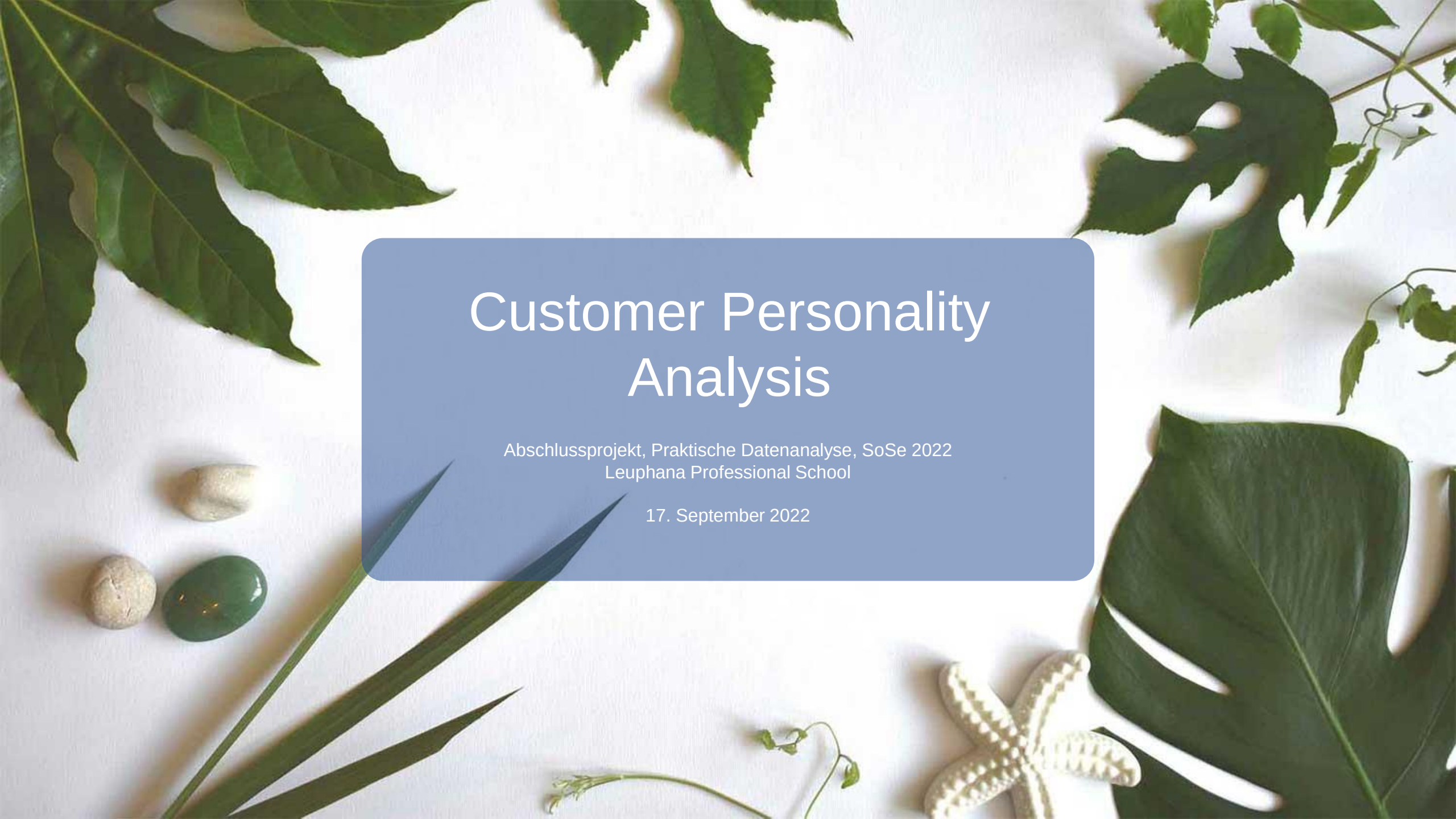


Customer Personality Analysis

Abschlussprojekt, Praktische Datenanalyse, SoSe 2022
Leuphana Professional School

17. September 2022

Eingangsfrage, Problemstellung und Zielsetzung
- Kundensegmentierung -

Hintergrund: Marketing

Viele Unternehmen verlassen sich heutzutage auf gezieltes Marketing, um ihre Kundschaft glücklich zu machen.

