

# Obligatorisk læringsaktivitet i introuger

## Flow 1, Dataanalyse

Afleveringsfrist:

fredag den 13. september 2024, kl. 16

Udarbejdet af:

Thorbjørn Baum, adjunkt Cphbusiness

Thorbjørn Wulf, adjunkt Cphbusiness


## Indholdsfortegnelse

<b>Opgave 1 – Data Science Modellen .....</b>	<b>3</b>
<i>Opgave 1.1 – find data .....</i>	<i>3</i>
<i>Opgave 1.2 - Vælg .....</i>	<i>3</i>
<i>Opgave 1.3 – Beskriv variabler .....</i>	<i>3</i>
<i>Opgave 1.4 – Research goal .....</i>	<i>3</i>
<b>Opgave 2 – Korrelation og simple lineær regression .....</b>	<b>4</b>
<i>Opgave 2.1 – Beskrivende statistik .....</i>	<i>4</i>
<i>Opgave 2.2 - Korrelation .....</i>	<i>4</i>
<i>Opgave 2.3 – Simple regressioner .....</i>	<i>4</i>
<i>Opgave 2.4 – Korrelation og lineær regression .....</i>	<i>4</i>
<b>Opgave 3 – Tilfældigheder og terninger .....</b>	<b>5</b>
<i>Opgave 3.1 – Funktion til terninger .....</i>	<i>5</i>
<i>Opgave 3.2 – Plot I .....</i>	<i>5</i>
<i>Opgave 3.3 – Plot II .....</i>	<i>5</i>
<i>Opgave 3.4 – Lav dine egne data .....</i>	<i>5</i>
<b>Opgave 4 – Danskernes forhold til alkohol .....</b>	<b>6</b>
<i>Opgave 4.1 – Hent data .....</i>	<i>6</i>
<i>Opgave 4.2 – Korrelation .....</i>	<i>6</i>
<i>Opgave 4.3 – Kritisk tænkning .....</i>	<i>6</i>
<b>Opgave 5 – Dataframes .....</b>	<b>7</b>
<i>Opgave 5.1 – Månedlige observationer .....</i>	<i>7</i>
<i>Opgave 5.2 – Kvartalsvise observationer .....</i>	<i>7</i>
<i>Opgave 5.3 - Pivot .....</i>	<i>7</i>

## Opgave 1 – Data Science Modellen

### Opgave 1.1 – find data

Data på ejendomme til salg i opgave 1 er fremkommet via webscapping af boligsiden.dk. Der ligger en csv-fil i mappen med OLA-opgaven. Her er to billeder fra boligsiden.dk. Find de to rækker i csv-filen, som matcher de to huse.




**Tousvej 106**  
8230 Åbyhøj

Villa **D**

**5.998.000 kr.** [Beregn boliglån](#)  
M<sup>2</sup>-pris 41.595 kr.

🏠 120 m <sup>2</sup>	📅 Ejerudg. 3.219 kr./md	📏 Grund 883 m <sup>2</sup>
🚗 5 Vær.	📅 4 dage	🏠 Opført 1938

[Se bolig](#)[danbolig](#)[Se hos mægler](#)



**Egevej 20**  
Uhrhøj, 7100 Vejle

Villa **A20**

**4.795.000 kr.** [Beregn boliglån](#)  
M<sup>2</sup>-pris 24.922 kr.

🏠 170 m <sup>2</sup>	📅 Ejerudg. 3.692 kr./md	📏 Grund 737 m <sup>2</sup>
🚗 6 Vær.	📅 19 dage	🏠 Opført 2021

[Se bolig](#)[Nybolig](#)[Se hos mægler](#)

### Opgave 1.2 - Vælg

Udvælg 2 ejendomme fra csv-filen og find dem på boligsiden.dk.

### Opgave 1.3 – Beskriv variabler

Forklar NA-værdierne i csv-filen ud fra, hvad I har observeret i opgave 1.1 og 1.2. Derudover gør rede for variable I mener mangler i csv-filen sammenlignet med boligsiden.dk. (Hint: Hvordan vil I unikt identificere en bolig til salg via boligsiden?)

### Opgave 1.4 – Research goal

Med udgangspunkt i Data Science Modellen skal I gøre rede for de skridt, der er blevet taget for at nå frem til csv-filen. I skal komme med et bud på et “research goal” som kunne have optimeret processen med at fremskaffe data fra boligsiden.

## Opgave 2 – Korrelation og simple lineær regression

### Opgave 2.1 – Beskrivende statistik

Lav beskrivende statistik for data på boliger til salg via boligsiden.dk fra opgave 1. Vær opmærksom på, hvilke variable I mener kan betragtes som x'er og y i en række af simpel lineære regressioner.

### Opgave 2.2 - Korrelation

Hvad er korrelationen mellem  $m^2$  og prisen for boliger lagt på boligsiden.dk? Giv en forklaring på begrebet korrelation.

### Opgave 2.3 – Simple regressioner

Lav minimum 5 simple regressioner mellem pris pr.  $m^2$  og 5 andre variable i csv-filen fra opgave 1. Giv en forklaring på, hvilken af de fem modeller, der bedst forklarer pris pr.  $m^2$ . Derudover skal I understøtte jeres 5 modeller med et bud på om, der er sammenhæng (korrelation) mellem jeres 5 udvalgte variable (hint: lav en korrelationsmatrix).

