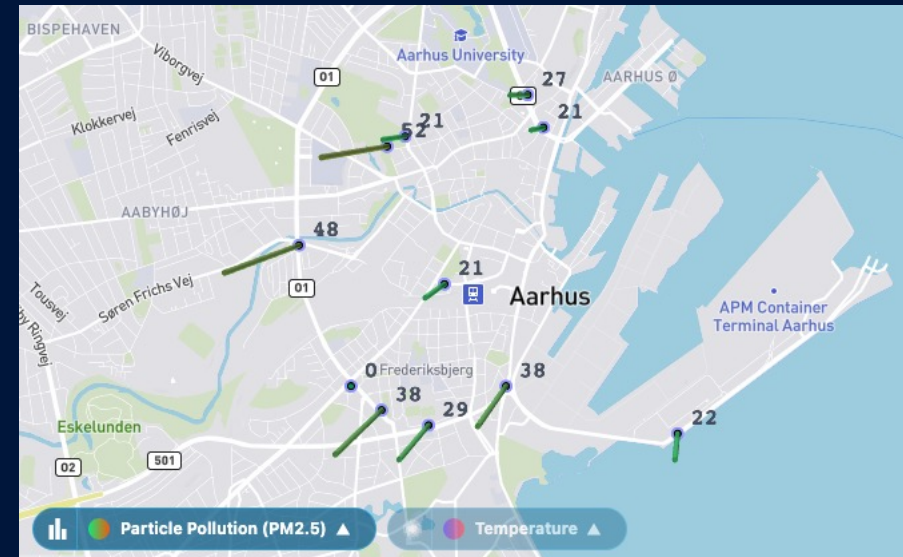
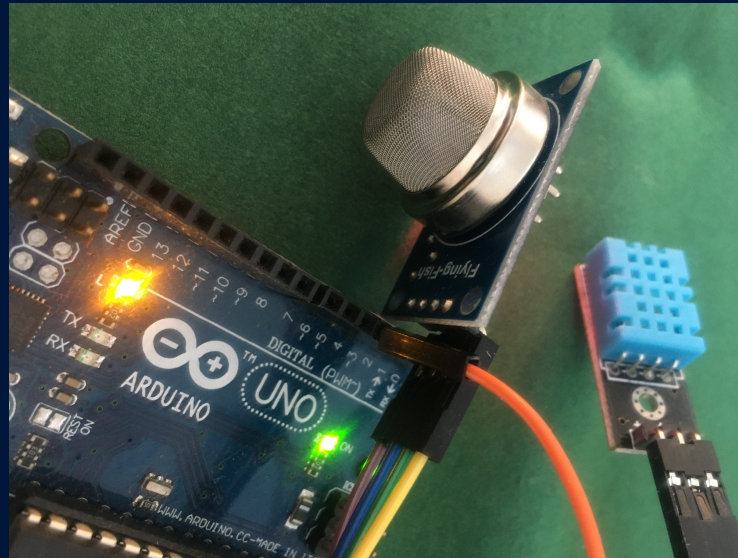


# COPENHAGEN BUSINESS ACADEMY



## DATA ENGINEERING



## FLOW 3 – Foreløbig plan

Uge 10	07.11.2022	intro til dataforespørgsler	Intro - API, Mongo, SQL og webscraping
	08.11.2022	Webscraping	Webscrapping: Case EDC, Bilbasen
	09.11.2022		
	10.11.2022	Webscraping / MongoDB	Mongoddb
	11.11.2022	Webscraping	Præsentation af OLA
Uge 11	14.11.2022	SQL	MySQL og R
	15.11.2022	SQL	MySQL: Case Northwind
	16.11.2022		
	17.11.2022	SQL	MySQL: Case Northwind
	18.11.2022	SQL	Arbejde med OLA
Uge 12	21.11.2022	Cloud Computing	AWS - server og services
	22.11.2022	Cloud Computing	API og Mongo: Casse smart city Aarhus
	23.11.2022		
	24.11.2022	Cloud Computing	Case: PR Flights, R & Mongo på AWS
	25.11.2022	Cloud Computing	ML på AWS
Uge 13	28.11.2022	IOT	Internet of Things
	29.11.2022	IOT	Case: Afstands-sensor
	30.11.2022		
	01.12.2022	IOT	Case:Afstands-sensor
	02.12.2022	OLA	
Uge 14	05.12.2022	Webscraping & NLP	Intro til NLP
	06.12.2022	Webscraping & NLP	Sentiment på boligannoncer
	07.12.2022		
	08.12.2022		
	09.12.2022	Opsamling	Præsentation af OLA, eksamensforberedelse

