



HATÉKONYSÁGNÖVEDELÉS ADATELŐKÉSZÍTÉS SEGÍTSÉGÉVEL

Data science képzés
Adatelőkészítés

2024.10.21.

Jónás Dániel,
data scientist



Áttekintés

adatelokeszites

```
X = df.loc[:, X_cols]
y = df.loc[:, y_col]
model = LinearRegression()
pred = cross_val_pred(model, X, y, cv=10)
```

kiertekeles

```
model.fit(X, y)
```

uj adat → adatelokeszites

```
X_uj = df_uj.loc[:, X_cols]
pred = model.predict()
```

Áttekintés

adatelokeszites

```
X = df.loc[:, X_cols]
```

```
y = df.loc[:, y_col]
```

```
model = LinearRegression()
```

```
pred = cross_val_pred(model, X, y, cv=10)
```

kiertekeles

Áttekintés

adatelokeszites

```
X = df.loc[:, X_cols]
```

```
y = df.loc[:, y_col]
```

```
model = GradientBoostingRegressor()
```

```
pred = cross_val_pred(model, X, y, cv=10)
```

kiertekeles

Áttekintés

adatelokeszites

```
X = df.loc[:, X_cols]
```

```
y = df.loc[:, y_col]
```

```
model = GradientBoostingRegressor(n_estimators=100, max_depth=3)
```

```
pred = cross_val_pred(model, X, y, cv=10)
```

kiertekeles

Áttekintés

adatelokeszites

```
X = df.loc[:, X_cols]
```

```
y = df.loc[:, y_col]
```

```
model = GradientBoostingRegressor(n_estimators=100, max_depth=3)
```

```
pred = cross_val_pred(model, X, y, cv=10)
```

kiertekeles



ADATELŐKÉSZÍTÉS

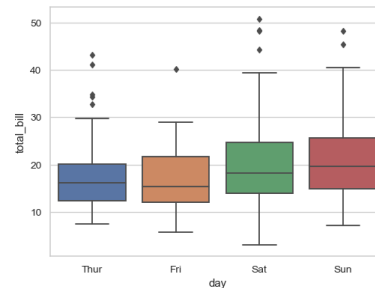
ADATELŐKÉSZÍTÉS LÉPÉSEI

- Hiányzó értékek kezelése

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0

ADATELŐKÉSZÍTÉS LÉPÉSEI

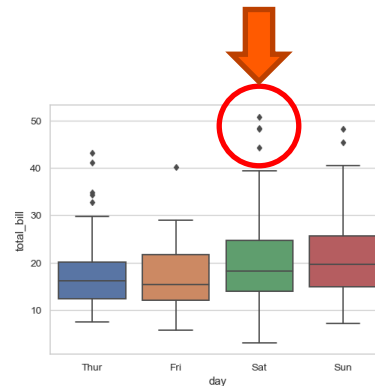
- Hiányzó értékek kezelése
- Kiugró értékek kezelése



Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0

ADATELŐKÉSZÍTÉS LÉPÉSEI

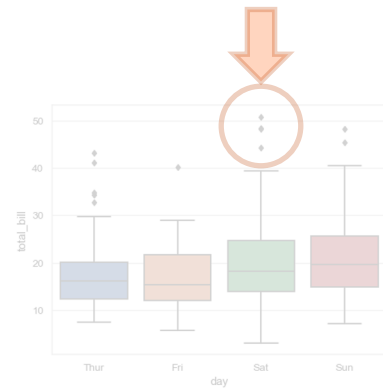
- Hiányzó értékek kezelése
- Kiugró értékek kezelése



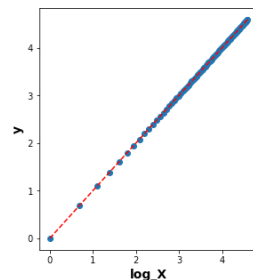
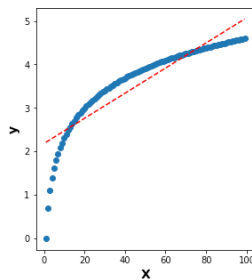
Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0

ADATELŐKÉSZÍTÉS LÉPÉSEI

- Hiányzó értékek kezelése
- Kiugró értékek kezelése
- Értékkészlet transzformáció



Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0

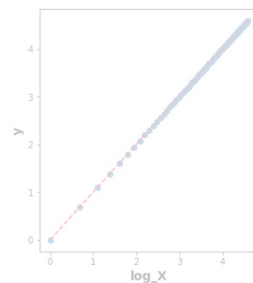
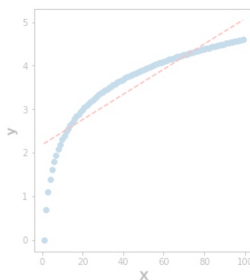


