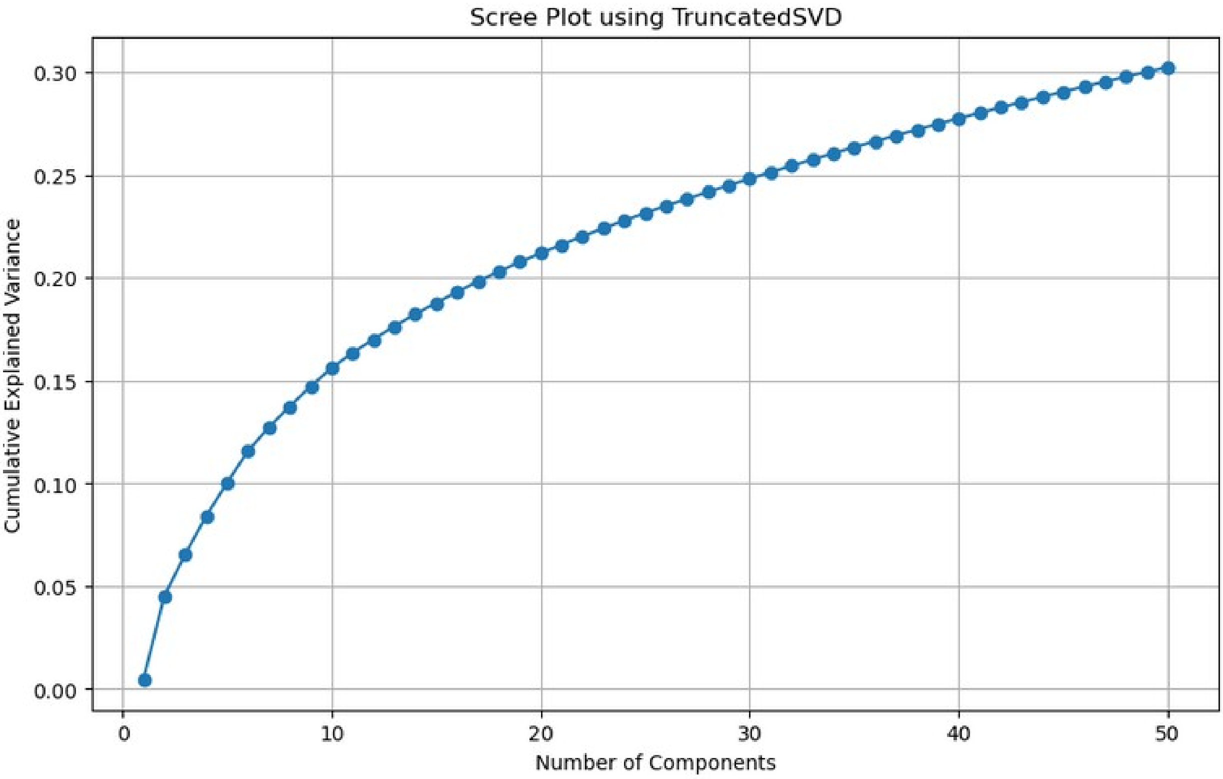
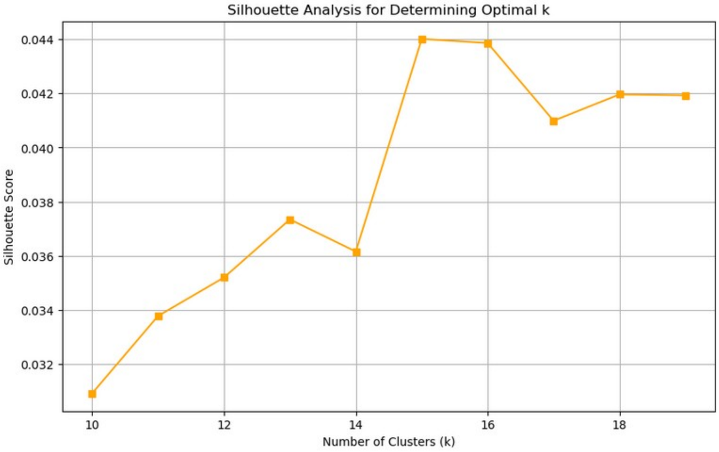
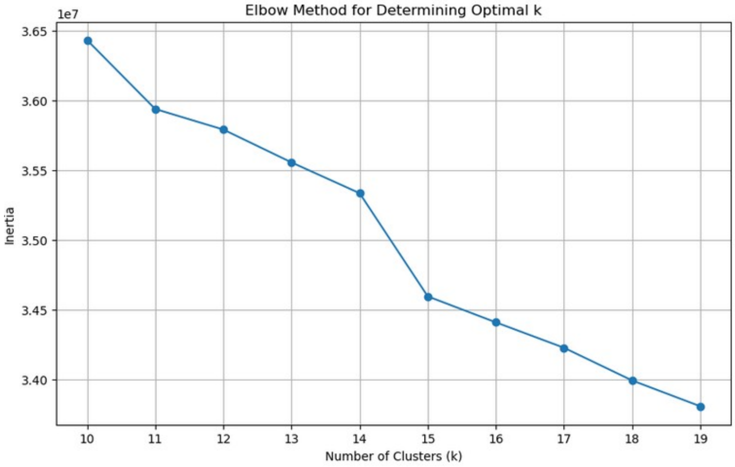
Customer Segmentation (Unsupervised Learning)

