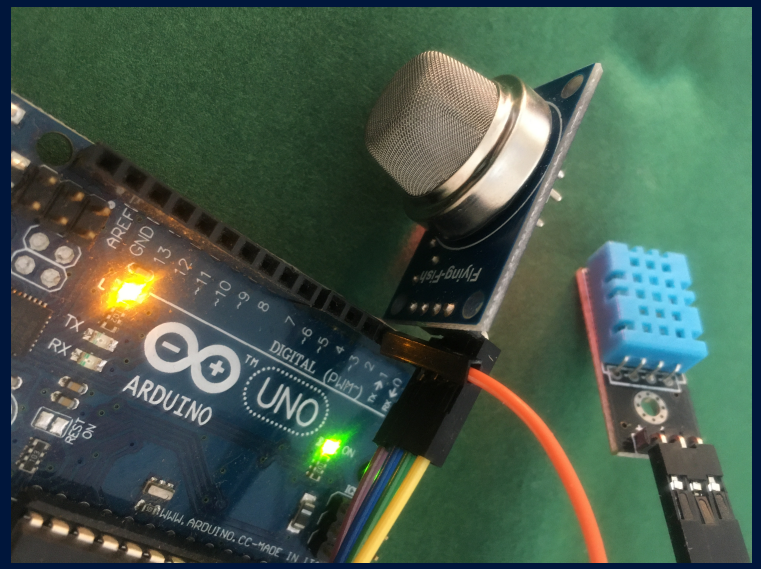
cphbusiness

COPENHAGEN BUSINESS ACADEMY

DATA ENGINEERING



FLOW 3 – Data Engineering

- Oplæg om de to cases i Flow 3
  - cityflow.dk – API
    - Json-dataformat
    - https request
    - python request
    - File-persistence (json, csv)
    - Database-persistence (MongoDB)

  - **Bilbasen.dk – WebScraping**
    - **HTML/CSS/JavaScript primer - jsfiddle**
    - **Scrape static (python bs4) - notebook**
    - Scrape dynamic (python selenium)
    - **MySQL install on server**
    - **MySQL WorkBench**
    - **PyMysql CRUD**

- Tooling
  - GitHub / GitBash
  - Python (Anaconda) & jupyter notebook
  - AWS EC2-instance (Ubuntu med MongoDB og MySQL)
  - Arduino with serial-port driver & HC-204R

