

# **Machine learning - DTU**

Rapport

**Anna Louise Hansen**

Fra Udviklings- og Forenklingsstyrrelsen

# Chapter 1

## Part I

### 1.1 Beskrivelse af data

Alle boligejere i Danmark betaler en skat, ejendomsværdiskat, som er baseret på værdien af deres ejendom. Dette vil sige hele ejendommen inkl. grunden som boligen ligger på. For at kunne gøre dette laver den danske stat offentlige ejendomsvurderinger som disse skatter bliver baseret på. Det er derfor vigtigt at disse vurderinger er retvisende og ikke mindst forklarbare, således at en borgers kan forstå hvilke parametre der ligger til grund for ejendomsvurderingen. Til dette projekt har jeg valgt at arbejde med anonymiseret data fra mit arbejde i udviklings- og forenklingsstyrrelsen, hvor jeg til dagligt arbejder med netop dette. Datasættet består af ejendomssalg fra en 6 årig periode. Ud over selve huspriserne består data også af en lang række attributter som beskriver karakteristika ved selve boligen. Det kan f.eks. være tagmateriale, boligens opførelsesår, information om størrelsen af huset og grunden eller bbr koder som dækker over boligens anvendelse. Der ud over består data også af en lang række attributter som fortæller noget om hvor boligens beliggenhed. Det kan f.eks. være boligens koordinater eller information om afstanden til kyst og skov eller afstand til motorvej og jernbane. Data kommer fra en række forskellige registre og offentlige styrrelser som eks. BBR og Styrrelsen for Dataforsyning og Effektivisering.

Til dette projekt vil jeg overordnet set prøve at se hvor godt man kan forudsige ejendomsværdier ud fra salgspriserne fra en 6-årig periode.

Jeg vil med Principal Component Analysis få et overblik over de data der er til rådighed og få et visuelt overblik over attributterne. Herefter vil jeg med en unsupervised learning forsøge at gruppere det data jeg har til at generer yderlige attributer som kan indgå i modellen. Jeg vil her specifikt prøve at se om det er muligt at gruppere salgene i forskellige boligtyper. Jeg vil i samme omgang også forsøge at frasorterer outliers i data med anomaly detection. Herefter vil jeg med regressions model forsøge at kaste lys over projektets overordnede problem ved at forsøge at forudsige huspriserne ud fra salgspriser. I tilfælde af at modellen ikke ikke kan komme med en god prædiktion af en given ejendom vil det være

muligt at denne ejendom bliver manuelt værdiansat af en sagsbehandler. Jeg vil derfor til slut med en classifikation forsøge at estimerer om en ejendom skal ud til manuel sagsbehandling baseret på dens estimerede ejendomsværdi.

## 1.2 Detaljeret beskrivelse af data

Det salgsdata som jeg har valgt at arbejde med dækker i udgangspunktet ‘r nrow(train)‘ observationer med ‘r ncol(train)‘ attributter. Inden jeg går i gang med at kigge på data har jeg valgt at lave en oprydning i data. Det har jeg gjort fordi mange af attributterne bliver i mit daglige arbejde brugt i forbindelse med imødekomme diverse forretningskrav. Desuden dækker observationerne mange forskellige typer af ejendomssalg. Det er en blanding af parcelhussalg, rækkehus-salg, sommerhussalg, salg af ejerlejligheder mm. og ud fra et forretningsmæssigt perspektiv giver det ikke mening at træne en model på alle salg og ejendomstypen vil påvirke salgsprisen. F.eks. vil der på sommerhuse være restriktioner på hvor meget om året man må bo i sommerhuset og der kan være i sommerhusområder være andre regler for hvad man må bruge sin grund til end der er i et parcelhusområde. Jeg har derfor ligeledes valgt at reducerer antallet af observationer således at de kun dækker almindelige parcelhus. Dette er gjort ved kun at beholde alle de ejendomssalg, hvor ejendommen i BBR er registreret med enheds- og bygningsanvendelsen 120.

Herefter er der ‘r nrow(train)‘ observationer tilbage og ‘r ncol(train)‘ attributter.

