In [50]:
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from IPython.display import VimeoVideo
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.metrics import mean_absolute_error
from sklearn.utils.validation import check_is_fitted
from sklearn.impute import SimpleImputer
from category_encoders import OneHotEncoder
from sklearn.pipeline import Pipeline, make_pipeline
import plotly.express as px
import plotly.graph_objects as go
from glob import glob
```

In [51]:
```
def wrangle(filename):
    df = pd.read_csv(filename)
    mask_apt = df['property_type'] == 'apartment'
    mask_cf = df['place_with_parent_names'].str.contains('Capital Federal')
    mask_usd = df['price_aprox_usd'] < 400_000

    #remove outliers for surface_covered_in_m2
    low, high = df['surface_covered_in_m2'].quantile([0.1, 0.9])
    mask_area = df['surface_covered_in_m2'].between(low, high)
    df = df[mask_apt & mask_cf & mask_usd & mask_area]
    #working with lat-lon
    df[['lat', 'lon']] = df['lat-lon'].str.split(',', expand=True).astype(float)
    df.drop(columns=['lat-lon'], inplace=True)
    df['neighbourhood'] = df['place_with_parent_names'].str.split('|', expand=True)[3]
    df.drop(columns=['place_with_parent_names'], inplace=True)


    return df
```

In [52]:
```
files = glob('buenos-aires-real-estate-*.csv')
files
```

Out[52]:
```
['buenos-aires-real-estate-1.csv', 'buenos-aires-real-estate-2.csv']
```

In [53]:
```
frames = []
for file in files:
    df = wrangle(file)
    frames.append(df)
```

In [54]:
```
type(frames[0])
```

Out[54]:
```
pandas.core.frame.DataFrame
```

In [55]:
```
df = pd.concat(frames, ignore_index=True)
df.shape
```

Out[55]:
```
(2788, 17)
```

In [56]:
```
features = ['neighbourhood']
X_train = df[features]
target = 'price_aprox_usd'
y_train = df[target] #actual data point
```

In [57]:
```
mean = y_train.mean()
y_pred_baseline = [mean] * len(y_train)
mae = mean_absolute_error(y_train, y_pred_baseline)
mae
```

Out[57]:
```
58958.12063234472
```

In [58]:
```
ohe = OneHotEncoder(use_cat_names=True)
ohe.fit(X_train)
#transform data
XT_train = ohe.transform(X_train)
XT_train.head()
```

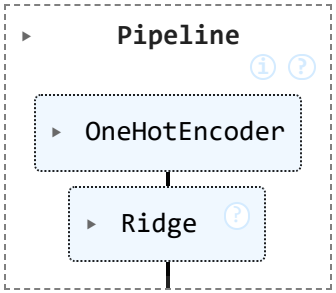
Out[58]:

	neighbourhood_Villa Crespo	neighbourhood_Chacarita	neighbourhood_Villa Luro	neighbourhood_Caballito	neighbourhood_Constitución
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	0
4	0	0	0	0	1

In [59]:
```
