Start with: Most columns + one-hot-encoded amenities

Ridge Regression + k-fold cross validation + PCA testing

Data Cleaning

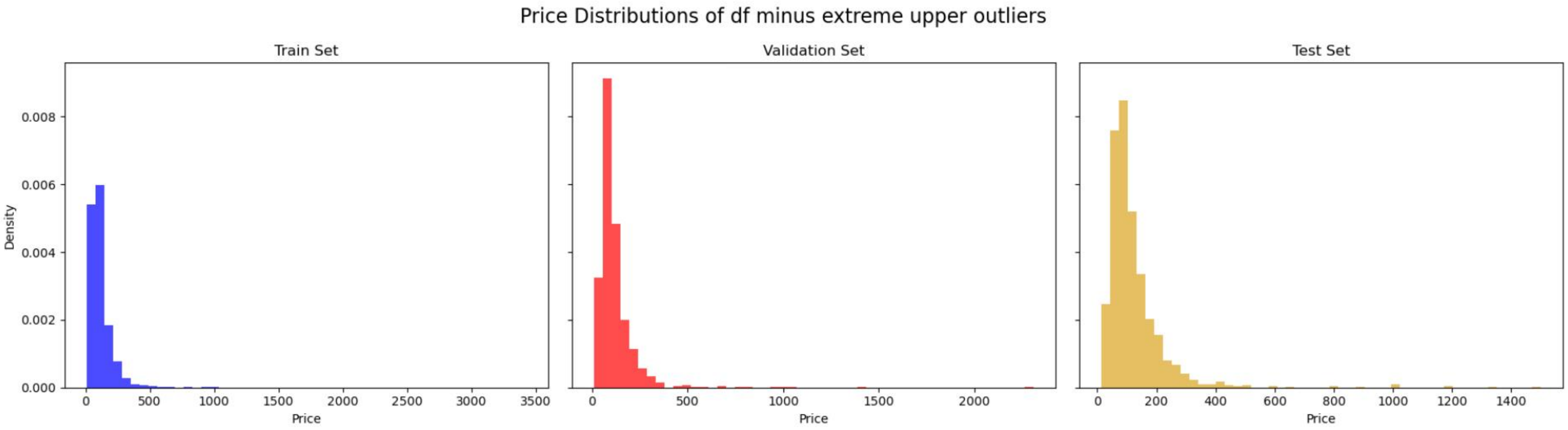
- Start with: listings_detailed_final_final_final.csv
  - All amenities are one-hot-encoded
  - Shape is (8898, 293)
- Remove lat-, lon_neighbourhood
- Add one-hot-encoded neighbourhood groups
- Fill missing values with reasonable estimates based on other data
  - Don't drop all rows with missing values as to not lose too much data
  - Drop all rows with missing prices
- Create csv's, each with different price outliers excluded