Hintergrund: Wirtschaften

Tech-Giganten wie Google, Meta Platforms (Facebook) & Co. haben ihre Geschäftsmodelle unter anderem rund um gezielte Werbung aufgebaut und verdienen damit hunderte Milliarden US-Dollar im Jahr.

Kundenstruktur

Kundensegmentierung ist eine wertvolle Methode zur Identifikation der verschiedenen Kundengruppen eines Unternehmens.

Kundenverständnis

Es werden Daten zu unterschiedlichen Kundenmerkmalen herangezogen, um die Gewohnheiten und Vorlieben der Kundschaft zu verstehen und zu analysieren

Kundenmerkmale

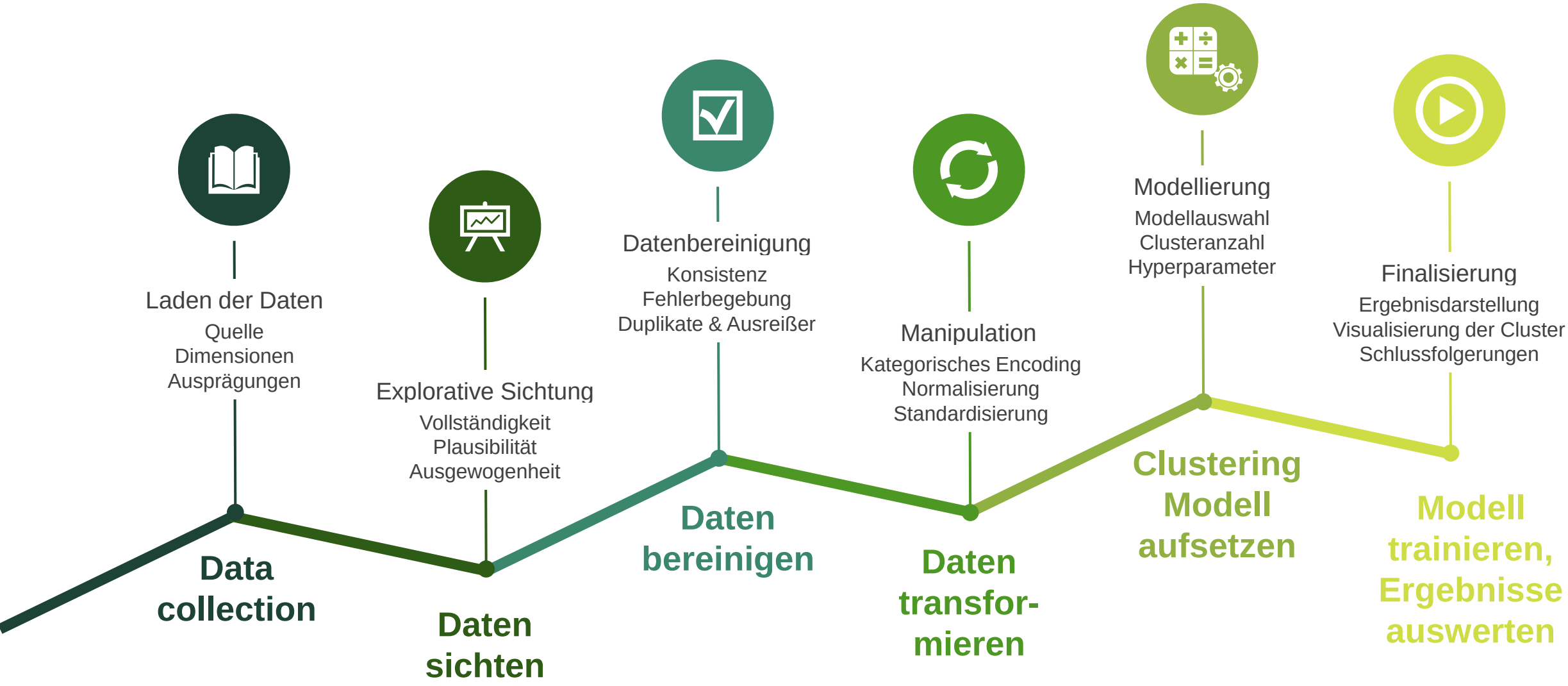
Merkmale wie Alter, Bildungsgrad, Familienstand, Kinder im Haushalt, Einkommen, die Höhe der Ausgaben für verschiedene Produkte und die Reaktion auf Rabattaktionen oder Marketingkampagnen

Methodik

Es werden Kundendaten herangezogen, bereinigt und vorverarbeitet, um darauf ein Clustering-Modell aufzubauen. Etwaige Erkenntnisse werden aus dem Clustering abgeleitet und bewertet.