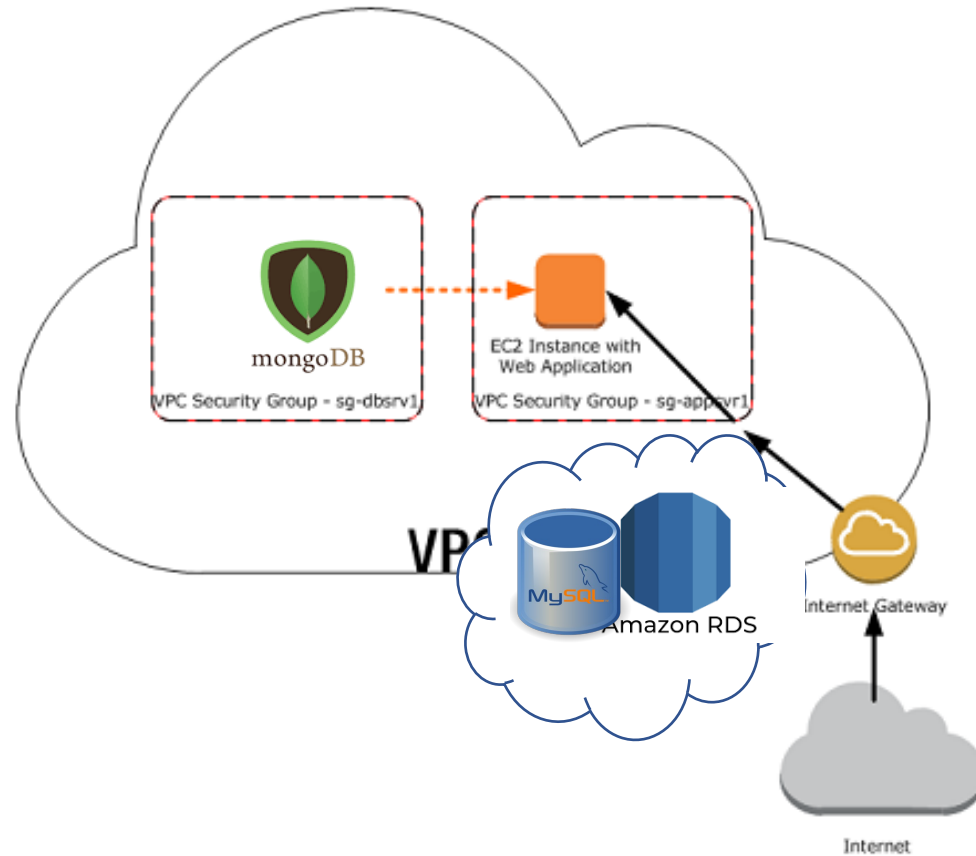
## FLOW 3 – User Stories

- Som ejer af en gammel Volvo vil jeg gerne kunne få en vurdering af hvor meget den er værd således at jeg kan planlægge hvornår jeg skal sælge den
- Som studerende vil jeg gerne vide hvor og hvornår det er bedst at lave en walk-and-talk i Århus i løbet af dagen på en hverdag
- Som underviser vil jeg gerne vide hvor meget den hvide bygning bliver brugt så jeg kan lægge aktiviteter i bygningen når den bliver mindst brugt

