Nama : Alfianri Manihuruk

NIM : 120450088

Matkul : Komputasi Statistik

1. **Lakukan eksplorasi data, tentukan variabel mana yang faktor, summary,dsb Apakah ada data pencilan. Visualisasikan data barchart dan boxplot.**

   a. Sebelum kita melakukan eksplrasi data, kita import data kita terlebih dahulu ke r studio dengan syntax sebagai berikut, tergantung lokasi kita menyimpan datanya.

```
# import data
dataku <- read.csv("D:/Sains Data Semester 5/Komputasi Statistik/3. Tugas Pertemuan 3.csv", header = TRUE)
```

   b. Melihat struktur data

```
str(dataku) # melihat str data yang kita miliki
```

```
> str(dataku) # melihat str data yang kita miliki
'data.frame':   1000 obs. of  11 variables:
 $ gender                     : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 1 2 2 ...
 $ race.ethnicity             : Factor w/ 5 levels "group A","group B",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ parent_education_level     : Ord.factor w/ 6 levels "some high school"<..: 4 3 6 6 4 1 4 4 5 3 ...
 $ lunch                      : Factor w/ 2 levels "free/reduced",..: 1 2 2 1 2 1 1 1 2 1 ...
 $ test_prep_course           : Factor w/ 2 levels "completed","none": 2 1 2 2 1 2 2 2 1 1 ...
 $ math                       : int  47 78 50 73 55 39 62 41 80 50 ...
 $ gender_code                : num  1 1 0 1 0 1 1 0 1 1 ...
 $ parent_education_level_code: num  4 3 6 6 4 1 4 4 5 3 ...
 $ lunch_code                 : num  1 2 2 1 2 1 1 1 2 1 ...
 $ nilai_huruf                : Factor w/ 7 levels "E","D","C","BC",..: 1 6 1 5 2 1 3 1 7 1 ...
 $ status_penilaian           : Factor w/ 2 levels "Tidak Lulus",..: 1 2 1 2 1 1 2 1 2 1 ...
```

Dari syntax tadi kita mendapatkkan informasi bahwa dataset ini terdiri dari 1000 pengamatan (baris) dari 6 variabel (kolom). Kemudian, kolom pertama sampai kelima adalah tipe character, dan kolom keenam bernama *math* yang bertipe integer.

   c. Tentukan mana kolom faktor

   Untuk mengecek apakah kolom tertentu berjenis faktor, dapat menggunakan `is.factor(x$y)`. Karena kita sudah melihat struktur data di atas, dan tidak ada tipe data faktor, artinya secara bawaan belum ada yang berjenis faktor. Akan tetapi jika kita lihat dengan teliti, kita dapat mengbah kolomnya menjadi jenis faktor.

   d. Tampilkan Summary

```
summary(dataku) # Menampilkan ringkasan data
```

```
> summary(dataku) # Menampilkan ringkasan data
    gender          race.ethnicity     parent_education_level    lunch          test_prep_course       math       
 Length:1000        Length:1000        Length:1000            Length:1000        Length:1000        Min.   :  0.00  
 Class :character   Class :character   Class :character       Class :character   Class :character   1st Qu.: 57.00  
 Mode  :character   Mode  :character   Mode  :character       Mode  :character   Mode  :character   Median : 66.00  
                                                                                                   Mean   : 66.09  
                                                                                                   3rd Qu.: 77.00  
                                                                                                   Max.   :100.00  
```

   Jika kita perhatian untuk summary datanya, tidak begitu lengkap hal ini dikarenakan type datanya kebanyakan character, akan tetapi hal ini dapat kita atasi dengan syntax sebagai berikut.

```
# mengubah type data character
dataku$gender <- factor(dataku$gender)
dataku$race.ethnicity <- factor(dataku$race.ethnicity)
dataku$parent_education_level <- factor(dataku$parent_education_level, order = TRUE,
                                        levels =c("some high school",
                                                  "high school",
                                                  "some college",
                                                  "associate\'s degree",
                                                  "bachelor\'s degree",
                                                  "master\'s degree"))
```