ADATELŐKÉSZÍTÉS LÉPÉSEI

- Hiányzó értékek kezelése
- Kiugró értékek kezelése
- Értékkészlet transzformáció
- Új jellemzők generálása



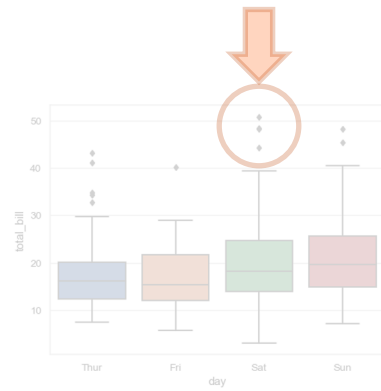
Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0



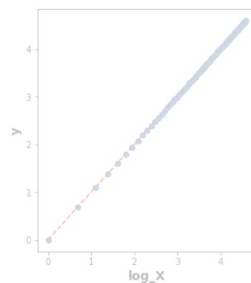
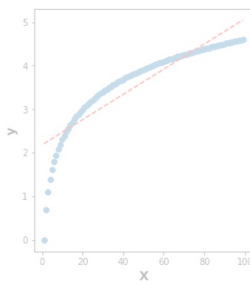
property_area	price_created_at	m2_price
35.0	9.5	2.714286e+05
35.0	11.2	3.200000e+05
50.0	14.5	2.900000e+05
50.0	15.3	3.060000e+05
50.0	17.9	3.580000e+05
...
30.0	11.0	3.666667e+05
45.0	20.5	4.555556e+05
40.0	8.5	2.125000e+05
65.0	18.9	2.907692e+05
40.0	42.0	1.050000e+06

ADATELŐKÉSZÍTÉS LÉPÉSEI

- Hiányzó értékek kezelése
- Kiugró értékek kezelése
- Értékkészlet transzformáció
- Új jellemzők generálása



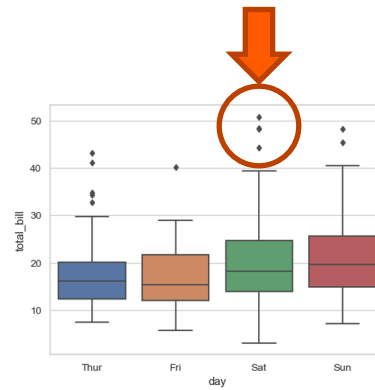
Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0



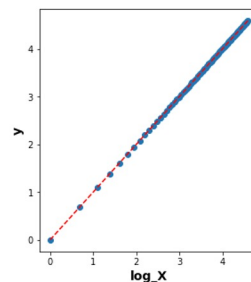
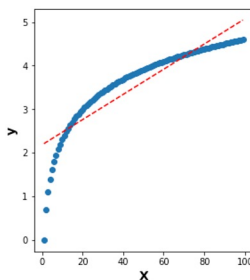
property_area	price_created_at	m2_price
35.0	9.5	2.714286e+05
35.0	11.2	3.200000e+05
50.0	14.5	2.900000e+05
50.0	15.3	3.060000e+05
50.0	17.9	3.580000e+05
...
30.0	11.0	3.666667e+05
45.0	20.5	4.555556e+05
40.0	8.5	2.125000e+05
65.0	18.9	2.907692e+05
40.0	42.0	1.050000e+06

ADATELŐKÉSZÍTÉS LÉPÉSEI

- Hiányzó értékek kezelése
- Kiugró értékek kezelése
- Értékkészlet transzformáció
- Új jellemzők generálása
- Bemeneti változók kiválasztása



Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0




property_area	price_created_at	m2_price
35.0	9.5	2.714286e+05
35.0	11.2	3.200000e+05
50.0	14.5	2.900000e+05
50.0	15.3	3.060000e+05
50.0	17.9	3.580000e+05
...
30.0	11.0	3.666667e+05
45.0	20.5	4.555556e+05
40.0	8.5	2.125000e+05
65.0	18.9	2.907692e+05
40.0	42.0	1.050000e+06

HIÁNYZÓ ÉRTÉKEK KEZELÉSE

- Oszlop/sor törlése

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0




Bath
1.0
1.0
2.0
2.0

HIÁNYZÓ ÉRTÉKEK KEZELÉSE

- Oszlop/sor törlése
- Értékek feltöltése (átlag, medián, okos módszer)

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0



Bath
1.0
1.0
2.0
2.0

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0




Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
2.0	2.0

HIÁNYZÓ ÉRTÉKEK KEZELÉSE


- Oszlop/sor törlése
- Értékek feltöltése (átlag, medián, okos módszer)
- Értékek feltöltése és plusz oszlopok generálása