Ziel: Dieses Projekt soll dabei unterstützen, die Kundschaft auf der Grundlage von Verhaltensweisen, Gewohnheiten, Bedürfnissen und Interessen gezielter, also individueller anzusprechen, um damit einen Mehrwert zu erzeugen. Unsere **Vision** sind **glückliche & zufriedene Kunden**!

Prozessschritte der Datenanalyse



Informationen zum Datensatz

Der Dataframe teilt sich in 2240 Zeilen und 29 Spalten auf.

Kundeninformationen

- **ID**: Eindeutige Kennung des Kunden
- **Year_Birth**: Geburtsjahr des Kunden
- **Education**: Bildungsgrad des Kunden
- **Marital_Status**: Familienstand des Kunden
- **Income**: Jährliches Haushaltseinkommen des Kunden
- **Kidhome**: Anzahl der Kinder im Haushalt des Kunden
- **Teenhome**: Anzahl der Teenager im Haushalt des Kunden
- **Dt_Customer**: Datum der ersten Registrierung des Kunden
- **Recency**: Anzahl der Tage seit dem letzten Kauf des Kunden
- **Complain**: 1, wenn sich der Kunde in den letzten 2 Jahren beschwert hat, sonst 0

Gekaufte Produkte

Betrag, der in den letzten 2 Jahren vom Kunden für die folgenden Produktkategorien ausgegeben wurde [MntWines, MntFruits etc.]:

- Weine
- Früchte
- Fleisch-Produkte
- Fisch-Produkte
- Süßwaren
- Gold-Produkte

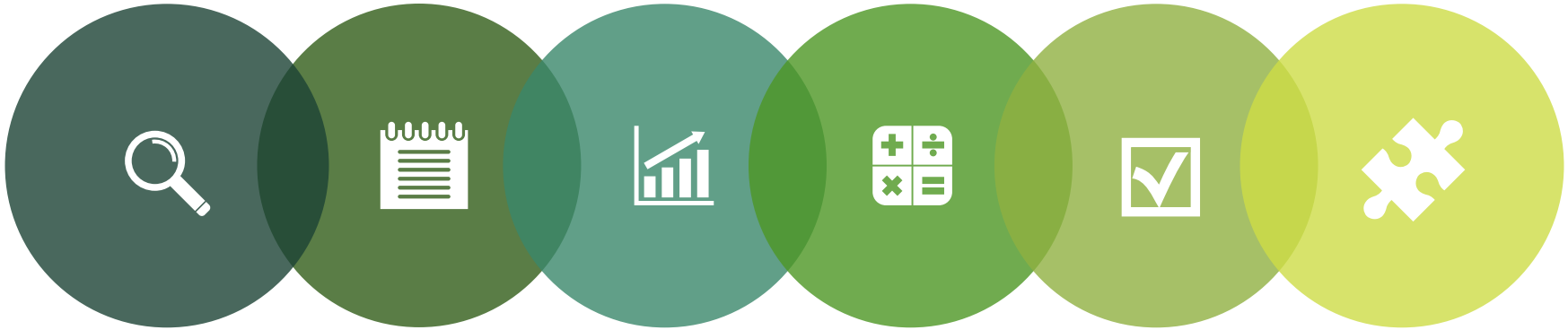
Rabatt-Aktionen & Marketing-Kampagnen

- **NumDealsPurchases**: Anzahl der Einkäufe mit einem Rabatt
- **AcceptedCmp1-5:** …jeweils: 1, wenn der Kunde das Angebot in der jeweiligen Kampagne angenommen hat, sonst 0
- **Response**: … 1, wenn der Kunde das Angebot in der letzten Kampagne angenommen hat, sonst 0

Orte des Einkaufs

- **NumWebPurchases**: Onlinekäufe
- **NumCatalogPurchases**: Katalogkäufe
- **NumStorePurchases**: Käufe in Geschäften
- **NumWebVisitsMonat**: Anzahl der Besuche auf der Website des Unternehmens im letzten Monat

Schritte zur Datenbereinigung und -transformation



01

Fehlende Werte

Income 24
ID 0
NumDealsPurchases 0 …

02

Datentypen

25 Features vom Datentyp "integer",
3 Features vom Datentyp "object" und
1 Feature vom Datentyp "float".

03

Duplikate

Geprüft und keine vorhanden

04

Outlier

- Income (identifiziert via Boxplot)
- Age (identifiziert durch Recherche)

239	1893	239	129
339	1899	339	123
192	1900	192	122
1950	1940	1950	82
424	1941	424	81

05

Feature Engineering (!)