After performing data preprocessing, date column transformation, some basic feature scaling/normalization. dropping virtually empty columns, encoding high dimentional features, and one-hot encoding we ended up with 962756 rows × 2390 columns for both the azdias and customers data combined. Therefore it is necessary to perform some kind of feature reduction. I went with Truncated SVD due to memory constraints. We perform the SVD on only the azdias dataset. Below we can see the plot for this



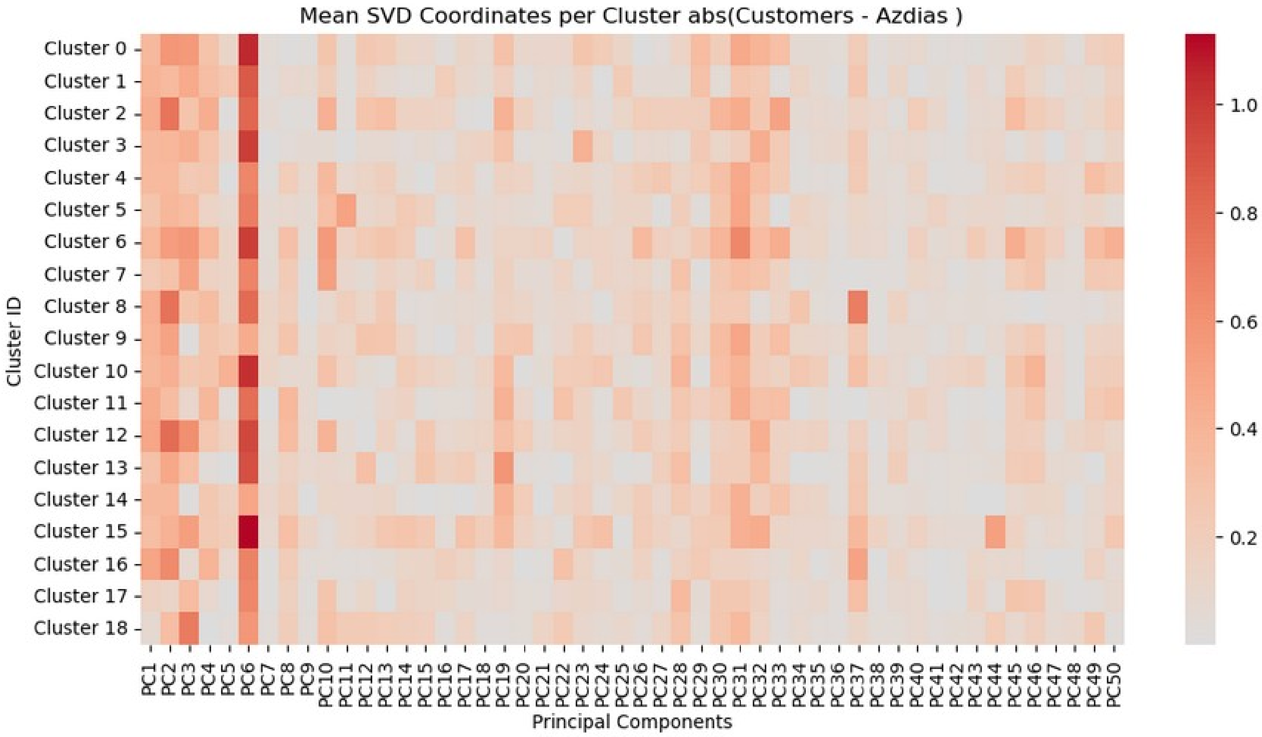
Next step is to cluser the data so we can analize clusters in the dataset.



