Anvendt Statistik for Erhvervsøkonomer

Louise Hviid

LEKTION 2: DATATYPER OG VISUALISERING

AALBORG UNIVERSITY

HEJ.

IGEN.

AALBORG UNIVERSITY

HVORFOR SIGER DIN STATISTIKLÆRER AT DU SKAL LÆRE STATISTIK?

Det er fordi, hun tror at du (som de fleste mennesker) har et nysgerrigt sind, der ønsker at besvare nye og spændende spørgsmål!

Hvorfor opfører folk sig, som de gør (psykologi)? Hvordan maksimerer virksomheder deres profit (erhvervsøkonomi)? Beskytter spisning af sild dig mod at få kræft (medicin eller biologi)?

Uanset hvad du studerer eller forsker i, er årsagen til, at du studerer det, nok fordi du er interesseret i at besvare spørgsmål.

For at besvare spørgsmål har vi brug for statistik. Mere præcist, for at besvare interessante spørgsmål har du brug for 2 ting: data og en forklaring på disse data.

NU HAR VI EN IDE OM HVORFOR DU ER HER I STATISTIK TIMEN?

Så vi er her for at besvare interessante spørgsmål, for hvilke vi har brug for data. I dag vil vi kigge på forskellige typer af data.

Men hvordan går vi egentlig til opgaven at besvare et interessant spørgsmål?

OBSERVATIONER

MÅSKE BEGYNDER VI MED EN OBSERVATION?

En observation kunne være anekdotisk (folk, der kører Tesla drikker meget kaffe) eller den kunne være baseret på data (du registrerer, hvor mange Tesla'er i forhold til andre biler er parkeret uden for Starbucks).

Men hvad gør vi så med disse observationer?

FORUDSIGELSER

MÅSKE BEGYNDER VI AT GENERERE FORKLARINGER OG TEORIER?

Herfra kunne vi generere forklaringer eller teorier på disse observationer, hvorfra vi laver forudsigelser (hypoteser).

Men hvad gør vi så med disse forudsigelser?

TESTNING

MÅSKE KAN VI BEGYNDE AT TESTE DEM?

For at teste vores forudsigelser har vi brug for data. Først indsamler vi nogle relevante data (og for at gøre dette skal vi identificere ting, der rent faktisk kan måles), og derefter visualiserer og analyserer vi disse data.

Men hvorfor skal vi visualisere og analysere vores data?

ANALYSE

Analyse af data kan enten støtte vores teori eller give os grund til at modificere den. Så teorier fører til indsamling/analyse af data, og indsamling/analyse af data informerer også teorier.

Ok, men kan vi gennemgå et eksempel på det her?

EKSEMPEL

OBSERVATION:

Jeg gør mig en tilfældig observation om verdenen (Tesla-kørere drikker meget kaffe), hvorefter jeg indsamler nogle data for at se, om denne observation er sand (og ikke bare en forudindtaget urealistisk idé jeg har).

For at gøre dette skal jeg definere en eller flere variabler, som jeg gerne vil måle. I dette eksempel: kaffedrikningen hos Tesla-kørere. Jeg kunne måle denne variabel ved at give dem en spørgeskemaundersøgelse, der beder dem om at rapportere deres drikkevaner.

Lad os sige, at jeg gjorde det, og fandt ud af, at 75% drak 6 eller flere kopper kaffe om dagen.

WOW… Temmelig meget!

EKSEMPEL

FORKLARING (EN TEORI):

Hvordan forklarer vi denne data?

Én forklaring kunne være, at kaffedrikkere er mere tilbøjelige til at købe en Tesla. Dette er en teori.

Én anden mulighed kunne være, at folk, der kører Tesla, bare kører mere, og derfor har brug for mere kaffe for at forblive vågne. Dette er en anden teori.

Vi kan indsamle mere data for at teste vores teorier!

EKSEMPEL

FORUDSIGELSE (EN HYPOTESE):

Men før vi indsamler mere data for at teste vores teories, kan vi lave to forudsigelser ud fra vores teorier, som kan testes.

FORUDSIGELSE 1: Folk, der køber en Tesla, vil have højere kaffeforbrug end den generelle befolkning (f.eks., 2% drikker 6 kopper eller mere). En forudsigelse fra en teori som denne kaldes en hypotese. Vi kunne teste denne hypotese ved at bede alle Tesla-forhandlere om at spørge deres købere om deres kaffeforbrug.

FORUDSIGELSE 2: Folk, der kører mere end den generelle befolkning (f.eks., 20.000km per år), køber en Tesla og drikker kaffe for at forblive vågne. Dette er en anden hypotese. Vi kunne indsamle data om kørselsmøster, bilkøb og kaffeindtag for at teste denne hypotese.

OK, NU HAR VI EN LILLE IDE OM HVORFOR VI ER HER.

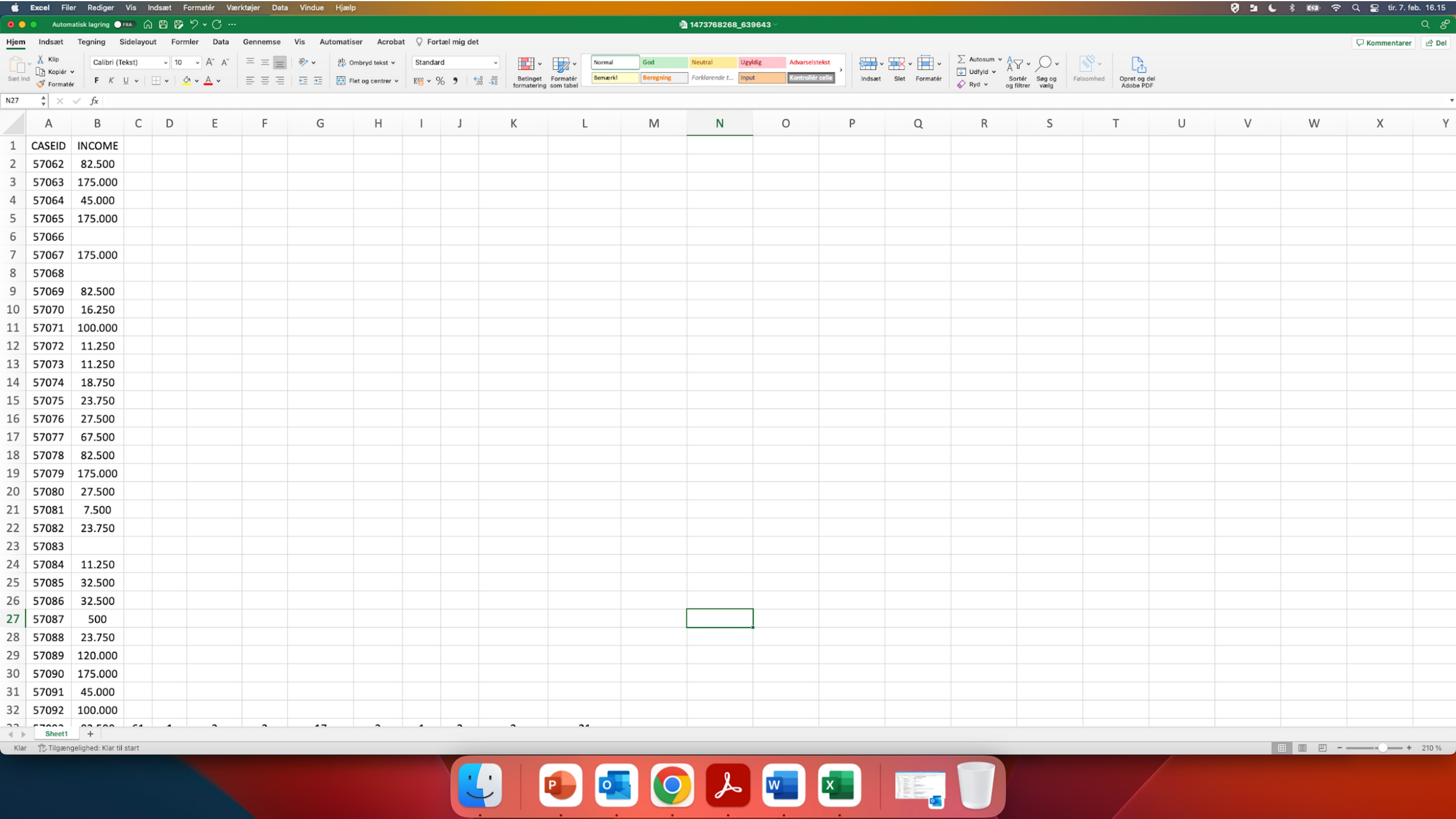
VI er her måske for at besvare interessante spørgsmål, for hvilke vi har brug for data og forklaringer.

I dag, vil vi kigge på forskellige typer af data og snakke om hvordan vi visualiserer disse.

HVORFOR SKAL VI LÆRE STATISTISK VISUALISERING?

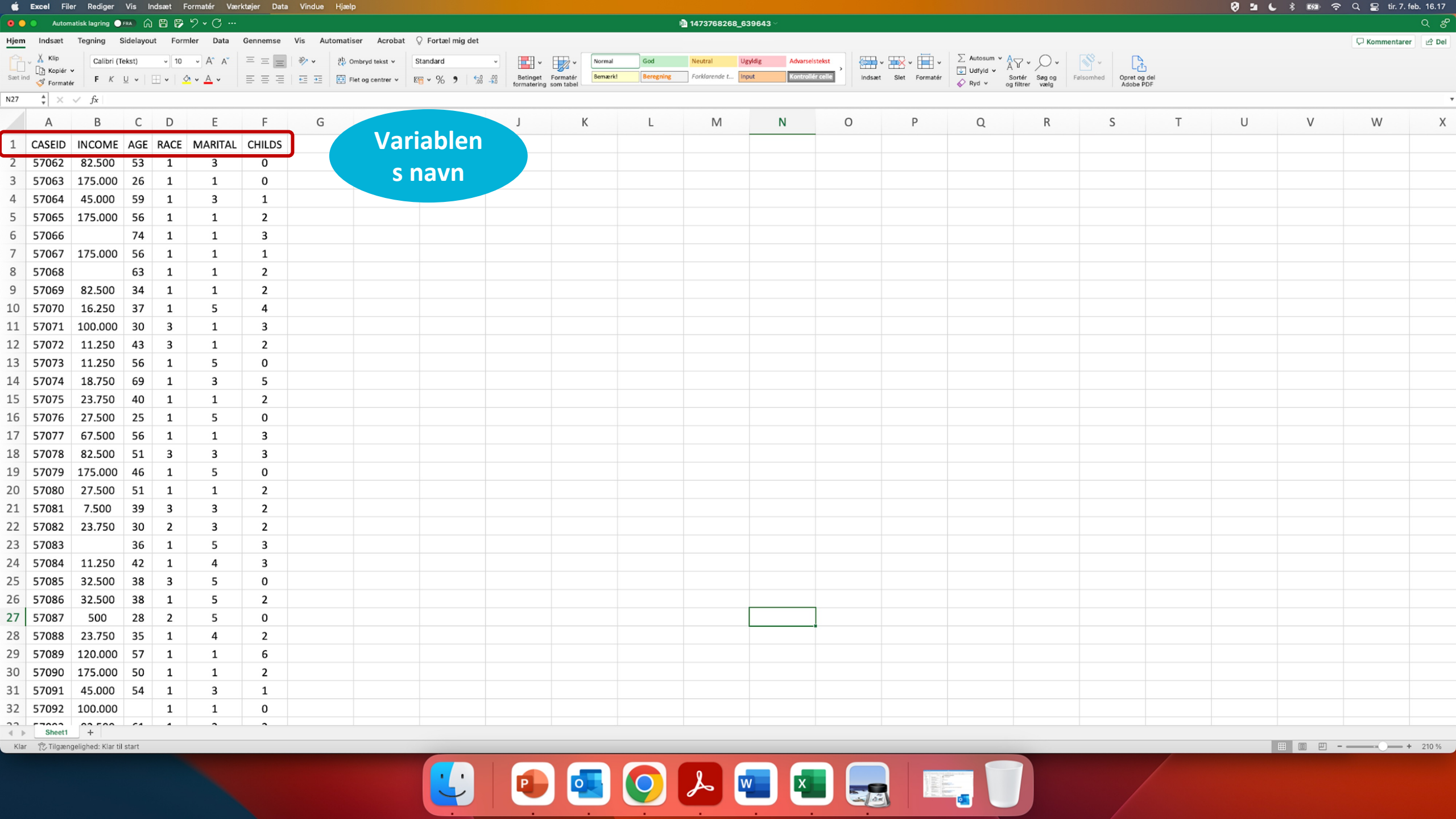
Visualisering er en fantastisk måde at komme fra data til information

Topledere tager beslutninger 25% hurtigere når de bliver præsenteret for materiale hvor grafikker indgår – *Wharton School of Business*

