

JODA 2019 harjoitustyö – vaihe 6: toimeenpano

Osku Mäkkylä osku.makkyla@tuni.fi

Tämän datatiedeprojektin tarkoituksen oli tehdä parempia suosituksia henkilöille, jotka haluavat saada omistamistaan asunnoista paremman hinnan Madridissa, kun he vuokraavat niitä Airbnb:n kautta. Vaihtoehtoisesti uusia sijoitusasuntoja tähän tarkoitukseen etsivälle tästä raportista olisi hyötyä. Tässä vaiheessa yhdistetään aikaisemmissa vaiheissa tehtyjä taulukoita ja kaavioita, joiden avulla voidaan tehdä suosituksia. Osittain tämä vaihe jää hieman vajaaksi edellisissä vaiheissa ilmenneiden ongelmien kautta, mutta pyrkimyksenä on korjata ne viimeiseen palautukseen mennessä.

Alla olevasta taulukosta pystyy alustavasti tekemään muutaman huomion.

```
In [30]: data1.describe()
```

```
Out[30]:
```

	price	minimum_nights	number_of_reviews	reviews_per_month
count	12462.000000	12462.000000	12462.000000	12462.000000
mean	86.540523	3.528487	41.776200	1.954675
std	311.892510	15.691286	60.966175	1.909732
min	8.000000	1.000000	1.000000	0.010000
25%	38.000000	1.000000	5.000000	0.460000
50%	60.000000	2.000000	17.000000	1.320000
75%	90.000000	3.000000	53.000000	2.970000
max	9999.000000	700.000000	559.000000	20.030000

Keskimääräinen asunnon tai huoneen päivävuokra on 86,54 euroa ja minimi yöpymisvaatimuksena on yleensä vähintään kolme yötä.

neighbourhood_group	
Centro	6645
Salamanca	815
Arganzuela	687
Chamberí	662
Tetuán	470
Retiro	417
Moncloa - Aravaca	369
Latina	333
Carabanchel	315
Ciudad Lineal	311
Chamartín	292
Puente de Vallecas	282
Usera	155
Hortaleza	148
San Blas - Canillejas	142
Fuencarral - El Pardo	136
Barajas	81
Villaverde	79
Moratalaz	62
Villa de Vallecas	42
Vicálvaro	19

Vieressä olevan taulukon perusteella eniten asuntoja on vuokralla keskustan kaupunginosassa, eli Centrossa. Viidenneksi eniten on Tetuánissa, joka sijaitsee hieman yllättäen kauempana kaupungin pohjoisosassa. Muuten taulukossa eteenpäin mentäessä sijainteja dominoi keskustan alueet tai kauempana rautatieasemien lähellä olevat asuinalueet. Näiden huomioiden perusteella voidaan päätellä, että eniten kysyntää sijaitsee siellä, minne asuntoja on tähän mennessä listattukin. Olemassa olevat asuntosijoittajat tai asunnon omistajat voivatkin päätellä oman asuntonsa menekkiä sijainnin perusteella.

```
print(data1.corr())
```

	price	minimum_nights	number_of_reviews
price	1.000000	0.009001	-0.036344
minimum_nights	0.009001	1.000000	-0.040973
number_of_reviews	-0.036344	-0.040973	1.000000
reviews_per_month	-0.042331	-0.079303	0.648108

	reviews_per_month
price	-0.042331
minimum_nights	-0.079303
number_of_reviews	0.648108
reviews_per_month	1.000000

Tästä taulukosta näkee eri muuttujien välisen korrelaation keskenään. Eniten vaikuttavuutta on arvostelujen kokonaismäärällä. Aiemmissa palautuksissa olen puhunut "reviews" kohdasta katselukertoina, joten korjataan se tässä.

```
In [71]: X = data1[['minimum_nights','price','reviews_per_month']]
y = data1['number_of_reviews']

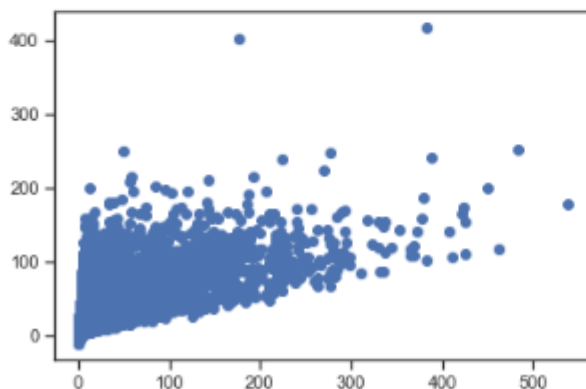
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=101)

from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train,y_train)

predictions = lm.predict(X_test)

plt.scatter(y_test,predictions)
```

Out[71]: <matplotlib.collections.PathCollection at 0xde68170>



Tämä kohta meni pieleen jo vaiheessa 5. Yritän korjata tähän paremman kuvaajan siitä, mitkä tekijät vaikuttavat vuokrattavan asunnon menekkiin.

Tässä vaiheessa voidaan kuitenkin mielestäni todeta, että sellaiset asunnot, jotka sijaitsevat keskustassa, sen välittömässä läheisyydessä tai isojen liikennesolmujen keskittymässä ovat suosittuja. Näiden lisäksi arvostelujen määrä korreloi asunnon tai huoneen menekin kanssa.

Helppoa ja vaikeaa:

- Helppoa on vetää yhtenäisiä johtopäätöksiä jo suoritettujen vaiheiden jälkeen.
- Muiden opiskelijoiden palautusten perusteella helpohkoa tehdä raportointi.
- Vaikeaa tehdä syvällisempää analyysiä näillä pohjatiedoilla.