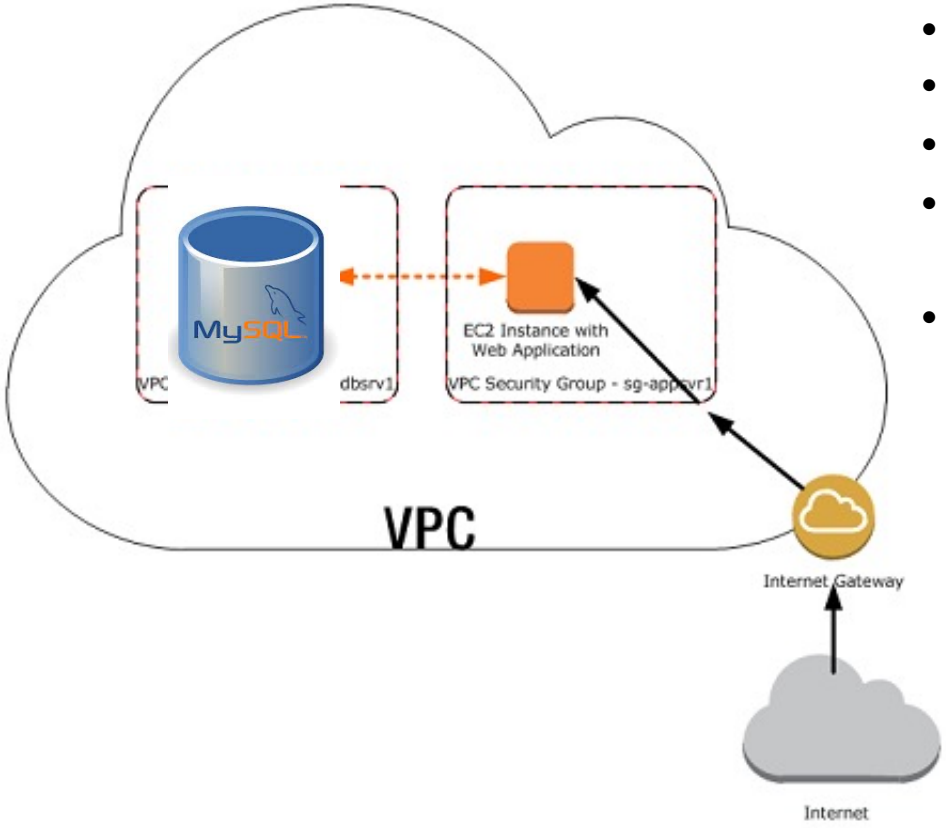
FLOW 3 – User Stories

- 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

- Som studerende vil jeg gerne vide hvor det er bedst at lave en walk-and-talk i Århus i løbet af dagen på en hverdag

- 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

Bilbasen – Save data Infrastructure



- Create EC2 instance (Windows)
- Get rdp-access
- Install mysql
- Client access
  - MySQL-Workbench
- Dev access
  - pymysql

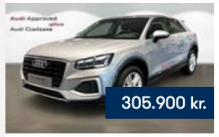
Bilbasen

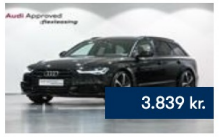
Log ind Menu

Søg på bil 21.538 biler Flere søgemuligheder Sælg din bil

Forside > Brugte biler

Sponsorerede resultater


305.900 kr.


3.839 kr.

259.900 kr.

324.800 kr.

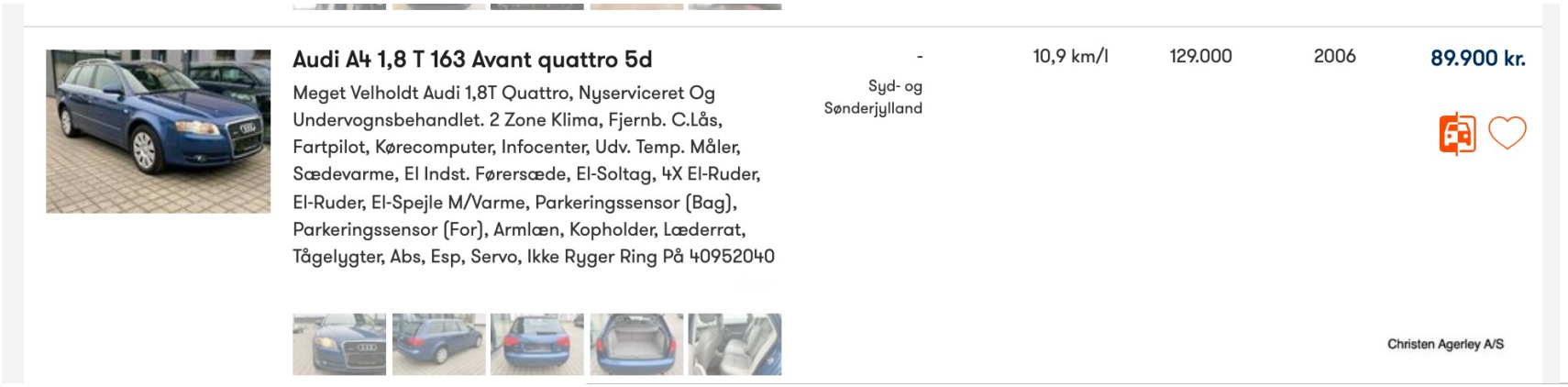
129.900 kr.

334.800 kr.

