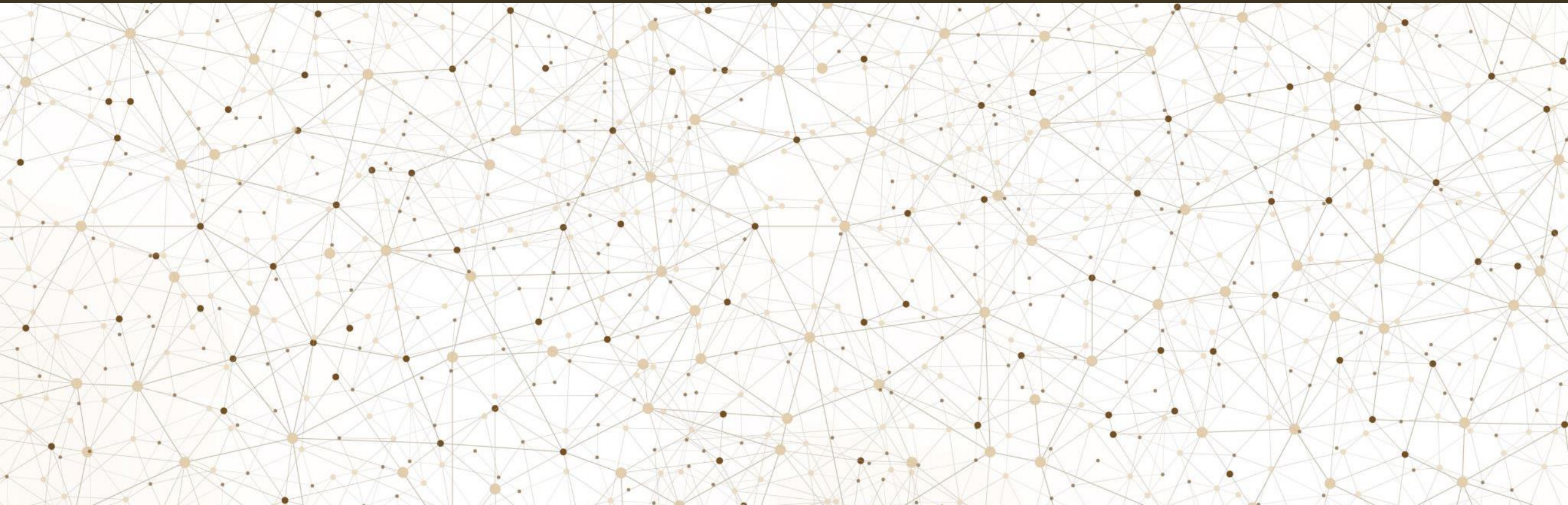
Kundensegmentierung

Julia Pfützer, Mya-Melissa Jahic, Hannah Laier



Agenda

- → Unser Datensatz
- → Zielsetzung
- → Datenvorverarbeitung
- → Die Daten visualisiert
- → Verwendete Methoden zur Analyse
- → Fazit

Datensatz nach Bereinigung

→ Kundendaten eines Supermarkts von Keggel

→ 2216 Zeilen und 19 Spalten

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchase
0	5524	1957	1	1	58138.0	0	0	635	88	546	172	88	88	3	
1	2174	1954	1	1	46344.0	1	1	11	1	6	2	1	6	2	
2	4141	1965	1	2	71613.0	0	0	426	49	127	111	21	42	1	
3	6182	1984	1	2	26646.0	1	0	11	4	20	10	3	5	2	
4	5324	1981	2	2	58293.0	1	0	173	43	118	46	27	15	5	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2235	10870	1967	1	2	61223.0	0	1	709	43	182	42	118	247	2	
2236	4001	1946	2	2	64014.0	2	1	406	0	30	0	0	8	7	
2237	7270	1981	1	1	56981.0	0	0	908	48	217	32	12	24	1	
2238	8235	1956	3	2	69245.0	0	1	428	30	214	80	30	61	2	
2239	9405	1954	2	2	52869.0	1	1	84	3	61	2	1	21	3	

