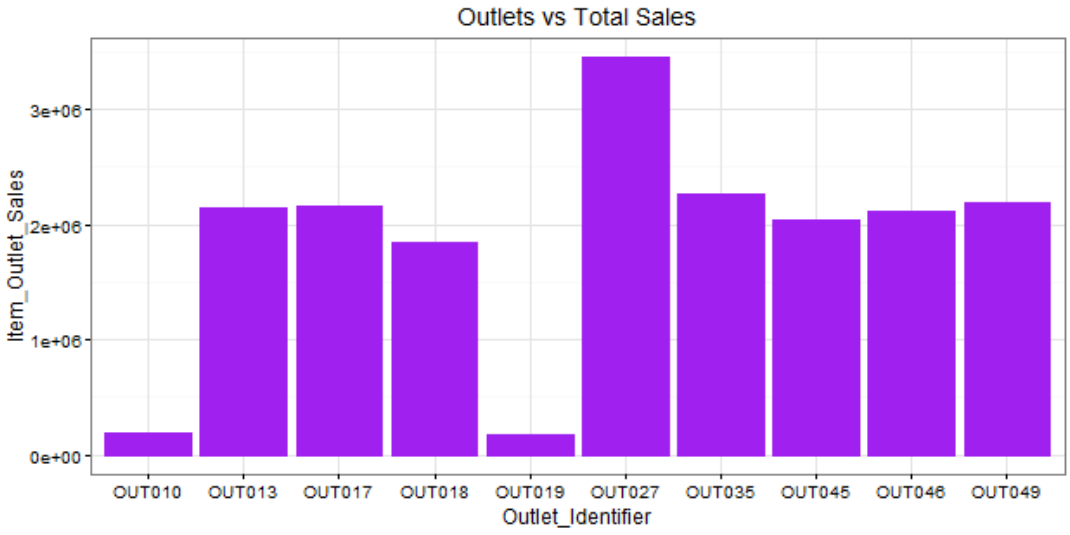


Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket	2097.27
FDX07	19.2	Regular	0	Fruits and V	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket	994.7052
FDP36	10.395	Regular	0	Baking Goo	51.4008	OUT018	2009	Medium	Tier 3	Supermarket	556.6088
FDO10	13.65	Regular	0.012741089	Snack Food	57.6588	OUT013	1987	High	Tier 3	Supermarket	343.5528
FDP10		Low Fat	0.127469857	Snack Food	107.7622	OUT027	1985	Medium	Tier 3	Supermarket	4022.7636
FDH17	16.2	Regular	0.016687114	Frozen Foo	96.9726	OUT045	2002		Tier 2	Supermarket	1076.5986
FDU28	19.2	Regular	0.09444959	Frozen Foo	187.8214	OUT017	2007		Tier 2	Supermarket	4710.535
FDY07	11.8	Low Fat	0	Fruits and V	45.5402	OUT049	1999	Medium	Tier 1	Supermarket	1516.0266
FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket	2187.153
FDX32	15.1	Regular	0.1000135	Fruits and V	145.4786	OUT049	1999	Medium	Tier 1	Supermarket	1589.2646
FDS46	17.6	Regular	0.047257328	Snack Food	119.6782	OUT046	1997	Small	Tier 1	Supermarket	2145.2076
FDF32	16.35	Low Fat	0.0680243	Fruits and V	196.4426	OUT013	1987	High	Tier 3	Supermarket	1977.426
FDP49	9	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket	1547.3192
NCB42	11.8	Low Fat	0.008596051	Health and	115.3492	OUT018	2009	Medium	Tier 3	Supermarket	1621.8888
FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket	718.3982
DRI11		Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket	2303.668
FDU02	13.35	Low Fat	0.10249212	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket	2748.4224
FDN22	18.85	Regular	0.138190277	Snack Food	250.8724	OUT013	1987	High	Tier 3	Supermarket	3775.086
FDW12		Regular	0.035399923	Baking Goo	144.5444	OUT027	1985	Medium	Tier 3	Supermarket	4064.0432

Gambar 2.7 Tabel yang ada pada dataset serta predictor variables dan response variable