Selanjutkan kita panggil lagi dan kita lihat perbedaannya

```
> str(dataku)
'data.frame':   1000 obs. of  6 variables:
 $ gender                : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...
 $ race.ethnicity        : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...
 $ parent_education_level: Ord.factor w/ 6 levels "some high school"<..: 5 3 6 4 3 4 3 3 2 2 ...
 $ lunch                 : chr  "standard" "standard" "standard" "free/reduced" ...
 $ test_prep_course      : chr  "none" "completed" "none" "none" ...
 $ math                  : int  72 69 90 47 76 71 88 40 64 38 ...
```

```
> summary(dataku)
   gender    race.ethnicity       parent_education_level    lunch          test_prep_course       math
female:518   group A: 89    some high school  :179       Length:1000        Length:1000        Min.   :  0.00
male  :482   group B:190    high school       :196       Class :character   Class :character   1st Qu.: 57.00
             group C:319    some college      :226       Mode  :character   Mode  :character   Median : 66.00
             group D:262    associate's degree:222                                           Mean   : 66.09
             group E:140    bachelor's degree :118                                           3rd Qu.: 77.00
                            master's degree   : 59                                           Max.   :100.00
```

Dari sini dapat kita lihat untuk dan ringkasan datanya sudah memberikan banyak informasi di bandingkan yang sebelumnya.

e. Data Pencilan

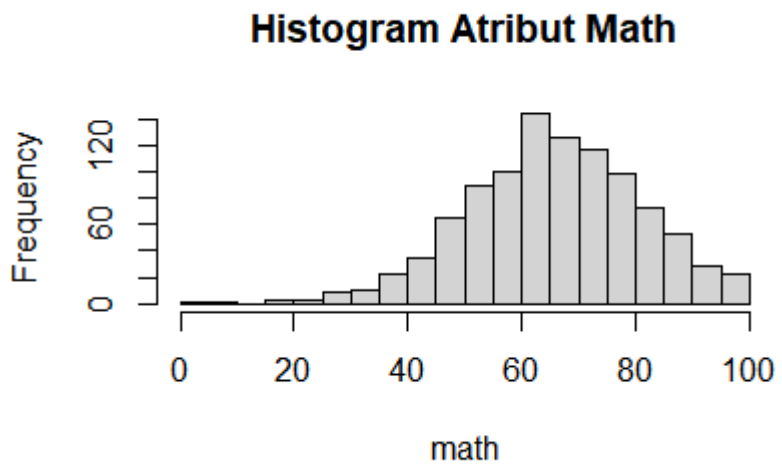
Pencilan adalah nilai atau pengamatan yang jauh dari pengamatan lain, yaitu titik data yang berbeda secara signifikan dari titik data lainnya. Saya menyajikan beberapa pendekatan untuk mendeteksi outlier di R, dari teknik sederhana seperti statistik deskriptif (termasuk minimum, maksimum, histogram, boxplot, dan persentil) hingga teknik yang lebih formal seperti filter Hampel, Grubbs, Dixon, dan tes Rosner untuk outlier .

Hal pertama yang kita lakukan adalah melihat statistik deskriptifnya

```
> summary(dataku$math)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   57.00   66.00   66.09   77.00  100.00
```

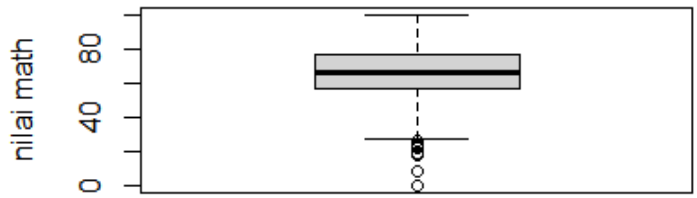
Selanjut kita dapat melakukan visualisasi terhadap data yang kita gunakan

```
hist(dataku$math,
     xlab = "math",
     main = "Histogram Atribut Math",
     breaks = sqrt(nrow(dataku)))
```



Kita juga dapat melihatnya dengan menggunakan boxplot

```
# membuat boxplot untuk mendeteksi outlier juga
boxplot(dataku$math,
        ylab = "nilai math")
```



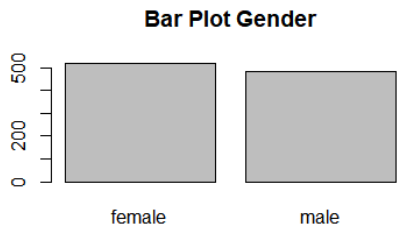

Jika kita perhatikan dengan jeli ada ada bebera nilai yang outlier, untuk lebih jelasnya kita dapat melihat nilai apa saya yang outlier dengan fungsi sebagai berikut.