2216 rows × 19 columns

Zielsetzung

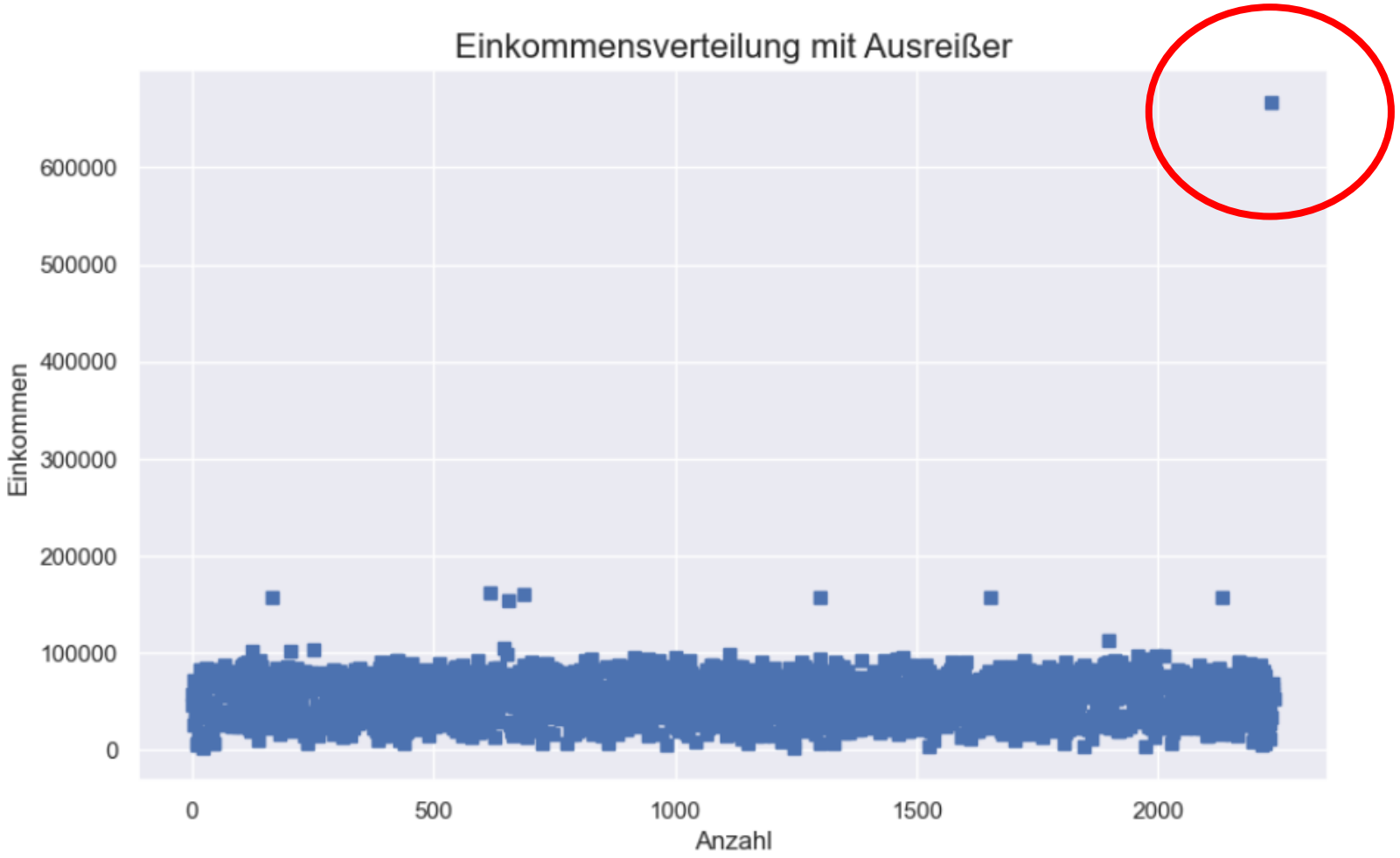
- → Ziel:
  - → Definieren von Zielgruppen
  - → mit Werbung gezielter eine spezifische Zielgruppe zu erreichen

- → 4 Fragen im Fokus
  - → Kaufen Kunden mit Kindern mehr Süßigkeiten?
  - → Kaufen Kunden mit höherem Einkommen mehr Fleisch/Fisch?
  - → Kaufen Kunden mit höherem Bildungsgrad mehr Wein?
  - → Präferieren Kunden bestimmte Werbekanäle?

Datenvorverarbeitung

- → Entfernung aller Zeilen mit fehlenden Werten
- → Beziehung - Status gruppieren in 1 und 2 für Single oder in einer Beziehung
- → Bildungsgrade ersetzen durch Nummern von 1-5
- → Nicht benötigte Spalten löschen
- → Ausreißer entfernen