Kita dapat melihat bahwa mayoritas penjualan diperoleh dari produk yang memiliki *visibility* kurang dari 0.2. Ini menghasilkan bahwa item_visibility < 2 harus merupakan faktor penting pada penentuan penjualan. Plot grafik untuk mengesplorasi data:

```
> ggplot(train, aes(Outlet_Identifier, Item_Outlet_Sales)) + geom_bar(stat = "identity", color = "purple") +theme(axis.text.x = element_text(angle = 70, vjust = 0.5, color = "black"))  + ggtitle("Outlets vs Total Sales") + theme_bw()
```



Gambar 2.8 Grafik outlet vs Total sales menjadi terlihat jelas

2.4 Human Resources Analytics dengan Big Data

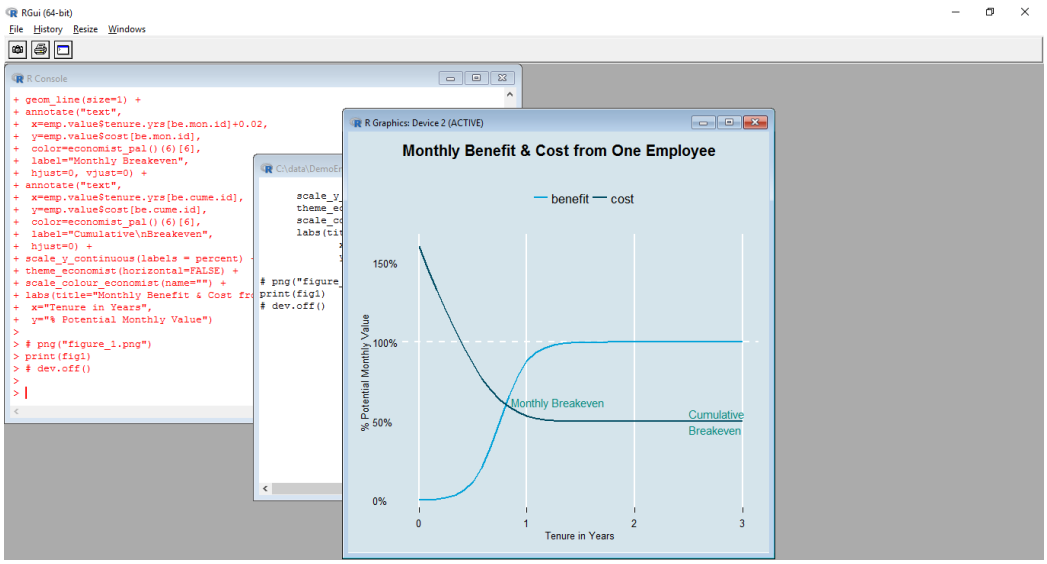
Pasha Roberts, ilmuan di Talent Analytics, membuat artikel dan demo Employee Churn untuk Predictive Analytics Times yang dikenal dengan Churn 201. Pada Churn 201, menampilkan konsep biaya, dan benefit pada pegawai. Berikut contohnya:

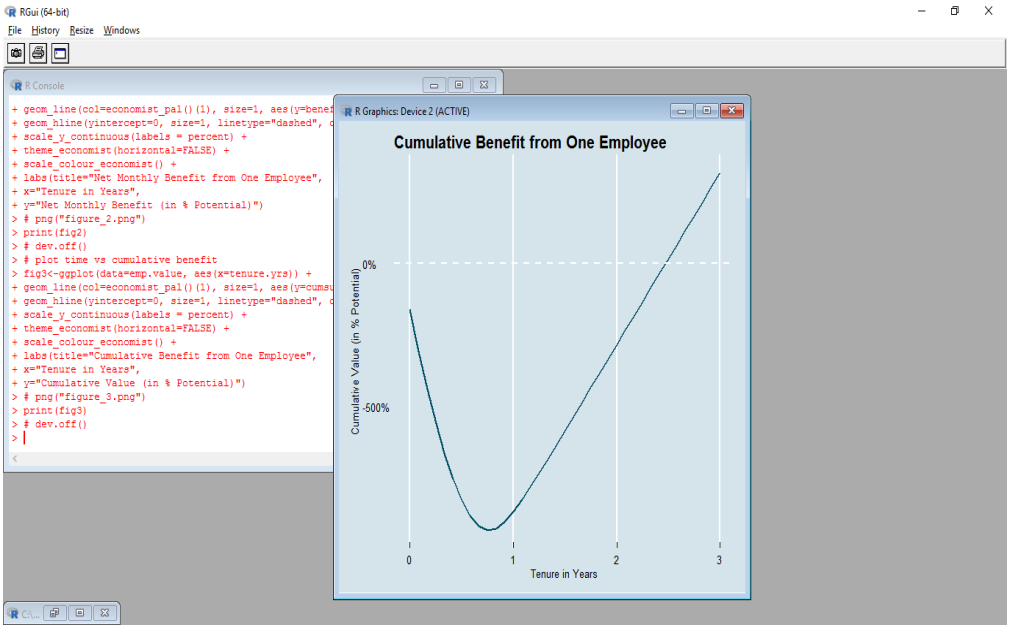
HREmployee.R:

```
library(reshape)
library(ggplot2)
library(ggthemes)
library(scales)

max.yrs <- 3              # max number of years to show on plot
max.benefit <- 1.5        # year at which employee delivers fully-trained value (asymptote)
cost.ramp <- 1.5          # higher numbers speed up time before costs = salary
cost.scale <- 3           # higher numbers increase height of original training costs
salary <- 0.5             # monthly salary as a percent of fully trained value delivered to company

# set up data frame with time series in months
emp.value <- as.data.frame(0:(max.yrs*12)/12)
names(emp.value)<- "tenure.yrs"

# set up benefit function, modeled as a sigmoid
emp.value$benefit <- 1/(1+exp(-(emp.value$tenure.yrs/max.benefit*12-6)))

# set up cost function, modeled as a gompertz
emp.value$cost <- exp(-exp(cost.ramp * emp.value$tenure.yrs)) * cost.scale + salary

# calc breakeven points
be.mon.id <- which.max((emp.value$benefit - emp.value$cost)>0)
be.mon <- emp.value$tenure.yrs[be.mon.id]
be.cume.id <- which.max(cumsum(emp.value$benefit - emp.value$cost)>0)
be.cume <- emp.value$tenure.yrs[be.cume.id]

writeLines(sprintf("At this rate net benefit begins at year %.2f, breakeven at year %.2f", be.mon,
be.cume))

# plot time vs. monthly cost, benefit
fig1<-ggplot(data=melt(emp.value, id.vars="tenure.yrs"), aes(x=tenure.yrs, y=value, col=variable)) +
        geom_hline(yintercept=1, size=1, linetype="dashed", col="white") +
        geom_line(size=1) +
        annotate("text",
                        x=emp.value$tenure.yrs[be.mon.id]+0.02,
                        y=emp.value$cost[be.mon.id],
                        color=economist_pal()(6)[6],
                        label="Monthly Breakeven",
                        hjust=0, vjust=0) +
```

```
        annotate("text",
                        x=emp.value$tenure.yrs[be.cume.id],
                        y=emp.value$cost[be.cume.id],
                        color=economist_pal()(6)[6],
                        label="Cumulative\nBreakeven",
                        hjust=0) +
        scale_y_continuous(labels = percent) +
        theme_economist(horizontal=FALSE) +
        scale_colour_economist(name="") +
        labs(title="Monthly Benefit & Cost from One Employee",
                x="Tenure in Years",
                y="% Potential Monthly Value")

# png("figure_1.png")
print(fig1)
# dev.off()

        # plot time vs. net monthly benefit
fig2<-ggplot(data=emp.value, aes(x=tenure.yrs)) +
geom_line(col=economist_pal()(1), size=1, aes(y=benefit-cost)) +
geom_hline(yintercept=0, size=1, linetype="dashed", col="white") +
scale_y_continuous(labels = percent) +
theme_economist(horizontal=FALSE) +
scale_colour_economist() +
labs(title="Net Monthly Benefit from One Employee",
x="Tenure in Years",
y="Net Monthly Benefit (in % Potential)")
# png("figure_2.png")
print(fig2)
# dev.off()
# plot time vs cumulative benefit
fig3<-ggplot(data=emp.value, aes(x=tenure.yrs)) +
geom_line(col=economist_pal()(1), size=1, aes(y=cumsum(benefit-cost))) +
geom_hline(yintercept=0, size=1, linetype="dashed", col="white") +
scale_y_continuous(labels = percent) +
theme_economist(horizontal=FALSE) +
scale_colour_economist() +
labs(title="Cumulative Benefit from One Employee",
x="Tenure in Years",
y="Cumulative Value (in % Potential)")
# png("figure_3.png")
print(fig3)
# dev.off()
```

Jalankan aplikasi tersebut, lalu lihat dan analisa hasilnya seperti gambar di bawah, terlihat bahwa penerapan Data Science dan Big Data untuk berbagai kasus Human Resources dapat dipecahkan dengan mudah menggunakan bahasa R.



