LAPORAN TUGAS 2

SELEKSI ASISTEN BASIS DATA 2016

disusun oleh

Edria Julianata 18214050

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

BANDUNG

2016

`

Daftar Isi

`

Daftar Gambar

`

I. Penjelasan Singkat *Dataset*

Dataset ini menyediakan data gaji penduduk kota San Francisco dari tahun 2011 sampai 2014, untuk berbagai macam job title. Kota San Francisco adalah salah satu kota di USA dengan standar gaji tinggi untuk penduduknya. Menurut Forbes, San Francisco adalah number 2 *Best-Paying City* dengan *overall median pay* sebesar $73,500 per tahun. Setiap *record* pada *dataset* ini terdiri dari *data ID, EmployeeName, JobTitle, BasePay, OvertimePay, OtherPay, Benefits, TotalPay, TotalPayBenefits, Year, Notes, Agency, Status (Full-time / Part-time).* Hal-hal yang dapat dianalisis lebih jauh dari dataset ini antara lain:

a. Tren rata-rata pendapatan penduduk San Francisco dari tahun ke tahun (dapat mencakup prediksi rata-rata pendapatan hingga tahun 2016 juga)

b. Pengelompokan penduduk San Francisco berdasarkan besar pendapatannya (rendah, sedang, tinggi)

c. Pengelompokan penduduk San Francisco berdasarkan bidang pekerjaannya (misal: Engineering, Politics, Public Services, Health Care, dll)

II. *Tools* yang Digunakan

Tools atau kakas yang penulis gunakan dalam melakukan analisis dan visualisasi data dari *dataset* yang diberikan adalah Bahasa R. Perangkat lunak yang mendukung analisis dan visualisasi data dengan menggunakan Bahasa R adalah RStudio beserta *packages* yang terkait, contohnya *ggplot2*, *dplyr*, dan *readr*.

III. Langkah Analisis

1. Memasukkan dataset *Salaries* ke dalam variabel pada RStudio
2. Membuat dataset baru yang berisi nilai rata-rata pendapatan penduduk dari tahun 2011-2014, dan prediksi rata-rata pendapatan penduduk pada tahun 2015 dan 2016.
3. Menunjukkan tren rata-rata pendapatan penduduk pada *console* dan histogram.

`

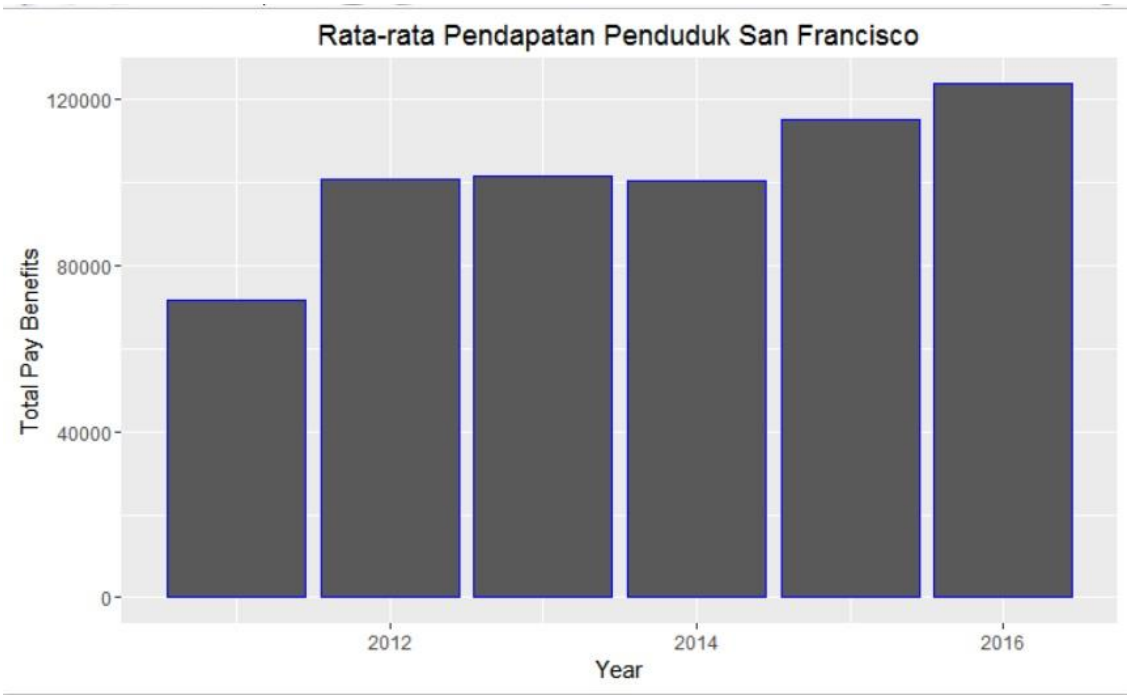
4. Menambahkan kolom *Group*. Kolom *Group* bernilai "Rendah" apabila total pendapatan penduduk kurang dari 65000, bernilai "Menengah" apabila total pendapatan penduduk diantara 65000-100000, dan bernilai "Tinggi" apabila total pendapatan penduduk lebih dari 100000. Angka tersebut didapat dari analisis penulis. Kemudian ditampilkan pada histogram.
5. Menambahkan kolom *Job*. Kolom *Job* ini bernilai jenis bidang pekerjaan dari masing-masing penduduk. Kolom ini bernilai jenis-jenis bidang pekerjaan seperti *Engineering, Public Works, Politics, Medical, Court*, dan lain sebagainya.
6. Penulis merapikan isi *dataset* terlebih dahulu dengan mengubah semua tulisan pada kolom *JobTitle* menjadi huruf besar semua (karena *case sensitive)* agar memudahkan menganalisa. Setelah dirapikan, ada lebih dari 1637 pekerjaan yang berbeda. Kemudian penulis mulai mengelompokkan pekerjaan-pekerjaan tersebut ke bidangnya.
7. Menunjukkan pengelompokkan penduduk San Fransisco berdasarkan bidang pekerjaannya secara keseluruhan pada histogram.