Die Einkommensverteilung vor Entfernung des Ausreißers



Der Ausreißer erschwert die Sicht auf die restlichen Daten

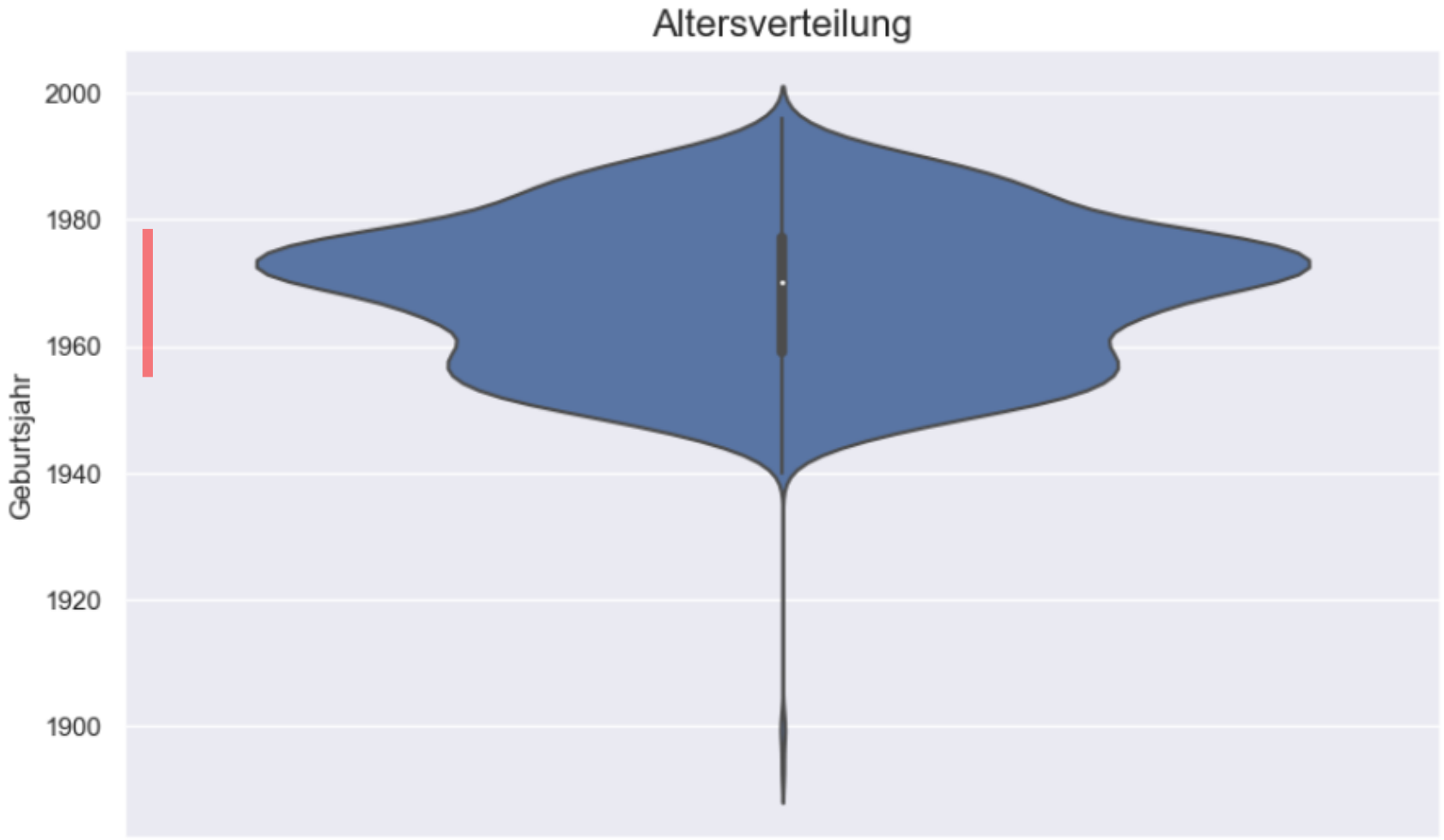
Korrelationsmatrix

Starke positive oder negative Korrelation

Kaum/keine positive oder negative Korrelation

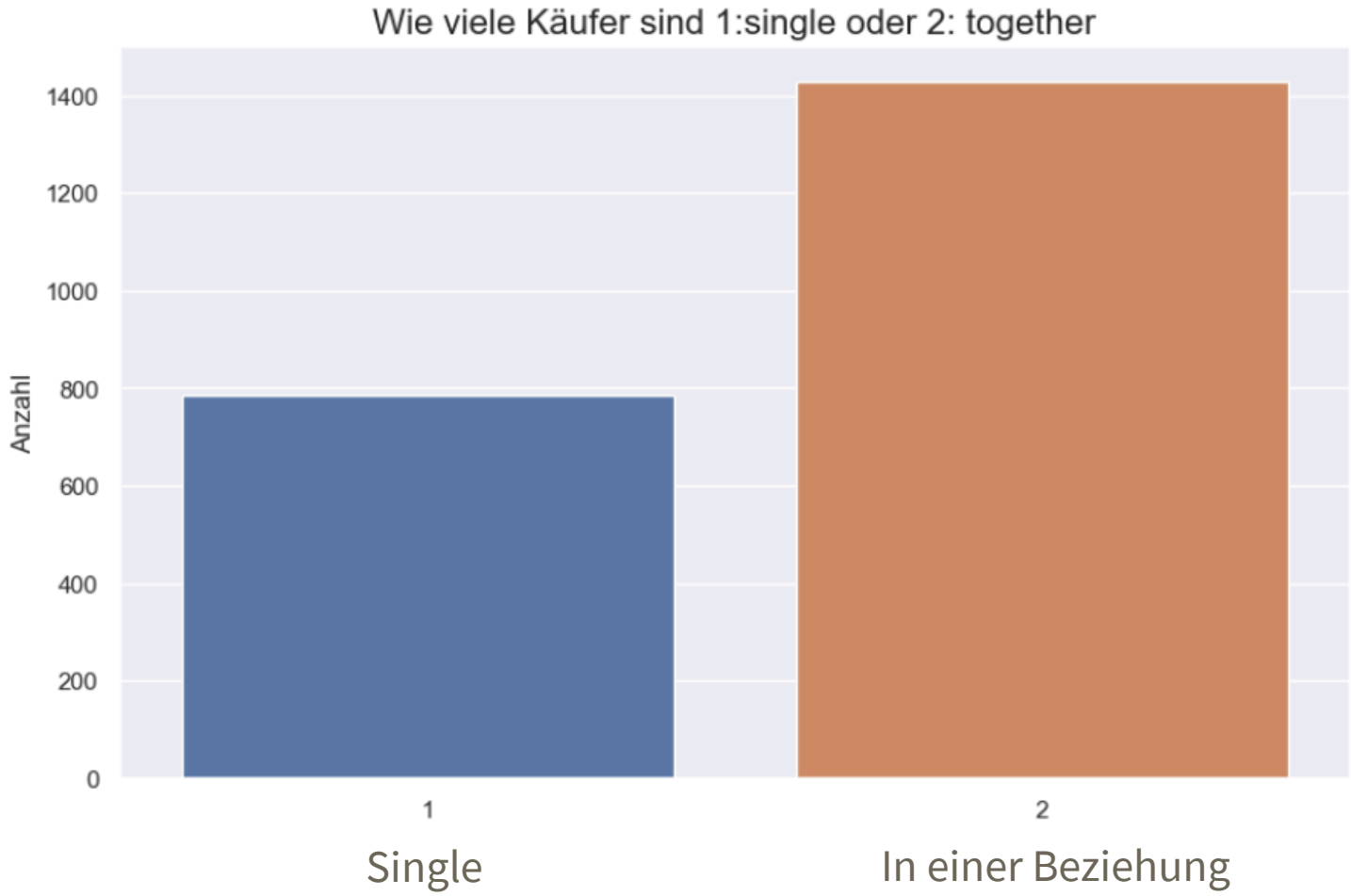
	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
Year_Birth		0,162	0,23	0,352	0,02	0,158	0,018	0,031	0,042	0,018	0,062	0,061	0,145	0,121	0,128	0,121
Income	0,162		0,429	0,019	0,004	0,579	0,431	0,585	0,439	0,441	0,326	0,083	0,388	0,589	0,529	0,553
Kidhome	0,23	0,429		0,036	0,009	0,496	0,373	0,437	0,388	0,371	0,35	0,222	0,362	0,502	0,5	0,448
Teenhome	0,352	0,019	0,036		0,016	0,005	0,177	0,261	0,204	0,162	0,022	0,388	0,155	0,111	0,051	0,135
Recency	0,02	0,004	0,009	0,016		0,016	0,004	0,023	0,001	0,023	0,017	0,001	0,011	0,025	0,001	0,021
MntWines	0,158	0,57	0,496	0,005	0,016		0,39	0,563	0,4	0,387	0,388	0,011	0,542	0,635	0,642	0,321