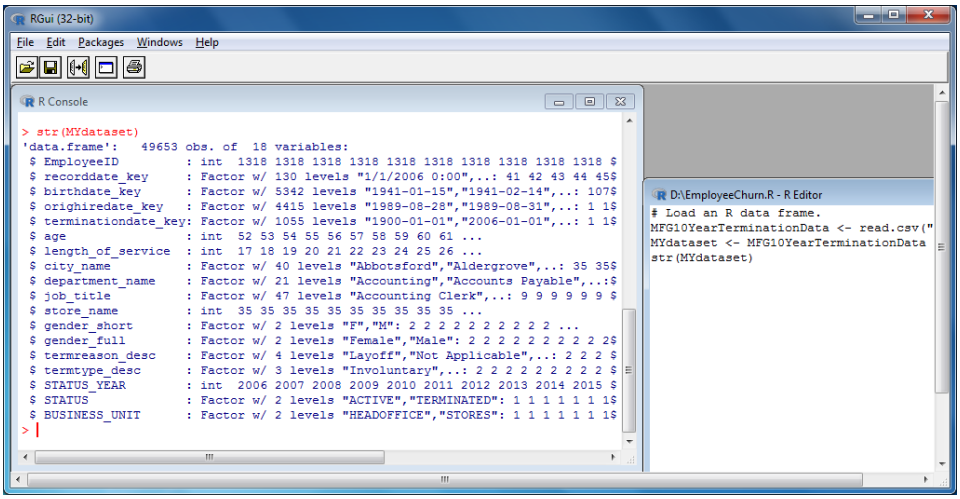


Gambar 2.9 Hasil Program Employee berbasis R

Berikut contoh Employee churn lainnya yang dapat dicoba, menggunakan dataset di:

https://www.analyticsinhr.com/wp-content/uploads/2016/09/Datasets-Tutorial-People-Analytics.zip

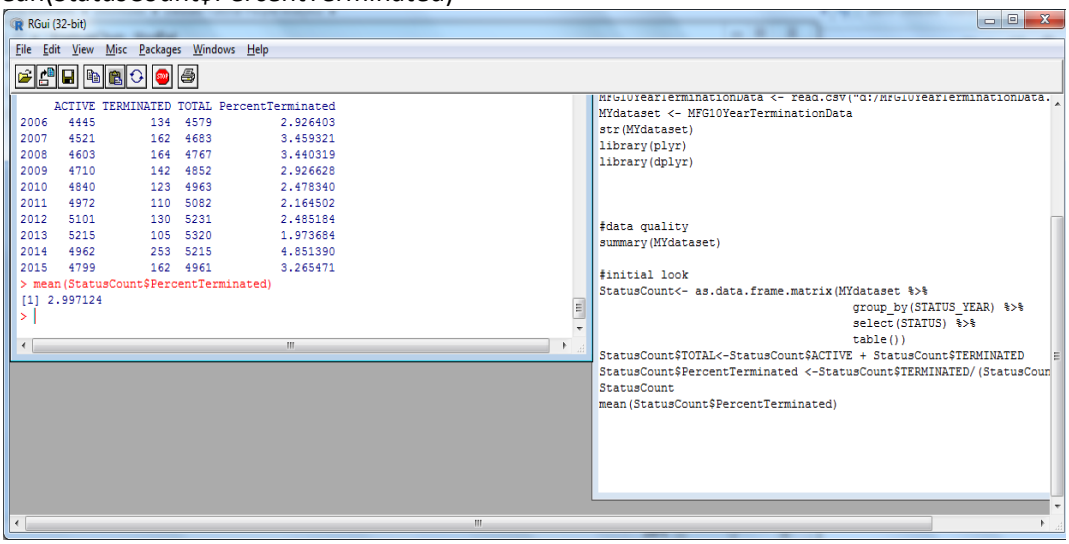
```
# Load an R data frame.
MFG10YearTerminationData <- read.csv("d:/MFG10YearTerminationData.csv")
MYdataset <- MFG10YearTerminationData
str(MYdataset)
```



Gambar 2.10 Data pegawai

Kita dapat mengetahui berapa status pegawai yang *leaving*?