## FLOW 3 – Data Engineering

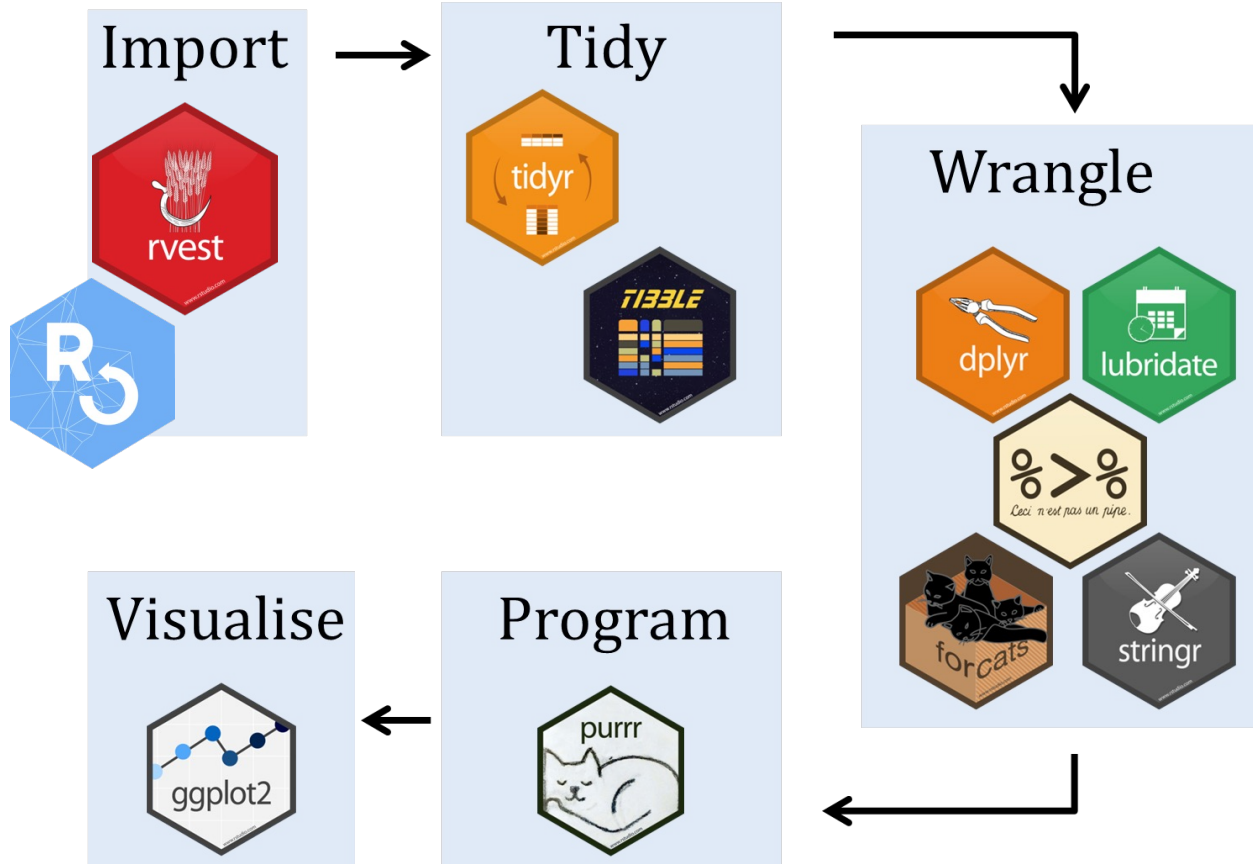
- Oplæg om de tre cases i Flow 3
  - Bilbasen.dk – WebScraping
    - HTML/CSS/JavaScript primer
      - talmedos.com
    - Scrape static (imdb)
    - Scrape dynamic
    - SQL intro
  - cityflow.dk – API
    - API refresher
    - Database-persistence (MongoDB)
    - AWS intro
      - EC2, S3, RDS
  - IoT i den hvide bygning
    - Arduino crash-kursus
    - Prototyping