MntFruits	0,018	0,431	0,373	0,177	0,004	0,39		0,543	0,595	0,567	0,393	0,132	0,297	0,488	0,462	0,418
MntMeatProducts	0,031	0,58	0,437	0,261	0,023	0,563	0,543		0,568	0,524	0,351	0,122	0,294	0,724	0,48	0,539
MntFishProducts	0,042	0,43	0,388	0,204	0,001	0,4	0,595	0,568		0,58	0,423	0,139	0,294	0,534	0,46	0,446
MntSweetProducts	0,018	0,441	0,371	0,162	0,023	0,387	0,567	0,524	0,58		0,37	0,12	0,349	0,491	0,449	0,423
MntGoldProds	0,062	0,326	0,35	0,022	0,017	0,388	0,393	0,351	0,423	0,37		0,049	0,422	0,438	0,382	0,251
NumDealsPurchases	0,061	0,083	0,222	0,388	0,001	0,011	0,132	0,122	0,139	0,12	0,049		0,234	0,009	0,069	0,348
NumWebPurchases	0,145	0,388	0,362	0,155	0,01	0,542	0,297	0,294	0,294	0,349	0,422	0,234		0,378	0,503	0,056
NumCatalogPurchases	0,121	0,589	0,502	0,111	0,02	0,635	0,488	0,724	0,534	0,491	0,438	0,009	0,378		0,519	0,52
NumStorePurchases	0,128	0,529	0,5	0,051	0,00	0,642	0,462	0,48	0,46	0,449	0,382	0,069	0,503	0,519		0,428
NumWebVisitsMonth	0,121	0,553	0,448	0,135	0,02	0,321	0,418	0,539	0,446	0,423	0,251	0,348	0,056	0,52	0,428	