Der er for en del af attributterne en stor andel af manglende værdier. Det er især i forhold til variable fra BBR, som beskriver forskellige karakteristika ved selve boligen. Her har jeg har valgt at fjerne alle de attributter som har mere end 95% manglende værdier.

|   | attributes | descrete | continous | att    |
|---|------------|----------|-----------|--------|
| :---  | :-----     | :-----   | :-----    | :----- |
| 28   aux.ice_info.adresse.afstand_mellem_soe    | descrete   |          | rat       |        |
| 85   bygning.afloeksforhold                     | descrete   |          | nom       |        |
| 27   aux.ice_info.adresse.afstand_lille_soe     | descrete   |          | rat       |        |
| 88   bygning.supplerendevarme                   | descrete   |          | nom       |        |
| 94   enhed.energiforsyning                      | descrete   |          | nom       |        |
| 87   bygning.opvarmningsmiddel                  | descrete   |          | nom       |        |
| 101   fremskreven_pris                          | continous  |          | int       |        |
| 102   fremskreven_pris_M2                       | continous  |          | int       |        |
| 100   aux.vurbanyttelseskode                    | descrete   |          | nom       |        |
| 34   aux.ice_info.adresse.udsigtslinjer_hav     | descrete   |          | rat       |        |
| 36   aux.ice_info.adresse.udsigtslaengde.samlet | descrete   |          | rat       |        |
| 37   aux.ice_info.adresse.udsigtslinjer_soe     | descrete   |          | rat       |        |
| 1   aux.adresse.etr89koordinatnord              | descrete   |          | int       |        |
| 2   aux.adresse.etr89koordinatoest              | descrete   |          | int       |        |
| 69   EV_NN_M2                                   |            |          |           |        |

|    |   |          |     |
|----|---|----------|-----|
| 72 | aux.adresse.regionkode  | descrete | nom |
| 93 | enhed.badeforhold   | descrete | nom |
| 97 | enhed.toiletforhold   | descrete | nom |
| 91 | bygning.varmeinstallation                                     | descrete | nom |
| 3  | aux.ice_info.adresse.afstand_kyst                             | descrete | rat |
| 4  | aux.ice_info.adresse.afstand_lokalvej                         | descrete | rat |
| 5  | aux.ice_info.adresse.afstand_trafikvej_fordeling              | descrete | rat |
| 6  | aux.ice_info.adresse.afstand_trafikvej_gennemfart             | descrete | rat |
| 7  | aux.ice_info.adresse.afstand_motorvej_motortrafikvej          | descrete | rat |
| 8  | aux.ice_info.adresse.afstand_jernbane_hovedbane               | descrete | rat |
| 9  | aux.ice_info.adresse.afstand_jernbane_any                     | descrete | rat |
| 10 | aux.ice_info.adresse.afstand_jernbane_lokalbane               | descrete | rat |
| 11 | aux.ice_info.adresse.afstand_jernbane_metro                   | descrete | rat |
| 12 | aux.ice_info.adresse.afstand_jernbane_regionbane              | descrete | rat |
| 13 | aux.ice_info.adresse.afstand_jernbane_region_privat           | descrete | rat |
| 14 | aux.ice_info.adresse.afstand_jernbane_s_bane                  | descrete | rat |
| 15 | aux.ice_info.adresse.afstand_stor_skov                        | descrete | rat |
| 16 | aux.ice_info.adresse.afstand_lille_skov                       | descrete | rat |
| 17 | aux.ice_info.adresse.afstand_lokalnet_hoejspaending           | descrete | rat |
| 18 | aux.ice_info.adresse.afstand_regionnet_hoejspaending          | descrete | rat |
| 19 | aux.ice_info.adresse.afstand_transmissionsnet_hoejspaending   | descrete | rat |
| 20 | aux.ice_info.adresse.afstand_lille_vandloeb                   | descrete | rat |
| 21 | aux.ice_info.adresse.afstand_mellem_vandloeb                  | descrete | rat |
| 22 | aux.ice_info.adresse.afstand_stort_vandloeb                   | descrete | rat |
| 23 | aux.ice_info.adresse.afstand_ukendt_vandloeb                  | descrete | rat |
| 24 | aux.ice_info.adresse.afstand_station_metro                    | descrete | rat |
| 25 | aux.ice_info.adresse.afstand_station_s_tog                    | descrete | rat |
| 26 | aux.ice_info.adresse.afstand_station_tog                      | descrete | rat |
| 29 | aux.ice_info.adresse.afstand_stor_soe                         | descrete | rat |
| 30 | aux.ice_info.adresse.afstand_lille_vindmoelle                 | descrete | rat |
| 31 | aux.ice_info.adresse.afstand_mellem_vindmoelle                | descrete | rat |
| 32 | aux.ice_info.adresse.afstand_stor_vindmoelle                  | descrete | rat |
| 33 | aux.ice_info.adresse.areal_samlet_skov                        | descrete | rat |
| 35 | aux.ice_info.adresse.udsigtslaengde_hav                       | descrete | rat |
| 38 | aux.ice_info.adresse.udsigtslaengde_soe                       | descrete | rat |
| 39 | aux.ice_info.bygning.etager.arealflovligeboelseikaelder       | descrete | rat |
| 40 | aux.ice_info.bygning.etager.arealafudnyttetdelaftagetage      | descrete | rat |
| 41 | aux.ice_info.bygning.etager.etagensadgangsareal               | descrete | rat |
| 42 | aux.ice_info.bygning.etager.kaelderareal                      | descrete | rat |
| 43 | aux.ice_info.bygning.etager.samletarealaftagetage             | descrete | rat |
| 44 | aux.ice_info.byzoneareal_opsummeret_delj                      | descrete | rat |
| 45 | aux.ice_info.landzoneareal_opsummeret_delj                    | descrete | rat |
| 46 | aux.ice_info.sommerhuszoneareal_opsummeret_delj               | descrete | rat |
| 47 | aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | descrete | rat |
| 48 | aux.ice_info.jordstykker.vejareal                             | descrete | rat |
| 49 | bolig_alder   | descrete | rat |