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0



Bath
1.0
1.0
2.0
2.0

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0



Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
2.0	2.0

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0



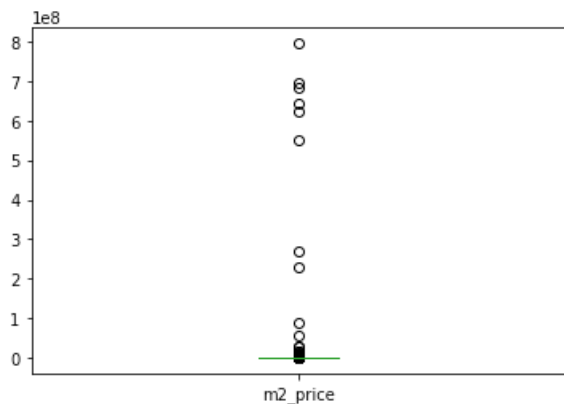
Bed	Bath	Bed_was_missing
1.0	1.0	FALSE
2.0	1.0	FALSE
3.0	2.0	FALSE
2.0	2.0	TRUE

KIUGRÓ ÉRTÉKEK KEZELÉSE

- Milyen kiugró érték? (valós vs. hibás adat)
- Sor/oszlop törlése
- Értékek kicserélése (pl. átlag, maximális érték)

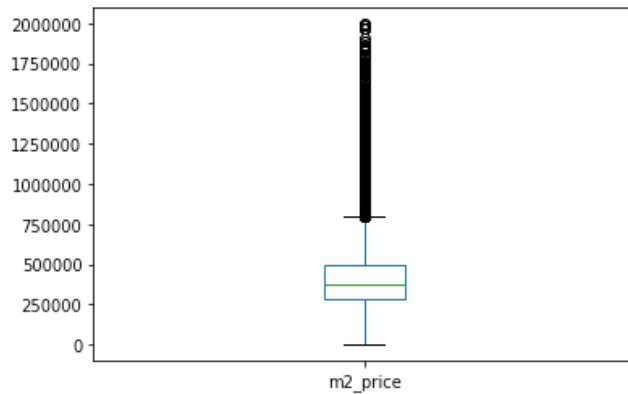
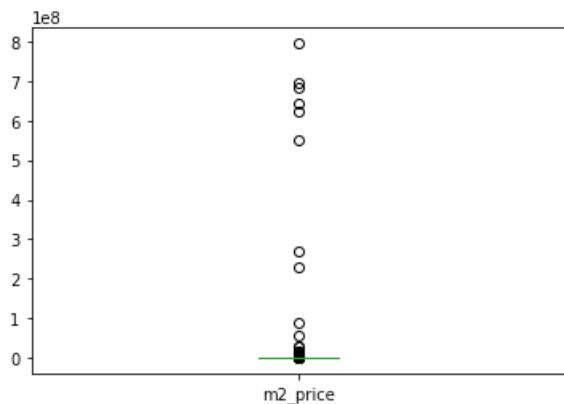
KIUGRÓ ÉRTÉKEK KEZELÉSE

➤ Milyen kiugró érték? Hibás?



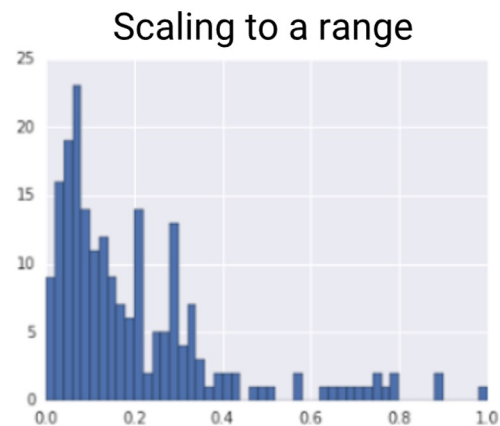
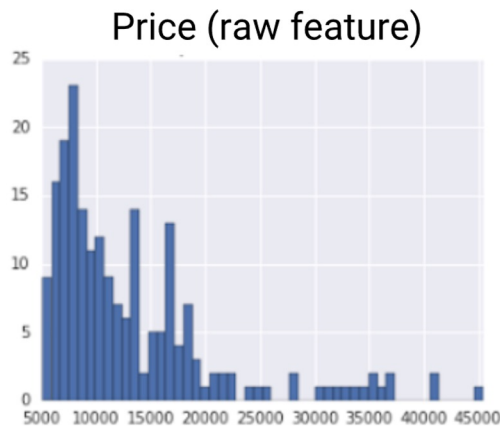
KIUGRÓ ÉRTÉKEK KEZELÉSE

➤ Milyen kiugró érték? Hibás? Valós?



