# Datalaboration 1 - 732G50

VT2021

## Introduktion

Denna datorövning behandlar visualisering av olika typer av variabler.

Denna laboration kommer att göras i R-studio.

Öppna ett nytt R-script **File** → **New file** → **R-script**.

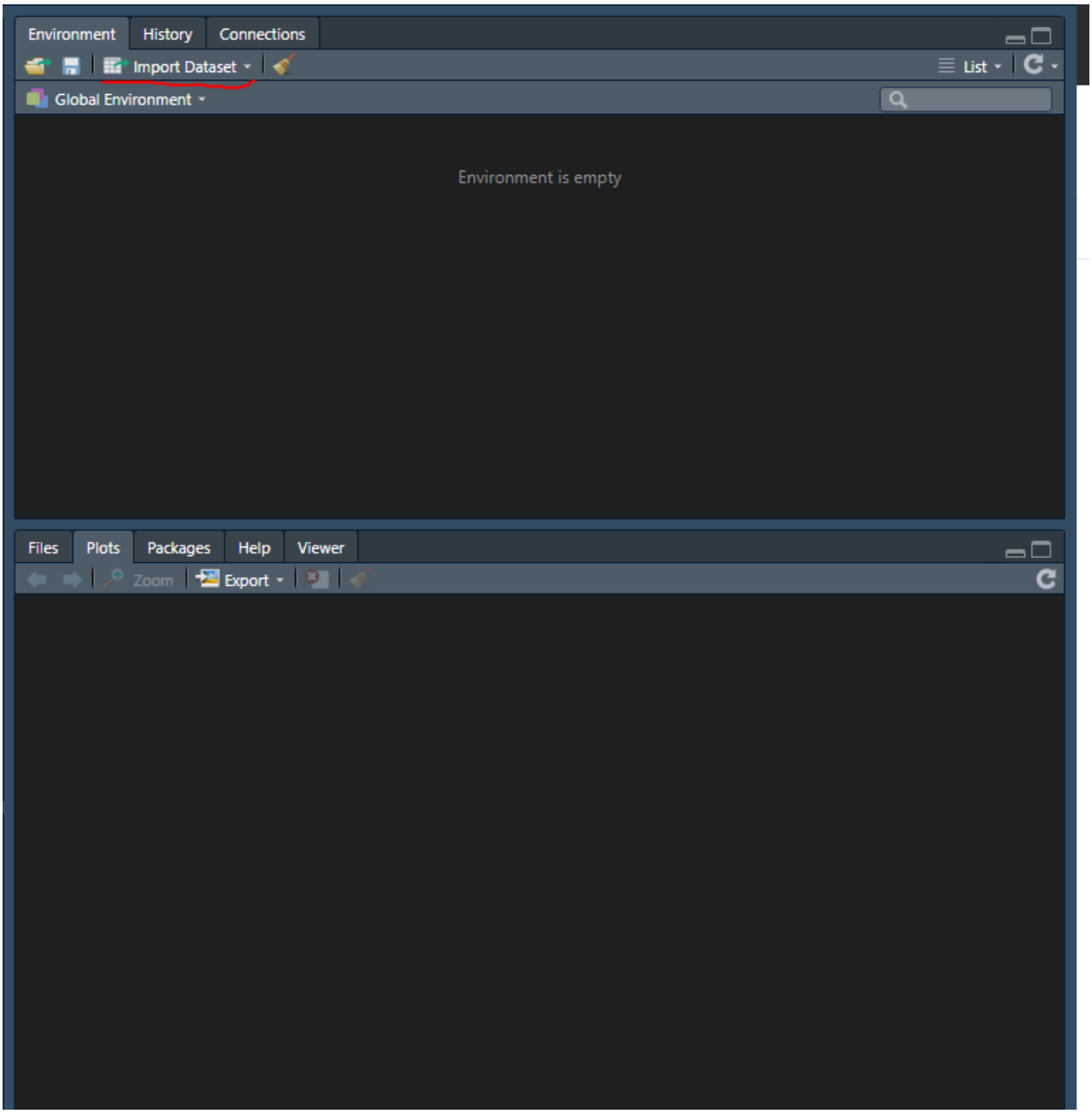
För att köra koden markerar du den koden du vill köra och trycker på **Run** uppe i högra hörnet. Du kan även trycka **Ctrl+Enter** för att köra den markerade koden.