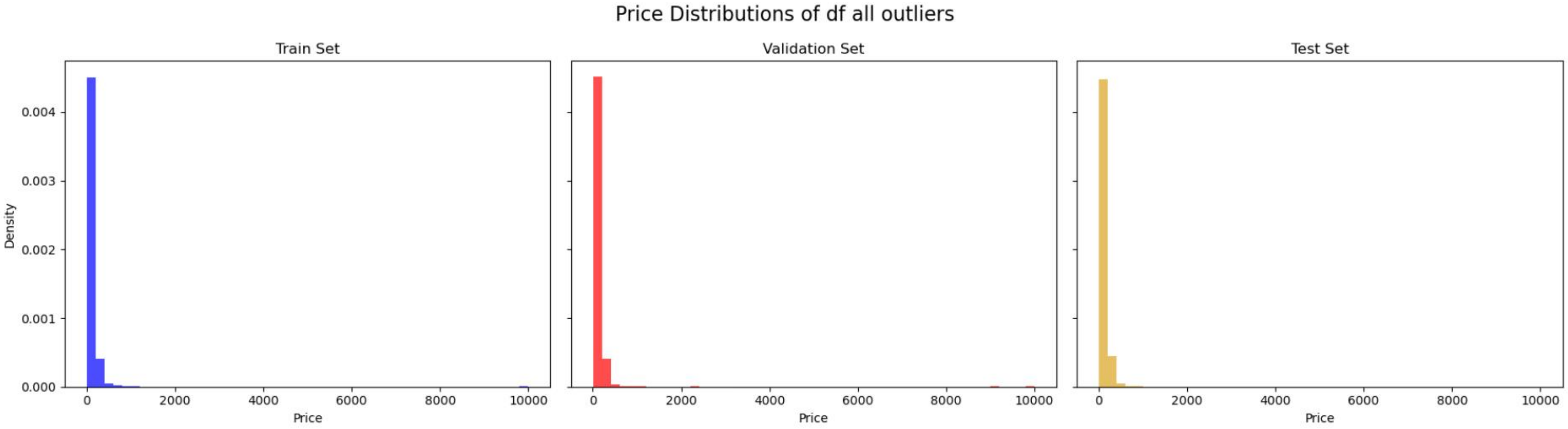
Modelling outliers

- Test performance of ridge regression when leaving out more and more price outliers (from both training and testing)
- Difference between 8th and 9th most expensive listing is

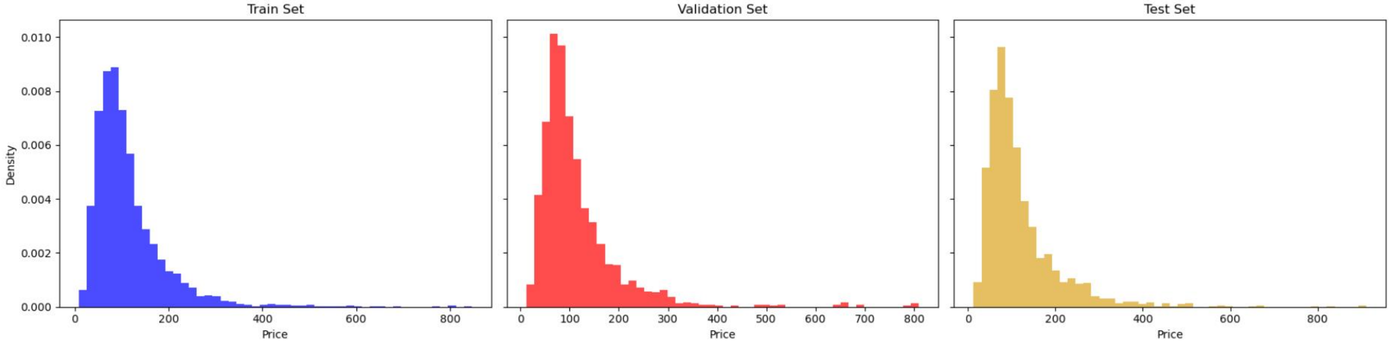
9000->3400, ie extreme outliers at the top

- Good compromise is removing upper 1% price outliers
    - RMSE = 45
    - Same Modell on data only without extreme outliers: RMSE roughly 89
- By removing upper 1%, RMSE goes from around 110 to 45 and still works on 99% of the real data