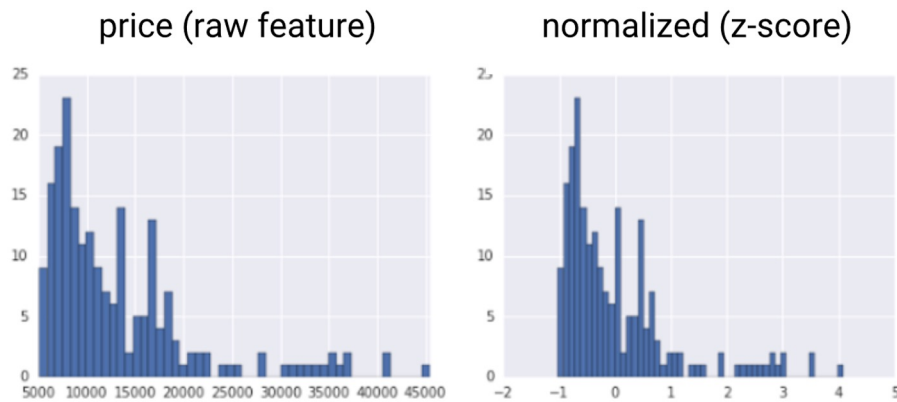
ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- **Két érték közé skálázás (normalizálás)**



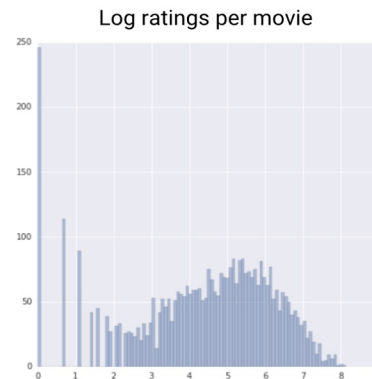
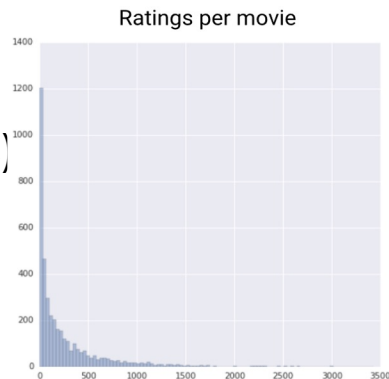
ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- **Standardizálás/ z-score normalizálás**
(pl. 0 várható értékű, 1 szórású)



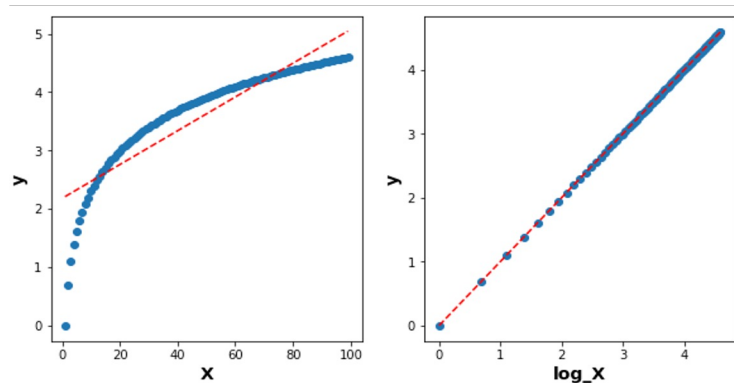
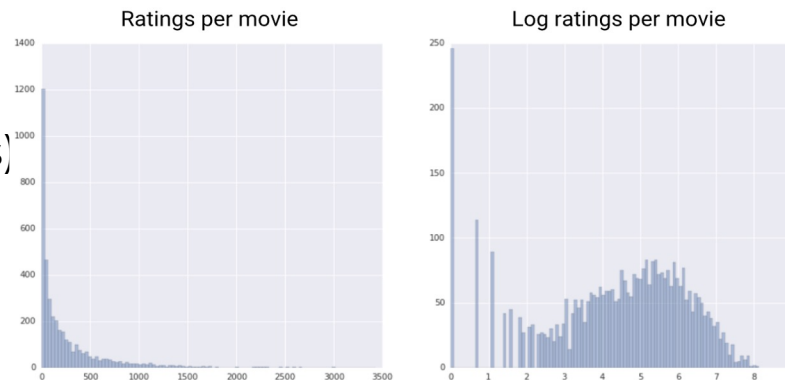
ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- Standardizálás/ z-score normalizálás (pl. 0 várható értékű, 1 szórású)
- **Nem lineáris transzformáció**



