Abschlusspräsentation

Vorhersage des Profits verschiedener E-Commerce Transaktionen

Agenda

- Der Datensatz
- Motivation / Ziel
- Wissenschaftliche Vorgehensweise
- Der Weg zum Ziel / Die Datenexploration
- Ergebnisse
- Fazit / kritische Reflexion

Der Datensatz

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	...	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority
0	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	...	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the-Head monaural Wir...	2309.650	7	0.0	762.1845	933.57	Critical
1	26341	IN-2013-77878	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales	...	FUR-CH-10003950	Furniture	Chairs	Novimex Executive Leather Armchair, Black	3709.395	9	0.1	-288.7650	923.63	Critical
2	25330	IN-2013-71249	17-10-2013	18-10-2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland	...	TEC-PH-10004664	Technology	Phones	Nokia Smart Phone, with Caller ID	5175.171	9	0.1	919.9710	915.49	Medium
3	13524	ES-2013-1579342	28-01-2013	30-01-2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin	...	TEC-PH-10004583	Technology	Phones	Motorola Smart Phone, Cordless	2892.510	5	0.1	-96.5400	910.16	Medium
4	47221	SG-2013-4320	05-11-2013	06-11-2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar	...	TEC-SHA-10000501	Technology	Copiers	Sharp Wireless Fax, High-Speed	2832.960	8	0.0	311.5200	903.04	Critical

Motivation

- Betriebswirtschaftlicher Zusammenhang:
  - Unternehmen testen welche Parameter ihres Produkts den höchsten Profit bringen
- Beispiel:
  - Fest steht: Es wird ein Handy zu festem Preis verkauft.
  - Offen ist: Welcher Absatzmarkt den höchsten Profit bringt

Das Ziel

Anhand eines Input-Vektors mit verschiedenen Parametern den Profit einer Transaktion vorhersagen

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_m x_m$$

Wissenschaftliche Vorgehensweise

- Vergleichbarkeit
- Messbarkeit
- Objektivität
- Repräsentativität
- Machine Learning Grundsätze
- Dokumentation / Nachvollziehbarkeit

Der Weg zum Ziel

1. Alleinstehende Aussagekräftigkeit
2. Test auf Korrelation mit dem Profit:
   - Visualisierung durch Diagramme
   - Rechnerische/statistische Bestimmung der Abhängigkeit beider Attribute
3. Interpretation der Ergebnisse
4. Attribut löschen oder beibehalten
5. Modell aufsetzen

Data Exploration – alleinstehende Aussage-kräftigkeit

```
numberOfTransactions("City")
```

City	Number of Transactions
Río Bravo	1
Montería	1
Sikasso	1
Chengjiang	1
Cheonan	1
...	...
Santo Domingo	443
San Francisco	510
Philadelphia	537
Los Angeles	747
New York City	915

3636 rows × 1 columns

```
howOftenDoAmountsAppear("City")
```

	Number of Transactions	counts
0	1	488
1	2	419
2	3	349
3	4	293
4	5	234
...	...	...
151	443	1
152	510	1
153	537	1
154	747	1
155	915	1

156 rows × 2 columns

```
df['Postal Code']
```

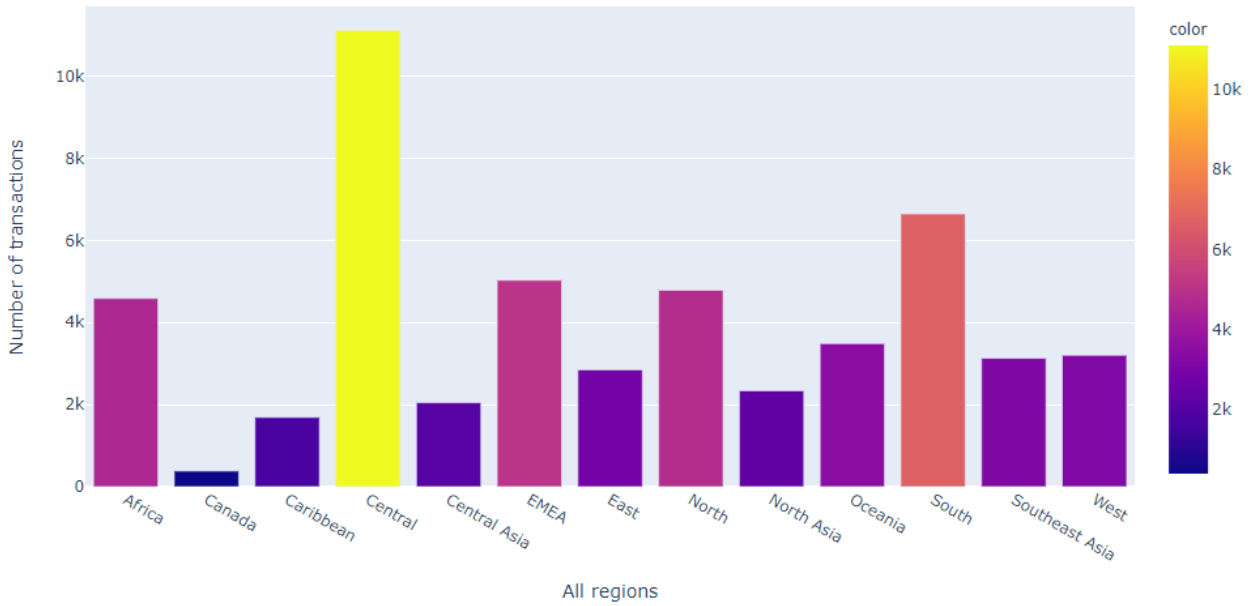
```
0        10024.0
1            NaN
2            NaN
3            NaN
4            NaN
          ...
51285        NaN
51286    77095.0
51287    93030.0
51288        NaN
51289        NaN
Name: Postal Code, Length: 51290, dtype: float64
```

```
df['Postal Code'].isna().sum()
```