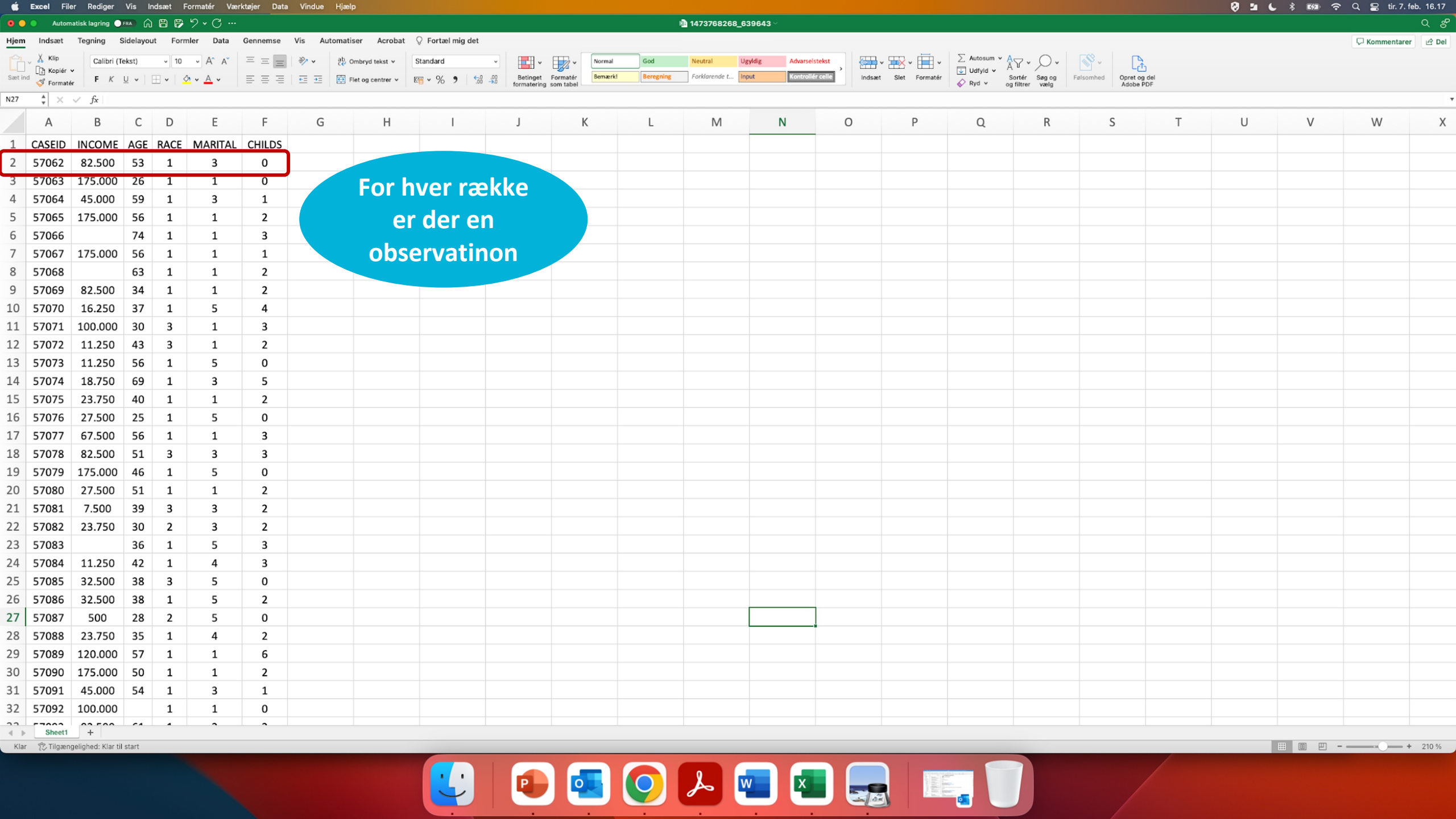


	A	B
1	CASEID	INCOME
2	57062	82.500
3	57063	175.000
4	57064	45.000
5	57065	175.000
6	57066	
7	57067	175.000
8	57068	
9	57069	82.500
10	57070	16.250
11	57071	100.000
12	57072	11.250
13	57073	11.250
14	57074	18.750
15	57075	23.750
16	57076	27.500
17	57077	67.500
18	57078	82.500
19	57079	175.000
20	57080	27.500
21	57081	7.500
22	57082	23.750
23	57083	
24	57084	11.250
25	57085	32.500
26	57086	32.500
27	57087	500
28	57088	23.750
29	57089	120.000
30	57090	175.000
31	57091	45.000
32	57092	100.000

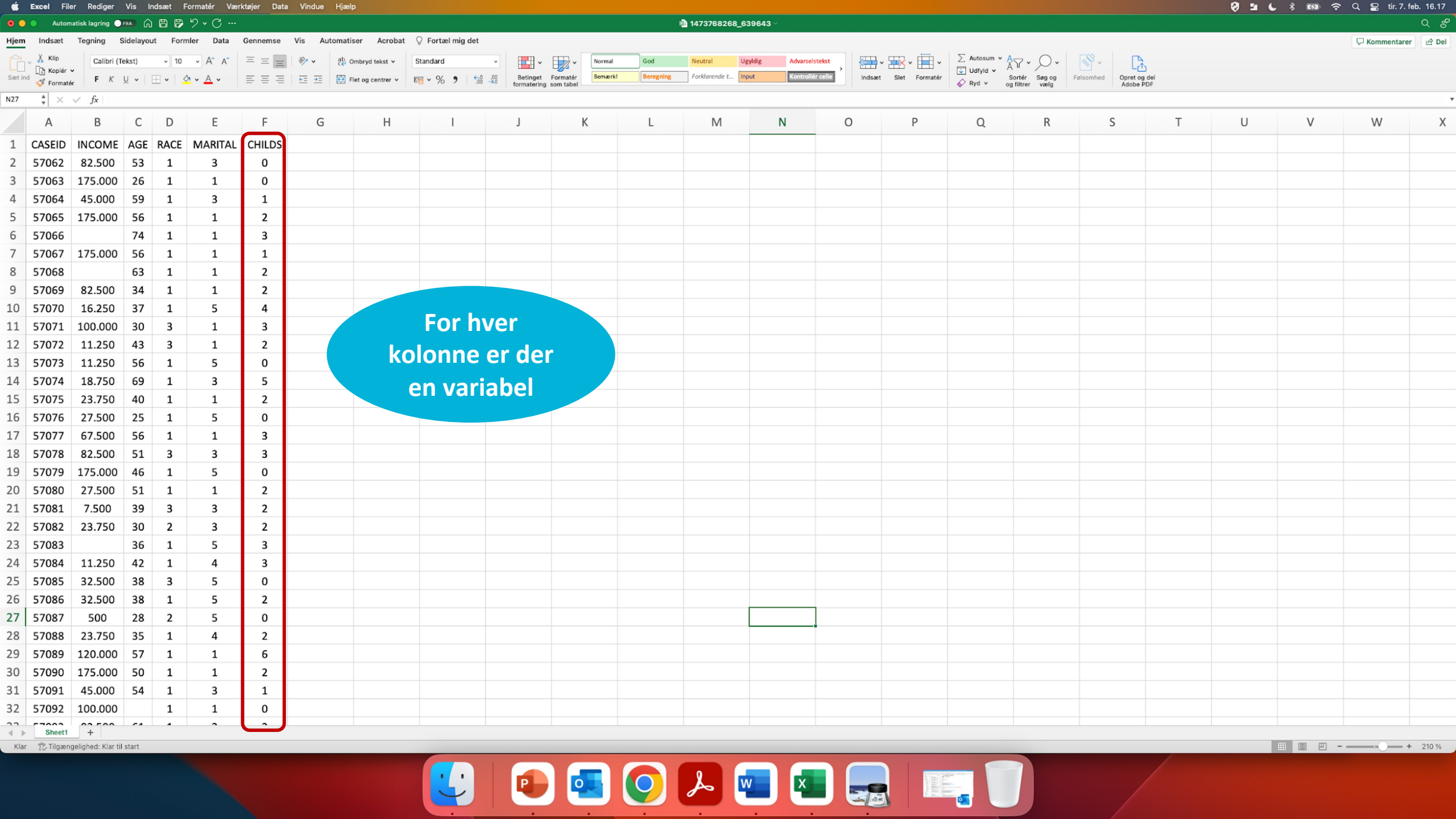
CASEID	INCOME	AGE	RACE	MARITAL	CHILDS
57062	82.500	53	1	3	0
57063	175.000	26	1	1	0
57064	45.000	59	1	3	1
57065	175.000	56	1	1	2
57066		74	1	1	3
57067	175.000	56	1	1	1
57068		63	1	1	2
57069	82.500	34	1	1	2
57070	16.250	37	1	5	4
57071	100.000	30	3	1	3
57072	11.250	43	3	1	2
57073	11.250	56	1	5	0
57074	18.750	69	1	3	5
57075	23.750	40	1	1	2
57076	27.500	25	1	5	0
57077	67.500	56	1	1	3
57078	82.500	51	3	3	3
57079	175.000	46	1	5	0
57080	27.500	51	1	1	2
57081	7.500	39	3	3	2
57082	23.750	30	2	3	2
57083		36	1	5	3
57084	11.250	42	1	4	3
57085	32.500	38	3	5	0
57086	32.500	38	1	5	2
57087	500	28	2	5	0
57088	23.750	35	1	4	2
57089	120.000	57	1	1	6
57090	175.000	50	1	1	2
57091	45.000	54	1	3	1
57092	100.000		1	1	0



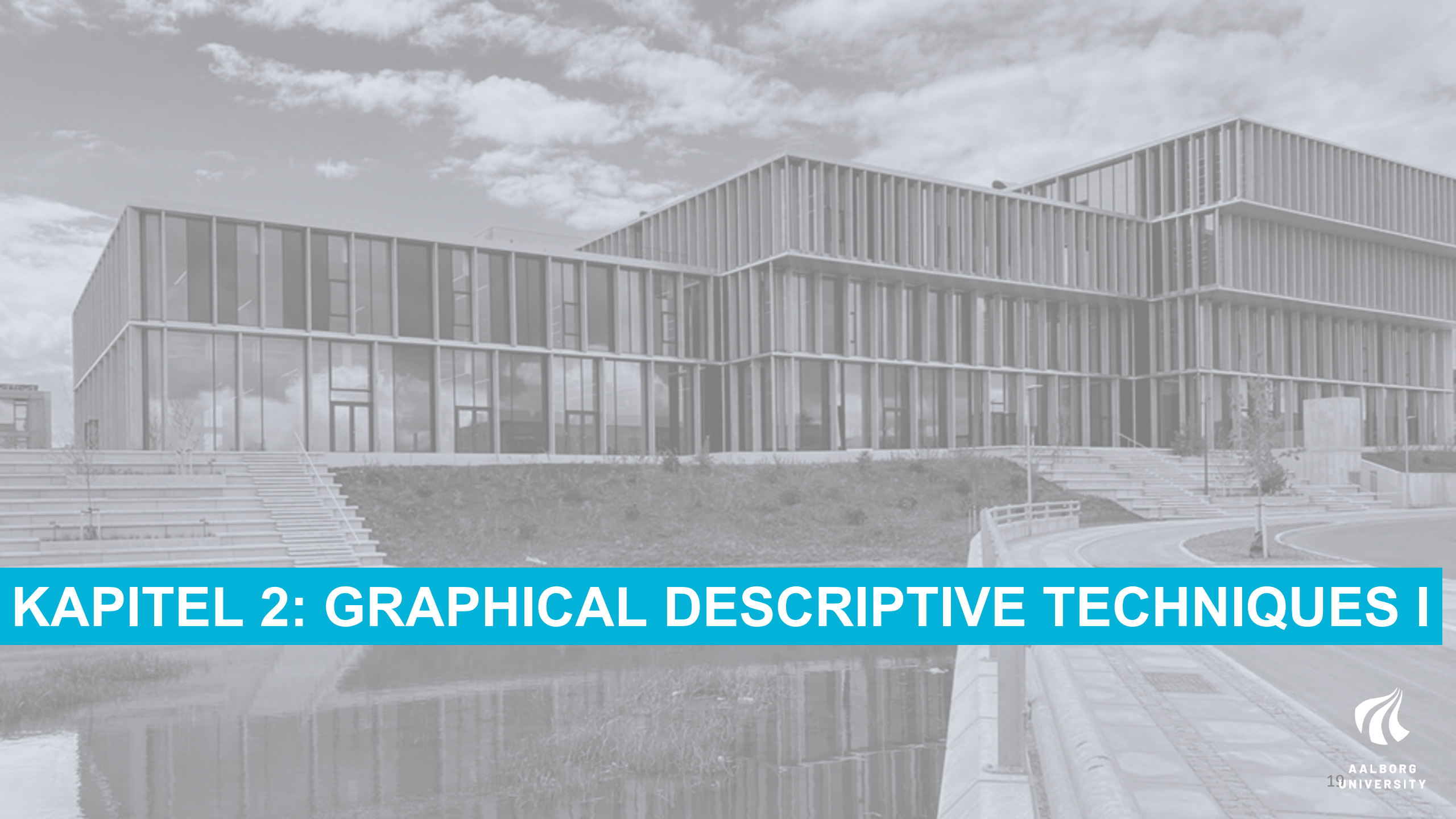
CASEID	INCOME	AGE	RACE	MARITAL	CHILDS
57062	82.500	53	1	3	0
57063	175.000	26	1	1	0
57064	45.000	59	1	3	1
57065	175.000	56	1	1	2
57066		74	1	1	3
57067	175.000	56	1	1	1
57068		63	1	1	2
57069	82.500	34	1	1	2
57070	16.250	37	1	5	4
57071	100.000	30	3	1	3
57072	11.250	43	3	1	2
57073	11.250	56	1	5	0
57074	18.750	69	1	3	5
57075	23.750	40	1	1	2
57076	27.500	25	1	5	0
57077	67.500	56	1	1	3
57078	82.500	51	3	3	3
57079	175.000	46	1	5	0
57080	27.500	51	1	1	2
57081	7.500	39	3	3	2
57082	23.750	30	2	3	2
57083		36	1	5	3
57084	11.250	42	1	4	3
57085	32.500	38	3	5	0
57086	32.500	38	1	5	2
57087	500	28	2	5	0
57088	23.750	35	1	4	2
57089	120.000	57	1	1	6
57090	175.000	50	1	1	2
57091	45.000	54	1	3	1
57092	100.000		1	1	0