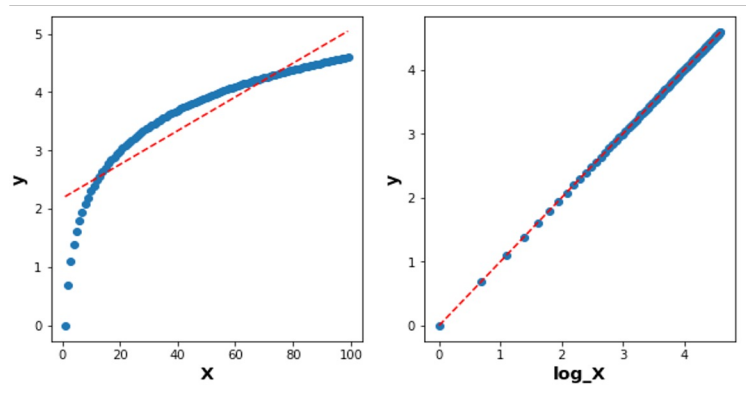
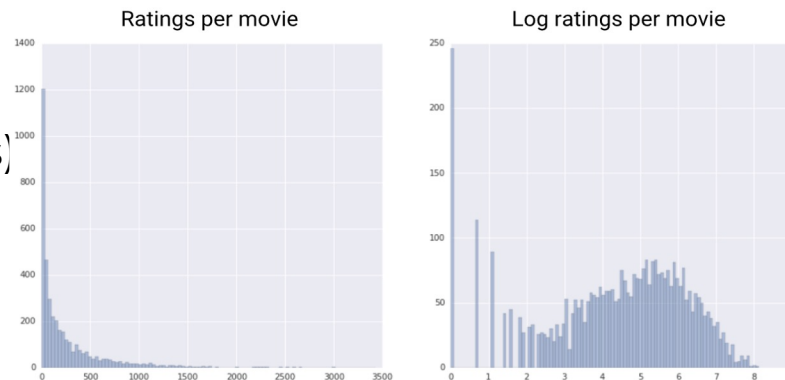
ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- Standardizálás/ z-score normalizálás (pl. 0 várható értékű, 1 szórású)
- **Nem lineáris transzformáció**



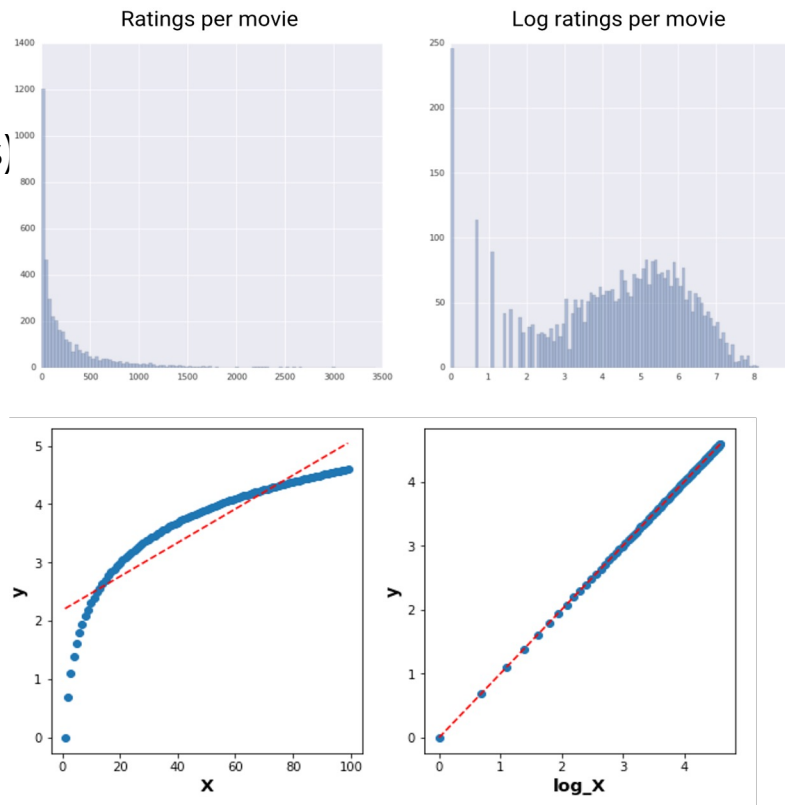
ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- Standardizálás/ z-score normalizálás (pl. 0 várható értékű, 1 szórású)
- **Nem lineáris transzformáció**
- - $\text{Pred} = b_0 + b_1 * x_1$



ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- Standardizálás/ z-score normalizálás (pl. 0 várható értékű, 1 szórású)
- **Nem lineáris transzformáció**
- - $\text{Pred} = b_0 + b_1 * x_1$
- - $\text{Pred} = b_0 + b_1 * (\log(x_1))$