|     |  |           |     |
|-----|--|-----------|-----|
| 50  | bolig_areal                                    | descrete  | rat |
| 51  | bygning.antaletager                            | descrete  | int |
| 52  | bygning.arealindbyggetcarport                  | descrete  | rat |
| 53  | bygning.arealindbyggetgarage                   | descrete  | rat |
| 54  | bygning.bebyggetareal                          | descrete  | rat |
| 55  | bygning.bygningenssamledeboligareal            | descrete  | rat |
| 56  | bygning.bygningenssamledeerhvervsareal         | descrete  | rat |
| 57  | bygning.omtilbygningsaar                       | descrete  | int |
| 58  | bygning.opfoerelsesaar                         | descrete  | int |
| 59  | bygning.samletbygningsareal                    | descrete  | rat |
| 60  | enhed.antalbadevaerelser                       | descrete  | rat |
| 61  | enhed.antalvaerelser                           | descrete  | rat |
| 62  | enhed.antalvandskylledetoiletter               | descrete  | rat |
| 63  | enhed.arealtilbeboelse                         | descrete  | rat |
| 64  | enhed.arealtilerhverv                          | descrete  | rat |
| 65  | enhed.enhedenssamledeareal                     | descrete  | rat |
| 66  | etage.areasafudnyttetdelaftagetagetage         | descrete  | rat |
| 67  | etage.etagensadgangsareal                      | descrete  | rat |
| 68  | etage.kaelderareal                             | descrete  | rat |
| 70  | ice_info.min_koteletratiobuff250               | continous | int |
| 71  | lombyg_alder                                   | descrete  | int |
| 73  | aux.ice_info.bb_anvendelse_0                   | descrete  | rat |
| 74  | aux.ice_info.bb_anvendelse_910                 | descrete  | rat |
| 75  | aux.ice_info.bb_anvendelse_920                 | descrete  | rat |
| 76  | aux.ice_info.bb_anvendelse_930                 | descrete  | rat |
| 77  | aux.ice_info.bb_anvendelse_940                 | descrete  | rat |
| 78  | aux.ice_info.bb_anvendelse_950                 | descrete  | rat |
| 79  | aux.ice_info.bb_anvendelse_960                 | descrete  | rat |
| 80  | aux.ice_info.bb_anvendelse_970                 | descrete  | rat |
| 81  | ice_info.bb_per_delgrund                       | descrete  | rat |
| 82  | ice_info.en_per_delgrund                       | descrete  | rat |
| 83  | ice_info.et_per_delgrund                       | descrete  | rat |
| 84  | ice_info.js_per_delgrund                       | descrete  | rat |
| 86  | bygning.bygningensanvendelse                   | descrete  | nom |
| 89  | bygning.tagdaekningsmateriale                  | descrete  | nom |
| 90  | bygning.vandforsyning                          | descrete  | nom |
| 92  | bygning.ydervaeggensmateriale                  | descrete  | nom |
| 95  | enhed.enhedensanvendelse                       | descrete  | nom |
| 96  | enhed.koekkenforhold                           | descrete  | nom |
| 98  | etage.bygningensetagebetegnelse                | descrete  | nom |
| 99  | region_nr                                      | descrete  | nom |
| 103 | aux.ice_info.adresse.afstand_station_any       | descrete  | rat |
| 104 | aux.ice_info.adresse.afstand_soe_any           | descrete  | rat |
| 105 | aux.ice_info.adresse.afstand_vandloeb_any      | descrete  | rat |
| 106 | aux.ice_info.adresse.afstand_hoejspaending_any | descrete  | rat |