CASEID	INCOME	AGE	RACE	MARITAL	CHILDS
57062	82.500	53	1	3	0
57063	175.000	26	1	1	0
57064	45.000	59	1	3	1
57065	175.000	56	1	1	2
57066		74	1	1	3
57067	175.000	56	1	1	1
57068		63	1	1	2
57069	82.500	34	1	1	2
57070	16.250	37	1	5	4
57071	100.000	30	3	1	3
57072	11.250	43	3	1	2
57073	11.250	56	1	5	0
57074	18.750	69	1	3	5
57075	23.750	40	1	1	2
57076	27.500	25	1	5	0
57077	67.500	56	1	1	3
57078	82.500	51	3	3	3
57079	175.000	46	1	5	0
57080	27.500	51	1	1	2
57081	7.500	39	3	3	2
57082	23.750	30	2	3	2
57083		36	1	5	3
57084	11.250	42	1	4	3
57085	32.500	38	3	5	0
57086	32.500	38	1	5	2
57087	500	28	2	5	0
57088	23.750	35	1	4	2
57089	120.000	57	1	1	6
57090	175.000	50	1	1	2
57091	45.000	54	1	3	1
57092	100.000		1	1	0



CASEID	INCOME	AGE	RACE	MARITAL	CHILDS
57062	82.500	53	1	3	0
57063	175.000	26	1	1	0
57064	45.000	59	1	3	1
57065	175.000	56	1	1	2
57066		74	1	1	3
57067	175.000	56	1	1	1
57068		63	1	1	2
57069	82.500	34	1	1	2
57070	16.250	37	1	5	4
57071	100.000	30	3	1	3
57072	11.250	43	3	1	2
57073	11.250	56	1	5	0
57074	18.750	69	1	3	5
57075	23.750	40	1	1	2
57076	27.500	25	1	5	0
57077	67.500	56	1	1	3
57078	82.500	51	3	3	3
57079	175.000	46	1	5	0
57080	27.500	51	1	1	2
57081	7.500	39	3	3	2
57082	23.750	30	2	3	2
57083		36	1	5	3
57084	11.250	42	1	4	3
57085	32.500	38	3	5	0
57086	32.500	38	1	5	2
57087	500	28	2	5	0
57088	23.750	35	1	4	2
57089	120.000	57	1	1	6
57090	175.000	50	1	1	2
57091	45.000	54	1	3	1
57092	100.000		1	1	0



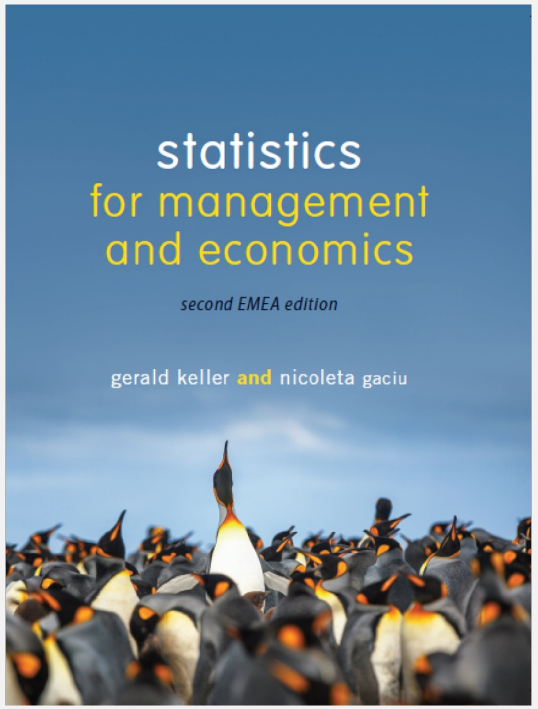
KAPITEL 2: GRAPHICAL DESCRIPTIVE TECHNIQUES I

AALBORG UNIVERSITY

KAPITEL 2

OVERVIEW

- Types of Data and Information
- Describing a Set of Nominal Data
- Describing the Relationship between Two Nominal Variables and Comparing Two or More Nominal Data Sets



TYPER AF DATA OG INFORMATION

DEFINITIONER

En **variabel** er et stykke information omkring en stikprøve eller en population

Fx. karakterer

Værdierne af en variable er spændet af de mulige værdier variable kan have

Fx. Karakterer (-3..12)

Data are the ***observerede værdier*** af en variabel

Fx. karakterer: {7, 7, 4, 0, 10, 7, 12}

TYPER AF DATA OG INFORMATION

HIERARKI FOR DATA

Intervaldata alle beregninger er muligt

Ordinaldata optælling og rangering

Nominaldata optælling

- Der findes tre overordnede datatyper som vi arbejder med i statistik
- Disse kan opstilles hierarkisk med den datatype hvor der er flest beregningsmæssige muligheder øverst og den hvor der er færrest muligheder nederst.

TYPER AF DATA OG INFORMATION

HIERARKI FOR DATA

Intervaldata alle beregninger er muligt

- Også kaldet kvantitativt eller numerisk data
- Denne form for data har ingen kategorier men er i stedet en kontinuér skala
- Eksempler: højde, vægt, karakterer, priser, indkomst, antal medarbejdere i virksomheden

TYPER AF DATA OG INFORMATION

HIERARKI FOR DATA

Ordinaldata optælling og rangering

- Ordinaldata minder om nominal data, men kategorierne i ordinaldata har en rangorden. Ordinal data fortæller os ikke kun, at ting er sket, men også i hvilken rækkefølge de er sket.
- Vi kan blot tælle hvor mange observationer der er i hver af de rangordnede kategorier
- Eksempel: hvad synes du om statistikundervisningen på AAU?
  Dårlig = 1, ok = 2, god = 3,
  meget god = 4, fremragende = 5

TYPER AF DATA OG INFORMATION

HIERARKI FOR DATA

Nominaldata optælling

- Også kaldet kategorisk eller kvalitativt data. Den enkleste kategoriske variabel er en binær variabel, for eksempel “død eller levende”, “gravid eller ej”, og du kan svare "ja" eller "nej", og en enhed kan placeres i kun den ene eller den anden af to kategorier.
- Værdierne i nominaldata repræsenterer flere kategorier
- Eksempel: hvad er du?
  Menneske = 1, Nisse = 2, Kat = 3,
  Hat = 4

Selvom tal ofte bliver brugt til at betegne kategorier, er det meningsløst at udføre udregninger på nominaldata; hvis du multiplicerer et menneske med en kat (1+3), får du ikke en hat (4).

Table 1: **Descriptive statistics for the main variables,** $N = 20,271$

	Mean	St. dev.	Min.	Max.
Innovation outcomes				
Sales from innovation	0.114	0.276	0	1
Sales from innovation new to the firm	0.054	0.181	0	1
Sales from innovation new to the market	0.043	0.162	0	1
Sales from innovation new to the world	0.017	0.101	0	1
Lagged employment shares				
Hires with founder experience	0.00550	0.010	0	0.273
... in *top management*	0.00007	0.001	0	0.053
... in *middle management*	0.00023	0.002	0	0.041
... in *non-management*	0.00520	0.010	0	0.273
Hires without founder experience	0.213	0.146	0	1
Stayers	0.781	0.149	0	1
Control variables				
Firm size	274.903	1,009.146	26	>33,500
Firm age	27.086	18.420	6	>100
Physical capital (thousand DKK)	377.525	2,599.477	0	>9,800,000
R&D workers	24.808	137.963	0	>7,000
University graduates	46.960	181.409	0	>8,800
R&D department	0.225	0.418	0	1
R&D intensity	0.023	0.102	0	1
Collaboration breadth	1.104	2.081	0	8
Applied for patent(s)	0.103	0.305	0	1
Acquired patent(s)	0.082	0.274	0	1
Sales growth	0.120	2.059	-0.996	>200
Investment intensity	0.041	0.095	0	1

Intervaldata numerisk/kvantitativt **VS.** **Nominaldata** kategorisk/kvalitativt

Firm size	274.903	1,009.146	26	>33,500
Firm age	27.086	18.420	6	>100
Physical capital (thousand DKK)	377.525	2,599.477	0	>9,800,000
R&D workers	24.808	137.963	0	>7,000
University graduates	46.960	181.409	0	>8,800
R&D department	0.225	0.418	0	1
R&D intensity	0.023	0.102	0	1
Collaboration breadth	1.104	2.081	0	8
Applied for patent(s)	0.103	0.305	0	1
Acquired patent(s)	0.082	0.274	0	1
Sales growth	0.120	2.059	-0.996	>200
Investment intensity	0.041	0.095	0	1

Table 2: **The effect of entrepreneur hires on firms' sales from innovation**

		Model I: Sales from innovation			Model II: Sales from innovation			Model III: Sales from innov. new to firm			Model IV: Sales from innov. new to market			Model V: Sales from innov. new to world		
		β	p	s.e.	β	p	s.e.	β	p	s.e.	β	p	s.e.	β	p	s.e.
	Lagged employment shares															
(1)	Hires with founder experience	0.502	0.005	0.180				0.256	0.036	0.122	0.163	0.125	0.106	0.083	0.249	0.072
(2)	... in *top management*				-2.081	0.126	1.358									
(3)	... in *middle management*				2.625	0.054	1.362									
(4)	... in *non-management*				0.236	0.223	0.194									
(5)	Hires without founder experience	-0.008	0.690	0.020	-0.002	0.921	0.020	0.004	0.755	0.014	-0.017	0.124	0.011	0.005	0.516	0.008
	Control variables															
	Log firm size	-0.012	0.181	0.009	-0.011	0.211	0.009	-0.009	0.170	0.006	0.002	0.747	0.005	-0.005	0.172	0.003
	Log physical capital	-0.003	0.287	0.002	-0.003	0.271	0.002	-0.003	0.063	0.001	-0.000	0.957	0.001	0.000	0.900	0.001
	Firm age	0.002	0.000	0.001	0.002	0.000	0.001	0.002	0.000	0.000	0.001	0.076	0.000	-0.000	0.432	0.000
	Log R&D workers	0.011	0.085	0.006	0.010	0.087	0.006	0.010	0.026	0.004	0.000	0.956	0.004	0.000	0.805	0.002
	Log university graduates	-0.008	0.263	0.007	-0.008	0.288	0.007	0.001	0.824	0.005	-0.009	0.061	0.005	-0.001	0.787	0.002
	R&D department	0.120	0.000	0.014	0.120	0.000	0.014	0.064	0.000	0.010	0.046	0.000	0.009	0.010	0.011	0.004
	R&D intensity	0.067	0.286	0.063	0.066	0.288	0.062	-0.012	0.773	0.043	0.025	0.490	0.036	0.054	0.281	0.050
	Collaboration breadth	0.018	0.000	0.002	0.018	0.000	0.002	0.009	0.000	0.001	0.007	0.000	0.001	0.002	0.015	0.001
	Applied for patent(s)	0.017	0.268	0.016	0.018	0.254	0.016	-0.009	0.429	0.011	0.017	0.077	0.010	0.009	0.097	0.005
	Acquired patent(s)	0.029	0.012	0.012	0.029	0.013	0.012	0.008	0.310	0.008	0.017	0.020	0.007	0.004	0.486	0.005
	Sales growth/investment intensity	Yes			Yes			Yes			Yes			Yes		
	Industry-year fixed effects	Yes			Yes			Yes			Yes			Yes		
	Firm fixed effects	Yes			Yes			Yes			Yes			Yes		
	Number of observations/firms	20,271/3,846			20,271/3,846			20,271/3,846			20,271/3,846			20,271/3,846		
	Adjusted R^2	0.310			0.310			0.198			0.206			0.267		
	F-tests															
	Hypothesis 1: (1)=(5)	7.82	0.005					4.21	0.040		2.79	0.095		1.14	0.286	
	Hypothesis 2: (3)=(2)				5.86	0.016										
	Hypothesis 2: (3)=(4)				3.03	0.082										
	Hypothesis 2: (3)=(2) *and* (4)				2.93	0.053										

Robust standard errors clustered by firm. Sample restricted to firms older than 5 years and with more than 25 employees, for years 2007-2016; stayers represent the baseline category. All models estimated by ordinary least squares with firm fixed effects.