IV. Hasil Analisis

1. Tren rata-rata pendapatan penduduk San Francisco dari tahun ke tahun (dapat mencakup prediksi rata-rata pendapatan hingga tahun 2016 juga)

- Rata-rata pendapatan penduduk San Fransisco pada tahun 2011 sebesar 71744.1
- Rata-rata pendapatan penduduk San Fransisco pada tahun 2012 sebesar 100250.9
- Rata-rata pendapatan penduduk San Fransisco pada tahun 2013 sebesar 100553.2
- Rata-rata pendapatan penduduk San Fransisco pada tahun 2014 sebesar 101440.5
- Prediksi rata-rata pendapatan penduduk San Fransisco pada tahun 2015 sebesar 115099.1
- Prediksi rata-rata pendapatan penduduk San Fransisco pada tahun 2016 sebesar 123739.9

`



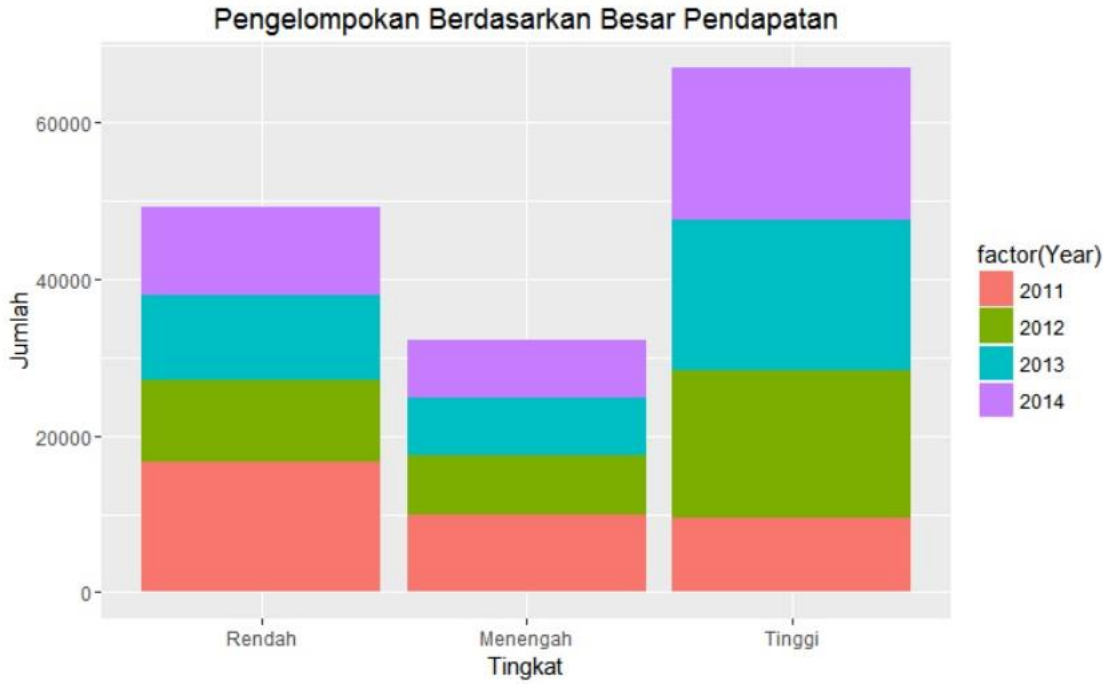
Gambar *1* Histogram Rata-rata Pendapatan Penduduk San Fransisco