# Bilbasen – slutmålet



- **MongoDB**
  - Create AWS EC2 instance
  - Get rdp-access
  - Client access
    - mongosh
  - Dev access
    - mongolite
- **SQL**
  - Create AWS RDS
  - Client access
    - MySQL-Workbench
  - Dev access
    - RMariaDB
- **Files**
  - Create AWS S3 bucket
  - Dev access
    - aws.s3

# Bilbasen



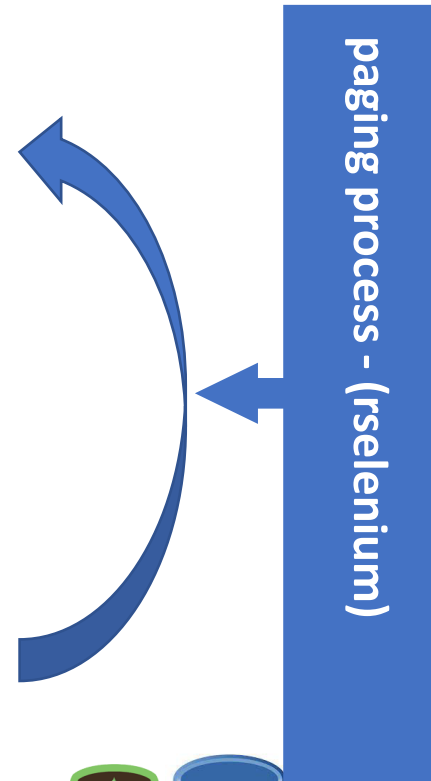
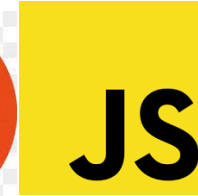
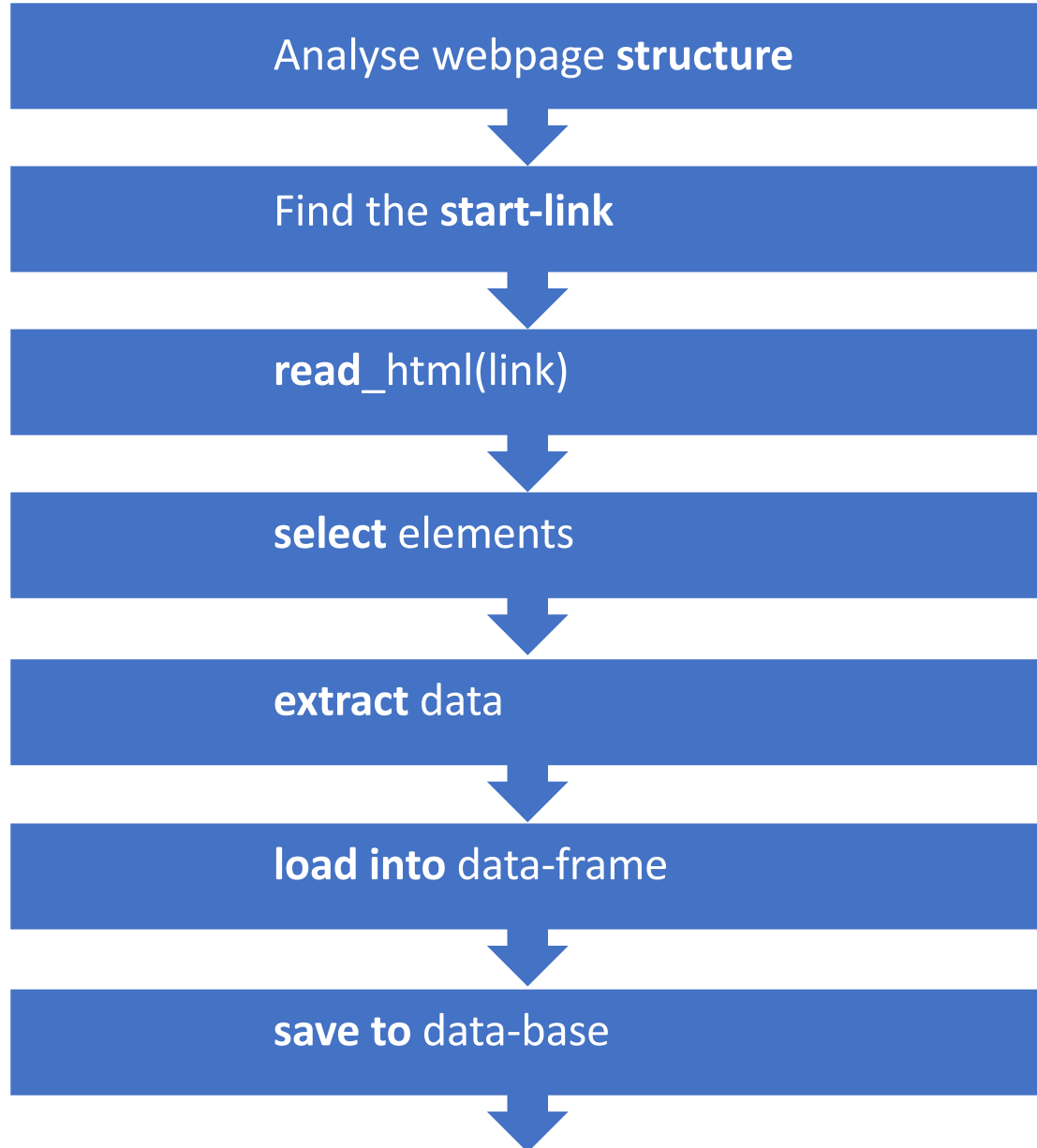
# Bilbasen