Alle (21.538) Forhandler (20.095) Privat (1.443) Vælg kolonne Km/l (nedc) Gem søgning

Dato	Mærke og model	Afstand (km)	Km/l (nedc)	Kilometer	Modelår	Pris (kr)
	FÅ FORHANDLERPRIS FOR DIN BIL PÅ POINTBI... Citroën C3 1,2 PureTech 110 Triumph EAT6 5d AUTOMATGEAR, Danmarks Bedste Carg ... Læs mere 	- Nordjylland	23,4 km/l	13.000	2020	179.321 kr. BILMESSEN
	FÅ FORHANDLERPRIS FOR DIN BIL PÅ POINTBI... Citroën C3 1,2 PureTech 110 Triumph EAT6 5d AUTOMATGEAR, Danmarks Bedste Carg ... Læs mere 	- Nordjylland	23,4 km/l	22.000	2020	174.321 kr. BILMESSEN
	Audi A4 1,8 T 163 Avant quattro 5d Meget Velholdt Audi 1,8T Quattro, Nyserviceret 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	89.900 kr.

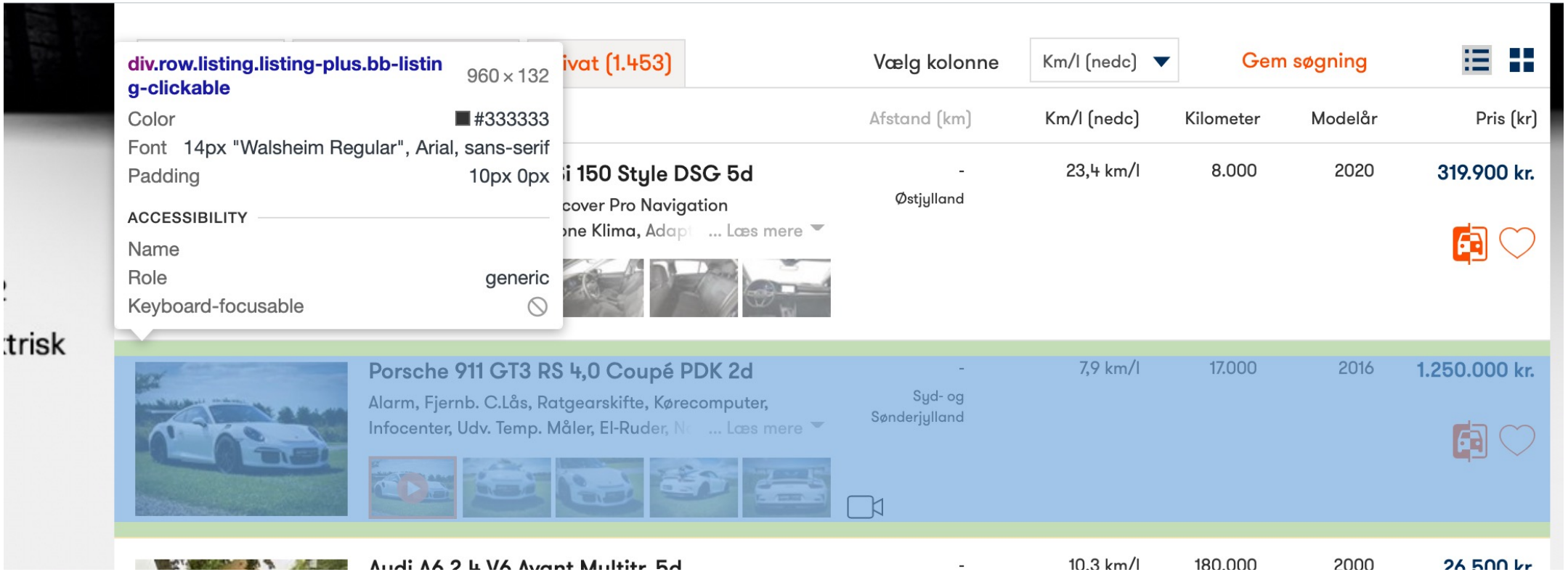
Bilbasen – hvilke data?