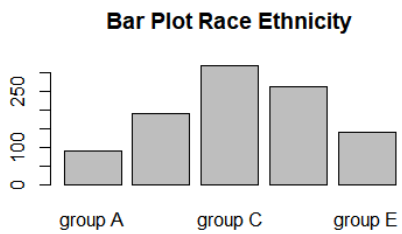
```
> boxplot.stats(dataku$math)$out
[1] 18  0 22 24 26 19 23  8
```

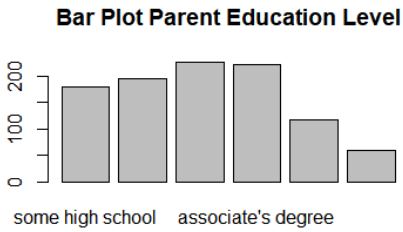
Dari hasil yang di tampilkan adalah merupakan nilai yang dianggap outlier.

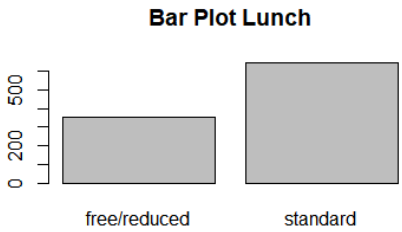
f. Visualisasi Data

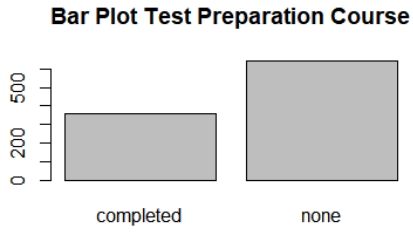
```
# Membuat visualisasi
# par(mfrow=c(3,2))
barplot(table(dataku$gender), main = "Bar Plot Gender")
barplot(table(dataku$race.ethnicity), main = "Bar Plot Race Ethnicity")
barplot(table(dataku$parent_education_level), main = "Bar Plot Parent Education Level")
barplot(table(dataku$lunch), main = "Bar Plot Lunch")
barplot(table(dataku$test_prep_course), main = "Bar Plot Test Preparation Course")
```






**2. Lakukan pengkodean terhadap Gender , "parent education level" and “lunch”**

Sebelum melakukan pengkodean kita memeriksa level faktornya terlebih dahulu agar memudahkan kita menentukan kode yang akan kita gunakan.

```
> sapply(dataku, levels)
$gender
[1] "female" "male"

$race.ethnicity
[1] "group A" "group B" "group C" "group D" "group E"

$parent_education_level
[1] "some high school"   "high school"        "some college"       "associate's degree" "bachelor's degree"
[6] "master's degree"

$lunch
[1] "free/reduced" "standard"

$test_prep_course
[1] "completed" "none"
```

Sebelumnya kita juga harus mengistall *library (tidyverse)*

```
# membuat recode gender menjadi numeric
dataku$gender_code <- recode(dataku$gender,
                      "female" = 0,
                      "male" = 1)
# membuat recode kolom Parent_education_level menjadi dummy/numeric
dataku$parent_education_level_code <- recode(dataku$parent_education_level,
                                    "some high school" = 0,
                                    "high school" = 1,
                                    "some college" = 2,
                                    "associate's degree" = 3,
                                    "bachelor's degree" = 4,
                                    "master's degree" = 5)

dataku$lunch_code <-  recode(dataku$lunch,
                      "free/reduced" = 1,
                      "standard"     = 2)
head(dataku)
```

Data sebelum di lakukan penkodean

| | gender | race.ethnicity | parent_education_level | lunch | test_prep_course | math |
|---|---|---|---|---|---|---|
| 1 | female | group B | bachelor's degree | standard | none | 72 |
| 2 | female | group C | some college | standard | completed | 69 |
| 3 | female | group B | master's degree | standard | none | 90 |
| 4 | male | group A | associate's degree | free/reduced | none | 47 |
| 5 | male | group C | some college | standard | none | 76 |

Setelah di lakukan pengkodean

| | gender | race.ethnicity | parent_education_level | lunch | test_prep_course | math | gender_code | parent_education_level_code | lunch_code |
|---|---|---|---|---|---|---|---|---|---|
| 1 | female | group B | bachelor's degree | standard | none | 72 | 0 | 4 | 2 |
| 2 | female | group C | some college | standard | completed | 69 | 0 | 2 | 2 |
| 3 | female | group B | master's degree | standard | none | 90 | 0 | 5 | 2 |
| 4 | male | group A | associate's degree | free/reduced | none | 47 | 1 | 3 | 1 |
| 5 | male | group C | some college | standard | none | 76 | 1 | 2 | 2 |

Dari sini kita dapat lihat bahwa telah muncul kolom baru yang mendefenisikan kode nilai dari kolom sebelumnya dari string jadi berupa numeric.