model = make_pipeline(
    OneHotEncoder(use_cat_names=True),
    Ridge(),
)
```

In [60]:
```
model.fit(X_train, y_train)
```

Out[60]:



In [61]:
```
y_pred_training = model.predict(X_train)
new_mae = mean_absolute_error(y_pred_training, y_train)
new_mae
```

Out[61]:
```
50600.96434149369
```

In [62]:
```
X_test = pd.read_csv('./apartment_data.csv')
X_test.rename(columns={'neighborhood': 'neighbourhood'}, inplace=True)
```

```
test_data = X_test['neighbourhood']
```

In [63]:
```
y_pred_test = model.predict(test_data)
y_pred_test
```

Out[63]:
```
array([255038.68977649, 186736.60628106, 125726.63749271, 117933.28778973,
       157766.43594483, 129573.58359618, 122910.85160978, 183576.15348919,
       230797.73830972, 136689.97755413, 115590.16550348, 157453.70201222,
       101676.72773178, 202187.79796439, 202187.79796439, 183576.15348919,
       128397.06657004, 136689.97755413, 183576.15348919, 136689.97755413])
```

In [65]:
```
intercept = model.named_steps['ridge'].intercept_
coef = model.named_steps['ridge'].coef_
```

In [66]:
```
coef[:5]
```

Out[66]:
```
array([  1071.42291844,  -4002.94359029,  -6044.97103951,  22147.88130914,
       -45107.51008272])
```

In [68]:
```
#features names
neibr_names = model.named_steps['onehotencoder'].get_feature_names_out()
neibr_names[:5]
```

Out[68]:
```
array(['neighbourhood_Villa Crespo', 'neighbourhood_Chacarita',
       'neighbourhood_Villa Luro', 'neighbourhood_Caballito',
       'neighbourhood_Constitución'], dtype=object)
```

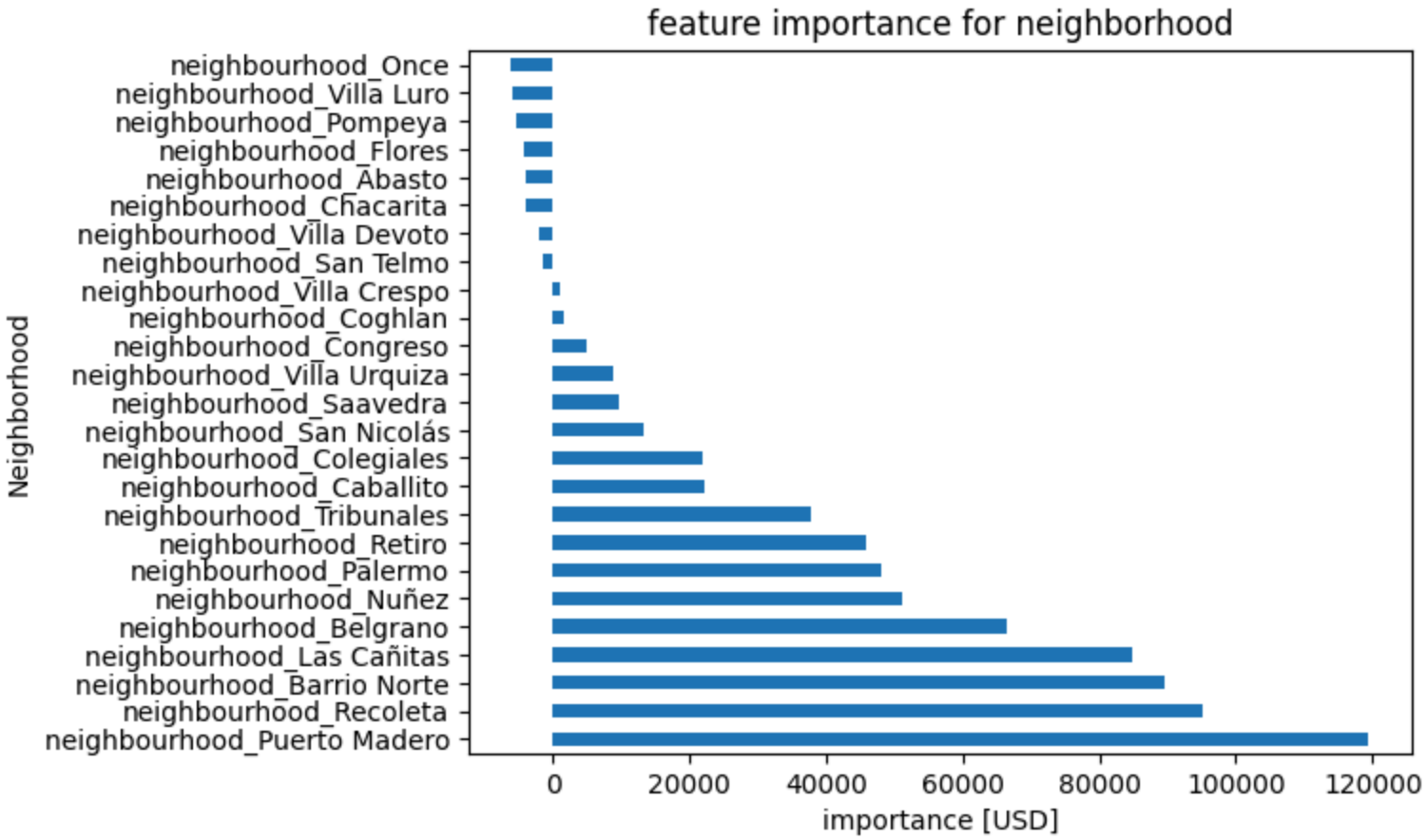
In [71]:
```
feat_imp = pd.Series(coef, index=neibr_names)
feat_imp
```

Out[71]:

neighbourhood_Villa Crespo	1071.422918
neighbourhood_Chacarita	-4002.943590
neighbourhood_Villa Luro	-6044.971040
neighbourhood_Caballito	22147.881309
neighbourhood_Constitución	-45107.510083
neighbourhood_Once	-6127.636668
neighbourhood_Almagro	-7221.488066
neighbourhood_Flores	-4173.702385
neighbourhood_Belgrano	66569.243329
neighbourhood_Liniers	-34752.943590
neighbourhood_San Cristobal	-19303.470624
neighbourhood_Congreso	5036.044836
neighbourhood_Saavedra	9625.274797
neighbourhood_Balvanera	-8882.969102
neighbourhood_Parque Avellaneda	-46856.699172
neighbourhood_Recoleta	95179.183674
neighbourhood_San Telmo	-1367.103584
neighbourhood_Nuñez	51118.051645
neighbourhood_Barrio Norte	89663.079948
neighbourhood_Parque Centenario	-33941.826904
neighbourhood_Abasto	-4062.322396