- Zusammenfassen von Merkmalen (zB Summen, Vereinheitlichung)
- Erstellen neuer Merkmale (KPIs)
- Löschen von nicht mehr benötigten Merkmalen (zB eindimensionale Merkmale)

06

Data Pre-Processing

- Categorical encoding
- Standard scaling

Gesucht & Gefunden:

…der “Average Customer”

Alter

~ 46 Jahre alt

Kinder und Jugendliche im Haushalt

- Durchschnittlich 1 Kind, und eine Famliengröße von 3 Personen

Ausgaben über alle Produktgruppen

- USD 607,06 pro Jahr

Einkommen

- im Durchschnitt USD 51.902

Bildungsgrad

- Mehrheitlich “Postgraduierte”
- …ist das plausibel?

Durchschnittliches Einkaufsvolumen

- USD 33,32 pro Einkauf

Der **Average-Customer** weist dabei die folgendes Profil auf:

```
overview.mean()
```

```
Age                 46.095878
Marital_Status       0.645609
Minors               0.952061
Family_size          2.597670
Income           51902.612007
Education            0.114247
D_engaged          540.006272
Spending           607.061380
Deals                2.330197
TotalNumPurch       14.899194
Recency             49.095878
No_Webview           5.315860
Cmp_total            0.448029
Response             0.149642
AOV                 33.325523
dtype: float64
```

Unsupervised Learning

- K-Means Algorithmus -

Modell(e)

- K-Means Clustering sowie hierarchisches Clustering
- “Silhoutte Score” als Vergleichs-/ Bewertungsmaßstab

Funktionsweise

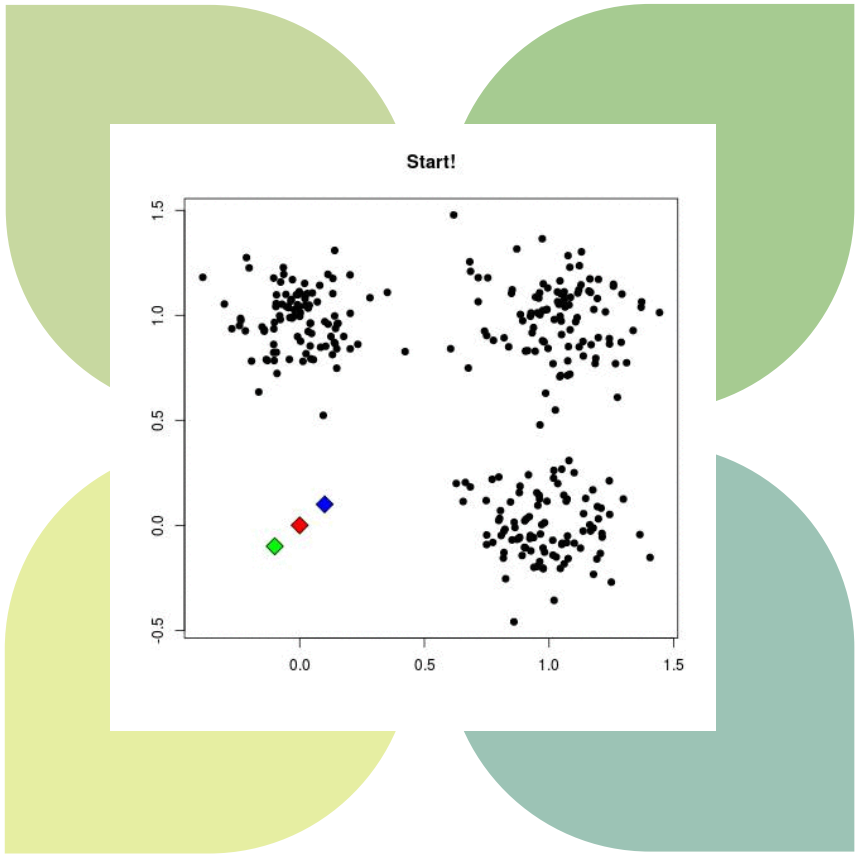
K-Means Clustering ist eine Methode des unüberwachten Lernens und versucht gleiche Daten zusammenzufassen und unterschiedliche Daten zu separieren. Dabei werden die Abstände gleichartiger Daten innerhalb eines Clusters möglichst minimiert und der Abstand des Datenpunktes gegenüber anderen Clustern möglichst maximiert.