3. **Kemudian Urutkan Data berdasarkan "Race"**

```
# mengurutkan data berdasarkan kolom "race"
dataku <- dataku[order(dataku$race.ethnicity) , ]
head(dataku)
```

Untuk mengurutkan data, kita dapat menggunakan fungsi order dan memilih berdasarkan kolom yang kita gunakan.

Data sebelum di urutkan

| | gender | race.ethnicity | parent_education_level | lunch | test_prep_course | math |
|---|---|---|---|---|---|---|
| 1 | female | group B | bachelor's degree | standard | none | 72 |
| 2 | female | group C | some college | standard | completed | 69 |
| 3 | female | group B | master's degree | standard | none | 90 |
| 4 | male | group A | associate's degree | free/reduced | none | 47 |
| 5 | male | group C | some college | standard | none | 76 |

Setelah di urutkan

| | gender | race.ethnicity | parent_education_level | lunch | test_prep_course | math |
|---|---|---|---|---|---|---|
| 4 | male | group A | associate's degree | free/reduced | none | 47 |
| 14 | male | group A | some college | standard | completed | 78 |
| 15 | female | group A | master's degree | standard | none | 50 |
| 26 | male | group A | master's degree | free/reduced | none | 73 |
| 47 | female | group A | associate's degree | standard | completed | 55 |

4. **Ubah nilai math menjadi Nilai Huruf dan Tambahkan kolom Status Penilaian sesuai dengan kriteria disamping ini.**
   Untuk konvensi nilai disini saya menggunakan fungsi cut dan kita dapat memilih kolom mana yang akan kita buat nilainya, dengan memasukkan batasan dan labelnya, seperti dibawah ini.

```
# konversi nilai ke nilai huruf
dataku$nilai_huruf = cut(dataku$math,
                         breaks = c(0, 54, 59,64, 69, 74, 79, Inf),
                         labels = c("E", "D", "C", "BC", "B", "AB", "A"),
                         right = TRUE,
                         include.lowest = TRUE)
```

   Untuk statusnya disini kita menggunakan fungsi recode dengan membuat nilai yang mana saja akan kita ubah atau kita buat statusnya.

```
#Tambahkan kolom Status Penilaian
dataku$status_penilaian <- recode(dataku$nilai_huruf,
                                  "A" = "Lulus",
                                  "AB" = "Lulus",
                                  "B" = "Lulus",
                                  "BC" = "Lulus",
                                  "C" = "Lulus",
                                  "D" = "Tidak Lulus",
                                  "E" = "Tidak Lulus")
```

   Untuk hasilnya dapat kita lihat sebagai berikut

| nilai_huruf | status_penilaian |
|---|---|
| E | Tidak Lulus |
| AB | Lulus |
| E | Tidak Lulus |
| B | Lulus |
| D | Tidak Lulus |
| E | Tidak Lulus |

5. **Tambahkan 5 data baru untuk melakukan reshaping**
   **- reshaping dengan Long to wide data format**
   **- reshaping dengan Wide to long data format**

   Sebelum di lakukan reshaping kita perlu menginstal library *(tidyr)*

   a. reshaping dengan Long to wide data format

```
# cek data
df<- read.csv("D:/Sains Data Semester 5/Komputasi Statistik/3. Tugas Pertemuan 3.csv", header = TRUE)
#pivot the data frame into a long format
dataku1<- df %>% pivot_longer(cols=c('gender', 'race.ethnicity'),
                              names_to='columns',
                              values_to='values')
head(dataku1)

# Reshaping dengan Long to Wide data format
#pivot the data frame into a wide format
```

pada reshaping ini, saya kembali mengimport datanya, dengan tujuan agar data yang awal tidak terganggu, dan untuk hasilnya sebagai berikut.

```
# A tibble: 6 x 6
  parent_education_level lunch    test_prep_course  math columns        values
  <chr>                  <chr>    <chr>            <int> <chr>          <chr>
1 bachelor's degree      standard none                72 gender         female
2 bachelor's degree      standard none                72 race.ethnicity group B
3 some college           standard completed           69 gender         female
4 some college           standard completed           69 race.ethnicity group C
5 master's degree        standard none                90 gender         female
6 master's degree        standard none                90 race.ethnicity group B
```