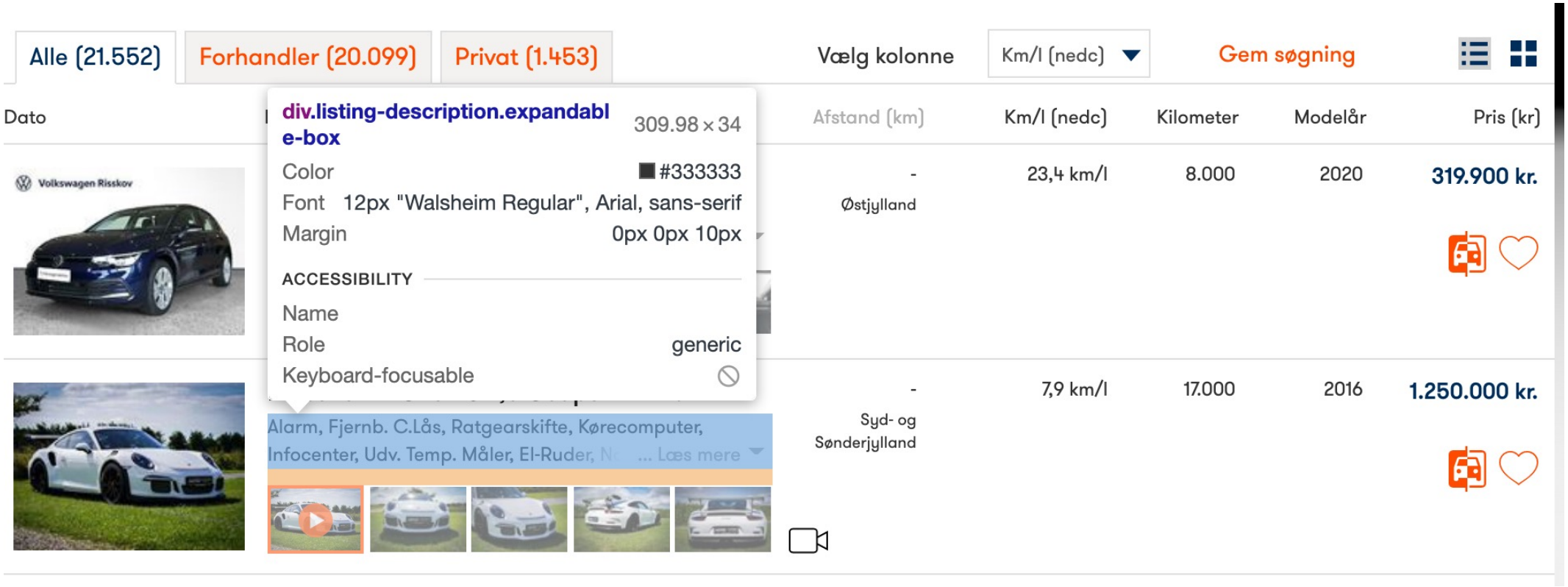
- Make - string
- Type - string
- Kmpl - int
- Km - int
- Year - int
- Price - int
- Color - string
- Doors - int
- Description - text
- Id - int
- Images – link?

DATE TYPE	SPEC	DATA TYPE	SPEC
CHAR	String (0 - 255)	INT	Integer (-2147483648 to 214748-3647)
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
LONGBLOB	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)