Table 2: **The effect of entrepreneur hires on firms' sales from innovation**

		Model I: Sales from innovation			Model II: Sales from innovation			Model III: Sales from innov. new to firm			Model IV: Sales from innov. new to market			Model V: Sales from innov. new to world		
		β	p	s.e.	β	p	s.e.	β	p	s.e.	β	p	s.e.	β	p	s.e.
	Lagged employment shares															
(1)	Hires with founder experience	0.502	0.005	0.180				0.256	0.036	0.122	0.163	0.125	0.106	0.083	0.249	0.072
(2)	... in *top management*				-2.081	0.126	1.358									
(3)	... in *middle management*				2.625	0.054	1.362									
(4)	... in *non-management*				0.236	0.223	0.194									
(5)	Hires without founder experience	-0.008	0.690	0.020	-0.002	0.921	0.020	0.004	0.755	0.014	-0.017	0.124	0.011	0.005	0.516	0.008
	Control variables															
	Log firm size	-0.012	0.181	0.009	-0.011	0.211	0.009	-0.009	0.170	0.006	0.002	0.747	0.005	-0.005	0.172	0.003
	Log physical capital	-0.003	0.287	0.002	-0.003	0.271	0.002	-0.003	0.063	0.001	-0.000	0.957	0.001	0.000	0.900	0.001
	Firm age	0.002	0.000	0.001	0.002	0.000	0.001	0.002	0.000	0.000	0.001	0.076	0.000	-0.000	0.432	0.000
	Log R&D workers	0.011	0.085	0.006	0.010	0.087	0.006	0.010	0.026	0.004	0.000	0.956	0.004	0.000	0.805	0.002
	Log university graduates	-0.008	0.263	0.007	-0.008	0.288	0.007	0.001	0.824	0.005	-0.009	0.061	0.005	-0.001	0.787	0.002
	R&D department	0.120	0.000	0.014	0.120	0.000	0.014	0.064	0.000	0.010	0.046	0.000	0.009	0.010	0.011	0.004
	R&D intensity	0.067	0.286	0.063	0.066	0.288	0.062	-0.012	0.773	0.043	0.025	0.490	0.036	0.054	0.281	0.050
	Collaboration breadth	0.018	0.000	0.002	0.018	0.000	0.002	0.009	0.000	0.001	0.007	0.000	0.001	0.002	0.015	0.001
	Applied for patent(s)	0.017	0.268	0.016	0.018	0.254	0.016	-0.009	0.429	0.011	0.017	0.077	0.010	0.009	0.097	0.005
	Acquired patent(s)	0.029	0.012	0.012	0.029	0.013	0.012	0.008	0.310	0.008	0.017	0.020	0.007	0.004	0.486	0.005
	Sales growth/investment intensity	Yes			Yes			Yes			Yes			Yes		
	Industry-year fixed effects	Yes			Yes			Yes			Yes			Yes		
	Firm fixed effects	Yes			Yes			Yes			Yes			Yes		
	Number of observations/firms	20,271/3,846			20,271/3,846			20,271/3,846			20,271/3,846			20,271/3,846		
	Adjusted R^2	0.310			0.310			0.198			0.206			0.267		
	F-tests															
	Hypothesis 1: (1)=(5)	7.82	0.005					4.21	0.040		2.79	0.095		1.14	0.286	
	Hypothesis 2: (3)=(2)				5.86	0.016										
	Hypothesis 2: (3)=(4)				3.03	0.082										
	Hypothesis 2: (3)=(2) *and* (4)				2.93	0.053										

Robust standard errors clustered by firm. Sample restricted to firms older than 5 years and with more than 25 employees, for years 2007-2016; stayers represent the baseline category. All models estimated by ordinary least squares with firm fixed effects.

Table 2: **The effect of entrepreneur hires on firms' sales from innovation**

		Model I: Sales from innovation		
		β	p	s.e.
	Lagged employment shares			
(1)	Hires with founder experience	0.502	0.005	0.180
(2)	... in *top management*			
(3)	... in *middle management*			
(4)	... in *non-management*			
(5)	Hires without founder experience	-0.008	0.690	0.020
	Control variables			
	Log firm size	-0.012	0.181	0.009
	Log physical capital	-0.003	0.287	0.002
	Firm age	0.002	0.000	0.001
	Log R&D workers	0.011	0.085	0.006
	Log university graduates	-0.008	0.263	0.007
	R&D department	0.120	0.000	0.014
	R&D intensity	0.067	0.286	0.063
	Collaboration breadth	0.018	0.000	0.002
	Applied for patent(s)	0.017	0.268	0.016
	Acquired patent(s)	0.029	0.012	0.012
	Sales growth/investment intensity	Yes		
	Industry-year fixed effects	Yes		
	Firm fixed effects	Yes		
	Number of observations/firms	20,271/3,846		
	Adjusted R^2	0.310		

p < 0,05
statistisk signifikant

R2
andel af **varians** i *Y* forklaret af *x*

TYPER AF DATA OG INFORMATION

INTERVAL

- Values are real numbers
- All calculations are valid
- Data may be treated as ordinal or nominal

ORDINAL

- Values must represent the ranked order of the data
- Calculations based on an ordering process are valid
- Data may be treated as nominal but not as interval

NOMINAL

- Values are the arbitrary numbers that represent categories
- Only calculations based on the frequencies of occurrence are valid
- Data may not be treated as ordinal or interval

BESKRIVENDE STATISTIK FOR NOMINAL DATA

HYPPIGHEDER

- Det eneste vi kan gøre med nominel data er at tælle hyppigheden af hver værdi for variablen
- Vi kan summere data i en tabel der præsenterer kategorierne og optælling kaldet en **frequency distribution** (hyppigheds distribution)
- En **relative frequency distribution** (relativ hyppigheds distribution) viser kategorierne og den proportion med hvilken hver kategori forekommer

BESKRIVENDE STATISTIK FOR NOMINAL DATA

EKSEMPEL: BESKÆFTIGELSE

- UK Census survey data er et national spørgeskema designet til at give et øjebliksbillede af den britiske befolkning og dets karakteristika.
- Census 2011 datasætter består af en *random sample* (tilfældigt udvalgt stikprøve) på 1% af befolkningen i England og Wales.
- Respondenterne er blandt anden blevet spurgt ind til deres arbejdsmarkedsstatus.

BESKRIVENDE STATISTIK FOR NOMINAL DATA

EKSEMPEL: BESKÆFTIGELSE

Which of the following characterized your employment situation:
1. Employed
2. Self-employed
3. Unemployed
4. Full time student
5. Retired
6. School
7. Looking after home and family
8. Long term sick or disabled
9. Other

BESKRIVENDE STATISTIK FOR NOMINAL DATA

EKSEMPEL: BESKÆFTIGELSE

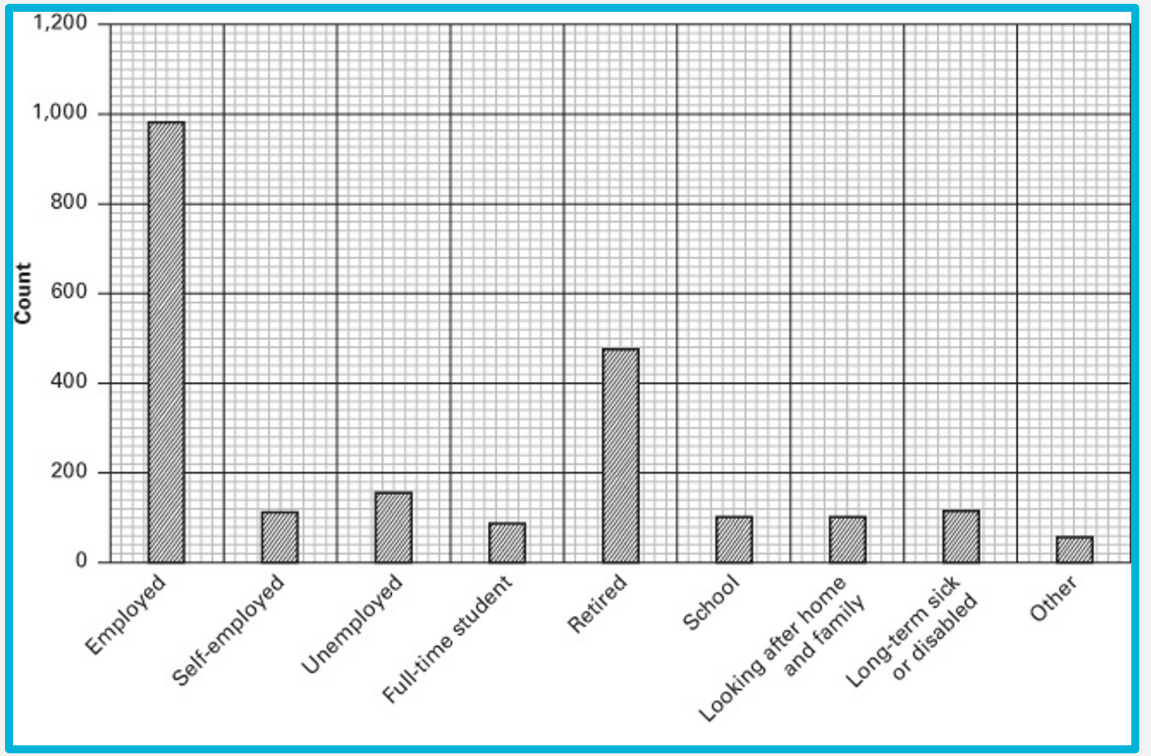
Frequency distributions table

Work Status	Code	Frequency	Relative Frequency (%)
Employed	1	978	45.4
Self-employed	2	108	5.0
Unemployed	3	155	7.2
Full time student	4	84	3.9
Retired	5	472	21.9
School	6	97	4.5
Looking after home and family	7	97	4.5
Long-term sick or disabled	8	112	5.2
Other	9	52	2.4
Total		2,155	100

BESKRIVENDE STATISTIK FOR NOMINAL DATA

EKSEMPEL: BESKÆFTIGELSE

- *Bar charts* (søjlediagrammer) bruges ofte til at præsentere *frequencies* (hyppigheder)



BESKRIVENDE STATISTIK FOR NOMINAL DATA

EKSEMPEL: BESKÆFTIGELSE

Frequency distributions table

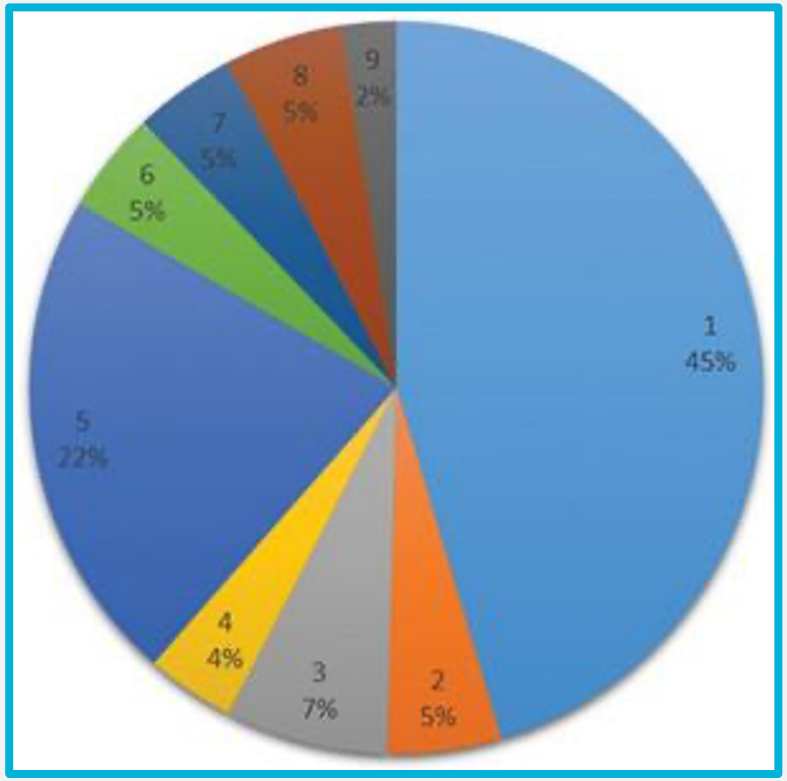
Pie chart

Work Status	Relative Frequency (%)	Slice of the Pie (°)
Employed	45.4	163.4
Self-employed	5.0	18.0
Unemployed	7.2	25.9
Full-time student	3.9	14.0
Retired	21.9	78.9
School	4.5	16.2
Looking after home and family	4.5	16.2
Long-term sick or disabled	5.2	18.7
Other	2.4	8.7
Total	100	360