# rvest

- Read link with `read_html(link)`
- Select parts of a document
  - CSS selectors: `html_nodes(doc, "table td")`
  - XPath selectors: `html_nodes(doc, xpath = "//table//td")`.
- Extract components with
  - `html_name()` (the name of the tag),
  - `html_text()` (all text inside the tag),
  - `html_attr()` (contents of a single attribute) and
  - `html_attrs()` (all attributes).
- Parse tables into data frames with `html_table()`.
- Navigate around with `html_session()`

# Webscrape – flow



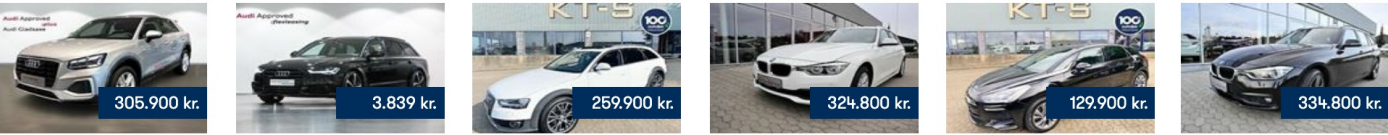






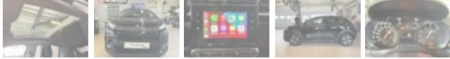



21.538 biler

Flere søgemuligheder

Sælg din bil

Forside > Brugte biler



Alle (21.538) Forhandler (20.095) Privat (1.443)			Vælg kolonne	Km/l (nede)	Gem søgning	
Dato	Mærke og model	Afstand (km)	Km/l (nede)	Kilometer	Modelår	Pris (kr)
	<b>FÅ FORHANDLERPRIS FOR DIN BIL PÅ POINTBI...</b> <b>Citroën C3 1,2 PureTech 110 Triumph EAT6 5d</b> AUTOMATGEAR, Danmarks Bedste Cargo ... Læs mere	- Nordjylland	23,4 km/l	13.000	2020	<b>179.321 kr.</b>
						
	<b>FÅ FORHANDLERPRIS FOR DIN BIL PÅ POINTBI...</b> <b>Citroën C3 1,2 PureTech 110 Triumph EAT6 5d</b> AUTOMATGEAR, Danmarks Bedste Cargo ... Læs mere	- Nordjylland	23,4 km/l	22.000	2020	<b>174.321 kr.</b>
						