Som udgangspunkt vil jeg træne en model som skal være i stand til at kunne prædiktere værdien af et standard parcelhus. Data som der skal trænes på skal derfor også være salg af standard parcelhuse. Data er derfor blevet ensrettet på følgende måde:

Slutteligt er der blevet taget et valg om kun at udvælge de rå attributter som menes at være de vigtigste i forhold til at forudsige værdien af et standad parcelhus.

| attributes  | descrete_continous | attribut |
|---|--------------------|----------|
| :-----  | :-----             | :-----   |
| fremskreven_pris_M2   | continous          | interval |
| bygning.ydervaeeggensmateriale                                | descrete           | nomial   |
| bygning.tagdaekningsmateriale                                 | descrete           | nomial   |
| ombyg_alder   | descrete           | interval |
| enhed.antalvaerelser  | descrete           | ratio    |
| enhed.antalbadevaerelser                                      | descrete           | ratio    |
| enhed.antalvandskylledetoiletter                              | descrete           | ratio    |
| bolig_areal   | descrete           | ratio    |
| bolig_alder   | descrete           | ratio    |
| aux.ice_info.jordstykker.registreretareal_fratrukket_vejareal | descrete           | ratio    |
| aux.ice_info.adresse.afstand_motorvej_motortrafikvej          | descrete           | ratio    |
| aux.ice_info.adresse.afstand_kyst                             | descrete           | ratio    |
| EV_NN_M2  | descrete           | interval |

De to attributter der dækker over tagtypematriale og ydervæggsmateriale er diskrete variable som fordel kan normaliseres med en one-out-of-k transformering.

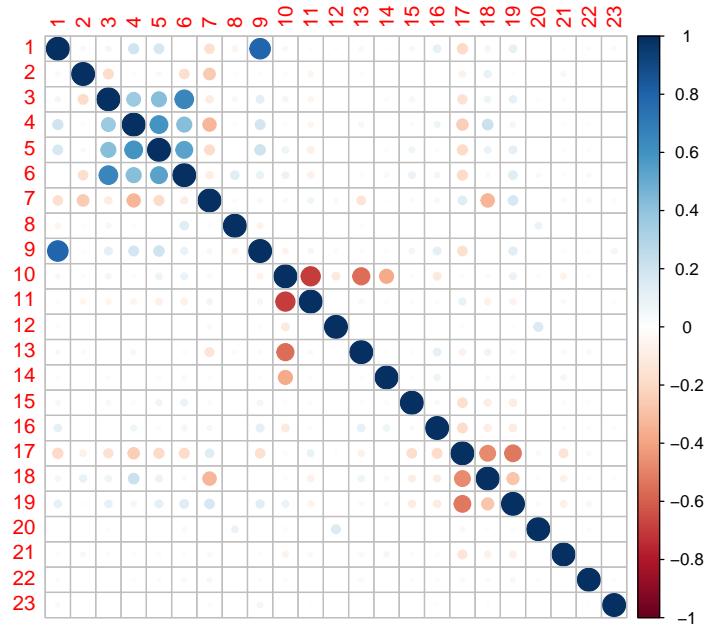
Afstand til kyst og afstand til motorvej er to variable som jeg har valgt at binariserer. Det har jeg da de to features forud for denne rapport er blevet imputeret. For afstand til kyst er afstanden op til 1500 meter målt. Alt herover er imputeret til 1501 meter. Ligeledes er gjort for afstand til motorvej. For at håndtere disse variable har jeg som sagt valgt at binariserer dem. For afstand til kyst har jeg ud fra et forretningsmæssigt synspunkt valgt at en ejendom ligger så tæt på kysten at det har en effekt i dens værdi hvis den ligger inden for 300 meter af kysten. Det samme er gjort for afstand til vej. Her er grænsen dog blot 100 meter.

