

# Einkaufsgewohnheiten in den USA

Ole Kepa, Fabian Elsner, Sören Bax

## Inhaltsverzeichnis

<b>1</b>	<b>Gendererklärung</b>	<b>2</b>
<b>2</b>	<b>Aufgabe und Daten verstehen</b>	<b>2</b>
<b>3</b>	<b>Beschreibung der Datenquelle</b>	<b>2</b>
<b>4</b>	<b>Untersuchung der Daten (Auf ausreißer)</b>	<b>7</b>
<b>5</b>	<b>Untersuchung der Thesen (mit Methoden der EDA)</b>	<b>7</b>
5.1	These 1 (Ole) . . . . .	7
5.2	These 2 (Ole) . . . . .	7
5.3	These 3 (Ole) . . . . .	7
5.4	These 4 (Fabian) . . . . .	7
5.5	These 5 (Fabian) . . . . .	7
5.6	These 6 (Fabian) . . . . .	7
5.7	These 7 (Sörn) . . . . .	7
5.8	These 8 (Sörn) . . . . .	7
5.9	These 9 (Sörn) . . . . .	7
<b>6</b>	<b>Anwendung und Beurteilung von Machine-Learning-Modellen</b>	<b>7</b>
6.1	Modell 1 (Ole) . . . . .	7
6.1.1	Anwendung . . . . .	7
6.1.2	Beurteilung . . . . .	7
6.2	Modell 2 (Ole) . . . . .	7
6.2.1	Anwendung . . . . .	7
6.2.2	Beurteilung . . . . .	7
6.3	Modell 3 (Fabian) . . . . .	7
6.3.1	Anwendung . . . . .	7
6.3.2	Beurteilung . . . . .	7

6.4	Modell 4 (Fabian)	7
6.4.1	Anwendung	7
6.4.2	Beurteilung	7
6.5	Modell 5 (Sörn)	7
6.5.1	Anwendung	7
6.5.2	Beurteilung	7
6.6	Modell 6 (Sörn)	7
6.6.1	Anwendung	7
6.6.2	Beurteilung	7
<b>7</b>	<b>Fazit</b>	<b>7</b>
7.1	Bewertung	7
7.2	Ideen für weitere Analysen	8
<b>8</b>	<b>Quellen</b>	<b>8</b>
<b>9</b>	<b>Ehrenwörtliche Erklärung</b>	<b>8</b>

## 1 Gendererklärung

Aus Lesbarkeitsgründen wird in dieser Studienarbeit auf die verschiedene Ansprechweisen, sei es divers, männlich oder weiblich verzichtet. Alle Formulierungen sprechen gleichermaßen alle Geschlechter an.

## 2 Aufgabe und Daten verstehen

## 3 Beschreibung der Datenquelle

Laden wir zuerst die beiden Datensets:

```
shopping_trends <- read_csv('./data/shopping_trends.csv')
shopping_behavior <- read_csv('./data/shopping_behavior_updated.csv')
```

Nun werfen wir einen kurzen Blick auf die Datenstrukturen. Zunächst einmal die Datenstruktur des Datensets “Shopping\_trends”:

```
shopping_trends
```

```
# A tibble: 3,900 x 19
  `Customer ID`   Age Gender `Item Purchased` Category Purchase Amount (USD~1
    <dbl> <dbl> <chr> <chr> <chr> <dbl>
1         1     55 Male   Blouse      Clothing      53
2         2     19 Male   Sweater     Clothing      64
3         3     50 Male   Jeans       Clothing      73
4         4     21 Male   Sandals     Footwear      90
5         5     45 Male   Blouse      Clothing      49
6         6     46 Male   Sneakers    Footwear      20
7         7     63 Male   Shirt       Clothing      85
8         8     27 Male   Shorts      Clothing      34
9         9     26 Male   Coat        Outerwear     97
10        10     57 Male   Handbag     Accessori~    31
# i 3,890 more rows
# i abbreviated name: 1: `Purchase Amount (USD)`
# i 13 more variables: Location <chr>, Size <chr>, Color <chr>, Season <chr>,
#   `Review Rating` <dbl>, `Subscription Status` <chr>, `Payment Method` <chr>,
#   `Shipping Type` <chr>, `Discount Applied` <chr>, `Promo Code Used` <chr>,
#   `Previous Purchases` <dbl>, `Preferred Payment Method` <chr>,
#   `Frequency of Purchases` <chr>
```

Jede Zeile steht für einen \*\*\*\*\*

Variable	Typ	Bedeutung
<b>Customer ID</b>	dbl	Eindeutige Kunden Identifikationsnummer
<b>Age</b>	dbl	Alter des Kunden
<b>Gender</b>	chr	Geschlecht des Kunden
<b>Item Purchased</b>	chr	Gekauftes Produkt
<b>Category</b>	chr	Kategorie des gekauften Produkts
<b>Purchase Amount</b>	dbl	Bezahlter Preis
<b>Location</b>	chr	
<b>Size</b>	chr	
<b>Color</b>	chr	
<b>Season</b>	chr	
<b>Review Rating</b>	dbl	
<b>Subscription Status</b>	chr	
<b>Payment Method</b>	chr	
<b>Shipping Type</b>	chr	
<b>Discount Applied</b>	chr	
<b>Promo Code Used</b>	chr	
<b>Previous Purchases</b>	dbl	
<b>Preferred Payment Method</b>	chr	