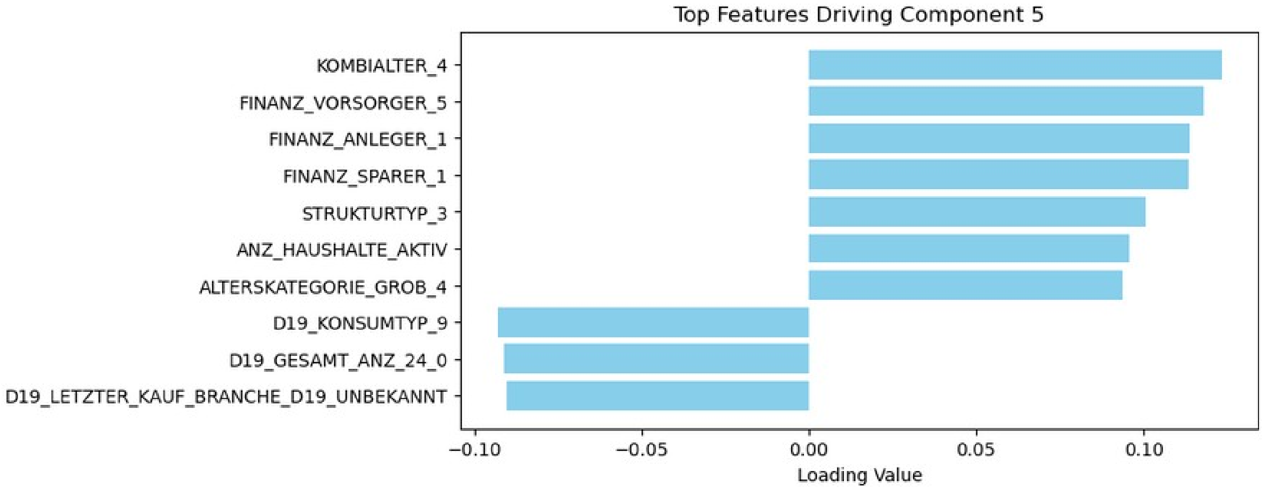


The above two graphs seem to indicate that K=15 is a good number of clusters.

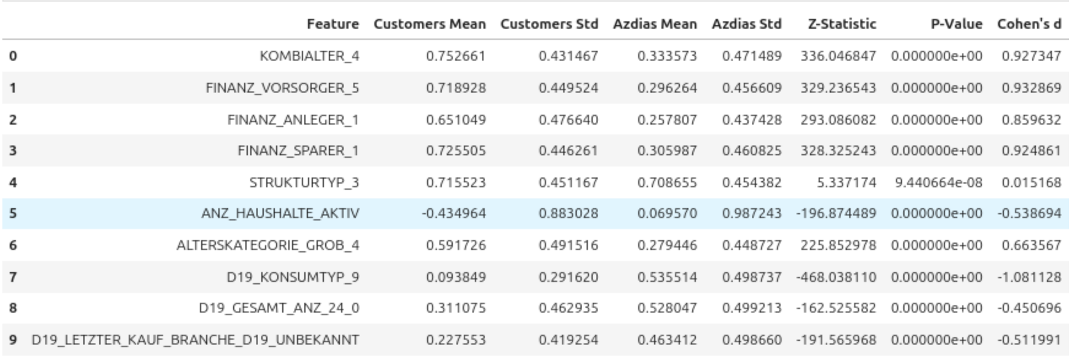
The next step is to calculate the centroids for the azdias and customers clusters. We find the absolute difference between the azdias and customers dataset, and plot it per-centroid.