Altersverteilung der Kunden nach Geburtsjahr



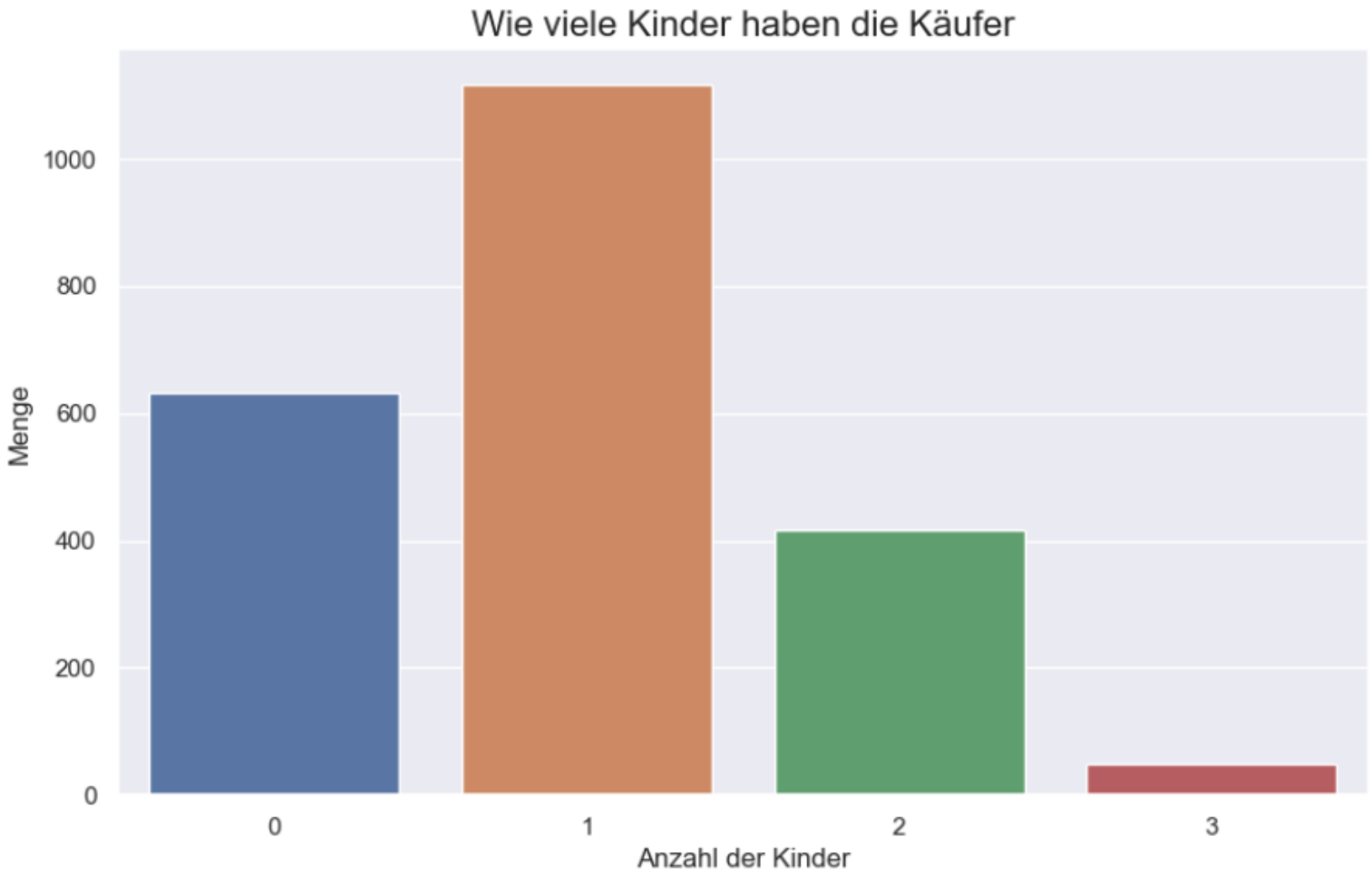
Die meisten Kunden wurden zwischen 1955-1980 geboren

Der Beziehungsstatus der Käufer



Deutlich mehr Kunden sind in einer Beziehung

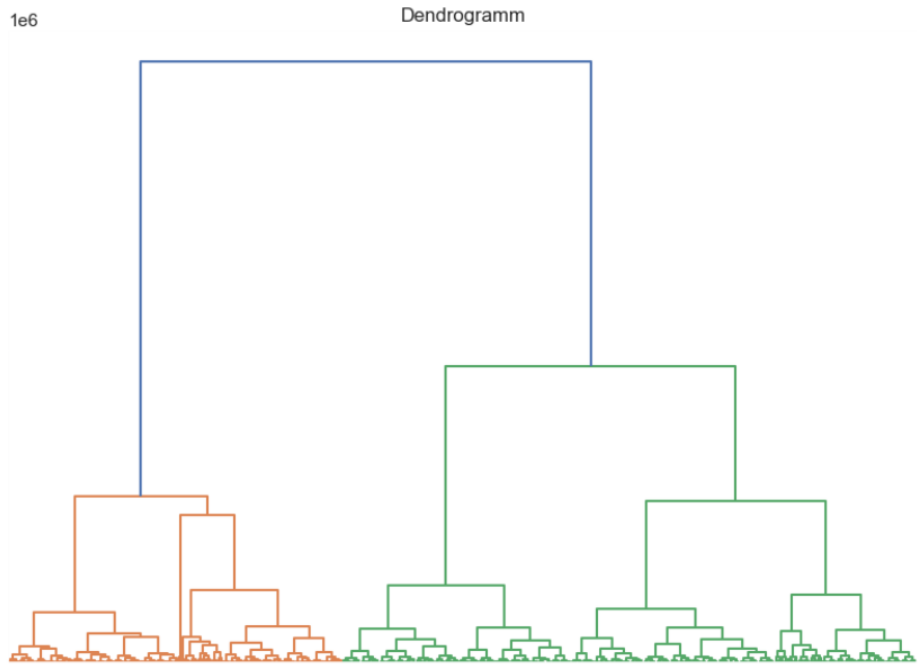
Die Anzahl an Kinder der Käufer



Kunden mit einem Kind sind am häufigsten vertreten

Verwendete Methode zur Analyse

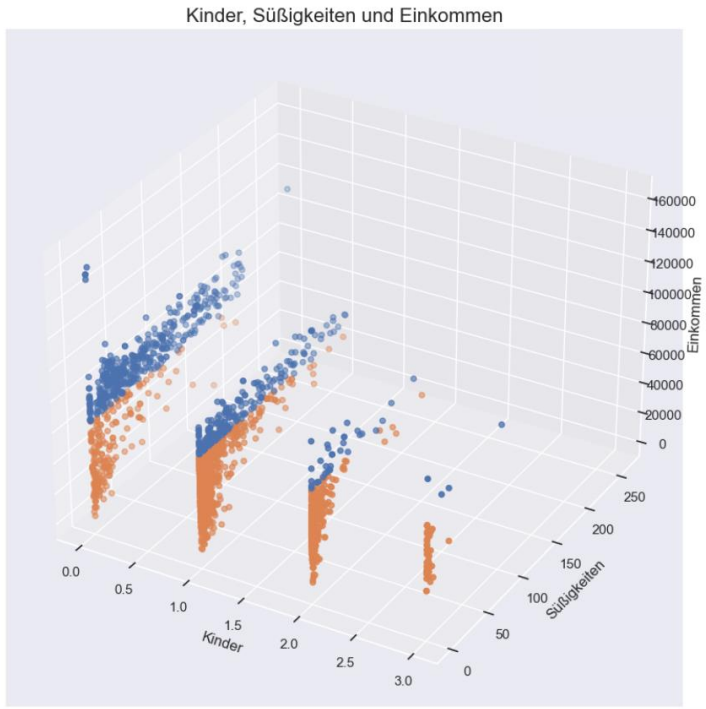
- → K-Means Algorithmus
  - → Clustering Algorithmus
  - → Clusteranzahl wird vorgegeben und mit Dendrogramm ermittelt
  - → Es wird die entsprechende Anzahl an Clustern gebildet und der Mittelpunkt berechnet
  - → Die Datenpunkte werden dem nächsten Clustermittelpunkt zugeordnet, danach wird der Mittelpunkt neu berechnet und so weiter