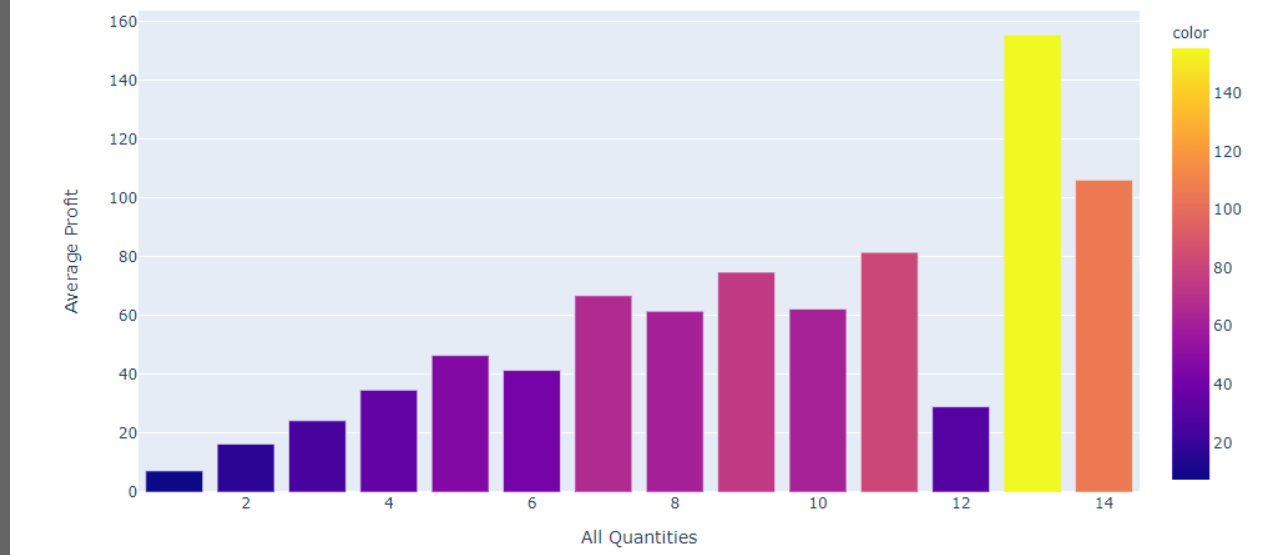
```
41296
```

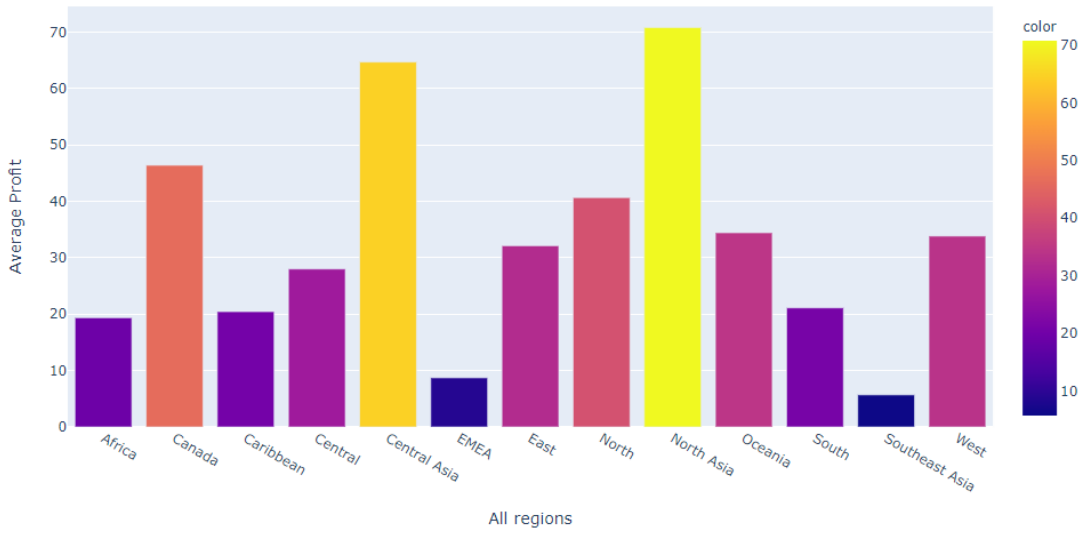
Von den 51290 haben 41296 Transaktionen keinen Wert bei Postal Code.

```
del df["Postal Code"]
```