2. Pengelompokan penduduk San Francisco berdasarkan besar pendapatannya (rendah, sedang, tinggi)

- Jumlah penduduk San Fransisco yang berpendapatan rendah sebanyak 49321 orang.
- Jumlah penduduk San Fransisco yang berpendapatan sedang sebanyak 32264 orang.
- Jumlah penduduk San Fransisco yang berpendapatan tinggi sebanyak 67069 orang.

Jumlah penduduk yang berpendapatan tinggi semakin tahun semakin meningkat.

`

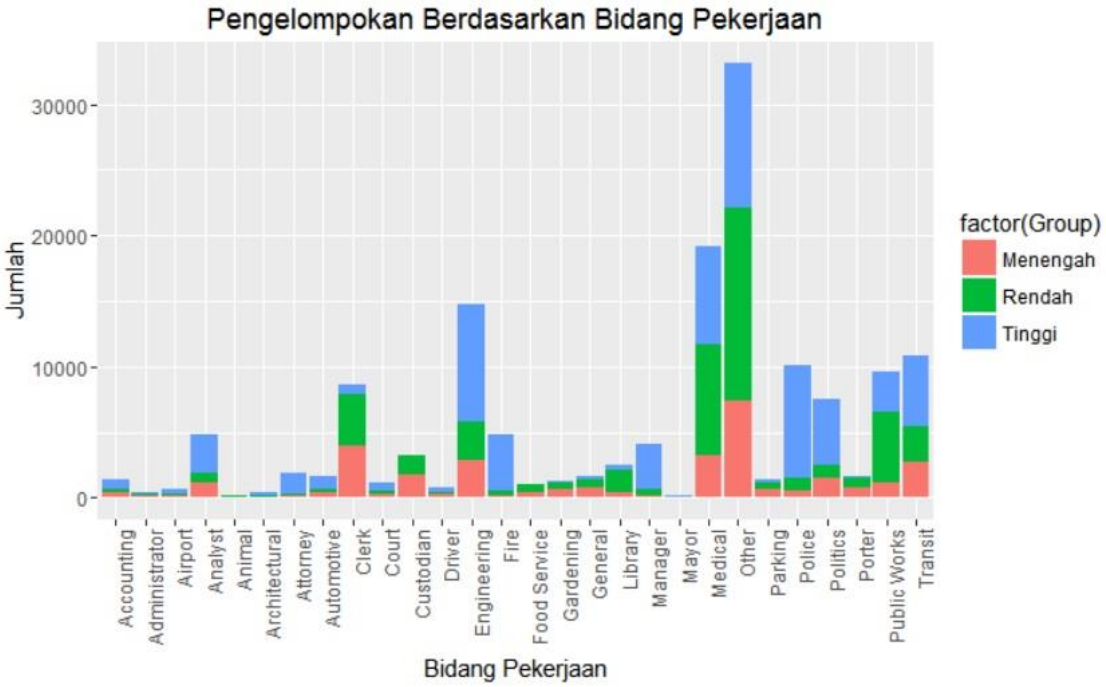


Gambar *2* Histogram Pengelompokan Berdasarkan Besar Pendapatan

3. Pengelompokan penduduk San Francisco berdasarkan bidang pekerjaannya (misal: Engineering, Politics, Public Services, Health Care, dll)

```
  Accounting Administrator       Airport       Analyst        Animal Architectural
        1331           443           637          4800           137           375
    Attorney    Automotive         Clerk         Court     Custodian        Driver
        1837          1680          8563          1130          3214           728
 Engineering          Fire  Food Service     Gardening       General       Library
       14738          4871          1063          1274          1593          2527
     Manager         Mayor       Medical         Other       Parking        Police
        4081           212         19136         33118          1412         10133
    Politics        Porter  Public Works       Transit
        7461          1695          9645         10820
```