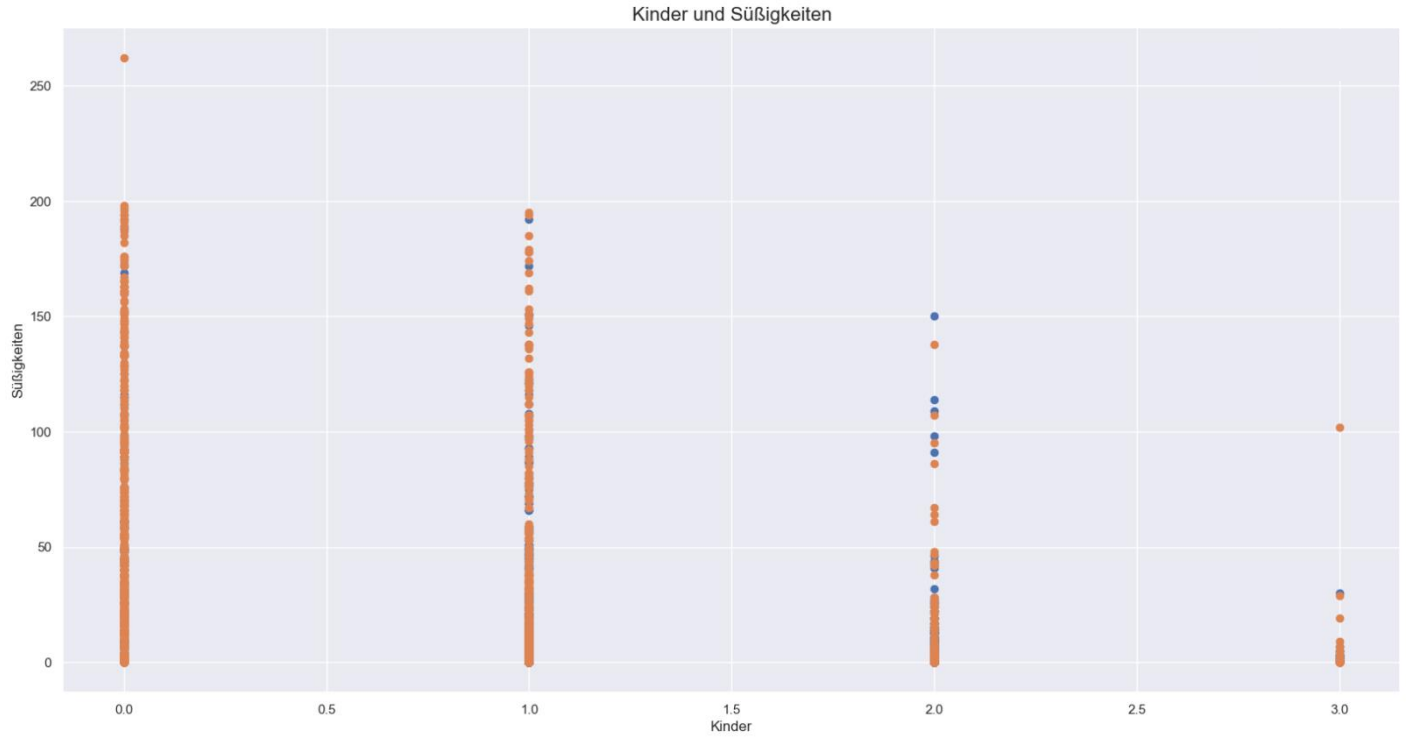


Frage 1 - Kaufen Kunden mit Kindern mehr Süßigkeiten?