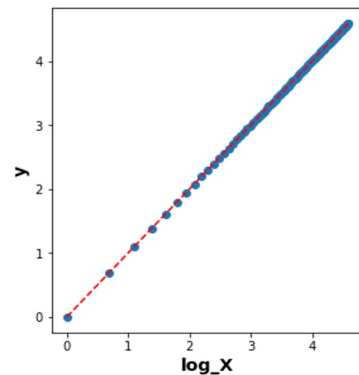
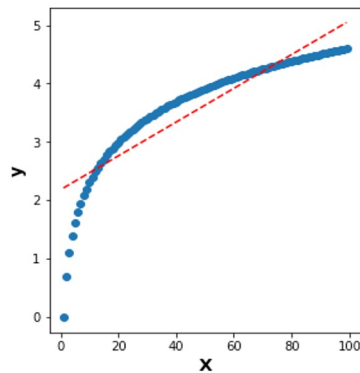


ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

➤ $\text{Pred} = b_0 + b_1 * x_1$

age	
37	
49	
32	
38	
51	
...	
32	
43	
55	
46	
35	

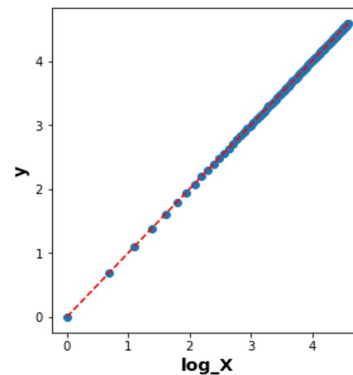
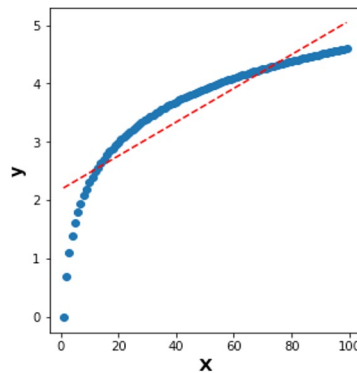
age	age_log
37	3.610918
49	3.891820
32	3.465736
38	3.637586
51	3.931826
...	...
32	3.465736
43	3.761200
55	4.007333
46	3.828641
35	3.555348



ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

➤ $\text{Pred} = b_0 + b_1 * x_1$ → $x_1 = \text{age_log}$

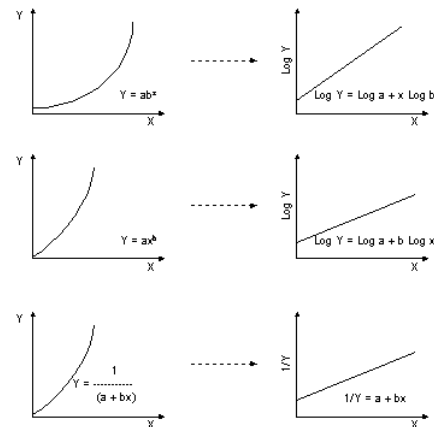
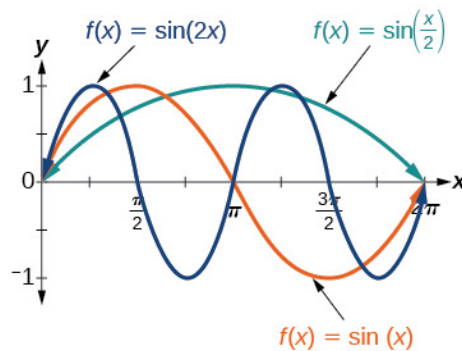
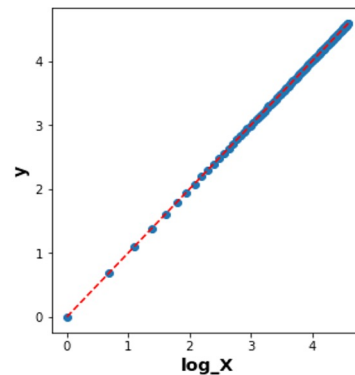
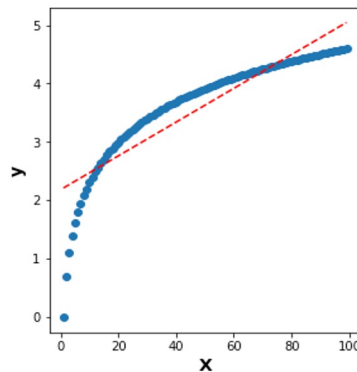
age	age	age_log
37	37	3.610918
49	49	3.891820
32	32	3.465736
38	38	3.637586
51	51	3.931826
...
32	32	3.465736
43	43	3.761200
55	55	4.007333
46	46	3.828641
35	35	3.555348



ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

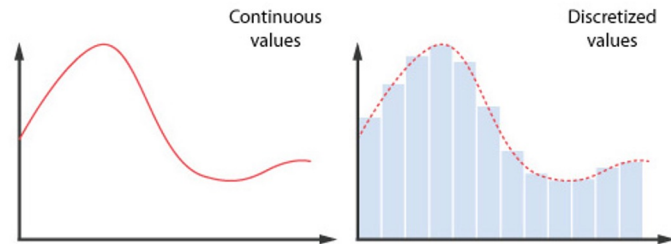
➤ $\text{Pred} = b_0 + b_1 * x_1$ → $x_1 = \text{age_log}$

age	age	age_log
37	37	3.610918
49	49	3.891820
32	32	3.465736
38	38	3.637586
51	51	3.931826
...
32	32	3.465736
43	43	3.761200
55	55	4.007333
46	46	3.828641
35	35	3.555348



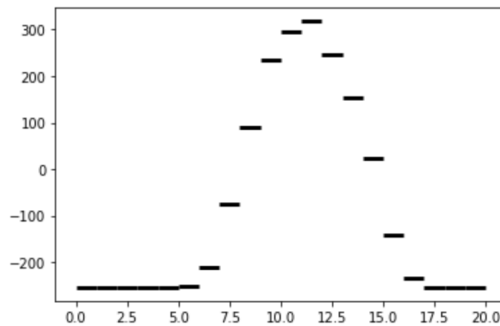
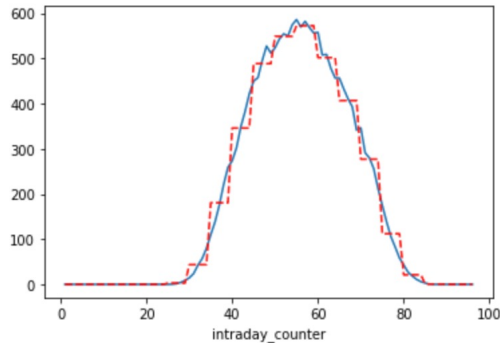
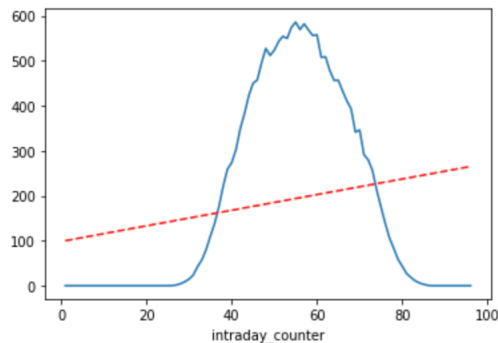
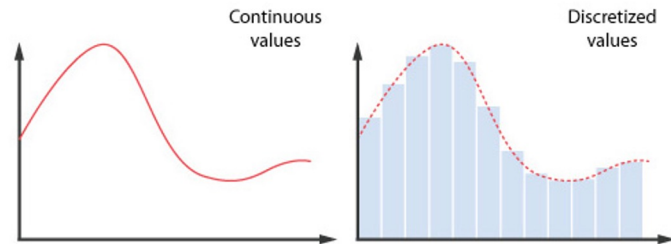
ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- Standardizálás/ z-score normalizálás (pl. 0 várható értékű, 1 szórású)
- Logaritmikus skálázás
- **Diszkrétizáció**



ÉRTÉKKÉSZLET TRANSZFORMÁLÁSA

- Két érték közé skálázás (normalizálás)
- Standardizálás/ z-score normalizálás (pl. 0 várható értékű, 1 szórású)
- Logaritmikus skálázás
- **Diszkrétizáció**



ÚJ JELLEMZŐK GENERÁLÁSA

FEATURE EXTRACTION

- Az eredeti adatot a megfelelő formátumba hozzuk

ÚJ JELLEMZŐK GENERÁLÁSA

FEATURE EXTRACTION

- Az eredeti adatot a megfelelő formátumba hozzuk
- Sorrendezhető szöveges változók átalakítása (*Label Encoding*)



The diagram illustrates the process of Label Encoding. It shows two tables connected by a blue arrow pointing from left to right. The left table, titled 'Breakfast', contains five rows of categorical data: 'Every day', 'Never', 'Rarely', 'Most days', and 'Never'. The right table, also titled 'Breakfast', shows the same data after encoding: 'Every day' is represented by the number 3, 'Never' by 0, 'Rarely' by 1, 'Most days' by 2, and the final 'Never' by 0.