### Ladda ned R-studio

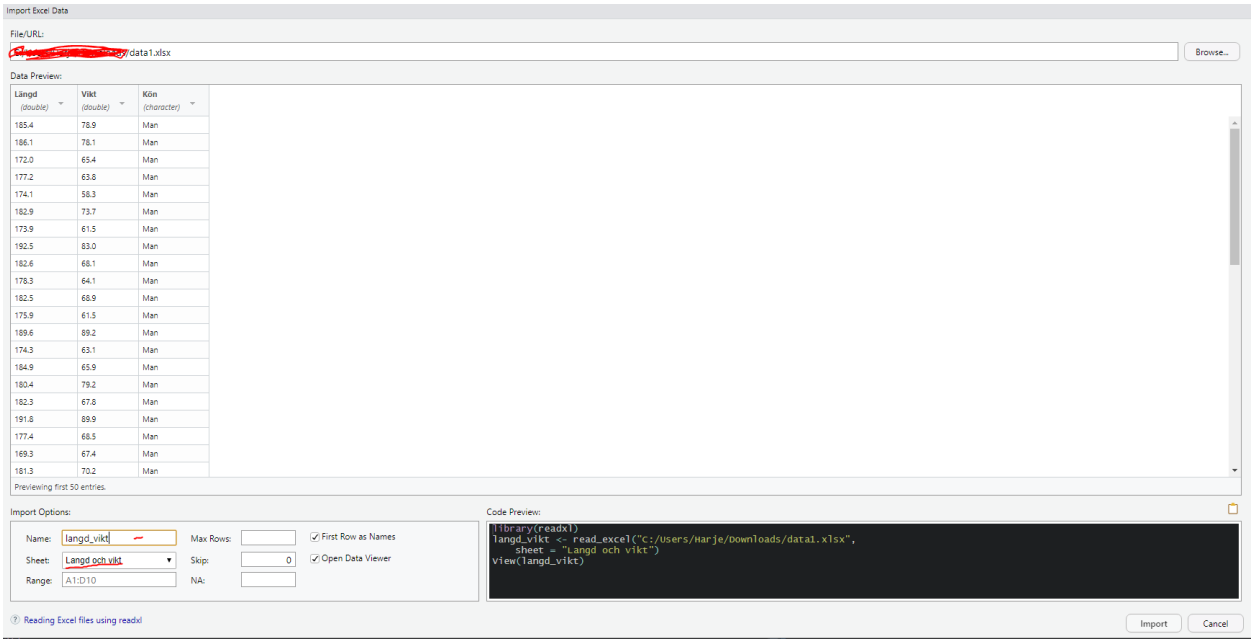
- **R-studio**

# Uppgift 1

Börja med att ladda ner det datamaterial som finns på kursens LISAM-sida. Datamaterialet innehåller tre flikar med information. Den första fliken behandlar data om längd och vikt hos 100 män och 100 kvinnor. Den andra fliken behandlar information om civilstånd och bilmärken på Påhittade gatan. Den tredje fliken behandlar information om tentaresultat och ålder hos ett antal påhittade studenter.




För att importera datamaterialet till R-studio tryck på **Import Dataset** → **From Excel...** välj sedan sheet **Langd och vikt**. Välj även namnet langd_vikt i Name:




Du kommer sedan att se att du infört datamaterialet till höger i Global Environment.

## 1

Identifiera $\overline{x}$, s samt kvartilerna för längd samt vikt.

Dollar tecknet $ skrivs efter din data.frame så väljer du sedan vilken variabel du är intresserad av att skriva ut, exempelvis Längd som nedan.