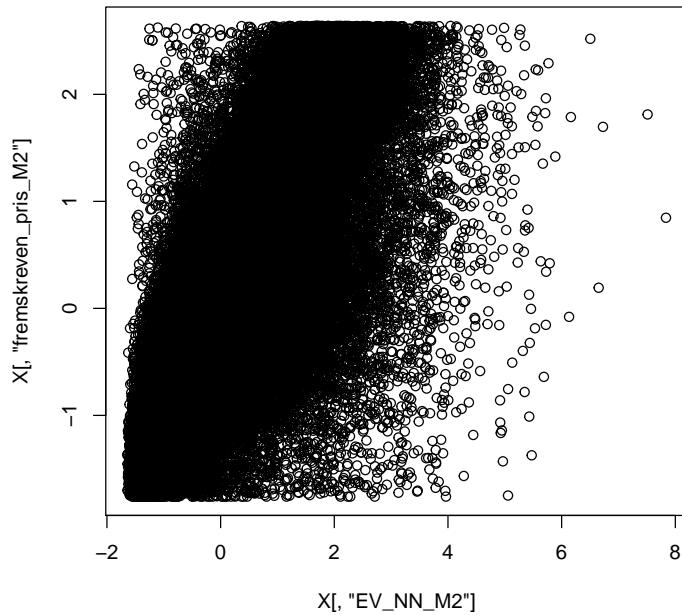
### 1.3 Data visualisering heriblandt Principal Component Analysis (PCA)

De forskellige attributter indeholder ikke værdier på samme skala og jeg har derfor valgt at standardisere data. Selve standardiseringen betyder at data for hver attribut reskaleres således at det får et gennemsnit på 1 med en standardafvigelse på 1.

Efter at have standardiseret data laves der en correlationsanalyse hvor resultaterne af visualiseres i et correlationsplot.

```
[1] "fremskrevens_pris_M2"
[2] "ombyg_alder"
[3] "enhed.antalvaerelser"
[4] "enhed.antalbadevaerelser"
[5] "enhed.antalvandskyllede toiletter"
[6] "bolig_areal"
[7] "bolig_alder"
[8] "aux.ice_info.jordstykke.registreretareal_fratrukket_vejareal"
[9] "EV_NN_M2"
[10] "ydervaegsmateriale_mursten"
[11] "ydervaegsmateriale_gasbeton"
[12] "ydervaegsmateriale_bindingsvaerk"
[13] "ydervaegsmateriale_traebeklaedning"
[14] "ydervaegsmateriale_andet"
[15] "tagtype_builtin"
[16] "tagtype_tagpap"
[17] "tagtype_fibercement"
[18] "tagtype_cementsten"
[19] "tagtype_tegl"
[20] "tagtype_straatag"
[21] "tagtype_andet"
[22] "taet_paa_kyst"
[23] "taet_paa_motorvej"
```