	<b>Audi A4 1,8 T 163 Avant quattro 5d</b> Meget Velholdt Audi 1,8T Quattro, Nyserveret Og Undervognsbehandlet. 2 Zone Klima, Fjernb. C.Lås, Fartpilot, Kørecomputer, Infocenter, Udv. Temp. Måler, Sædevarme, El Indst. Førersæde, El-Soltag, 4X El-Ruder, El-Ruder, El-Spejle M/Varme, Parkeringssensor (Bag), Parkeringssensor (For), Armlæn, Kopholder, Læderrat, Tågelygter, Abs, Esp, Servo, Ikke Ryger Ring På 40952040	- Syd- og Sønderjylland	10,9 km/l	129.000	2006	<b>89.900 kr.</b>
						

div.row.listing.listing-plus.bb-listing-clickable

960 x 132

Color #333333  
Font 14px "Walsheim Regular", Arial, sans-serif  
Padding 10px 0px

ACCESSIBILITY

Name  
Role generic  
Keyboard-focusable

ivat (1.453)

Vælg kolonne

Km/l (nedc) ▼

Gem søgning



Afstand (km)	Km/l (nedc)	Kilometer	Modelår	Pris (kr)
--------------	-------------	-----------	---------	-----------

i 150 Style DSG 5d

-

23,4 km/l

8.000

2020

319.900 kr.

cover Pro Navigation

Østjylland

one Klima, Adapt ... Læs mere ▼



Porsche 911 GT3 RS 4,0 Coupé PDK 2d

Alarm, Fjernb. C.Lås, Ratgearskifte, Kørecomputer, Infocenter, Udv. Temp. Måler, El-Ruder, N ... Læs mere ▼

-

7,9 km/l

17.000

2016

1.250.000 kr.

Syd- og Sønderjylland



Audi A6 2.4 V6 Avant Multit 5d

-

10.3 km/l

180.000

2000

26.500 kr.



## Volvo V60 CC 2,0 D3 150 Plus aut. 5d

. GRATIS JUBILÆUMSTILBUD ÅRET UD! • Pakken Består Af: •  
Beskyttelse / Lakforseglingspakke • 1 Års ... Læs mere ▼



Syd- og  
Sønderjylland

21,7 km/l

154.000

2018

269.900 kr.



kvalitetsbiler.dk

```
▶ <div id="fixed-nav-wrapper" style="height: 70.425px;">...</div>
▶ <div class="row listing listing-plus bb-listing-clickable" data-track-content-id="5585622">...</div>
▶ <div class="row listing listing-plus bb-listing-clickable" data-track-content-id="5585236">...</div>
▶ <div class="row listing listing-plus bb-listing-clickable" data-track-content-id="5580937">...</div>
▶ <div class="row listing listing-plus bb-listing-clickable" data-track-content-id="5549817">...</div>
▶ <div class="row listing listing-plus bb-listing-clickable" data-track-content-id="5581019">...</div>
▶ <div class="row listing listing-plus bb-listing-clickable" data-track-content-id="5465392">...</div>
```

```
<li></li>
<li>
  <a class="next" href="https://www.bilbasen.dk/brugt/bil?includeengroscvr=true&pricefrom=0&includeleasing=false
    &page=2"></a> == $0
</li>
```

```
e
}
.
cl
```

# Bilbasen rselenium

- get version of chrome
  - `chrome://version/` -> 106.0.5249.103
- get the matching chrome-driver
  - <https://sites.google.com/chromium.org/driver/downloads>
- Install rselenium-package (and java JDK)
- Instantiate the driver: `rD = rsDriver()`
- get the client: `rcD = rD[["client"]]`
- navigate to site: `rcD$navigate(link)`
- get the page-source: `source=rcD$getPageSource()`

# Bilbasen

- `html = read_html(source[[1]])`

Leave JS (selenium) for CSS & HTML

## Type & make

**a.listing-heading.darkLink** 301.98 × 21.42  
Color ■ #202020  
Font 15px "Walsheim Medium", Arial, sans-s...  
Margin 0px 8px 2px 0px  
ACCESSIBILITY  
Contrast Aa 16.29 ✓  
Name Audi A4 1,8 T 163 Avant quattro 5d  
Role link  
Keyboard-focusable ✓