**Exemplet som visas har jag namnet på data.frame till langd_vikt**

```
langd_vikt$Längd
head(langd_vikt) # Skriver ut de första 5 observationerna.
```

Ni kommer anända följande funktioner som är inbyggda i R:

- `summary()` skriver ut kvartilerna min och max.
- `sd()` skriver ut standardavvikelsen.
- `mean()` skriver ut medelvärdet.
- `median()` skriver ut medianen.

```
summary(langd_vikt$Längd) #  summary() kvartilerna
sd(langd_vikt$Längd)#        sd() standardavvikelsen
mean(langd_vikt$Längd) #     mean()  medelvärdet
median(langd_vikt$Längd) #   median()  medianen
```

Genom att byta ut **Längd** till **Vikt**

Grupera datamaterialet efter kön.

```
data_kvinna <- langd_vikt[which(langd_vikt$Kön == "Kvinna"),]
data_man <- langd_vikt[which(langd_vikt$Kön == "Man"),]
```

```
summary(data_kvinna$Längd) #  summary() kvartilerna
sd(data_kvinna$Längd)#        sd() standardavvikelsen
mean(data_kvinna$Längd) #     mean()  medelvärdet
median(data_kvinna$Längd) #   median()  medianen
```

Gör nu samma för männen genom att skriva `data_man$Längd` istället för `data_kvinna$Längd`. Är det någon skilland mellan man och kvinna?

## 2

Nästa steg är att undersöka fördelningen av Längd och Vikt med histogram. Genom att använda funktionen `hist()` som är inbyggd i R.

`main = "text"` bestämmer titlen för ploten är.

`ylab = "text"` bestämmer vad som ska stå på y-axel

`xlab = "text"` bestämmer vad som ska stå på x-axel

`col = "red", col = 2, col= c(1:8)` sätter olika färger på ploten.

```
hist(langd_vikt$Längd, main= "Histogram över längd båda könen")
```

```r
hist(data_kvinna$Längd, main= "Histogram över längd för Män")
```

```r
hist(data_man$Längd, main= "Histogram över längd för Kvinna")
```

Vad kan vi säga om fördelningen av de olika variablerna? Skillnader mellan könen?

# Uppgift 2

Vi ska nu gå till bladet som heter Bilmärke och fortsätta med uppgifterna. Importera Bilarken från samma datamaterial från tidigare uppgift genom att välja sheet **Bilmarke**.

**Exemplet som visas har jag namnet på data.frame till Bilmarke**

## 1

Skapa en frekvenstabell över variablerna Bil och Civilstånd med hjälp av funtionen `table()`.

```
table(Bilmarke$Bil) # table() skriver ut en frekvenstabell av vald variabel
```

```
table(Bilmarke$Civilstånd, Bilmarke$Bil) # Du kan även skriva in fler variabler in i funktionen
table( Bilmarke$Bil, Bilmarke$Civilstånd) # Du kan välja den ordningen som passar dig bäst.
```

Vilket bilmärke är den vanligaste? Hur stor andel av bilarna är fordar? Hur många har civilstånd par?

## 2.

Att presentera tabeller är inte alltid det enklaste sättet för läsaren att få en bild av data. Det är oftast enklare (och snyggare) att skapa ett diagram.

Här kommer vi måsta kombinera två olika functioner `table()` samt `barplot()` testa gärna att expremententera med olika färger

```
barplot(table(Bilmarke$Bil))

barplot(table(Bilmarke$Bil), col=1:8)# col="red", col=8:1
```

```
barplot(table(Bilmarke$Civilstånd), col=1:2) # Två grupper blir det två olika fäger.
```

```
Bil_par <- Bilmarke[which(Bilmarke$Civilstånd == "Par"),]
Bil_singel <- Bilmarke[which(Bilmarke$Civilstånd != "Par"),]


barplot(table(Bil_par$Bil),main="Bilmärken för par")
barplot(table(Bil_singel$Bil),main="Bilmärken för singlar")
```

Om ni vill ha dem i samma bild kan ni sätta genomskinlighet på detta sätt.

```
blue <- rgb(0, 0, 1, alpha=0.2)
red <- rgb(1, 0, 0, alpha=0.2)
barplot(table(bilmarke$Civilstånd,bilmarke$Bil),main="Bilmärken för alla kön", col=c(blue,red))
legend("topleft", c("Par", "Singel"), col=c(blue,red), pch=19)
```

# Uppgift 3

Läs nu in det sista bladet som heter Elevålder som ni gjort tidigare med att välja sheet **Elevalder**.