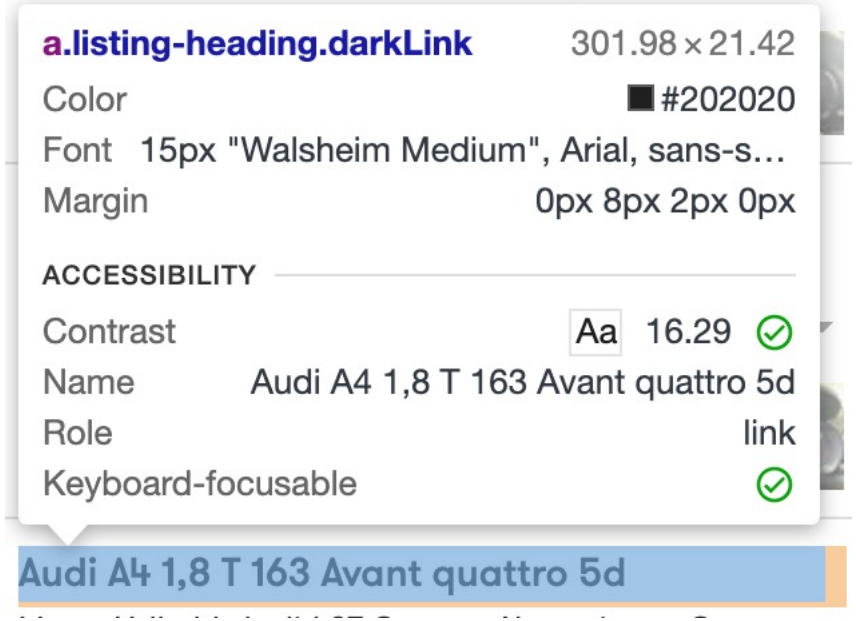
Copyright © mysqltutorial.org. All rights reserved.


div.row.listing.listing-plus.bb-listing-clickable
960 × 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
cover Pro Navigation
ne Klima, Adap
... Læs mere
-
Østjylland
23,4 km/l
8.000
2020
319.900 kr.
trisk
Porsche 911 GT3 RS 4,0 Coupé PDK 2d
Alarm, Fjernb. C.Lås, Ratgearskifte, Kørecomputer,
Infocenter, Udv. Temp. Måler, El-Ruder,
... Læs mere
-
Syd- og Sønderjylland
7,9 km/l
17.000
2016
1.250.000 kr.
Audi A6 2.4 V6 Avant Multitr. 5d
-
10,3 km/l
180.000
2000
26.500 kr.

