Task:

Write 5 programs («machine learning models») to classify or predict a target feature for new data points.

At least one of the models need to be a neural network.

The goal is to predict the Sales Price for new data points.

The models must work on the original data set.

We need to perform preprocessing on the data before we use it in our models.

Preprocessing must be done in Python and included in the Jupyter Notebook, in order for it to work on the original dataset.

We must write a text < 3000 words explaining why we have done what we have done and what our findings were.

Features in Original Dataset   
(red = removed for all models.   
Yellow = maybe remove, or remove for only some models):

- index. Begynner på 4 og går oppover  
- Borough. «strøk». 5 strøk: Manhattan(1), Bronx(2), Brooklyn (3), Queens (4), staten Island(5)  
- Neighbourhood. Alphabet city, Chelsea, Chinatown, Civic Center, Clinton, East Village,   
 Fashion, Financial, Flatiron, Gramercy, Greenwich Village, Harlem-Central, …  
- Building Class Category. Rentals, Coops, Condo – Rental, Condo – Elevator, … .  
- Tax Class at Present.   
- # Block. Nummer på blokken som blir solgt.   
- # Lot  
- Ease-Ment.   
- Building Class At Present. A = leilighet I bra stand. B = leilighet I middles stand. C = leilighet I   
 “dårlig» stand. Etc.   
- Address  
- Apartment Number   
- # Zip Code  
- # Residential units. Antall leiligheter i bygningen. Flere leiligheter = (høy)blokk.   
- # Commercial Units. Antall leiligheter som blir brukt til å generere «profitt», som regel utleie.   
- # Total Units. Sier hvor mange leiligheter totalt som er i blokken.   
- Land Square Feet. Hvor mye land som blir solgt.   
- Gross Square Feet. Hvor mye “brukbar” plass det er i leiligheten (eller på tomten?)  
- # Year Built. Når leiligheten ble bygget.   
- Tax Class at Time of Sale  
- Building Class at Time of Sale. Hvor “bra” leiligheten + beliggenheten etc er estimert å være   
 ved salg. <https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html>.  
- Sale Price. Det eiendommen ble solgt for. Fjern rader med – og 0.   
- Sale Date. Når eiendommen ble solgt.

Preprocessing:

- Fjern «Index». Det sier oss vel ikke noen ting? Har ikke noe med salg å gjøre  
- Fjern «Tax Class at Present», for det har ingenting å si for den endelige salgsprisen.   
- Fjern «#Lot». Denne Featuren sier noe om hvor mye land som blir solgt – men det gjør også  
 land square feet og gross square feet. Unødvendig å ta den med.   
- Fjern Ease-Ment, da denne er tom.   
- Fjern Block. Det er bare nummeret på blokken og har ingenting å si for salgsprisen.   
- Fjern Sale Date? Dette har vel ikke så mye med salgsprisen å gjøre?   
- Kanskje fjern Commerical, eller bruk subset >= 5? hvor relevant er utleieaktiviteter for endelig  
 salgspris av en leilighet? (bare et få hundre av totalt 80,000 leiligheter har Commercial units.)  
- Fjern total units dersom vi fjernet Commercial Units, da Total Units – Commercial Units =  
 Residental units   
- Hent ut representative data fra hvert nabolag eller zipcode – hvis hvert nabolag har 100   
 registrerte adresser, bruk bare x (x = 20? 30? 40? 50?) av de. Dette er for å minimere risikoen  
 for overfitting. (pre - processing eller for hver modell?)

Idéer til datasett:

* Regn ut average apartment pris per blokk. Lagre og fjern Apartment Number-feature. Dette reduserer mengden data og minimerer risiko for overfitting, og kan fortsatt hjelpe til pris-estimering.   
   \* Er dette relevant? Om en blokk er på average 2,4 mill per leilighet og en annen er på  
   average 2.5m per leilighet, er det ikke likevel det overordnede området blokken   
   befinner seg i, som bestemmer pris?
* Vi trenger til enhver tid bare enten neighborhood eller bare zipcode? Addresse blir for snevert og mange leiligheter har ikke adresse.

Modell 1 – Linear Classification:

INCLUDE:  
- Borough  
- Neighborhood   
- Building Class Category  
- #Residental Units  
- Land Square Feet  
- Gross Square Feet  
- #Year Built  
- Tax Class at Time of Sale  
- Building Class at Time of Sale  
- Sale Price

EXCLUDE:   
- #Zip Code  
- #Commercial units  
- #Total Units

IDK: <-- merge these to Average Block Price?  
- Address?   
- Apartment Number?  
- Block?

Modell 2 – kNN:

INCLUDE:  
-

EXCLUDE:  
-

Modell 3 – Decision Trees:

INCLUDE:  
-

EXCLUDE:  
-

Modell 4 – Linear Regression:

INCLUDE:  
-

EXCLUDE:  
-

Modell 5 – Forward Feeding Neural Network:

INCLUDE:  
-

EXCLUDE:  
-