BESKRIVENDE STATISTIK FOR NOMINAL DATA

EKSEMPEL: BESKÆFTIGELSE

- *Pie charts* (cirkeldiagrammer) bruges til at præsentere *relative frequencies* (relative hyppigheder)



FORHOLDET MELLEM TO NOMINELLE VARIABLE

CROSS-CLASSIFICATION TABLE

Når vi skal beskrive forholdet mellem to nominelle variable, kan vi stadig blot tælle hyppigheder på tværs af kategorier. Vi kan visualisere hyppigheder på tværs variable via at *cross-classification* (krydsklassificeringstabel) og illustrere dem i et *bar chart* (søjlediagram)

FORHOLDET MELLEM TO NOMINELLE VARIABLE

EKSEMPLEL: AVISLÆSERE

- I en større amerikansk by findes der fire konkurrerende aviser: the *Globe and Mail (G&M), Post, Sun,* og *Star.*
- For at kunne designe markedsføringskampagner for aviserne, er marketingledere nødt til at vide hvilke segmenter i avismarkedet der læser deres avis.
- En spørgeskemaundersøgelse blev gennemført for at analysere forholdet mellem avislæsning og beskæftigelse. En *sample* (stikprøve) af avislæsere blev spurgt om hvilken avis de læser: *Globe and Mail* (1) *Post* (2), *Star* (3), *Sun* (4) samt om de er blue-collar worker (1), white-collar worker (2), eller professional (3).

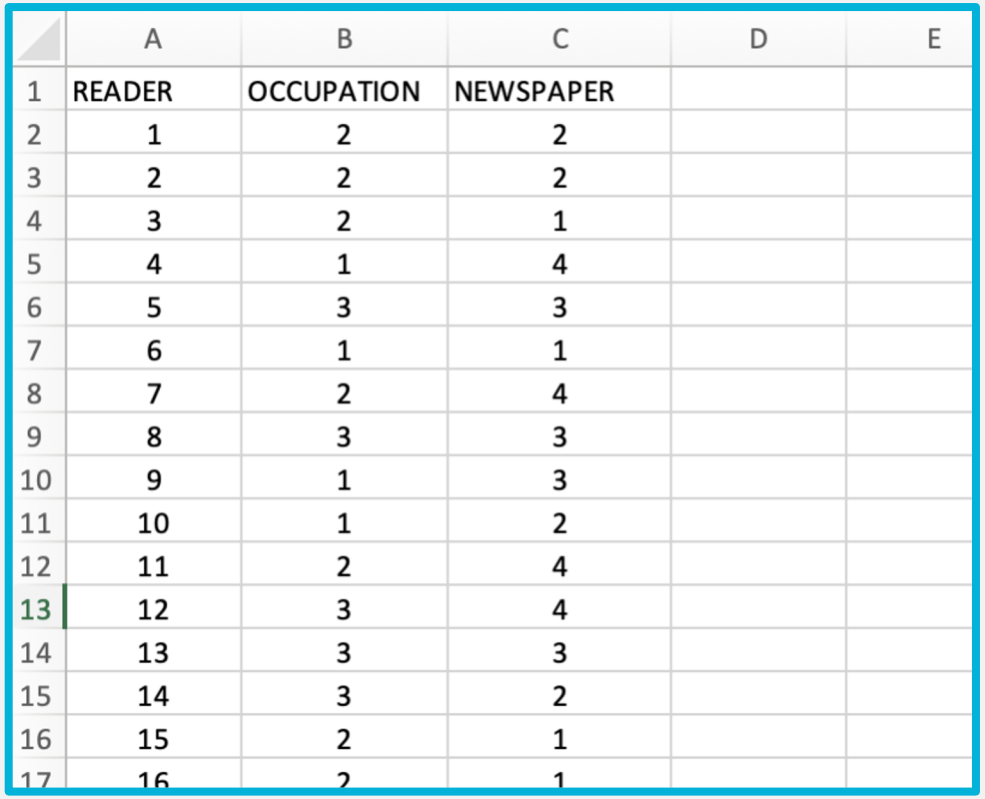
FORHOLDET MELLEM TO NOMINELLE VARIABLE

EKSEMPLEL: AVISLÆSERE

- Data fra spørgeskemaet kunne se sådan her ud

- Occupation
  1. blue-collar worker
  2. white-collar worker
  3. professional

- Newspaper
  1. Globe and Mail
  2. Post
  3. Star
  4. Sun



	A	B	C	D	E
1	READER	OCCUPATION	NEWSPAPER		
2	1	2	2		
3	2	2	2		
4	3	2	1		
5	4	1	4		
6	5	3	3		
7	6	1	1		
8	7	2	4		
9	8	3	3		
10	9	1	3		
11	10	1	2		
12	11	2	4		
13	12	3	4		
14	13	3	3		
15	14	3	2		
16	15	2	1		
17	16	2	1		

FORHOLDET MELLEM TO NOMINELLE VARIABLE

EKSEMPLEL: AVISLÆSERE

- Data fra spørgeskemaet kunne se sådan her ud

- Occupation
  1. blue-collar worker
  2. white-collar worker
  3. professional

- Newspaper
  1. Globe and Mail
  2. Post
  3. Star
  4. Sun

	A	B	C	D	E
1	READER	OCCUPATION	NEWSPAPER		
2	1	2	2		
3	2	2	2		
4	3	2	1		
5	4	1	4		
6	5	3	3		
7	6	1	1		
8	7	2	4		
9	8	3	3		
10	9	1	3		
11	10	1	2		
12	11	2	4		
13	12	3	4		
14	13	3	3		
15	14	3	2		
16	15	2	1		
17	16	2	1		

FORHOLDET MELLEM TO NOMINELLE VARIABLE

EKSEMPLEL: AVISLÆSERE

Cross-Classification Table of Frequencies

	Newspaper				
Occupation	G&M	Post	Star	Sun	*Total*
Blue collar	27	18	38	37	120
White collar	29	43	21	15	108
Professional	33	51	22	20	126
Total	89	112	81	72	354

FORHOLDET MELLEM TO NOMINELLE VARIABLE

EKSEMPLEL: AVISLÆSERE

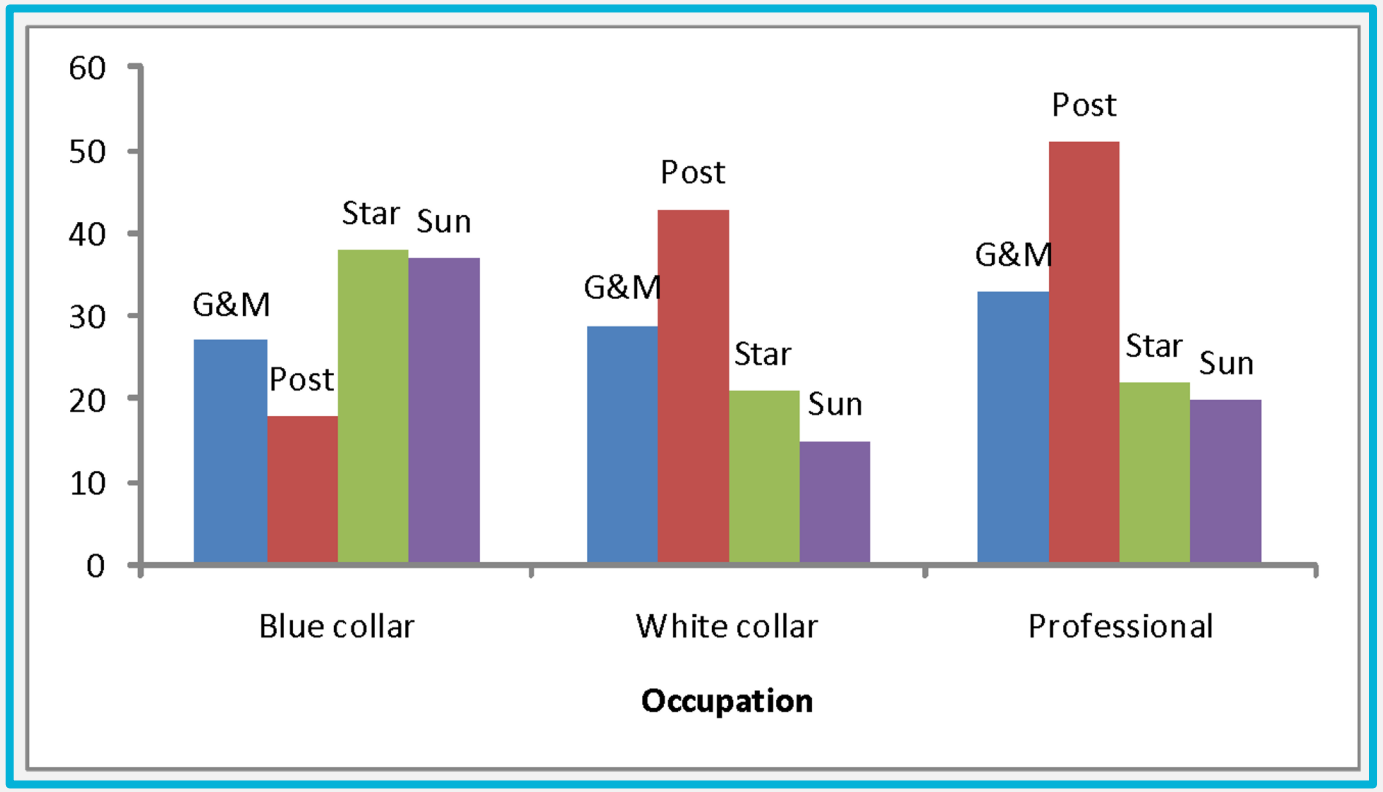
Cross-Classification Table of relative Frequencies

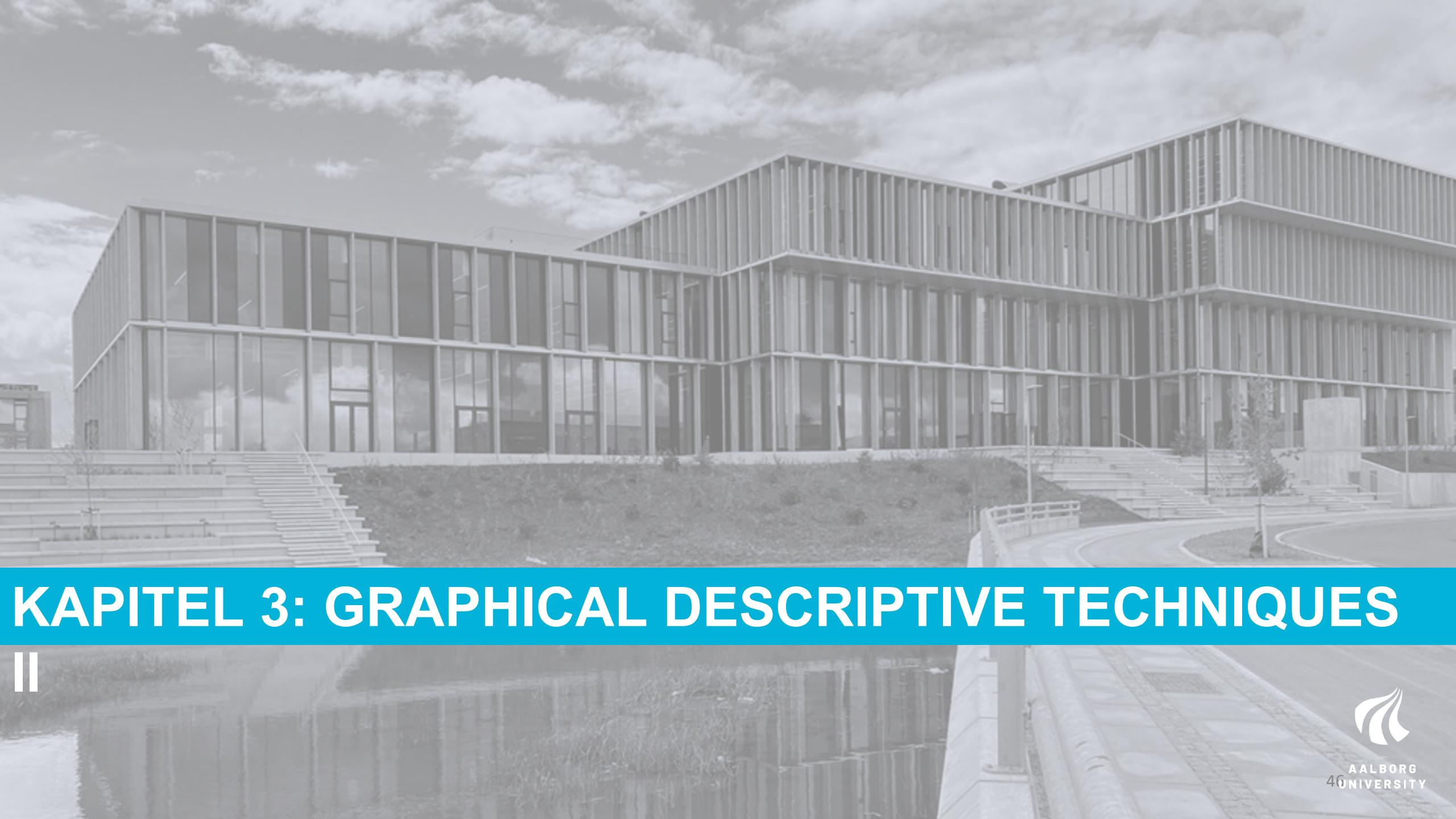
	Newspaper				
Occupation	G&M	Post	Star	Sun	*Total*
Blue collar	0,23	0,15	0,32	0,31	1,00
White collar	0,27	0,40	0,19	0,14	1,00
Professional	0,26	0,40	0,17	0,16	1,00
Total	0,25	0,32	0,23	0,20	1,00