b. - reshaping dengan Wide to long data format

```
#pivot the data frame into a wide format
dataku1 %>% pivot_wider(names_from = columns, values_from = values)

head(dataku1)
```

```
# i Use `print(n = ...)` to see more rows
Warning message:
Values from `values` are not uniquely identified; output will contain list-cols.
* Use `values_fn = list` to suppress this warning.
* Use `values_fn = {summary_fun}` to summarise duplicates.
* Use the following dplyr code to identify duplicates.
  {data} %>%
    dplyr::group_by(parent_education_level, lunch, test_prep_course, math, columns) %>%
    dplyr::summarise(n = dplyr::n(), .groups = "drop") %>%
    dplyr::filter(n > 1L)
> head(dataku1)
# A tibble: 6 x 6
  parent_education_level lunch    test_prep_course  math columns        values
  <chr>                  <chr>    <chr>            <int> <chr>          <chr>
1 bachelor's degree      standard none                72 gender         female
2 bachelor's degree      standard none                72 race.ethnicity group B
3 some college           standard completed           69 gender         female
4 some college           standard completed           69 race.ethnicity group C
5 master's degree        standard none                90 gender         female
6 master's degree        standard none                90 race.ethnicity group B
>
```

6. **Menurut Anda, Apakah dengan mengambil Kursus persiapan ujian Dapat menentukan** seseorang lulus Atau tidak pada kasus disamping.
   a. Melihat perbandingan mahasiswa mengikuti kursus dan tidak mengikuti kursus

```
# perbandingan mahasiswa mengikuti kursus dan tidak mengikuti kursus
tidak_kursus= subset(dataku, dataku$test_prep_course == "none")
table(tidak_kursus$status_penilaian)

sum(is.na(tidak_kursus$status_penilaian))
```

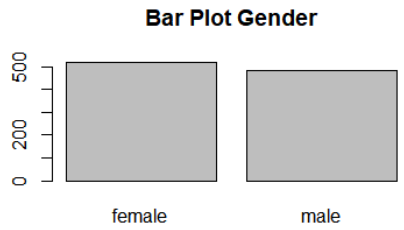
Untuk melihat perbandingan nilainya, disini saya menggunkan fungsi subset dengan Atribut " test_prep_course" yang bernilai "none", dan hasilnya sebagai berikut.

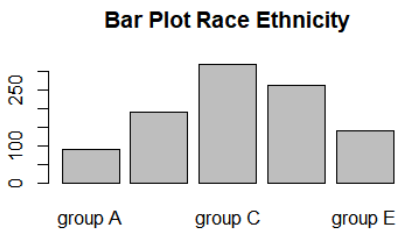
```
> table(tidak_kursus$status_penilaian)

Tidak Lulus       Lulus
        236         406
> sum(is.na(tidak_kursus$status_penilaian))
[1] 0
```

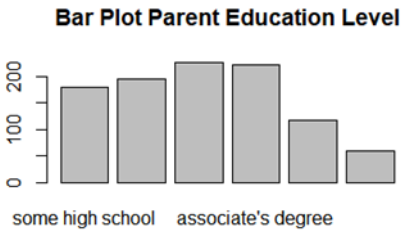
   b. Mahiswa ikut kursus

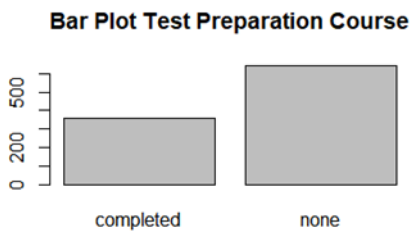
```
# mahasiswa ikut kursus
data_kursus = subset(dataku, dataku$test_prep_course=="completed")
table(data_kursus$status_penilaian)

sum(is.na(data_kursus$status_penilaian))
```

```
> table(data_kursus$status_penilaian)

Tidak Lulus       Lulus
         87         271
> sum(is.na(data_kursus$status_penilaian))
[1] 0
```



Kesimpulan:

Jika kita llihat di grafik perbandingan jumlah laki-laki dan perempuan tidak terlihat ada perbendaan yang sangat singnifikan, dimana hal ini menunjukkan bahwa gender tidak mempengaruhi kelulusan masiswa, begitu juga dengan education level dan enthnicity
Jika kita lihat plot berdistribusi normal.
Jika kita hitung persentasenya, mahasiswa yang tidak mengikuti kursus lulus ujian dengan persentase 63.2%. Sedangkan, mahasiswa yang mengikuti kursus lulus ujian dengan persentase 75.69%. Perbedaan yang tidak terlalu signifikan. Hanya memiliki selisih 12, 49 % hal ini menunjukkan kursus tidak terlalu mempengaruhi kelulusan dan tidak di pengaruhi oleh gender, jadi hasil analisis saya tidak perlu melakukan kursus dan semua gender bisa mencobanya.