`



Gambar *3* Histogram Pengelompokan Berdasarkan Bidang Pekerjaan

V. *Script*

```
#library yang digunakan
library(ggplot2)
library(dplyr)
library(readr)

#Memasukan data ke dalam variabel
Salaries <- read.csv("Salaries.csv", header = TRUE)

#####
#Mencari rata2 (point a)
mean(Salaries$TotalPayBenefits[which(Salaries$Year == "2011")])
mean(Salaries$TotalPayBenefits[which(Salaries$Year == "2012")])
mean(Salaries$TotalPayBenefits[which(Salaries$Year == "2013")])
mean(Salaries$TotalPayBenefits[which(Salaries$Year == "2014")])

#Mencari prediksi
rata <- aggregate(TotalPayBenefits ~ Year, Salaries, mean)
reg <- lm(rata$TotalPayBenefits~rata$Year, data = rata)
rata2 <- rata
rata <- data.frame(Year=c(2015,2016), TotalPayBenefits = 0)
rata$TotalPayBenefits <- predict(reg, rata)

rata2 <- rbind(rata2, rata) #menggabungkan

#histogram nomor 1
ggplot(data=rata2,aes(x = Year, y = TotalPayBenefits)) +
  geom_bar(colour = "blue",stat = "identity") +
```

`

```
  ggtitle("Rata-rata Pendapatan Penduduk San Francisco") +
  ylab("Total Pay Benefits") +
  xlab("Year")

#########
#Nomor 2
#Pengelompokan pendapatan dari yang rendah hingga tinggi
Salaries.Job <- Salaries
Salaries.Job$Group = "Rendah"
Salaries.Job[which(Salaries.Job$TotalPayBenefits > 65000), 'Group'] = "Menengah"
Salaries.Job[which(Salaries.Job$TotalPayBenefits > 100000), 'Group'] = "Tinggi"

#Histogram
Salaries.Job %>%
  ggplot(aes(x=factor(Group, level=c('Rendah', 'Menengah', 'Tinggi')))) +
  geom_bar() + aes(fill=factor(Year)) +
  ggtitle("Pengelompokan Berdasarkan Besar Pendapatan") +
  ylab("Jumlah") +
  xlab("Tingkat")

table(Salaries.Job$Group)

######
#No 3 (poin c)
#Mengubah semua tulisan menjadi huruf besar
Salaries.Job$JobTitle <- toupper(Salaries.Job$JobTitle)

#Mengecek atribut yang unik
#unique(as.character(Salaries.Job$JobTitle))


Salaries.Job$Job = "Other"
Salaries.Job[which(grepl('FIRE',Salaries.Job$JobTitle)), 'Job'] = "Fire"
Salaries.Job[which(grepl('POLICE',Salaries.Job$JobTitle)), 'Job'] = "Police"
Salaries.Job[which(grepl('SHERIF',Salaries.Job$JobTitle)), 'Job'] = "Police"
Salaries.Job[which(grepl('PROBATION',Salaries.Job$JobTitle)), 'Job'] = "Police"
Salaries.Job[which(grepl('SERGEANT',Salaries.Job$JobTitle)), 'Job'] = "Police"
Salaries.Job[which(grepl('INSPECTOR',Salaries.Job$JobTitle)), 'Job'] = "Police"
Salaries.Job[which(grepl('LIEUTENANT',Salaries.Job$JobTitle)), 'Job'] = "Police"
Salaries.Job[which(grepl('MTA',Salaries.Job$JobTitle)), 'Job'] = "Transit"
Salaries.Job[which(grepl('TRANSIT',Salaries.Job$JobTitle)), 'Job'] = "Transit"
Salaries.Job[which(grepl('ANESTH',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('MEDICAL',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('NURS',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('PHYSICIAN',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('HEALTH',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('ORTHOPEDIC',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('HEALTH',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('PHARM',Salaries.Job$JobTitle)), 'Job'] = "Medical"
Salaries.Job[which(grepl('DENTIST',Salaries.Job$JobTitle)), 'Job'] = "Medical"

Salaries.Job[which(grepl('AIRPORT',Salaries.Job$JobTitle)), 'Job'] = "Airport"
Salaries.Job[which(grepl('ANIMAL',Salaries.Job$JobTitle)), 'Job'] = "Animal"
Salaries.Job[which(grepl('ARCHITECT',Salaries.Job$JobTitle)), 'Job'] = "Architectural"
Salaries.Job[which(grepl('COURT',Salaries.Job$JobTitle)), 'Job'] = "Court"
```