FORHOLDET MELLEM TO NOMINELLE VARIABLE

EKSEMPLEL: AVISLÆSERE

Two-dimensional bar chart



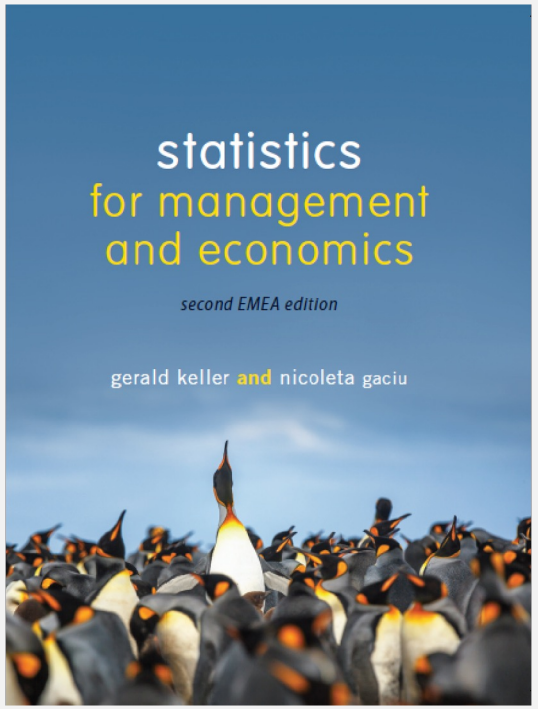


KAPITEL 3: GRAPHICAL DESCRIPTIVE TECHNIQUES II

KAPITEL 3

OVERVIEW

- Graphical Techniques to Describe a Set of Interval Data
- Describing Time-Series Data
- Describing the Relationship between Two Interval Variables
- Art and Science of Graphical Presentations



BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

- Markedsværdien for gaming stiger kraftigt år for år. I 2019 er værdien forventet at nå 33,6 milliarder US dollars med Asien- og Stillehavsregionen som det største gamingmarked.
- I 2018 var der mere en 2,5 milliarder gamers på verdensplan og 35% af disse var i aldres 21-35 år.
- Som et led i et større studie, har en gaming virksomhed, som for nyligt lancerede et nyt onlinespil, gerne ville indhente information omkring alderen på spillerne af deres nye online spil.

BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

- Virksomhedens marketing manager har udvalgt en *random sample* (tilfældig stikprøve) på 200 subscribers og noteret alderen på spillerne
- Hvilken information kan udledes fra disse data?
- *Range* (spændvidde): yngste spiller er 8 og ældre er 53

BESKRIVENDE STATISTIK FOR INTERVAL DATA

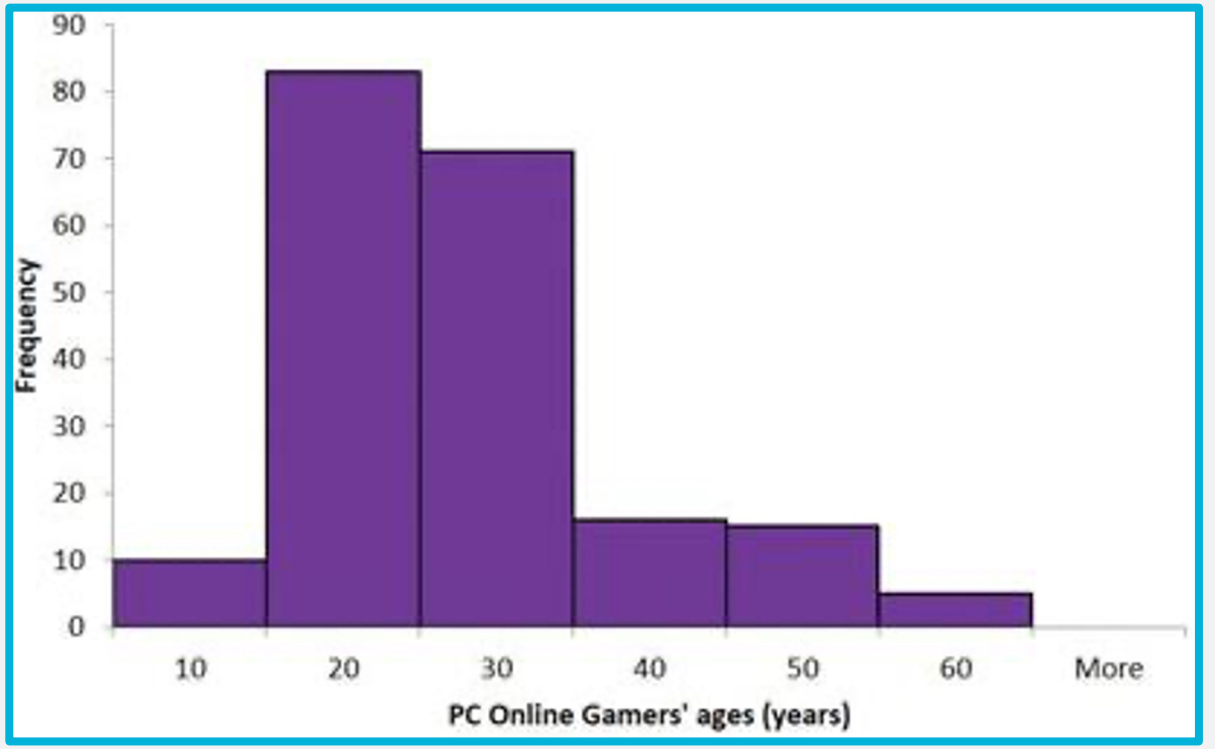
EKSEMPEL: ONLINE GAMERS

- Virksomhedens marketing manager har udvalgt en *random sample* (tilfældig stikprøve) på 200 subscribers og noteret alderen på spillerne
- Hvilken information kan udledes fra disse data?

BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

Histogram baseret på frequency distributions

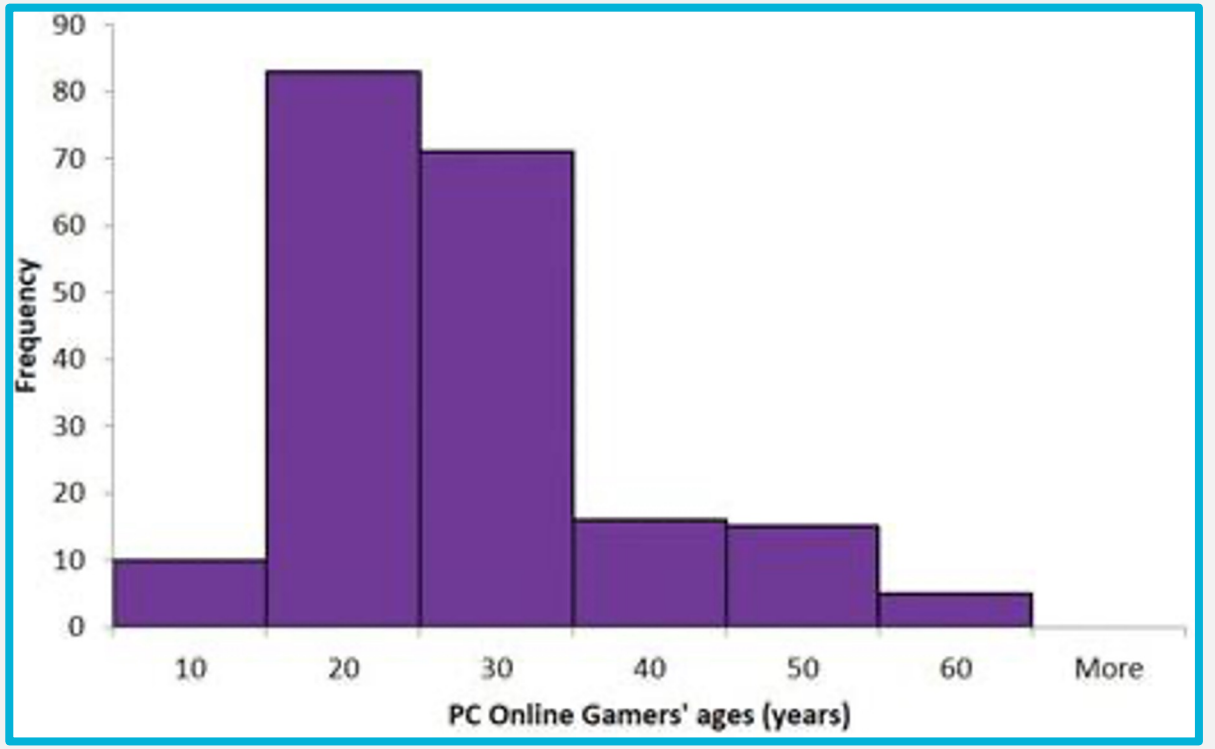


AALBORG UNIVERSITY

BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

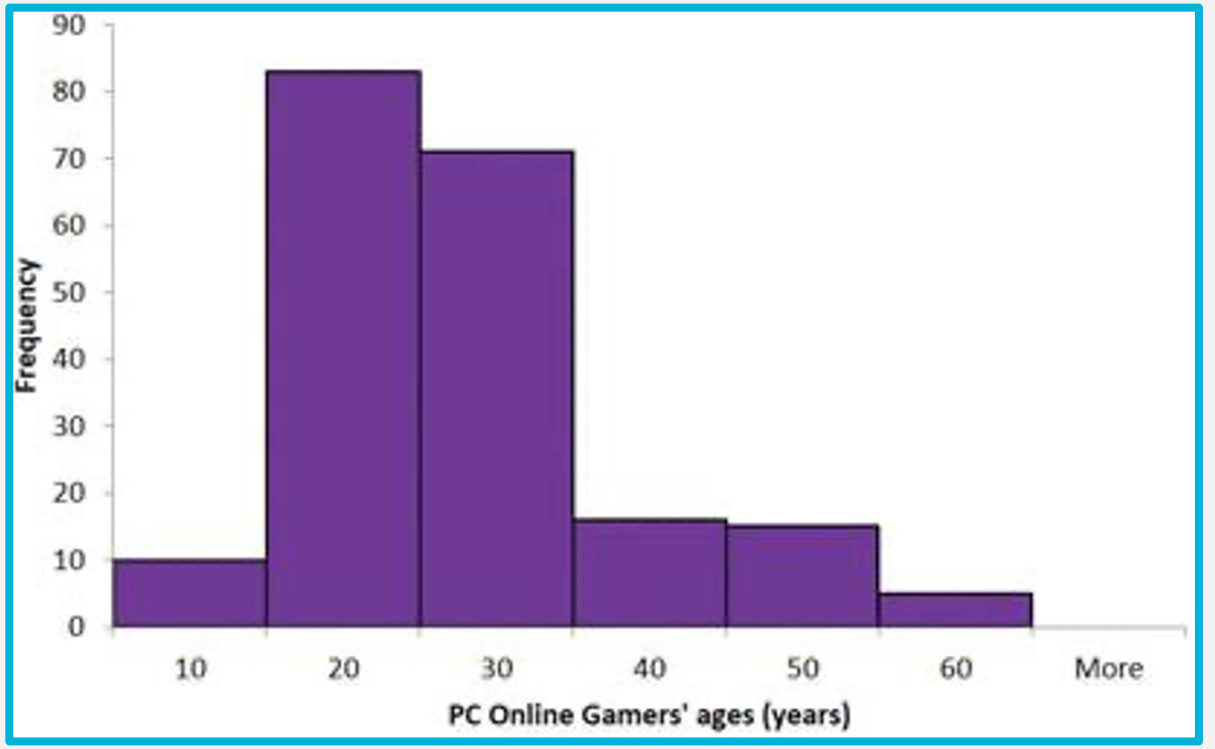
Hvad fortæller histogrammet os?



BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

Hvordan laver vi et histogram?



BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

- Data skal inddeles i *classes* (intervaller) defineres således at hver observation falder ind i én og kun én class (interval): *mutually exclusive* (gensidigt udelukkende)

Classes

Ages that are more than 1 but less than or equal to 10
Ages that are more than 10 and less than or equal to 20
Ages that are more than 20 but less than or equal to 30
Ages that are more than 30 but less than or equal to 40
Ages that are more than 40 but less than or equal to 50
Ages that are more than 50 but less than or equal to 60

BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

- Hvordan afgør vi hvor mange *classes* (intervaller) der skal være?
- Antallet af *classes* bestemmes på baggrund af antal observationer
- *Class* størrelsen bestemmes ud fra hvad der giver den bedste datapræsentation og fortolkningsmuligheder og har noget med *rangen* (spændvidden) på data at gøre
- Et *interval* bliver til en *bin* i histogrammet

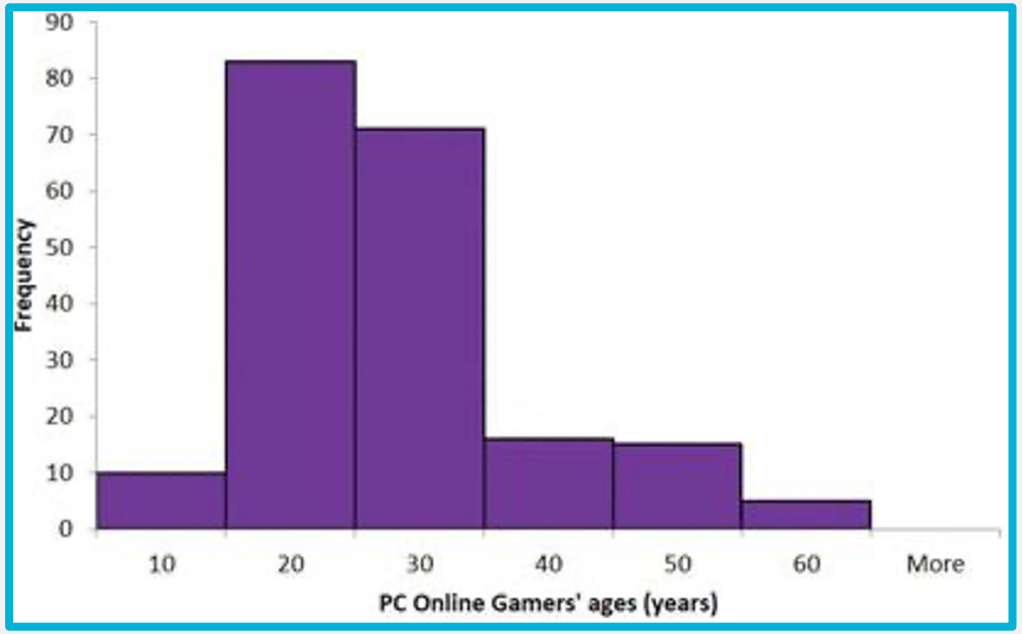
Approximate Number of Classes in Frequency Distributions

Number of Observations	Number of Classes
Less than 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1,000	10 - 11
1,000 - 5,000	11 - 13
5,000 - 50,000	13 - 17
More than 50,000	17 - 20

BESKRIVENDE STATISTIK FOR INTERVAL DATA

EKSEMPEL: ONLINE GAMERS

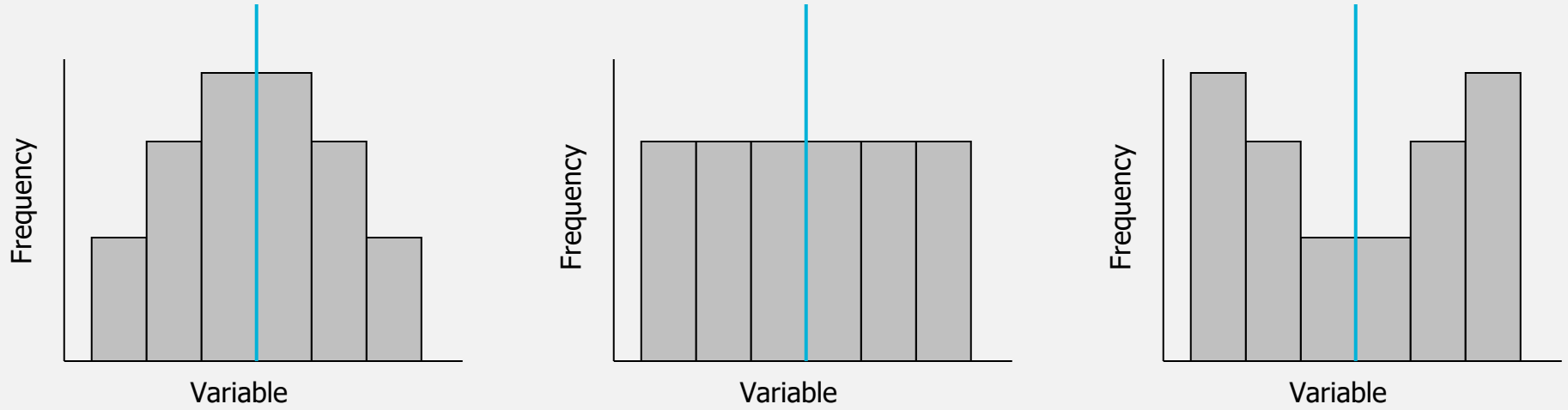
Class Limits	Frequency
0–10	10
11–20	83
21–30	71
31–40	16
41–50	15
51–60	5
Total	200



AALBORG UNIVERSITY

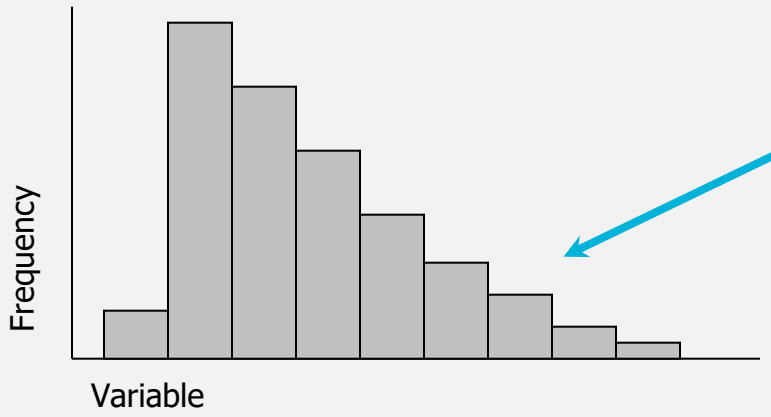
HISTOGRAMMETS FORM

- **Symmetri:** hvis vi tegner en linje midt igennem vores histogram og hver side er identiske i form og størrelse, så er histogrammet symmetrisk

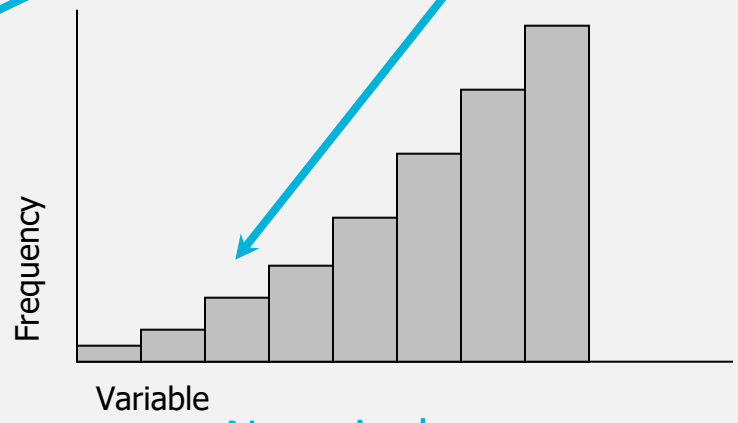


HISTOGRAMMETS FORM

- **Skewness** (skævhed)**:** et skævt histogram har en lang hale mod højre eller venstre



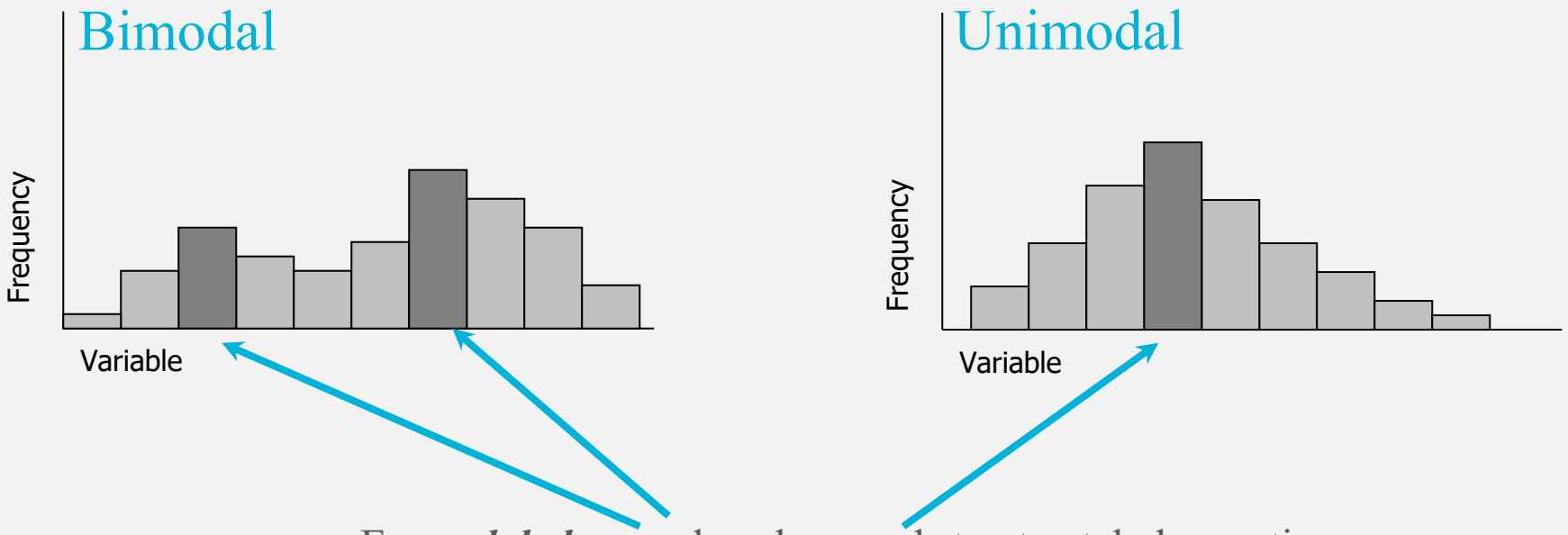
Positively Skewed



Negatively Skewed

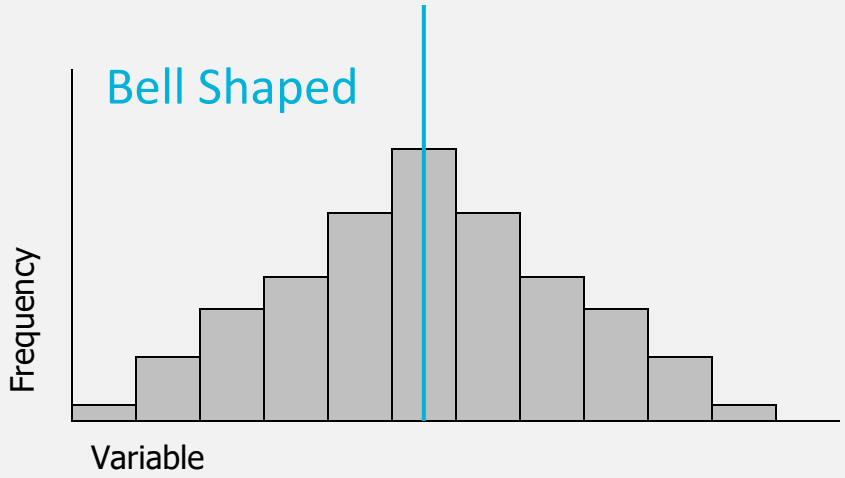
HISTOGRAMMETS FORM

- **Modularity:** et unimodalt histogram har et peak og et bimodalt histogram har to peaks