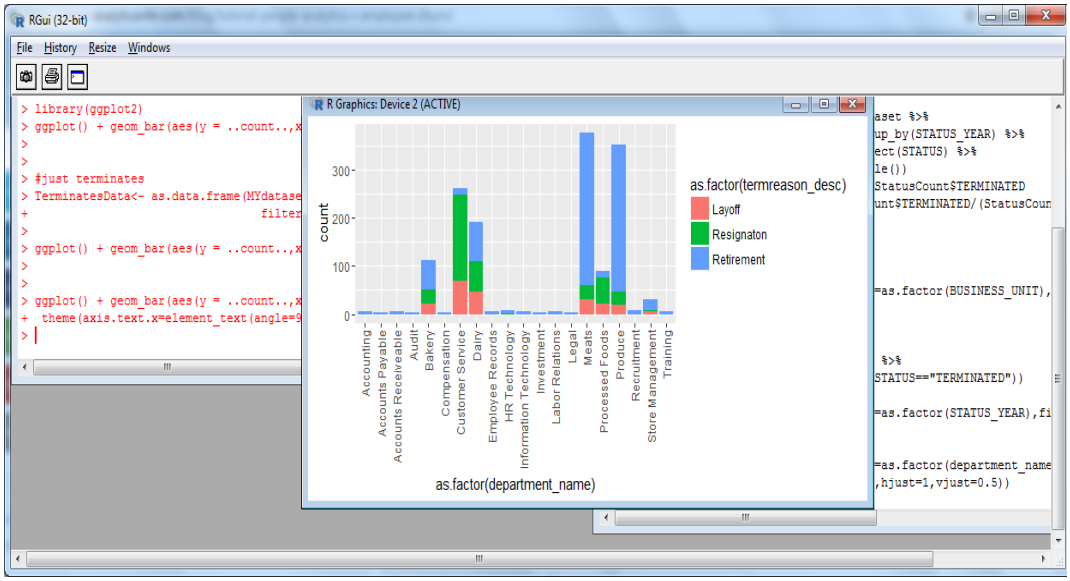
```
library(plyr)
library(dplyr)
#data quality
summary(MYdataset)
#initial look
StatusCount<- as.data.frame.matrix(MYdataset %>%
                        group_by(STATUS_YEAR) %>%
                        select(STATUS) %>%
                        table())
StatusCount$TOTAL<-StatusCount$ACTIVE + StatusCount$TERMINATED
StatusCount$PercentTerminated <-StatusCount$TERMINATED/(StatusCount$TOTAL)*100
StatusCount
mean(StatusCount$PercentTerminated)
```



Gambar 2.11 Data pegawai yang leaving rata rata 2.9%

Siapakah yang berhenti bekerja dapat dijawab sebagai berikut:

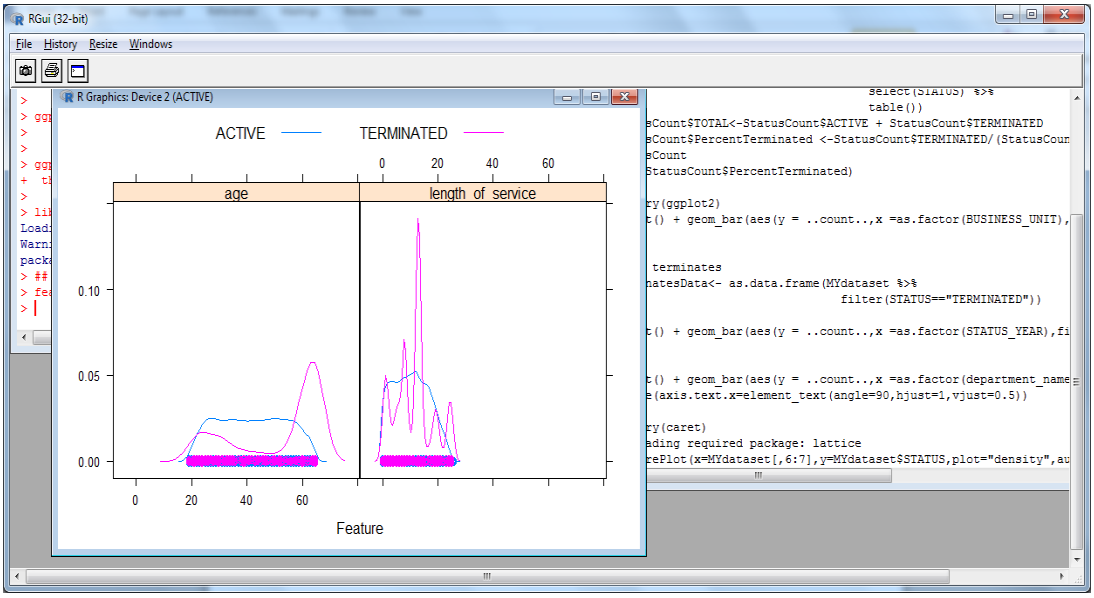
```
ggplot() + geom_bar(aes(y = ..count..,x =as.factor(department_name),fill =
as.factor(termreason_desc)),data=TerminatesData,position = position_stack())+

 theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



Gambar 2.11 Yang berhenti bekerja

Bagaimana umur dan lama kerja memengaruhi termination ?