neighbourhood_	-32021.571131
neighbourhood_Paternal	-22439.994842
neighbourhood_Mataderos	-31516.376904
neighbourhood_Palermo	47957.598853
neighbourhood_Villa Lugano	-58988.657876
neighbourhood_Coghlan	1589.561299
neighbourhood_Las Cañitas	84767.574917
neighbourhood_Villa Urquiza	8727.965978
neighbourhood_Monserrat	-9891.917143
neighbourhood_Villa Pueyrredón	-10039.626904
neighbourhood_Floresta	-20028.389132
neighbourhood_Parque Patricios	-8877.703970
neighbourhood_San Nicolás	13292.822316
neighbourhood_Villa del Parque	-13522.789229
neighbourhood_Boedo	-17874.883159
neighbourhood_Centro / Microcentro	-6358.326686
neighbourhood_Parque Chacabuco	-17685.266846
neighbourhood_Barracas	-12707.703026
neighbourhood_Parque Chas	-30661.443709
neighbourhood_Colegiales	21835.147377
neighbourhood_Villa General Mitre	-8256.699172

```
neighbourhood_Villa Ortuzar          -16682.128863
neighbourhood_Villa Devoto            -2054.178118
neighbourhood_Retiro                  45907.373096
neighbourhood_Versalles              -13079.036424
neighbourhood_Boca                   -28597.646803
neighbourhood_Puerto Madero          119420.135141
neighbourhood_Agronomía              -30044.475402
neighbourhood_Monte Castro           -14783.159676
neighbourhood_Tribunales              37766.324917
neighbourhood_Villa Santa Rita       -33544.270787
neighbourhood_Velez Sarsfield        -18294.843709
neighbourhood_Pompeya                 -5309.277318
neighbourhood_Villa Soldati          -36568.732318
dtype: float64
```

In [78]:
```
feat_imp.sort_values(ascending=False).head(25).plot(kind='barh')
plt.xlabel('importance [USD]')
plt.ylabel('Neighborhood')
plt.title('feature importance for neighborhood');
```


feature importance for neighborhood
neighbourhood_Once
neighbourhood_Villa Luro
neighbourhood_Pompeya
neighbourhood_Flores
neighbourhood_Abasto
neighbourhood_Chacarita
neighbourhood_Villa Devoto
neighbourhood_San Telmo
neighbourhood_Villa Crespo
neighbourhood_Coghlan
neighbourhood_Congreso
neighbourhood_Villa Urquiza
neighbourhood_Saavedra
neighbourhood_San Nicolás
neighbourhood_Colegiales
neighbourhood_Caballito
neighbourhood_Tribunales
neighbourhood_Retiro
neighbourhood_Palermo
neighbourhood_Nuñez
neighbourhood_Belgrano
neighbourhood_Las Cañitas
neighbourhood_Barrio Norte
neighbourhood_Recoleta
neighbourhood_Puerto Madero
Neighborhood
0
20000
40000
60000
80000
100000
120000
importance [USD]