Hyperparameter-Tuning

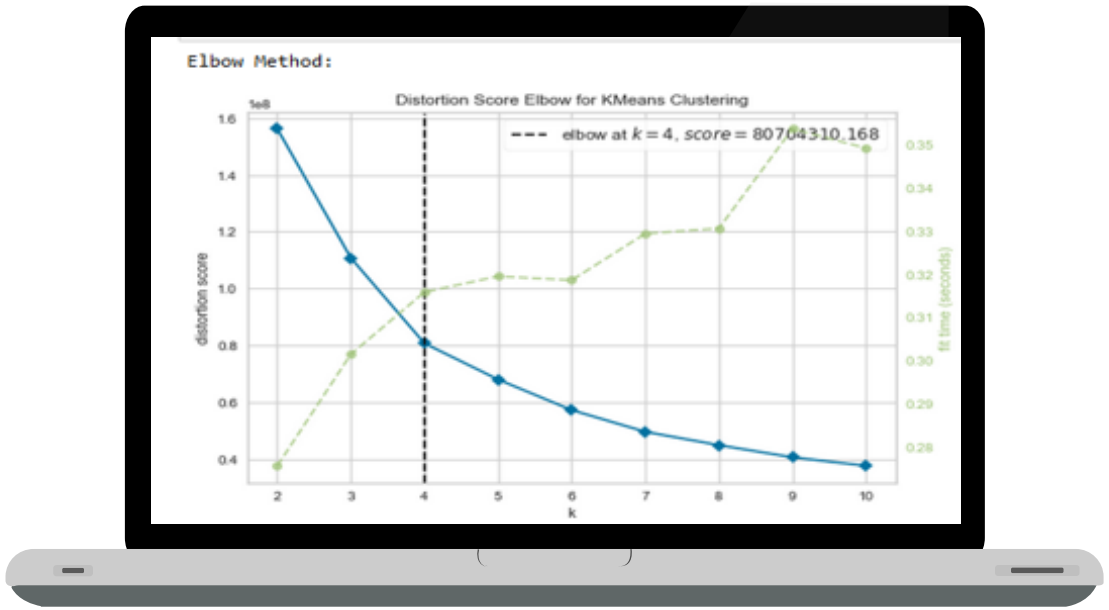
- n_clusters wurde optimiert (Elbow-Method)
- “n_init” sowie “max_iter” wurden verdoppelt
- Komplexitätsreduktionen im Dataframe
- Vergleich von Standardisierung und Normalisierung der Daten vor Modellberechhnung
- Dimensionsreduktion (PCA) als Ansatz

Voraussetzungen

Da der Dataframe keine fehlenden Werte beinhaltet, dabei über 2200 Datensätze zählt und die Werte miteinander insgesamt nicht all zu stark korrelieren, ist das K-Means Clustering als sehr gut geeignetes Modell für dieses Projekt zu bewerten. Sehr viele Dimensionen könnten störend sein.



Elbow Method



“Elbow Plot”

Der Punkt, ab dem die zusätzliche erklärte Varianz bei Zunahme eines weiteren Clusters stark sinkt, nennt sich „elbow“.

Die passende Anzahl an Clustern finden

Mit Hilfe der Elbow-Method und dem KElbowVisualizer wird die passende Anzahl an Cluster für den vorliegenden Dataframe auf **4** festgelegt.

„Jedes Mal wenn wir ein weiteres Cluster hinzufügen, also K + 1, verringert sich die Varianz. Die Reduktion der Varianz wird immer kleiner. Wenn wir K = N hätten , wobei N für die Anzahl der Datenpunkte steht (also jeder Datenpunkt ist ein eigener Cluster), wäre die Varianz = 0. Dann wäre der Sinn des Clusterings allerdings verfehlt, da es ja gerade um eine Zusammenfassung gleichartiger Daten geht (Daten mit dem geringsten euklidischen Abstand zueinander finden).“