PC6 (index 5) seems to be the largest componet. This is interesting so further exploration is done. We list the top features contributing to PC6 (componet 5)



Finally some statisical analysis is done to show that the features are significantly different between azdias and customers.



	Feature	Customers Mean	Customers Std	Azdias Mean	Azdias Std	Z-Statistic	P-Value	Cohen's d
0	KOMBIALTER_4	0.752661	0.431467	0.333573	0.471489	336.046847	0.000000e+00	0.927347
1	FINANZ_VORSORGER_5	0.718928	0.449524	0.296264	0.456609	329.236543	0.000000e+00	0.932869
2	FINANZ_ANLEGER_1	0.651049	0.476640	0.257807	0.437428	293.086082	0.000000e+00	0.859632
3	FINANZ_SPARER_1	0.725505	0.446261	0.305987	0.460825	328.325243	0.000000e+00	0.924861
4	STRUKTURTYP_3	0.715523	0.451167	0.708655	0.454382	5.337174	9.440664e-08	0.015168
5	ANZ_HAUSHALTE_AKTIV	-0.434964	0.883028	0.069570	0.987243	-196.874489	0.000000e+00	-0.538694
6	ALTERSKATEGORIE_GROB_4	0.591726	0.491516	0.279446	0.448727	225.852978	0.000000e+00	0.663567
7	D19_KONSUMTYP_9	0.093849	0.291620	0.535514	0.498737	-468.038110	0.000000e+00	-1.081128
8	D19_GESAMT_ANZ_24_0	0.311075	0.462935	0.528047	0.499213	-162.525582	0.000000e+00	-0.450696
9	D19_LETZTER_KAUF_BRANCHE_D19_UNBEKANNT	0.227553	0.419254	0.463412	0.498660	-191.565968	0.000000e+00	-0.511991

The above table indicates that there are significant and meaningful differences between the Customers and Azdias groups across most features.
These differences vary in magnitude. Diving deeper and understanding these distinctions can be crucial for targeted strategies, marketing, product development, or other business decisions based on the characteristics that differentiate Customers from Azdias.

The exception is STRUKTURTYP_3, which, despite being statistically significant, has a negligible effect size, indicating that the difference between the groups for this feature is practically insignificant..

Most features have large to medium effect sizes, suggesting that the differences are not only statistically significant but also practically meaningful. The exception is STRUKTURTYP_3, which, despite being statistically significant, has a negligible effect size, indicating that the difference between the groups for this feature is practically insignificant.

Predicting Customer Acquisition (Supervised Learning)

Feature selection was similar to the unsupervised learning, but after one hot encoding VarianceThreshold is done to remove instances that do not frequently occur. This helps by reducing the feature from 2222 to 1846.

There is a massive class imballance of Counter({0: 42430, 1: 532}) so we use compute_class_weight to apply to the loss function during learning. A CatBoostClassifier model is used. Optuna is used for hyperparameter tuning and handling k-fold CV training.

Overall the top features were

|

feature	importance
D19_SOZIALES_1	27.787978
D19_KONSUMTYP_MAX_2	3.193066
EINGEZOGENAM_HH_JAHR	1.422330
KBA13_CCM_2500_3	1.030702
D19_SONSTIGE_6	0.931474

But if we look at the classification report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.77      0.87      8487
           1       0.04      0.69      0.07       106

    accuracy                           0.77      8593
   macro avg       0.52      0.73      0.47      8593
weighted avg       0.98      0.77      0.86      8593
```

Our model is not performing well with the rare positive classes. Next steps would to use something like SMOTE to further help with the class imballance. And to test other models.