Audi A4 1,8 T 163 Avant quattro 5d

## Region

**div.col-xs-2.listing-region** 78.33 × 14.28  
Color ■ #333333  
Font 10px "Walsheim Regular", Arial, sans-serif  
Padding 0px 0px 0px 10px  
ACCESSIBILITY  
Contrast Aa 12.63 ✓  
Name  
Role generic  
Keyboard-focusable

Nordjylland

## Beskrivelse

**div.listing-description.expandabl**  
**e-box** 309.98 × 34  
Color ■ #333333  
Font 12px "Walsheim Regular", Arial, sans-serif  
Margin 0px 0px 10px  
ACCESSIBILITY  
Name  
Role generic  
Keyboard-focusable

Mineral Grey-Metallic, Sportline-Model, Navigation Via  
Apple Carplay, Xenonlygter, El-Sportss... [Læs mere](#)

## Km/l, km og year

20,4 km/l 38.000 2018  
**div.col-xs-3.listing-data** 117.5 × 17.14  
Color ■ #333333  
Font 12px "Walsheim Regular", Arial, sans-serif  
Padding 0px 0px 0px 10px  
ACCESSIBILITY  
Name  
Role generic  
Keyboard-focusable

## pris

**div.col-xs-3.listing-price** 117.5 × 20  
Color ■ #002E5C  
Font 14px "Walsheim Medium", Arial, sans-s...  
Padding 0px 10px  
ACCESSIBILITY  
Contrast Aa 13.58 ✓  
Name  
Role generic  
Keyboard-focusable

174.321 kr.

## Car\_id

**div.compare-cars-icon** 28 × 28  
ACCESSIBILITY  
Name Tilføj til sammenligning  
Role generic  
Keyboard-focusable





## Bilbasen – hvilke data?

- make - string
- type - string
- mpg - int
- milage - int
- year - int
- price – int
- EngineType
- width
- length
- height
- cyl
- color - string
- doors - int
- description - text
- Id - int
- Images – link



### Audi A4 1,8 T 163 Avant quattro 5d

Meget Velholdt Audi 1,8T Quattro, Nyserveret Og Undervognsbehandlet. 2 Zone Klima, Fjernb. C.Lås, Fartpilot, Kørecomputer, Infocenter, Udv. Temp. Måler, Sædevarme, El Indst. Førersæde, El-Soltag, 4X El-Ruder, El-Ruder, El-Spejle M/Varme, Parkeringssensor (Bag), Parkeringssensor (For), Armlæn, Kopholder, Læderrat, Tågelygter, Abs, Esp, Servo, Ikke Ryger Ring På 40952040



- 10,9 km/l 129.000 2006 **89.900 kr.**

Syd- og Sønderjylland



Christen Agerley A/S

DATE	TYPE	SPEC	DATE	TYPE	SPEC
CHAR		String (0 - 255)	INT		Integer (-2147483648 to 2147483647)
VARCHAR		String (0 - 255)	BIGINT		Integer (-9223372036854775808 to 9223372036854775807)
TINYTEXT		String (0 - 255)	FLOAT		Decimal (precise to 23 digits)
TEXT		String (0 - 65535)	DOUBLE		Decimal (24 to 53 digits)
BLOB		String (0 - 65535)	DECIMAL		"DOUBLE" stored as string
MEDIUMTEXT		String (0 - 16777215)	DATE		YYYY-MM-DD
MEDIUMBLOB		String (0 - 16777215)	DATETIME		YYYY-MM-DD HH:MM:SS
LONGTEXT		String (0 - 4294967295)	TIMESTAMP		YYYYMMDDHHMMSS
LOBLOB		String (0 - 4294967295)	TIME		HH:MM:SS
TINYINT		Integer (-128 to 127)	ENUM		One of preset options
SMALLINT		Integer (-32768 to 32767)	SET		Selection of preset options
MEDIUMINT		Integer (-8388608 to 8388607)	BOOLEAN		TINYINT(1)