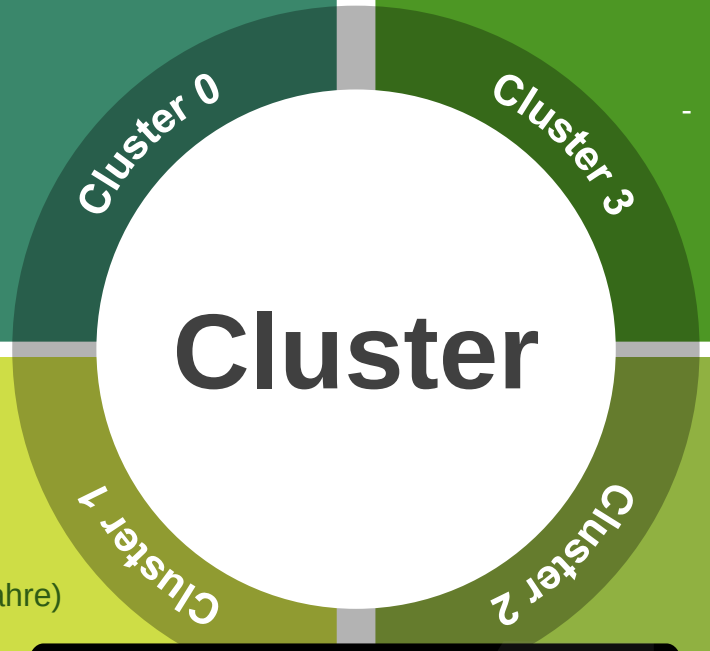
Die identifizierten Clusters

Cluster: “Gesundes Mittelfeld”

- definitiv mindestens ein Elternteil
- zwischen 2 und 4 Haushaltsmitgliedern
- Alleinerziehende sind eine Teilmenge
- die meisten haben einen Teenager zu Hause
- relativ gesehen älter (49 Jahre)
- Mittleres Einkaufsvolumen

Cluster: “STAR-Kundschaft”

- sind eher keine Eltern
- maximal 2 Familienmitglieder
- leichte Mehrheit von Paaren gegenüber Singles
- alle Altersgruppen - lieben Wein
- Gruppe mit höchsten Einkommen
- Höchstes Einkaufsvolumen
- - kaufen viel übers Internet



Cluster: “Pfennigfuchser”

- Sie sind definitiv ein Elternteil
- Eher größere Haushalte mit bis zu 5 Mitgliedern
- Vergleichsweise jünger im Schnitt (44 Jahre)
- einkommensschwächere Gruppe
- Geringstes Einkaufsvolumen
- Geringe Umsätze insgesamt über alle Produktgruppen

Cluster: “High Potentials”

- Die meisten dieser Gruppe haben keine abhängigen Kinder (mehr) im Haushalt
- Eher 2 Familienmitglieder
- Vergleichsweise älter
- Höheres Einkommen und Einkaufsvolumen
- Sowohl im Web als auch im Store unterwegs

Ergebnisse der Datenanalyse

- Optimierungsbedarfe & Ausblick -

01 - Unser Ziel sollte es sein, dass Cluster mit den **Star-Kunden** stärker zu fokussieren. Diese haben die höchsten Umsätze und die stärkste Kaufkraft.

03 - Der **Donnerstag ist der Wochentag**, der meisten Neukunden. Um das weiter zu forcieren, könnte man an diesem Tag besonders viel Werbung schalten.

05 - Es gibt ein Cluster mit "**High Potentials**" die sich zur Star-Kundschaft entwickeln könnten. Dies sollte gezielt mit Marketingaktionen herbeigeführt werden.

02 - Die Star-Kunden geben gerne Geld für bestimmte Produkte wie **Wein** aus. Dieses Potential sollten wir stärker nutzen.

04 - Die **Sommermonate** Juli und August sind ebenfalls starke Monate, was die Neukundengewinnung angeht. Vielleicht hilft ein Sommerfest oder eine Sommer-Rabatt-Aktion dies noch mehr zu fördern.

06 - Wir sollten die Verkaufskanäle nutzen, um mehr Kundschaft in den **Store** zu locken, wo die höchsten Umsätze gemacht werden.

Limitationen/Optimierungspotenziale:

- Weitere **Informationen zum Datensatz** und dessen Ursprung sowie dahinterstehendem Unternehmen fehlten
- Eine **Dimensionsreduktion** könnte hilfreich sein, um die Ergebnisqualität weiter zu verbessern

Ausblick:

- Zukünftig könnten noch **ergänzende Clustering-Techniken** etabliert werden
- Es könnten noch **mehr Features und KPIs** zu besseren Steuerung gebildet werden