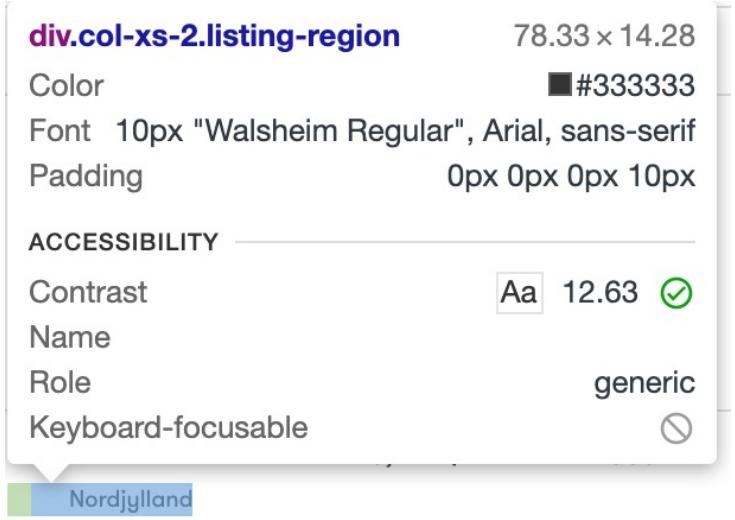
Description



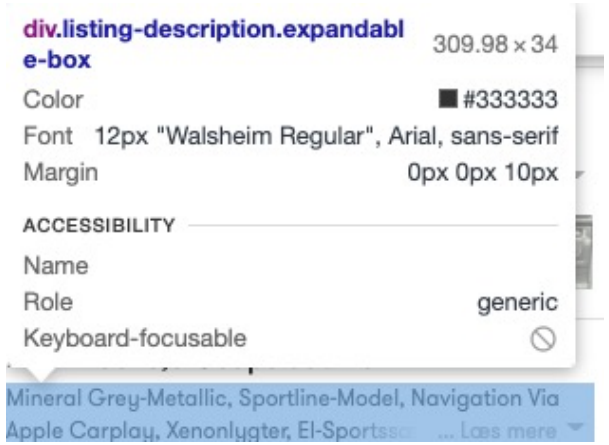
Type & make



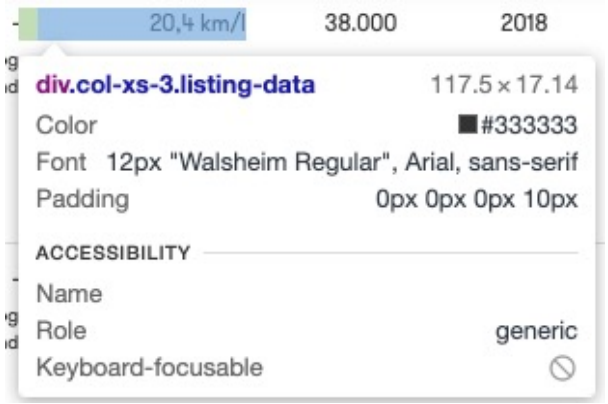
Region



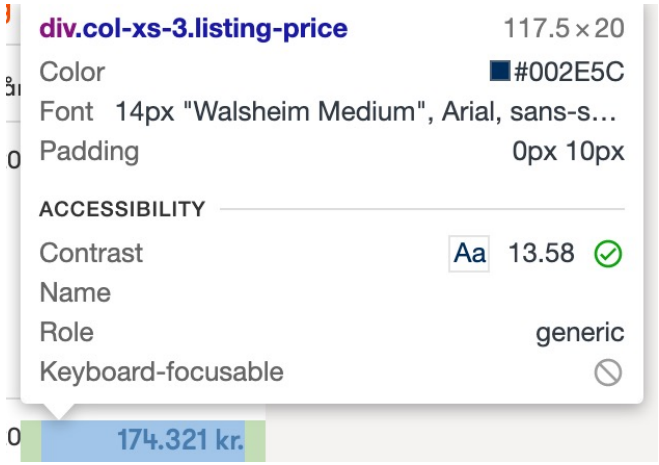
Beskrivelse



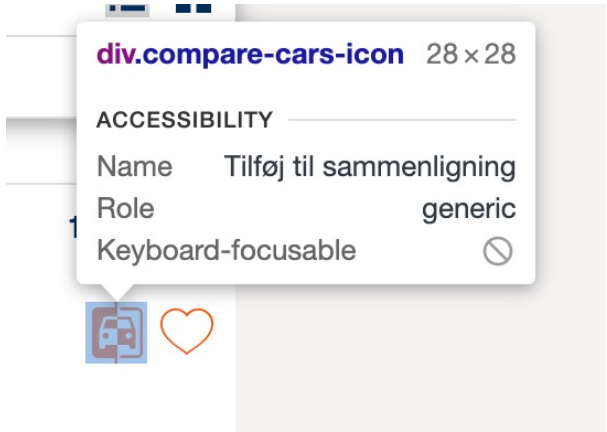
Km/l, km og year



pris



Car_id



Python – Beautiful Soup

Kinds of objects
- `Tag`
  - Name
  - Attributes
    - Multi-valued attributes
- `NavigableString`
- `BeautifulSoup`
- Comments and other special strings

Navigating the tree
- Going down
  - Navigating using tag names
  - `.contents` and `.children`
  - `.descendants`
  - `.string`
  - `.strings` and `stripped_strings`
- Going up
  - `.parent`
  - `.parents`
- Going sideways
  - `.next_sibling` and `.previous_sibling`
  - `.next_siblings` and `.previous_siblings`
- Going back and forth
  - `.next_element` and `.previous_element`
  - `.next_elements` and `.previous_elements`

tabcontent=elements[30].table.find_all('td’)
(bs4.element.ResultSet)

elements[30]
(bs4.element.Tag)

elements[30]).name
(str)

elements[30]).attrs
(dict)

elements[30]).contents
(list)