**Exemplet som visas har jag namnet på data.frame till Elevalder**

## 1

Börja med att visualisera fördelningen av dessa två variabler. Här kan vi antingen välja ett histogram eller stolpdiagram på Ålder eftersom man kan tänka sig att värdena kan anta decimalvärden.

Svara på följande frågor

(a) Hur stor andel fick full pott, alltså 20 poäng, på tentan?

(b) Vilka åldrar är mest frekventa?

```
barplot(table(Elevalder$Ålder),xlab = "Ålder",main="Stolpdiagram över åldrar", ylab="Antal")
```

```
barplot(table(Elevalder$Resultat),xlab = "Resultat",main="Stolpdiagram över resultat", ylab="Antal")
```

### a)

Detta kan svaras på att summera hur många som fick resultatet 20.

`sum()` summerar en vektor. Om du skriver `Elevalder$Resultat == 20` så kommer det upp två olika värden `TRUE` och `FALSE` detta värde kallas `Boolean`, sant eller falskt. Med funktionen `sum()` kan man summera alla sanna värden.

```
sum(Elevalder$Resultat == 20)
```

### b)

Du kan antingen skriva ut en tabell som vi gjorde i 2.1 eller kolla stolpdiagrammet.

## 2

R har inga grundfunktioner för dotplot så i stället testar vi att skapa boxplot och försök att dra slutsatser för ålder och resultat. Har vi några `outliers` dvs några observationer som sticker ut?

```
boxplot(Elevalder$Ålder, col="lightblue", main="Boxplot över ålder",
        ylab="Ålder")
boxplot(Elevalder$Resultat, col="lightblue", main="Boxplot över resultat",
        ylab="Resultat")
```

## 3

Ibland kan man vilja dela upp en kvantitativ variabel, t.ex. ålder, i olika klassindelningar. Detta kallas med ett annat ord diskretisering.

Vi ska nu gruppera åldrarna till 6 olika åldersgrupper.

`0-19`, `20-24`, `25-29`, `30-34`, `35-39`, `40-`

I R-studio använder vi följande kommandon för att göra detta.

- $A == B$ betyder $A = B$, A är lika med B.
- $A < B$ betyder $A < B$, A är mindre än B.
- $A <= B$ betyder $A \leq B$, A är mindre eller lika med B.
- $A >= B$ betyder $A \geq B$, A är större eller lika med B.
- $A! = B$ betyder $A \neq B$, A är inte lika med B.

I detta exempel räcer det med att använda <= och >= tillsammans med funktionen `which()`.

Börja med att skapa en tom variabel `NA` står för Not Applicable eller Not Available, kan även tolkas som Null.

```
Elevalder$Alderklass <- NA
head(Elevalder) # Skriver automatiskt ut de första 5 observationerna i datamaterialet.
```

```
Elevalder$Alderklass[which(Elevalder$Ålder <= 19)] <- "0-19"
# Alla under 20 kommer ha gruppnamnet 0-19
```

```
Elevalder$Alderklass[which(Elevalder$Ålder >= 20 & Elevalder$Ålder <= 24)] <- "20-24"
# Alla med åldern mellan 20 och 24 kommer få gruppnamnet 20-24
Elevalder$Alderklass[which(Elevalder$Ålder >= 25 & Elevalder$Ålder <= 29)] <- "25-29"
# Alla med åldern mellan 2 och 29 kommer få gruppnamnet 20-24
Elevalder$Alderklass[which(Elevalder$Ålder >= 30 & Elevalder$Ålder <= 34)] <- "30-34"
# Alla med åldern mellan 30 och 34 kommer få gruppnamnet 30-34

Elevalder$Alderklass[which(Elevalder$Ålder >= 35 & Elevalder$Ålder <= 39)] <- "35-39"
# Alla med åldern mellan 35 och 39 kommer få gruppnamnet 35-39

Elevalder$Alderklass[which(Elevalder$Ålder >= 40)] <- "40-"
# Alla med åldern 40 och uppåt kommer få gruppnamnet 40-
```

```
barplot(table(Elevalder$Alderklass),main="Stolpdiagram ålderklasser",xlab="Ålder",ylab="Antal")
```

Saknas det en grupp i stolpdiagrammet och i så fall varför?