HISTOGRAMMETS FORM

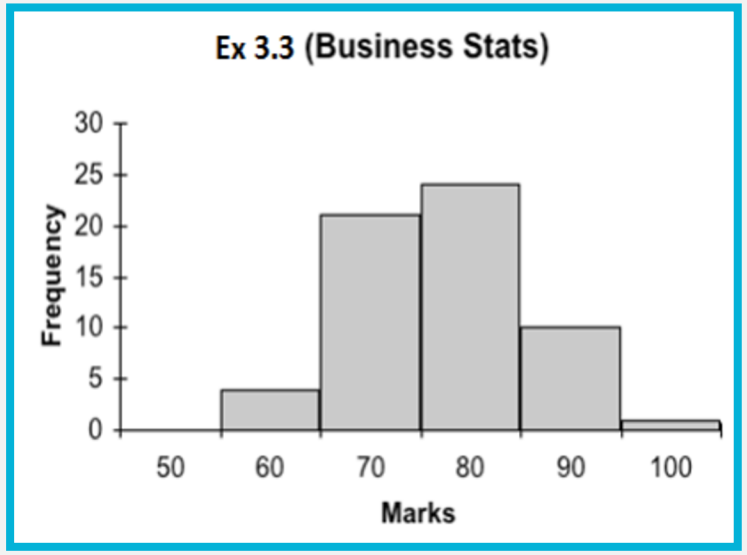
- **Bell shape** (klokkeformen) er en særlig type et af symmetrisk unimodalt histogram. Dette er en vigtig antagelse for mange statistiske teknikker. Vi kommer tilbage til det senere på kurset

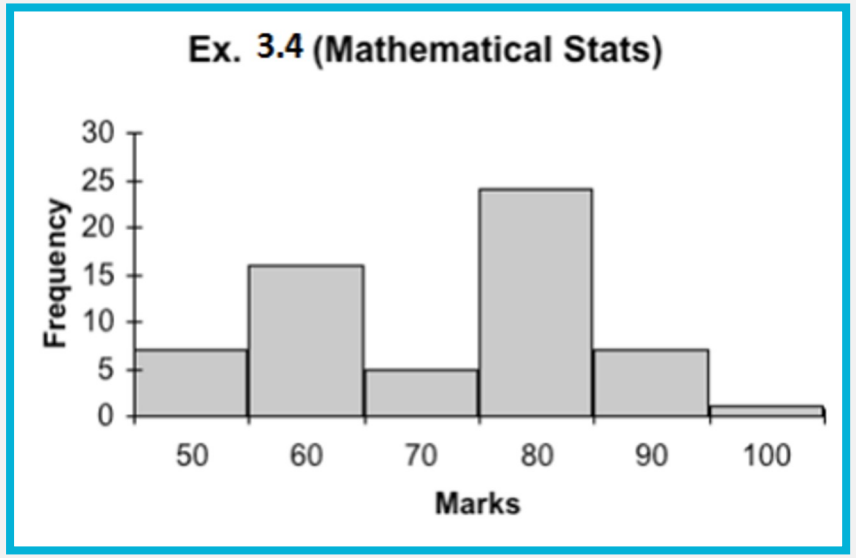


HISTOGRAMMETS FORM

TO EKSEMPLER

- Sammenlign de to histogrammer der viser eksamenskarakterer fra to forskellige statistikfag





TIDSSERIEDATA

SAMME VARIABEL MÅLT OVER TID

- Observationer målt på ét tidspunkt kaldes *cross-sectional* (tværsnit) data
- Observationer målt på flere på hinanden påfølgende tidspunkter kaldes *time-series* (tidsserie) data
- Tidsserie data plottes ofte på en *line chart* (linjeskala) som plotter værdien af variablen på den vertikale akse og tiden på den horisontale akse

TIDSSERIEDATA

EKSEMPEL: BENZINPRISER

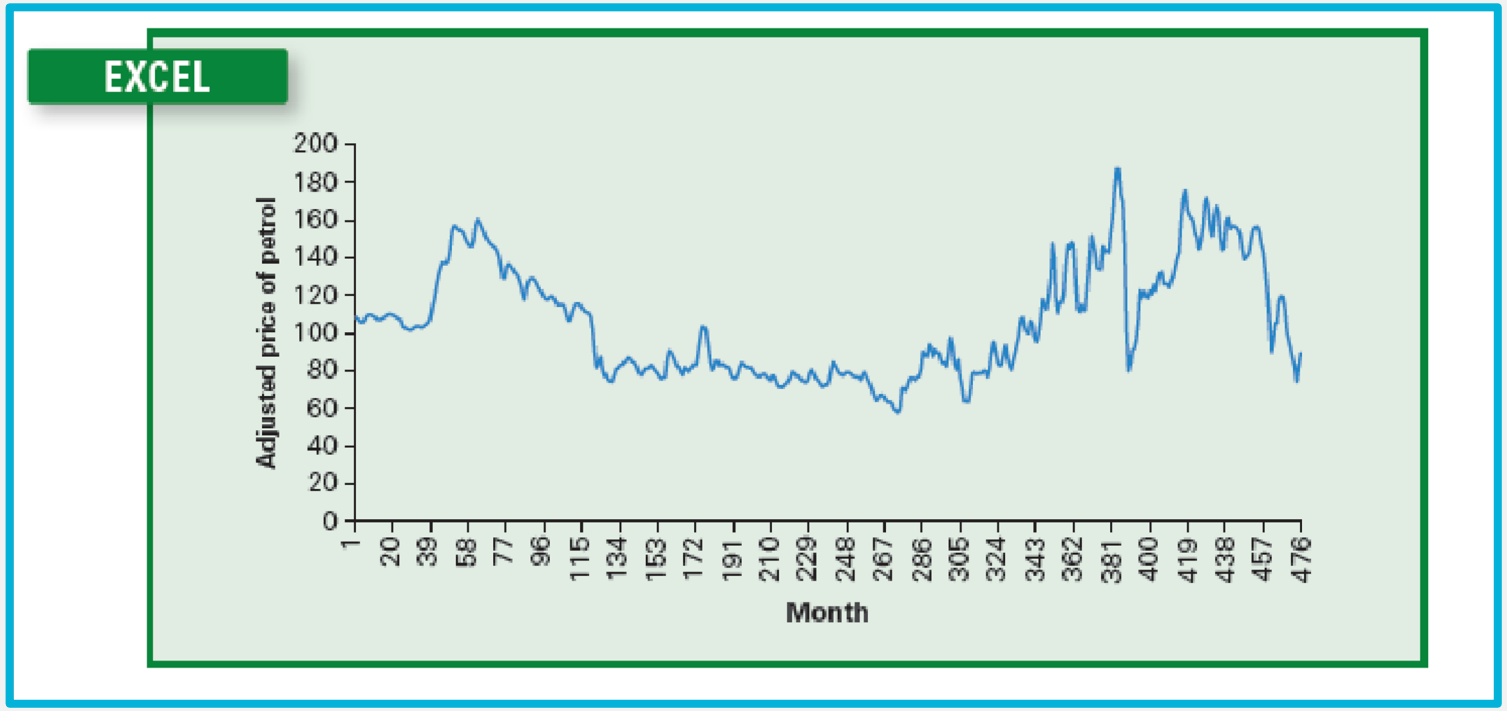
- Den månedlige gennemsnitlige detailpris på benzin er registreret siden 1976
- Når man har prisdata der strækker sig over længere tid er det vigtigt at justere for inflation

Year	Month	Price of Petrol	CPI	Adjusted Price of Petrol
1976	1	60.5	55.8	108.4
1976	2	60.0	55.9	107.3
1976	3	59.4	56.0	106.1
1976	4	59.2	56.1	105.5
2016	1	196.7	238.1	82.6
2016	2	176.7	237.7	74.3
2016	3	195.8	237.9	82.3
2016	4	213.4	238.9	89.3

TIDSSERIEDATA

EKSEMPEL: BENZINPRISER

Line chart

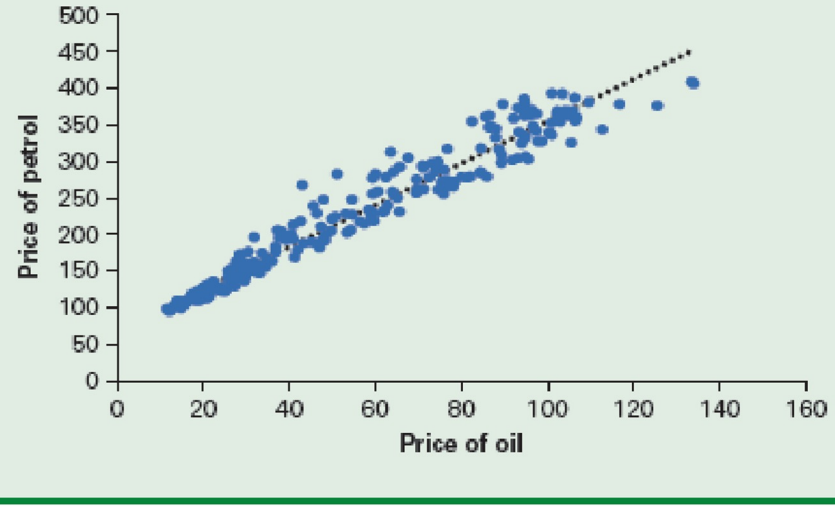


SAMMENHÆNG MELLEM TO VARIABLE

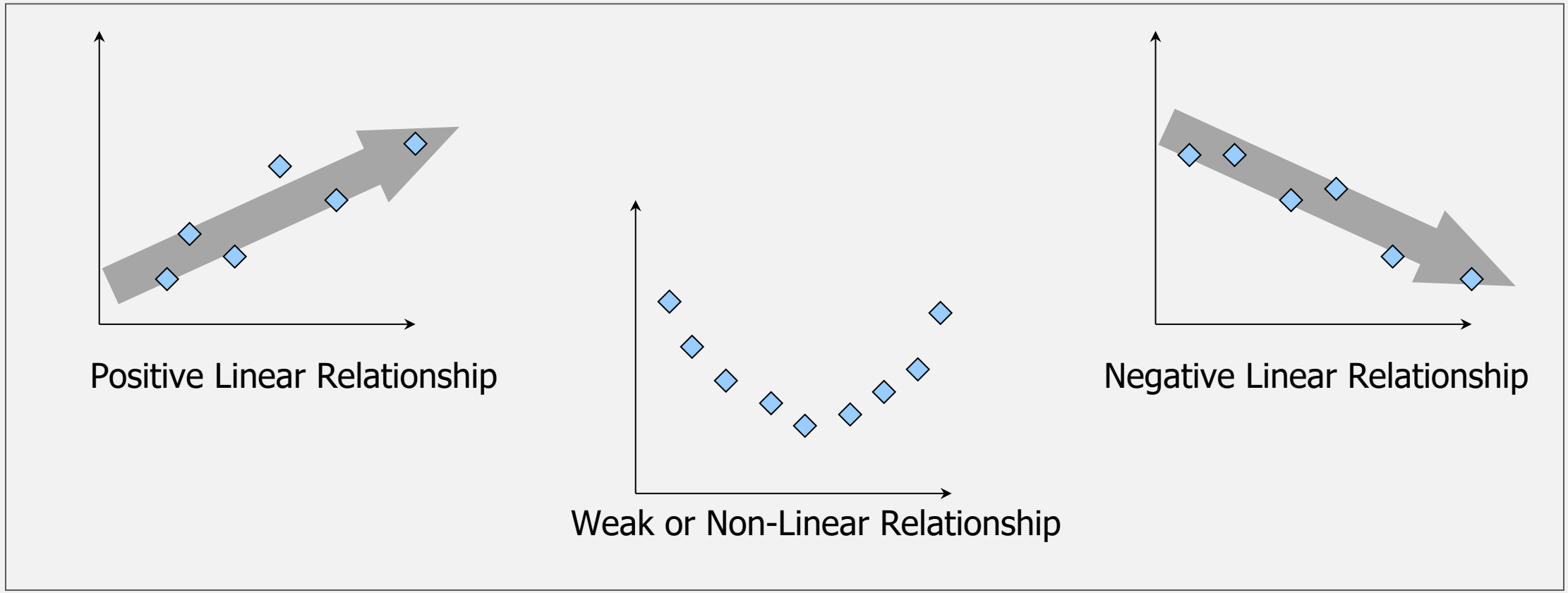
EKSEMPEL: BENZINPRISER

Scatterplot

EXCEL



MØNSTRE I SCATTERPLOTS



OVERBLIK

AALBORG UNIVERSITY

HVILKEN VISUALISERING SKAL VI VÆLGE?

	Interval Data	Nominal Data
Single Set of Data	Histogram	Frequency and Relative Frequency Tables, Bar and Pie Charts
Relationship Between Two Variables	Scatter Diagram	Cross-classification Table, Bar Charts


Q&A


THANK YOU
AALBORG
UNIVERSITY