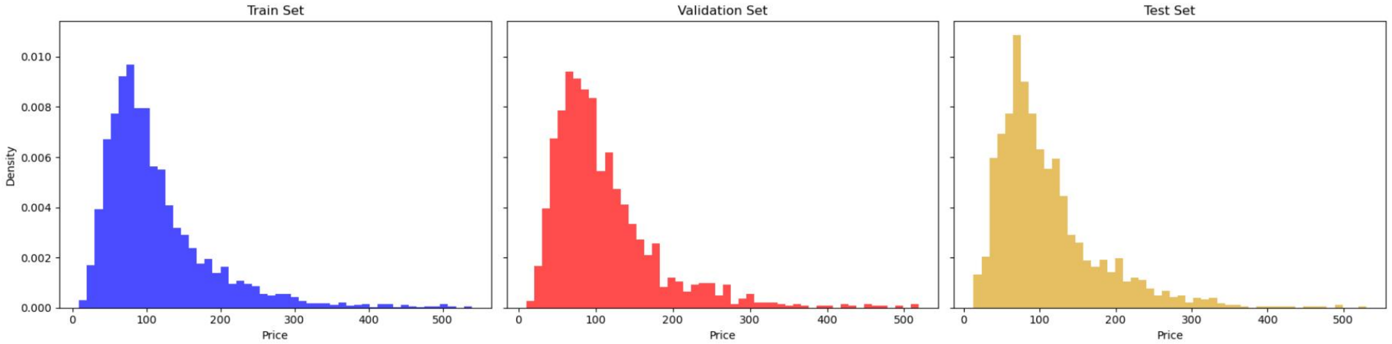

Price Distributions of df minus extreme upper outliers
Train Set
Validation Set
Test Set
Density
0.000
0.002
0.004
0.006
0.008
Price
0
500
1000
1500
2000
2500
3000
3500
Price
0
500
1000
1500
2000
Price
0
200
400
600
800
1000
1200
1400


Price Distributions of df all outliers
Train Set
Validation Set
Test Set
Density
0.000
0.001
0.002
0.003
0.004
Price
0
2000
4000
6000
8000
10000
Price
0
2000
4000
6000
8000
10000
Price
0
2000
4000
6000
8000
10000

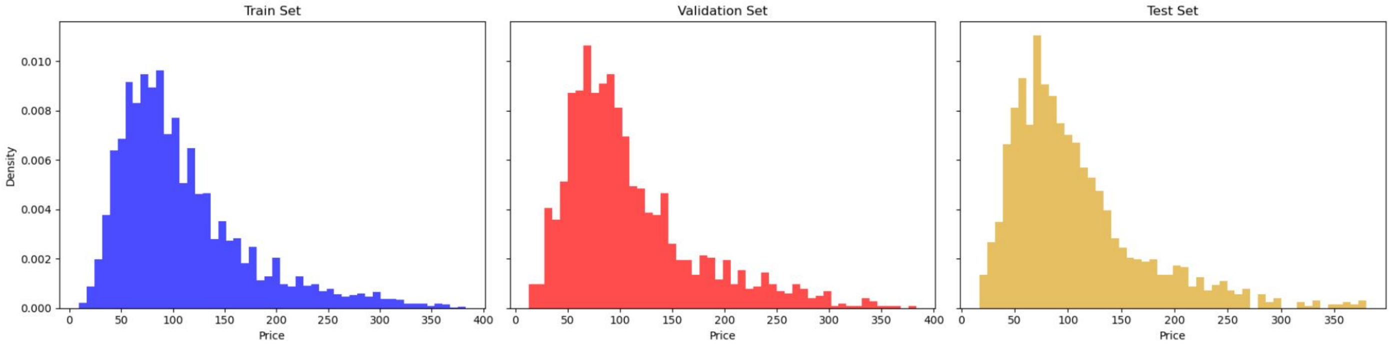
Price Distributions of df minus upper 0.5%