→ Bei Kindern und Süßigkeiten besteht eine negative Korrelation von -0,37

	Kidhome	Teenhome	MntSweetProducts
Kidhome	1.000000	-0.039000	-0.378000
Teenhome	-0.039000	1.000000	-0.163000
MntSweetProducts	-0.378000	-0.163000	1.000000



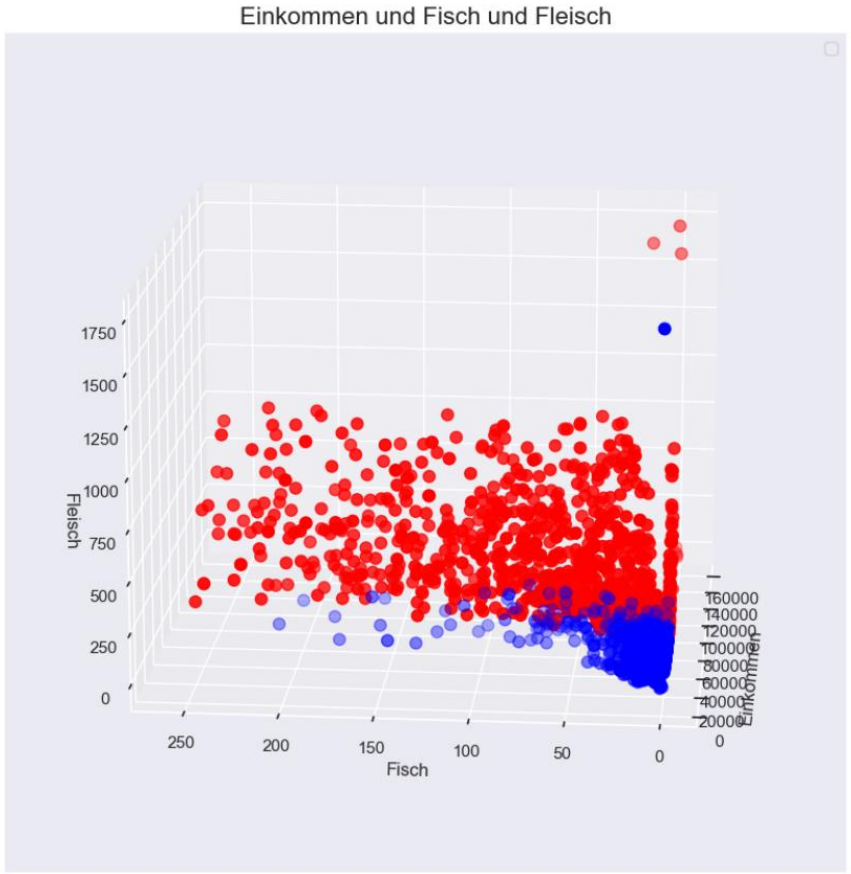


Zielgruppe: Kunden mit 0/1 Kindern

Frage 2 - Kaufen Kunden mit höherem Einkommen mehr Fleisch/Fisch?

→ Bei Fleisch und Fisch besteht eine Korrelation von 0,43 und 0,58 mit dem Einkommen

	Income	MntMeatProducts	MntFishProducts
Income	1.000000	0.585000	0.439000
MntMeatProducts	0.585000	1.000000	0.568000
MntFishProducts	0.439000	0.568000	1.000000



Zielgruppe: Kunden mit eher mehr Einkommen

Frage 3 - Kaufen Kunden mit höherem Bildungsgrad mehr Wein?

→ Keine Korrelation

→ Keine zufriedenstellenden Ergebnisse

]: Slide Type

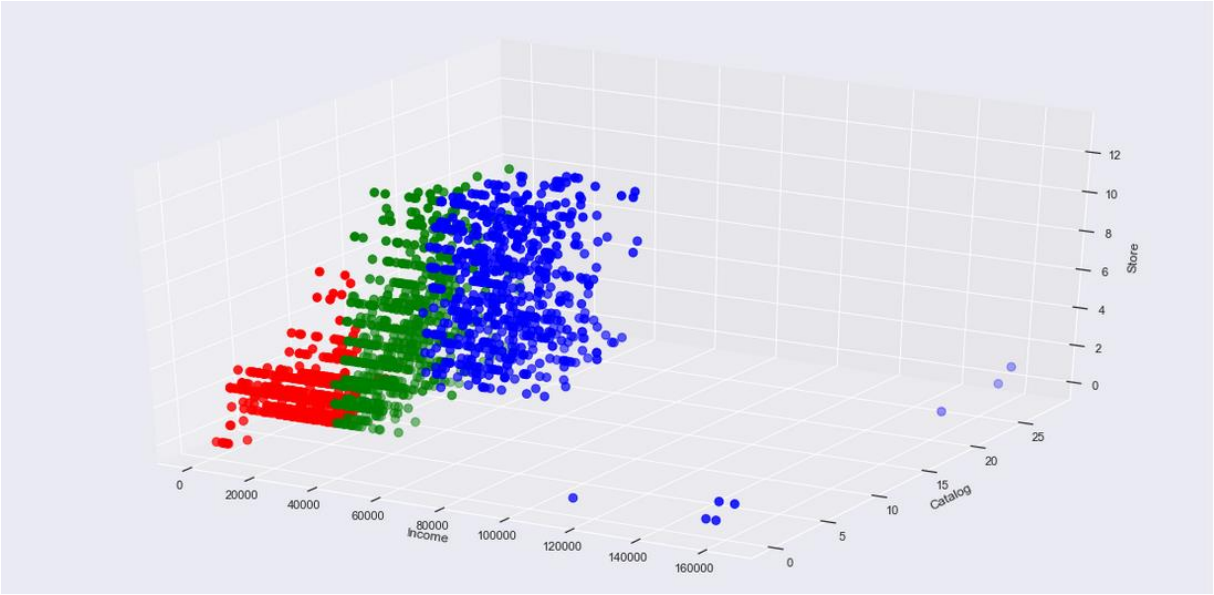
```
km = KMeans(n_clusters=2)
clusters = km.fit_predict(df1.iloc[:,1:])
df1["label"] = clusters

fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df1.Year_Birth[df1.label == 0], df1["MntWines"][df1.label == 0], df1["Education"][df1.label == 0], c='blue', s=60)
ax.scatter(df1.Year_Birth[df1.label == 1], df1["MntWines"][df1.label == 1], df1["Education"][df1.label == 1], c='red', s=60)
#ax.scatter(df1.Year_Birth[df1.label == 2], df1["MntWines"][df1.label == 2], df1["Education"][df1.label == 2], c='green', s=60)
#ax.scatter(df1.Year_Birth[df1.label == 3], df1["Income"][df1.label == 3], df1["Menge"][df1.label == 3], c='orange', s=60)
#ax.scatter(df1.Year_Birth[df1.label == 4], df1["Income"][df1.label == 4], df1["Menge"][df1.label == 4], c='purple', s=60)
#ax.view_init(15, 350)
plt.xlabel("Geburtsjahr")
plt.ylabel("Catalog")
ax.set_zlabel('Store')
plt.show()
```