Breakfast
Every day
Never
Rarely
Most days
Never

Breakfast
3
0
1
2
0

Label Encoding

ÚJ JELLEMZŐK GENERÁLÁSA

FEATURE EXTRACTION

- Az eredeti adatot a megfelelő formátumba hozzuk
- Sorrendezhető szöveges változók átalakítása (*Label Encoding*)
- Kategorikus változók átalakítása (*One-Hot Encoding*)

Breakfast	Breakfast
Every day	3
Never	0
Rarely	1
Most days	2
Never	0

Label Encoding

Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	1	0

One-Hot Encoding

ÚJ JELLEMZŐK GENERÁLÁSA

FEATURE EXTRACTION

- Az eredeti adatot a megfelelő formátumba hozzuk
- Sorrendezhető szöveges változók átalakítása (*Label Encoding*)
- Kategorikus változók átalakítása (*One-Hot Encoding*)

Breakfast	Breakfast
Every day	3
Never	0
Rarely	1
Most days	2
Never	0

Label Encoding

Color	Red	Yellow
Red	1	0
Red	1	0
Yellow	0	1
Green	0	0
Yellow	0	1

One-Hot Encoding

ÚJ JELLEMZŐK GENERÁLÁSA

FEATURE EXTRACTION

- Az eredeti adatot a megfelelő formátumba hozzuk
- Sorrendezhető szöveges változók átalakítása (*Label Encoding*)
- Kategorikus változók átalakítása (*One-Hot Encoding*)

Breakfast	Breakfast
Every day	3
Never	0
Rarely	1
Most days	2
Never	0

Label Encoding

Color	Red	Yellow
Red	1	0
Red	1	0
Yellow	0	1
Green	0	0
Yellow	0	1

One-Hot Encoding

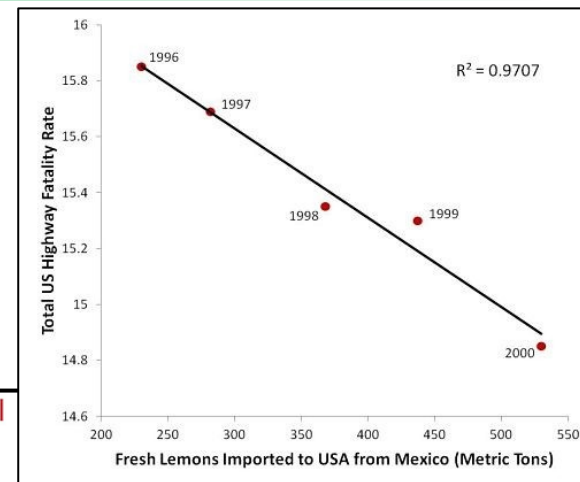
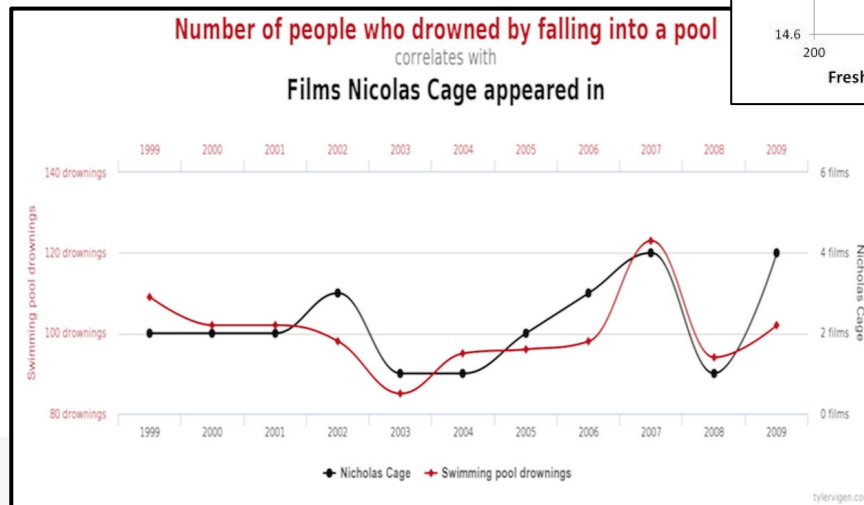
- Új változókat generálunk

property_area	price_created_at	m2_price
35.0	9.5	2.714286e+05
35.0	11.2	3.200000e+05
50.0	14.5	2.900000e+05
50.0	15.3	3.060000e+05
50.0	17.9	3.580000e+05
...
30.0	11.0	3.666667e+05
45.0	20.5	4.555556e+05
40.0	8.5	2.125000e+05
65.0	18.9	2.907692e+05
40.0	42.0	1.050000e+06

JELLEMZŐK KIVÁLASZTÁSA

MIÉRT?

- A zaj csökkentésével védekezhetünk a túltanulás ellen
- A felesleges változók eltávolításával növelhető a pontosság
- Tanítási idő csökkentése



JELLEMZŐK KIVÁLASZTÁSA

MIÉRT?

- A zaj csökkentésével védekezhetünk a túltanulás ellen
- A felesleges változók eltávolításával növelhető a pontosság
- Tanítási idő csökkentése

HOGYAN?

- Korreláció alapján legjobb k darab változó
- Rekurzívan (forward/backward selection)
- Modellre bízunk (pl. Lasso regularizáció)