DRAFT

Searching the tree
- Kinds of filters
  - A string
  - A regular expression
  - A list
  - `True`
  - A function
- `find_all()`
  - The `name` argument
  - The keyword arguments
  - Searching by CSS class
  - The `string` argument
  - The `limit` argument
  - The `recursive` argument
- Calling a tag is like calling `find_all()`
- `find()`
- `find_parents()` and `find_parent()`
- `find_next_siblings()` and `find_next_sibling()`
- `find_previous_siblings()` and `find_previous_sibling()`
- `find_all_next()` and `find_next()`
- `find_all_previous()` and `find_previous()`
- CSS selectors

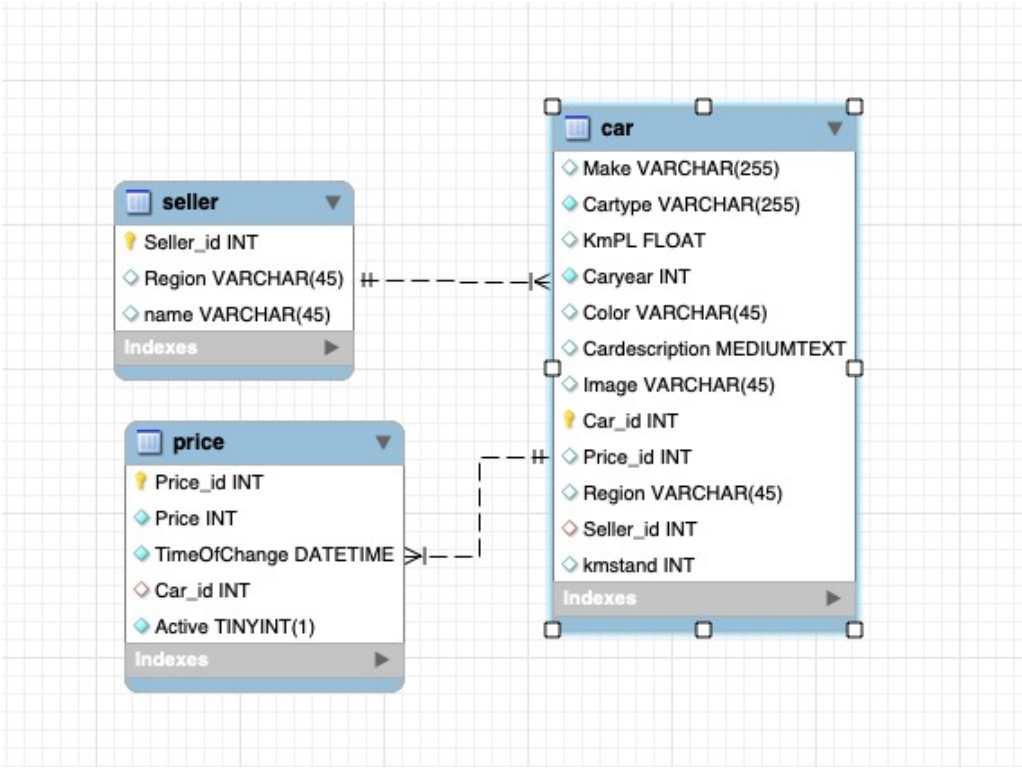
BilBasen – databasens skema i MySQL Workbench

```
13 •  select * from carprice;
14
```

100% 19:13

Result Grid Filter Rows: Search Export:

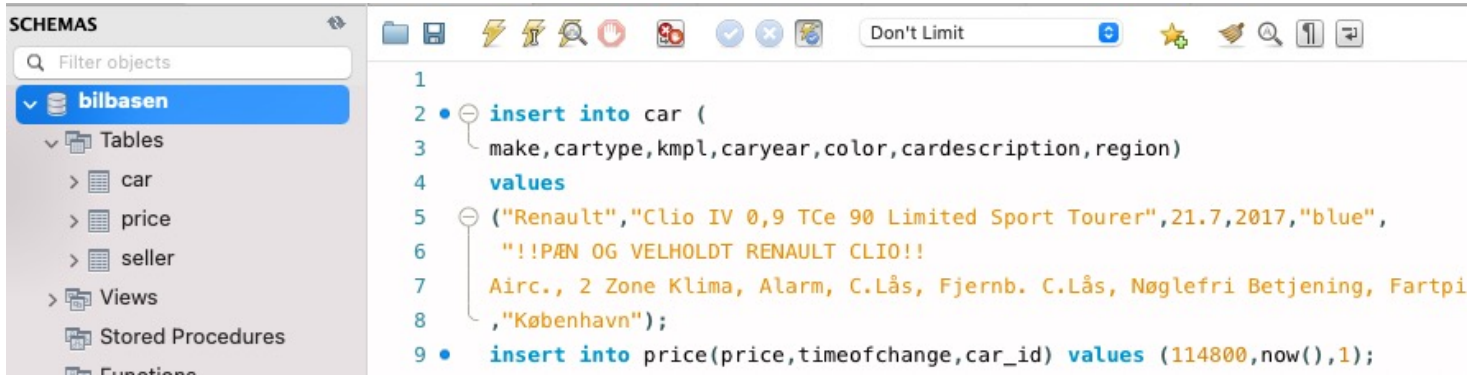
make	cartype	caryear	color	price	timeofchange
Renault	Clio IV 0,9 TCe 90 Limited Sport Tourer	2017	blue	113800	2021-11-21 18:24:57
Porsche	Boxster 2,5	1997	Silver	139800	2021-11-21 18:25:00



Bilbasen & MySQL Primer – manually inserting cars

SQL SELECT STATEMENTS
SELECT * FROM t
SELECT c1,c2 FROM t
SELECT c1,c2 FROM t WHERE *conditions*
SELECT c1,c2 FROM t WHERE *conditions* ORDER BY c1 ASC,c2 DESC

SQL UPDATE DATABASE
INSERT INTO t (c1,c2...) VALUES (v1,v2...)
INSERT INTO t1 (c1,c2...) SELECT c1,c2... FROM t2 WHERE *conditions*
UPDATE t SET c1 = v1, c2 = v2,... WHERE conditions
DELETE FROM t WHERE *conditions*



Cityflow – PyMysql CRUD primer

C	• insert()
R	• find()
U	• update()
D	• remove()

```
In [24]: import pymysql

In [25]: url="3.138.135.109"
         user="thor"
         passwd="thor123"
         db="bilbasen"
         conn=pymysql.connect(host=url,user=user,passwd=passwd,db=db,autocommit=True)

In [ ]: print(conn.get_host_info())
        cursor=conn.cursor()

In [ ]: query="select * from carprice2"
        cursor.execute(query)

In [29]: resSet=cursor.fetchall();
```

```
"""
sqlString="INSERT INTO car (car_id,make,cartype,kmpl,kmstand,caryear,color,cardescription,region) \
VALUES(%s,%s,%s,%s,%s,%s,%s,%s,%s)"
data=(rcid,mmake,ctype,rkml,rkm,cyear,"none",desc,rreg)
"""
sqlString2="INSERT INTO price (price,timeofchange,car_id) VALUES(%s,now(),%s)"
data2=(rprice,rcid)
print(data2)
counter=counter+1
try:
    cursor.execute(sqlString2,data2)
    pass
except conn.Error as e:
    print(e)
    print("Error at "+str(counter)+" Will continue ...")
    continue
```

Scraping dynamically - Selenium

```
#importing modules
from selenium import webdriver
from bs4 import BeautifulSoup

options = webdriver.ChromeOptions()
options.add_argument('--ignore-certificate-errors')
options.add_argument('--incognito')
options.add_argument('--headless')

driver = webdriver.Chrome(chrome_options=options)
source =driver.get('https://analyticsindiamag.com/')
source_code=driver.page_source

soup = BeautifulSoup(source_code,'lxml')
article_block =soup.find_all('div',class_='post-title')

for titles in article_block:
        title =titles.find('span').get_text()
        print(title)
```

DRAFT