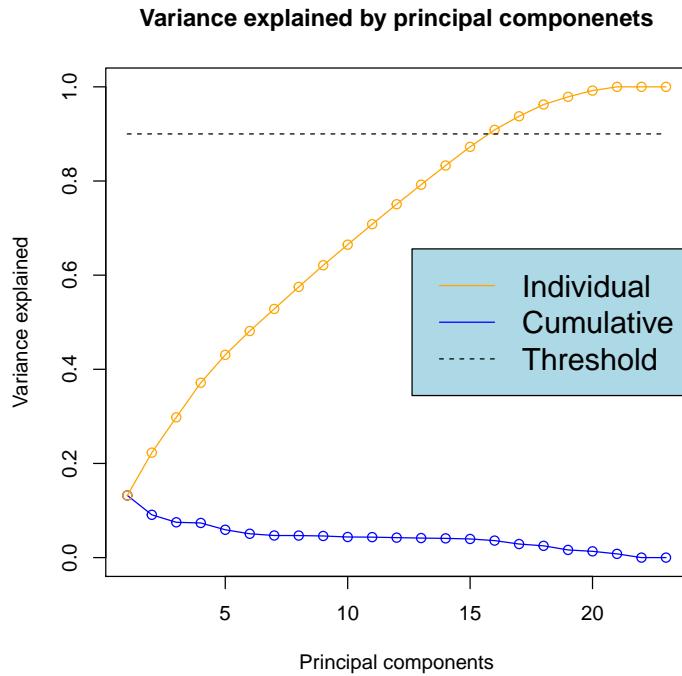




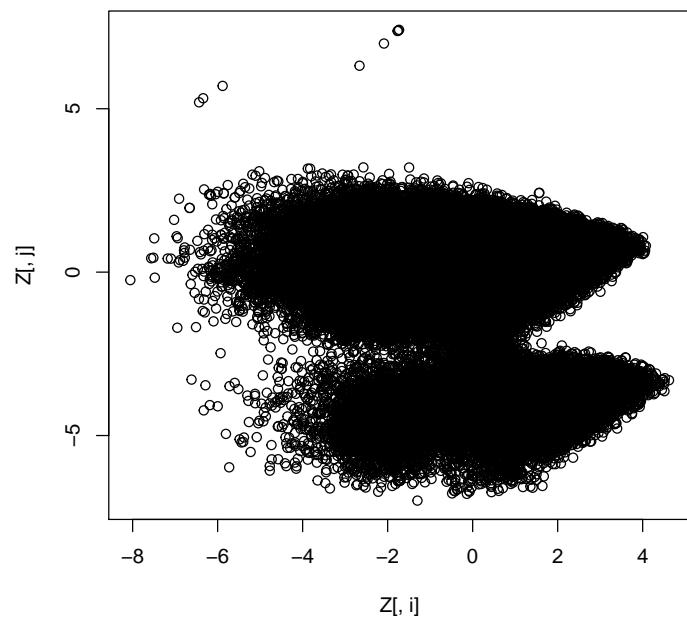
Salgsdata består efter alle datatransformationerne af ‘r N‘ observationer (N) med ‘r M‘ features (M). En af de stærkere correleerde variable i data er EV\_NN\_M2 og de fremskrevne handelspriser. EV\_NN\_M2 er en variabel som dækker over områdeprisen. Den dækker over et vægtet gennemsnit over de nærmeste 15 solgte naboers kvadratmeterpriser. Det giver altså riktig god mening at der er en stor korrelation mellem de solgte naboers kvadratmeter priser og ejendommens egen kvadratmeterpris. Når de to variable plottes mod hinanden ser det således ud:

Der er en stærk korrelation mellem naboernes kvadratmeterpriser og ejendommens egen kvadratmeterpris. Det er fordi at priserne for en given lokation i høj grad er med til at bestemme prisen for en ejendom. Udover områdeprisen er det boligens egen karakteristika som er med til at udgøre den samlede ejendomsværdi.

Som en del af PCA udregnes herefter Singular Value Decomposition (SVD).



De føste 16 principal components kan man forklare 90% af variationen i data. For at komme over 95% skal man have de 18 første komponenter. Der er i alt 23 mulige komponenter. Det som det umiddelbart synes at betyde er at der ikke kan reduceres mange attributter væk uden at det betyder at man mister variation. PCA analysen kan bruges til at visualiserer data, feature extraction og compression. Principal component analyse kan bruges til at reducerer dimensionerne i data.



## **Chapter 2**

## **Part II**

## **Chapter 3**

### **Part III**