### Opgave 2.4 – Korrelation og lineær regression

Forklar den teoretiske sammenhæng mellem korrelation og simple lineær regression.

## Opgave 3 – Tilfældigheder og terninger

### Opgave 3.1 – Funktion til terninger

Lav et script (en funktion) i R-studio, der kan slå med 25.000 terninger. Hvor mange 5'ere har jeres terning slået? Hvad er sandsynligheden, givet jeres resultatet med de 25.000 terninger, for, at jeres script slår en 5'er?

### Opgave 3.2 – Plot I

Lav et script (en funktion) i R-studio, der kan slå med 6 terninger og vise summen. Slå nu 10.000 gange med de 6 terninger og lav et barplot af jeres resultat. (hint: barplot kræver fx pakken "ggplot2"). Forklar om jeres resultat giver mening i forhold til jeres funktion (hint: I bruger sample funktionen).

### Opgave 3.3 – Plot II

Brug jeres script fra 3.2 og slå nu 1.000.000 gange med de 6 terninger. Lav igen et barplot og sammenlign med jeres plot fra 3.2.

### Opgave 3.4 – Lav dine egne data

Lav et script i R-studio, der viser en tilfældigt opstillet række af tallene 1, 2, 3, 5, 6. Lav en matrix med to kolonner og fem rækker, hvor den første kolonne skal være tallene 2 til 6 og den anden kolonne skal være jeres tilfældige række af tallene 1, 2, 3, 5, 6. (hint: Google funktionen cbind(), der sætter lige lange kolonner sammen i en matrix).

## Opgave 4 – Danskernes forhold til alkohol

### Opgave 4.1 – Hent data

Hent data fra tabel FU02, alle forbrugsgrupper under 02.1 (alkoholiske drikkevarer) i faste priser for perioden 2000 til 2022 og indlæs i R. Illustrer udviklingen i de enkelte grupper.

### Opgave 4.2 – Korrelation

Lav en korrelationsmatrix over forbrugsgrupperne under 02.1 og konkluder på resultaterne

### Opgave 4.3 – Kritisk tænkning

Forhold jer kritisk til de indlæste data når I reflekter over jeres resultater.

## Opgave 5 – Dataframes

### Opgave 5.1 – Månedlige observationer

Lav en 36 x 3 dataframe med kolonnenavne "Klasse", "Uge", "Score". Første kolonne skal fyldes med A,B,C,D så der startes med 9 A'er, derpå 9 B'er osv. Anden kolonne skal fyldes med tallene 1 til 9, der gentages for hvert bogstav. Sidste kolonne skal fyldes med observationer. Det er op til jer, hvilke værdier I vil putte i framen. (hint: benyt R-funktionen *seq()*)

### Opgave 5.2 – Kvartalsvise observationer

I skal lave en ny dataframe, der er 9x3 og bygger på den dataframe I lavede opgave 5.1. I skal tage udgangspunkt i framen fra opgave 5.1 og loope igennem. I loopet skal I hver tredje gang lave en dataframe 1x3 dataframe med samme navne som i 5.1. (Hint: brug modulo-operatoren til at ramme hver tredje). Indholdet skal være som følger: 1 element og 2 element henter I fra 5.1-framen. Det sidste element skal være gennemsnittet af de forrige tre observationer.

Fx I går fra denne (vær opmærksom på det er et eksempel på kvartal):

INDIKATOR	TID	value	Over til:	INDIKATOR	TID	value
1	A	2000-01-01		1	A	2000-04-01
2	A	2000-02-01		2	A	2000-08-01
3	A	2000-03-01		3	A	2000-12-01
4	A	2000-04-01		4	A	2001-04-01
5	A	2000-05-01		5	A	2001-08-01
6	A	2000-06-01		6	A	2001-12-01
7	A	2000-07-01		7	B	2000-04-01
8	A	2000-08-01		8	B	2000-08-01
				9	B	2000-12-01
24	A	2001-12-01		10	B	2001-04-01
25	B	2000-01-01		11	B	2001-08-01
26	B	2000-02-01		12	B	2001-12-01
27	B	2000-03-01		13	C	2000-04-01
28	B	2000-04-01		14	C	2000-08-01
				15	C	2000-12-01
				16	C	2001-04-01
				17	C	2001-08-01
				18	C	2001-12-01

### Opgave 5.3 - Pivot

I skal nu konvertere denne nye dataframe til en ny dataframe som har følgende navne på kolonnerne:

"Uge", "A", "B", "C", "D" og rækkerne indeholder de gennemsnit som I har beregnet. (Hint: Brug funktionen *pivot-wider* fra pakken *tidyr*).

TID	A	B	C
1	2000-04-01	4.861433	4.687455
2	2000-08-01	5.415208	5.366524
3	2000-12-01	4.530170	4.474418
4	2001-04-01	4.738251	4.240166
5	2001-08-01	4.790563	4.178769
6	2001-12-01	3.779561	4.485748