```
library(caret)

## Loading required package: lattice

featurePlot(x=MYdataset[,6:7],y=MYdataset$STATUS,plot="density",auto.key = list(columns = 2))
```



Gambar 2.12 Hubungan umur dan lama bekerja yang memengaruhi *termination*

EmployeeChurn.R:

```
# Load an R data frame.
MFG10YearTerminationData <- read.csv("d:/MFG10YearTerminationData.csv")
MYdataset <- MFG10YearTerminationData
str(MYdataset)
library(plyr)
library(dplyr)
#data quality
summary(MYdataset)
#initial look
StatusCount<- as.data.frame.matrix(MYdataset %>%
                          group_by(STATUS_YEAR) %>%
                          select(STATUS) %>%
                          table())
StatusCount$TOTAL<-StatusCount$ACTIVE + StatusCount$TERMINATED
StatusCount$PercentTerminated <-StatusCount$TERMINATED/(StatusCount$TOTAL)*100
StatusCount
mean(StatusCount$PercentTerminated)
library(ggplot2)
ggplot() + geom_bar(aes(y = ..count..,x =as.factor(BUSINESS_UNIT),fill =
as.factor(STATUS)),data=MYdataset,position = position_stack())
#just terminates
TerminatesData<- as.data.frame(MYdataset %>%
                        filter(STATUS=="TERMINATED"))
ggplot() + geom_bar(aes(y = ..count..,x =as.factor(STATUS_YEAR),fill =
as.factor(termtype_desc)),data=TerminatesData,position = position_stack())
ggplot() + geom_bar(aes(y = ..count..,x =as.factor(STATUS_YEAR),fill =
as.factor(termreason_desc)),data=TerminatesData, position = position_stack())
ggplot() + geom_bar(aes(y = ..count..,x =as.factor(department_name),fill =
as.factor(termreason_desc)),data=TerminatesData,position = position_stack())+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
library(caret)
featurePlot(x=MYdataset[,6:7],y=MYdataset$STATUS,plot="density",auto.key = list(columns = 2))
```

Latihan:

1. Buatlah program control struktur untuk pengecekan kondisi sebagai berikut

Ifelse.R:

```
#initialize a variable
N <- 10
#check if this variable * 5 is > 40
if (N * 5 > 40){
```

```
    print("This is easy!")
} else {
    print ("It's not easy!")
}
[1] "This is easy!"
```

2. Buatlah program looping menggunakan *For* sebagai berikut:

Forloop.R:

```
#initialize a vector
y <- c(99,45,34,65,76,23)
#print the first 4 numbers of this vector
for(i in 1:4){
   print (y[i])
}
```

Hasilnya:

```
[1] 99
[1] 45
[1] 34
[1] 65
```

3. Buatlah program looping menggunakan while sebagai berikut:

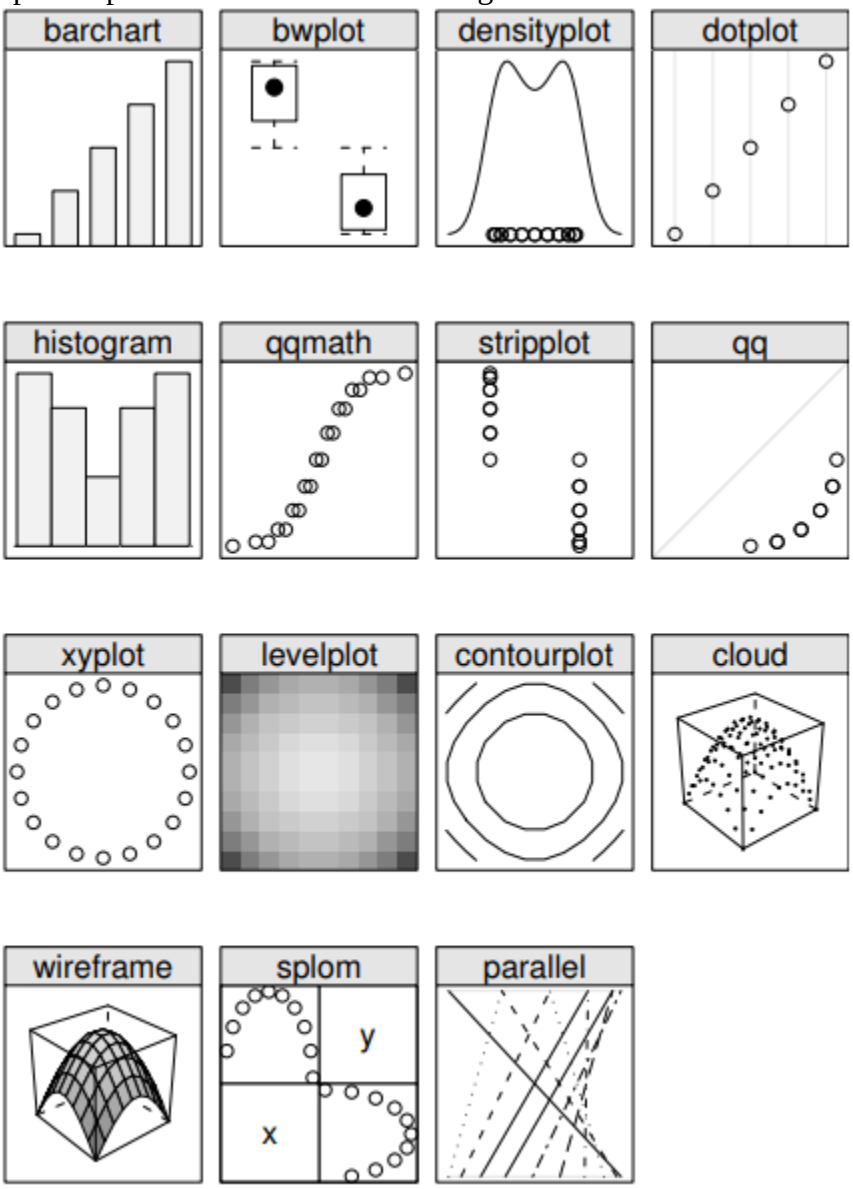
Whileloop.R:

```
#initialize a condition
Age <- 12
#check if age is less than 17
while(Age < 17){
    print(Age)
    Age <- Age + 1 #Once the loop is executed, this code breaks the loop
}
```

Hasilnya:

```
[1] 12
[1] 13
[1] 14
[1] 15
[1] 16
```

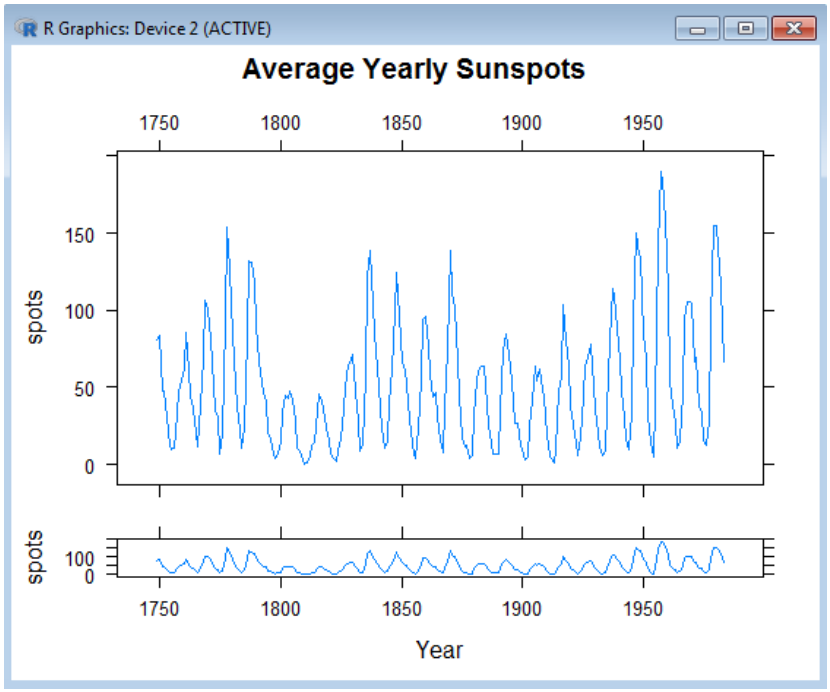
4. Eksplorari paket lattice untuk membuat grafik:



Gambar 2.13 Paket lattice untuk membuat berbagai model grafik

DemoLattice.R:

```
library(lattice)
spots <- by(sunspots, gl(235, 12, lab=1749:1983), mean)
plot1 <- xyplot(spots ~ 1749:1983, xlab="", type="l",
main="Average Yearly Sunspots",
scales=list(x=list(alternating=2)))
plot2 <- xyplot(spots ~ 1749:1983, xlab="Year", type="l")
print(plot1, position=c(0, 0.2, 1, 1), more=TRUE)
print(plot2, position=c(0, 0, 1, 0.33))
```

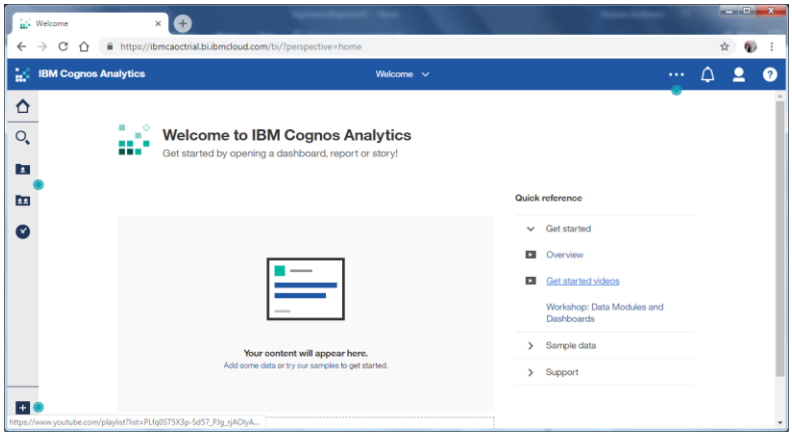


Gambar 2.14 hasil program grafik

5. Buatlah program grafik yang membaca data dari tabel di Ms. Excel

```
> install.packages("XLConnect")
> library("XLConnect")
> excel.file <- file.path("~/data.xlsx")
```

6. Cobalah menggunakan IBM Cognos Analytics 11.1, akses alamat https://ibmcaoctrial.bi.ibmcloud.com/bi/?perspective=home, lalu login



Gambar 2.15 halaman muka Congos Analytics[7]

Setelah login dan masuk, maka Anda dapat meload data examples, dengan klik try our samples, lalu pilihlah data yang sesuai. Berbagai tools dapat digunakan untuk analisa data yang ada.

Referensi

1. Introduction to Big Data, accessed at https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
2. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. (2011). Big data: The next frontier for innovation, competition and productivity. McKinsey & Co.
3. Widodo Budiharto. (2016). Machine Learning dan Computational Intelligence, Andi Offset Publisher, Yogyakarta. isbn: 978-979-29-5849-2
4. Widodo Budiharto dan Anggita Dian Cahyani. (2017), Pemrograman R untuk Sains dan Psikologi, Deepublish Publisher, Yogyakarta. Isbn: 978-602-401-759-0.
5. Vignesh Prajapati (2013). Big Data using R and Hadoop. Pack Publishing.
6. W. Budiharto and Meiliana. (2019) "Prediction and Analysis of Indonesia Presidential Election from Twitter using Sentiment Analysis, *Journal of Big Data*, Springer Publisher.
7. IBM Cognos Analytics 11.1, accessed at https://www.ibm.com/products/cognos-analytics

Tentang Penulis:

Prof. Dr. Widodo Budiharto *adalah guru besar bidang Kecerdasan Buatan di School of Computer Science, BINUS University Jakarta. Beliau dapat dihubungi di* wbudiharto@binus.edu*. HP: 08569887384.*