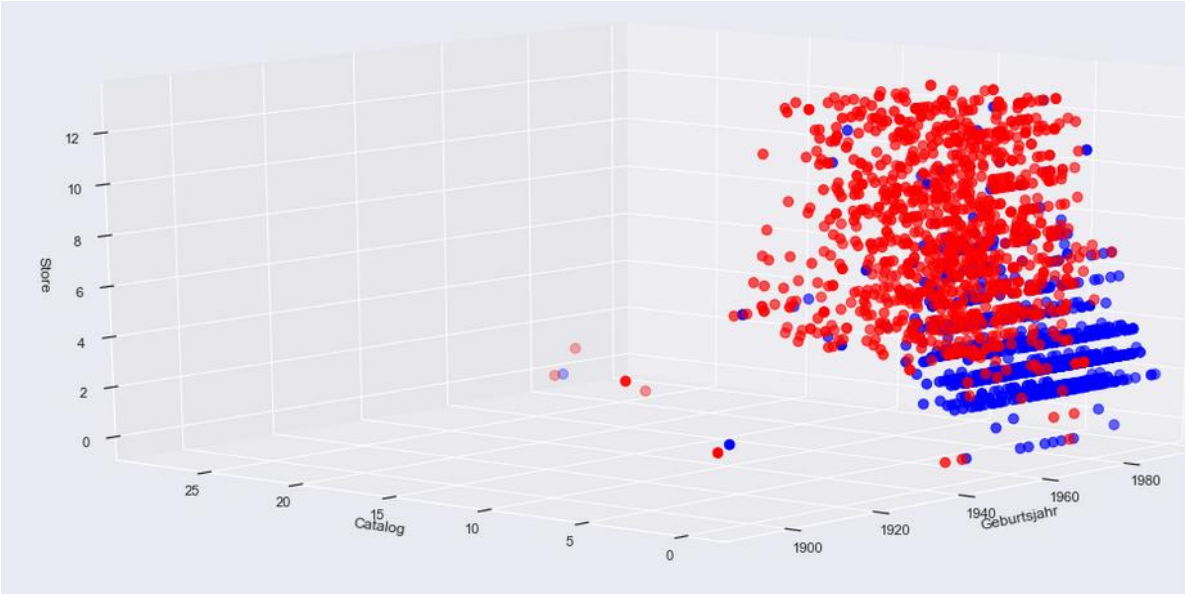
```
    184     if rgba is None:  # Suppress exception chaining of cache lookup failure.
--> 185         rgba = _to_rgba_no_colorcycle(c, alpha)
    186         try:
    187             _colors_full_map.cache[c, alpha] = rgba

~\anaconda\lib\site-packages\matplotlib\colors.py in _to_rgba_no_colorcycle(c, alpha)
    266         # float)` and `np.array(...).astype(float)` all convert "0.5" to 0.5.
    267         # Test dimensionality to reject single floats.
--> 268         raise ValueError(f"Invalid RGBA argument: {orig_c!r}")
    269     # Return a tuple to prevent the cached value from being modified.
    270     c = tuple(c.astype(float))

ValueError: Invalid RGBA argument: masked_array(data=[0.0, 0.0, 1.0, 0.609910042730091],
             mask=False,
       fill_value='?',
            dtype=object)


<Figure size 1440x720 with 1 Axes>
```

Frage 4 - Präferieren Kunden bestimmte Werbekanäle?

- → Es besteht eine positive Korrelation zwischen
  - → Web- , Katalog- und Ladenkäufen und der Menge
  - → Einkommen und der gekauften Menge

	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	Menge	Income
NumWebPurchases	1.000000	0.387000	0.516000	0.529000	0.459000
NumCatalogPurchases	0.387000	1.000000	0.518000	0.780000	0.697000
NumStorePurchases	0.516000	0.518000	1.000000	0.675000	0.630000
Menge	0.529000	0.780000	0.675000	1.000000	0.793000
Income	0.459000	0.697000	0.630000	0.793000	1.000000





Werbekanal: Catalog

Fazit mit Bezug auf Ziel und Businesskontext

→ Generelle Zielgruppe: höheres Einkommen und wenige Kinder

→ Supermarkt für Zielgruppe attraktiver gestalten und mehr in Werbung im Catalog investieren

K-Means erfüllte nicht alle unserer Erwartungen.

→ Frage 3 bereitete Schwierigkeiten und Clusterdarstellung teilweise kompliziert

Danke für Eure Aufmerksamkeit!