mean, median, standard deviation und percentiles von allen nuerischen columns im datensatz können mittels data.describe() gesehen werden.

Column header und daten typen mit data.info.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Doppelte entries herausfinden kann man mittels data.duplicated und weiters kann man einstellen ob man diese behalten will mit z.B. keep=‘last‘

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

reduce outliers with quantiles

max in range mit dataframe["column name"].quantile(0.95)

min in range mit dataframe["column name"].quantile(0.05)

die daten zwischen den 2 werten werden behalten. Falls der Datensatz auf outlier angewiesen ist dann soll dies anders gelöst werden oder gefragt werden aber im Durchschnitt sollte das so passen.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Correkte Typen checkt man mit: mit data.dtypes

No int64

X1 transaction date float64

X2 house age float64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Gibt es Wertunterschiede wie 1g und 0,0001 kg:

Kann mittels min und max wert herausgefunden werden oder mit einem richtigen typ wie is\_numeric\_dtpe von pandas falls es um string vs int geht z.B.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

wurde es bearbeitet?

muss man nachfragen kann man schlecht im code herausfinden

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

gibt es fehlende werte?

Ist mit data.isnull() beantwortbar weiters kann dropna die columns droppen die null werte beinhalten. Man sollte aber im besten fall nachfragen warum ein wert fehlt und ob das gewollt ist.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Wie groß ist das datenset und ist das genug?

data.shape gibt uns die größe:

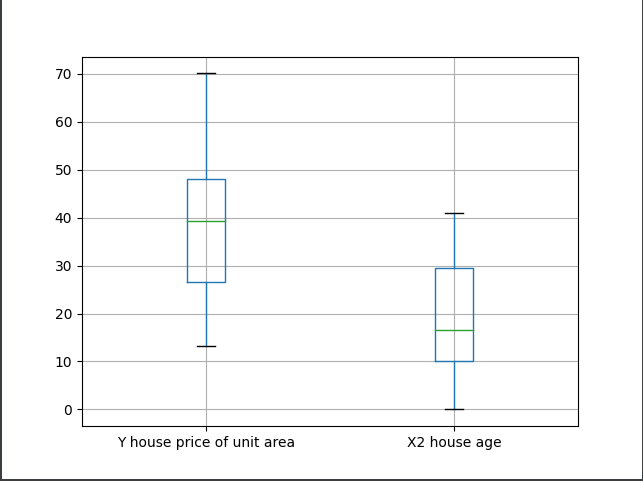
(414, 8)

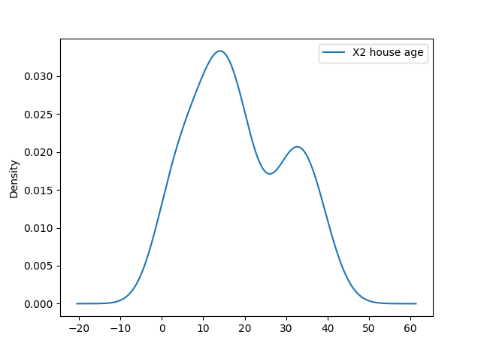
Obs genug ist kommt drauf an für was es verwendet wird

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Plots:

Ein Boxplot zeigt uns die quantile der Werterange einer column.



Ein density plot kann gut genutzt werden um, den zusammenhang zweier columns zu erkennen. 

Ein bar plot ist gut darin outlier zu visualisieren.