Price Distributions of df minus upper 1%



Price Distributions of df minus upper 2%



Modelling columns

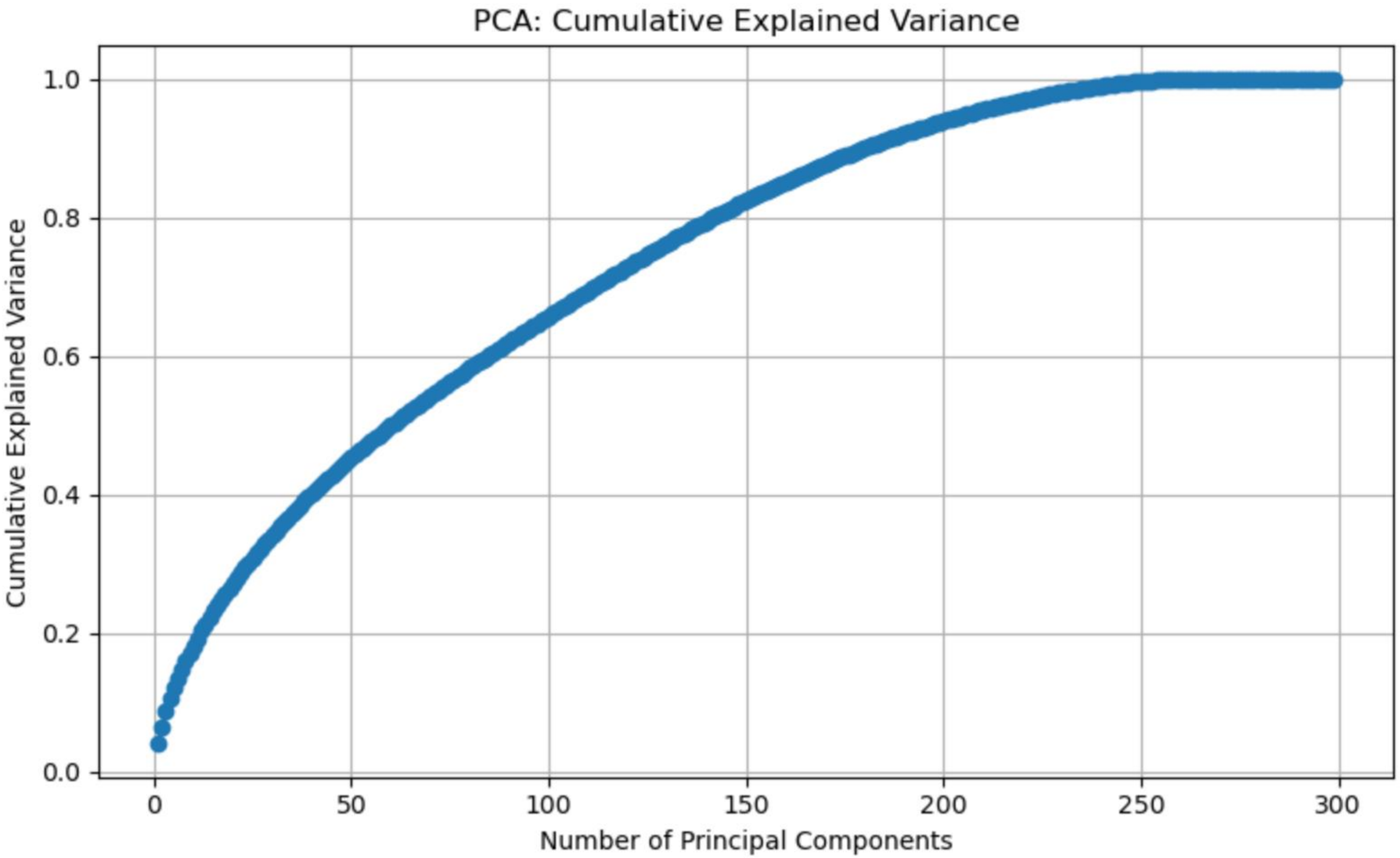
- Id and latitude, longitude columns do not significantly affect the MSE for the ridge regression modell, since id should be more or less random and lat, Lon effect is likely not linear -> remove them
- In comparison, removing the amenities does affect the MSE visibly
- Modell performs only slightly worse (about 5% more MSE) on training data than on validation. -> slight, but acceptable amount of underfitting via hyperparameters on the training data for better performance on general data

Modelling k-fold cross validation

- Use dataset without upper 1% of prices and without lat, lon, id columns
- Ideal alpha ridge parameters from 5-fold and 10-fold ridge regression perform almost the same as "single fold" on test set, ie. RMSE of 45
- Same for the dataset with specific ids excluded (RMSE 65 originally -> 65.5)
- Result for df without upper 1%:
    - RMSE: 45, Mean Absolute Error: 32, MRE: 36%

PCA

- Start again without upper 1% and without lat, lon, id -> 299 features
- PCA
  - 240 principal components to retain 99% of var
  - 271 PCs dim to retain 100% of var
- Modell with only 240 dimensions still has RMSE of about 46, only slightly worse than modell with full dimensions



```
Number of components to retain 95% variance: 207
Number of components to retain 98% variance: 229
Number of components to retain 99% variance: 240
Number of components to retain 100% variance: 271
```