Variable	Typ	Bedeutung
Frequency of Purchases	chr	

```
describe_tbl(shopping_trends)
```

```
3 900 (3.9k) observations with 19 variables
0 observations containing missings (NA)
0 variables containing missings (NA)
0 variables with no variance
```

Im Datensatz “Shopping\_trends” gibt es 3.900 Instanzen (Beobachtungen). Keine dieser Instanzen enthalten Werte ohne Angabe (NA), daher müssen wir das Datenset nicht aufgrund fehlender Variablen aufbereiten. Nun schauen wir auf die Datenstruktur des Datensets “Shopping\_behavior”:

```
shopping_behavior
```

```
# A tibble: 3,900 x 18
  `Customer ID`   Age Gender `Item Purchased` Category Purchase Amount (USD~1
    <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
1             1    55 Male   Blouse         Clothing         53
2             2    19 Male   Sweater        Clothing         64
3             3    50 Male   Jeans          Clothing         73
4             4    21 Male   Sandals        Footwear         90
5             5    45 Male   Blouse         Clothing         49
6             6    46 Male   Sneakers       Footwear         20
7             7    63 Male   Shirt          Clothing         85
8             8    27 Male   Shorts         Clothing         34
9             9    26 Male   Coat           Outerwear        97
10            10    57 Male   Handbag        Accessori~       31
# i 3,890 more rows
# i abbreviated name: 1: `Purchase Amount (USD)`
# i 12 more variables: Location <chr>, Size <chr>, Color <chr>, Season <chr>,
#   `Review Rating` <dbl>, `Subscription Status` <chr>, `Shipping Type` <chr>,
#   `Discount Applied` <chr>, `Promo Code Used` <chr>,
#   `Previous Purchases` <dbl>, `Payment Method` <chr>,
#   `Frequency of Purchases` <chr>
```

Jede Zeile steht für \*\*\*\*\*.

Variable	Typ	Bedeutung
<b>Customer ID</b>	dbl	Eindeutige Kunden Identifikationsnummer
<b>Age</b>	dbl	Alter des Kunden
<b>Gender</b>	chr	Geschlecht des Kunden
<b>Item Purchased</b>	chr	Gekauftes Produkt
<b>Category</b>	chr	Kategorie des gekauften Produkts
<b>Purchase Amount</b>	dbl	Bezahlter Preis
<b>Location</b>	chr	
<b>Size</b>	chr	
<b>Color</b>	chr	
<b>Season</b>	chr	
<b>Review Rating</b>	dbl	
<b>Subscription Status</b>	chr	
<b>Payment Method</b>	chr	
<b>Shipping Type</b>	chr	
<b>Discount Applied</b>	chr	
<b>Promo Code Used</b>	chr	
<b>Previous Purchases</b>	dbl	
<b>Payment Method</b>	chr	
<b>Frequency of Purchases</b>	chr	

```
describe_tbl(shopping_behavior)
```

```
3 900 (3.9k) observations with 18 variables
0 observations containing missings (NA)
0 variables containing missings (NA)
0 variables with no variance
```

Im Datensatz “Shoping\_behavior” gibt es ebenfalls 3.900 Instazen (Beobachtungen). Zudem gibt wa qiwswe bwi jwswe Instanz keine enthalten Werte ohne Angabe (NA). Daher müssen wir, wie beim vorherigen Datensatz keine fehlenden Variablen aufbereiten.



## **4 Untersuchung der Daten (Auf ausreißer)**

## **5 Untersuchung der Thesen (mit Methoden der EDA)**

5.1 These 1 (Ole)

5.2 These 2 (Ole)

5.3 These 3 (Ole)

5.4 These 4 (Fabian)

5.5 These 5 (Fabian)

5.6 These 6 (Fabian)

5.7 These 7 (Sörn)

5.8 These 8 (Sörn)

5.9 These 9 (Sörn)

## **6 Anwendung und Beurteilung von Machine-Learning-Modellen**

6.1 Modell 1 (Ole)

6.1.1 Anwendung

6.1.2 Beurteilung

6.2 Modell 2 (Ole)

6.2.1 Anwendung

6.2.2 Beurteilung

6.3 Modell 3 (Fabian)

6.3.1 Anwendung

6.3.2 Beurteilung

6.4 Modell 4 (Fabian)

6.4.1 Anwendung

6.4.2 Beurteilung

6.5 Modell 5 (Sörn)

6.5.1 Anwendung

## 7.2 Ideen für weitere Analysen

## 8 Quellen

Datenset: <https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/data>

## 9 Ehrenwörtliche Erklärung

Hiermit erklären wir, dass wir die vorliegende Studienarbeit (Produktstudie) selbständig angefertigt haben und die Bearbeiter der einzelnen Abschnitte wahrheitsgemäß angegeben haben. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut haben wir als solches kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form ganz oder teilweise noch keiner Prüfungsbehörde vorgelegen.

*Ole Kepa Ole Kepa Ole Kepa*