`

```
Salaries.Job[which(grepl('LEGAL',Salaries.Job$JobTitle)), 'Job'] = "Court"
Salaries.Job[which(grepl('DEFENDER',Salaries.Job$JobTitle)), 'Job'] = "Court"
Salaries.Job[which(grepl('CRIMINAL',Salaries.Job$JobTitle)), 'Job'] = "Court"
Salaries.Job[which(grepl('VICTIM',Salaries.Job$JobTitle)), 'Job'] = "Court"

Salaries.Job[which(grepl('MAYOR',Salaries.Job$JobTitle)), 'Job'] = "Mayor"
Salaries.Job[which(grepl('LIBRAR',Salaries.Job$JobTitle)), 'Job'] = "Library"
Salaries.Job[which(grepl('PARKING',Salaries.Job$JobTitle)), 'Job'] = "Parking"
Salaries.Job[which(grepl('PUBLIC WORKS',Salaries.Job$JobTitle)), 'Job'] = "Public Works"
Salaries.Job[which(grepl('PUBLIC SERVICE',Salaries.Job$JobTitle)), 'Job'] = "Public Works"
Salaries.Job[which(grepl('INSPECTOR',Salaries.Job$JobTitle)), 'Job'] = "Public Works"

Salaries.Job[which(grepl('ATTORNEY',Salaries.Job$JobTitle)), 'Job'] = "Attorney"
Salaries.Job[which(grepl('MECHANIC',Salaries.Job$JobTitle)), 'Job'] = "Automotive"
Salaries.Job[which(grepl('AUTOMOTIVE',Salaries.Job$JobTitle)), 'Job'] = "Automotive"
Salaries.Job[which(grepl('CUSTODIAN',Salaries.Job$JobTitle)), 'Job'] = "Custodian"
Salaries.Job[which(grepl('ENGINEER',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('ENGR',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('ACCOUNT',Salaries.Job$JobTitle)), 'Job'] = "Accounting"
Salaries.Job[which(grepl('AUDITOR',Salaries.Job$JobTitle)), 'Job'] = "Accounting"
Salaries.Job[which(grepl('GARDENER',Salaries.Job$JobTitle)), 'Job'] = "Gardening"
Salaries.Job[which(grepl('GENERAL LABORER',Salaries.Job$JobTitle)), 'Job'] = "General"
Salaries.Job[which(grepl('FOOD SERV',Salaries.Job$JobTitle)), 'Job'] = "Food Service"
Salaries.Job[which(grepl('CLERK',Salaries.Job$JobTitle)), 'Job'] = "Clerk"
Salaries.Job[which(grepl('PORTER',Salaries.Job$JobTitle)), 'Job'] = "Porter"
#others
Salaries.Job[which(grepl('DEPUTY',Salaries.Job$JobTitle)), 'Job'] = "Politics"
Salaries.Job[which(grepl('MANAGER',Salaries.Job$JobTitle)), 'Job'] = "Manager"
Salaries.Job[which(grepl('TECHNICIAN',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('CONTROLLER',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('ELECTRIC',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('PLUMBER',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('TECH',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('PROGRAM',Salaries.Job$JobTitle)), 'Job'] = "Engineering"



Salaries.Job[which(grepl('DEPARTMENT',Salaries.Job$JobTitle)), 'Job'] = "Politics"
Salaries.Job[which(grepl('LEGISLATIVE',Salaries.Job$JobTitle)), 'Job'] = "Politics"
Salaries.Job[which(grepl('SECRETARY',Salaries.Job$JobTitle)), 'Job'] = "Politics"
Salaries.Job[which(grepl('DEPT',Salaries.Job$JobTitle)), 'Job'] = "Politics"
Salaries.Job[which(grepl('EMPLOYEE',Salaries.Job$JobTitle)), 'Job'] = "Public Works"
Salaries.Job[which(grepl('PROTECTIVE',Salaries.Job$JobTitle)), 'Job'] = "Public Works"
Salaries.Job[which(grepl('SOCIAL',Salaries.Job$JobTitle)), 'Job'] = "Public Works"
Salaries.Job[which(grepl('MAINTENANCE',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('ANALYST',Salaries.Job$JobTitle)), 'Job'] = "Analyst"
Salaries.Job[which(grepl('ADMINISTRATOR',Salaries.Job$JobTitle)), 'Job'] = "Administrator"
Salaries.Job[which(grepl('PLANNER',Salaries.Job$JobTitle)), 'Job'] = "Engineering"
Salaries.Job[which(grepl('DRIVER',Salaries.Job$JobTitle)), 'Job'] = "Driver"

#Histogram point c
Salaries.Job %>%
  ggplot(aes(x=Job)) +
  geom_bar() + aes(fill=factor(Group)) +
  ggtitle("Pengelompokan Berdasarkan Bidang Pekerjaan") +
```

`

```
  ylab("Jumlah") +
  xlab("Bidang Pekerjaan") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

table(Salaries.Job$Job)

#Menuliskan table ke file csv
write.csv(Salaries.Job, file = "Salaries_new.csv")
write.csv(rata2, file = "ratarata.csv")
View(Salaries.Job)
View(rata2)
```