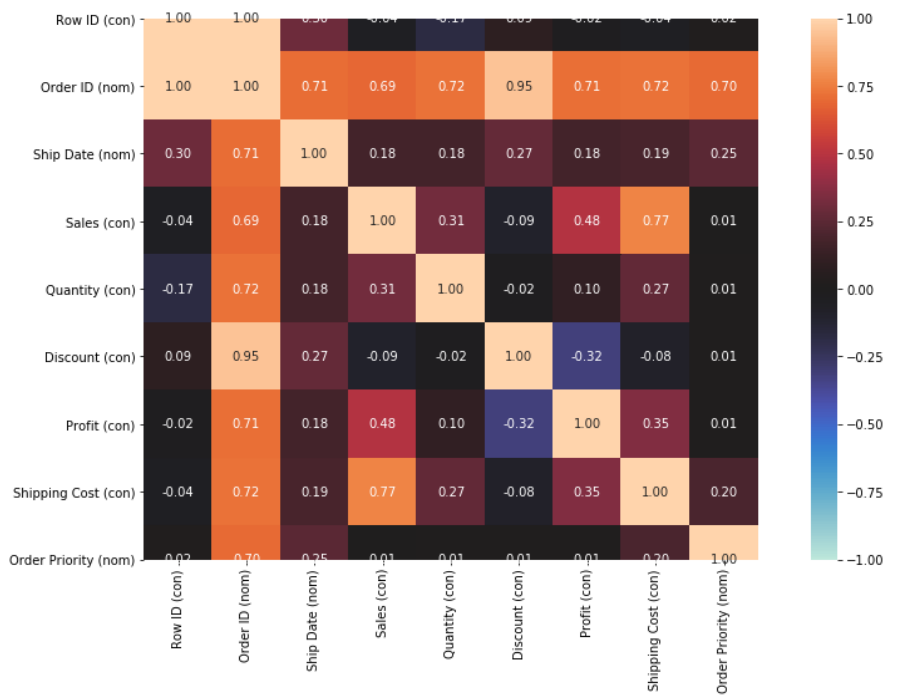


Data Exploration – Test auf grafische Korrelation mit Profit





Data Exploration – Test auf rechnerische Korrelation mit Profit



Wert aus der Heatmap: 0,09

```
calculateAnovaScore("Region","Profit")
```

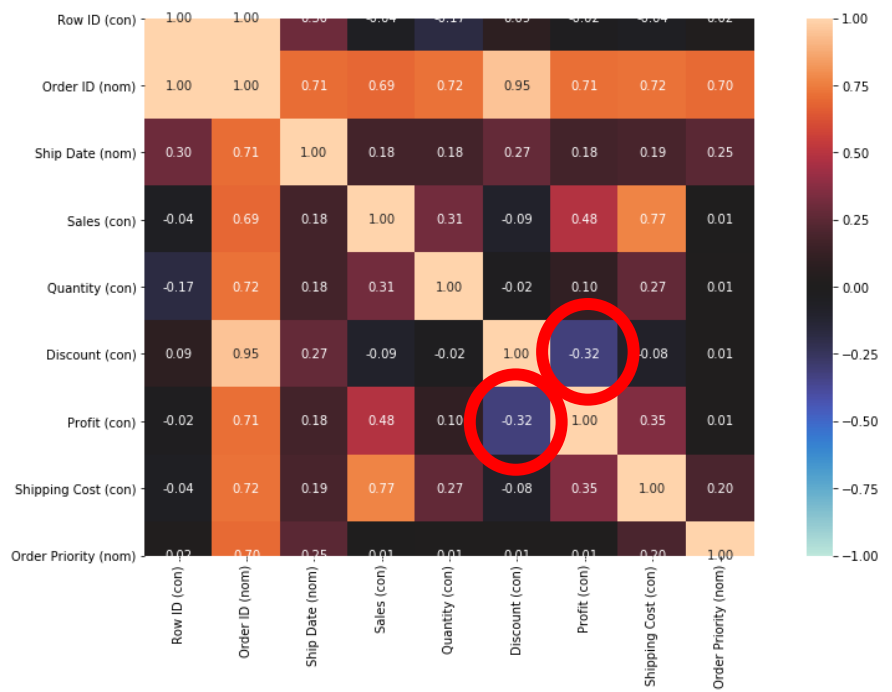
```
P-Wert für Anova ist:  3.617471983600004e-80
F Score für Anova ist:  34.261215682356564
Der Unterschied der Profit-Mittelwerte zwischen den verschiedenen Gruppen von Region ist signifikant
```

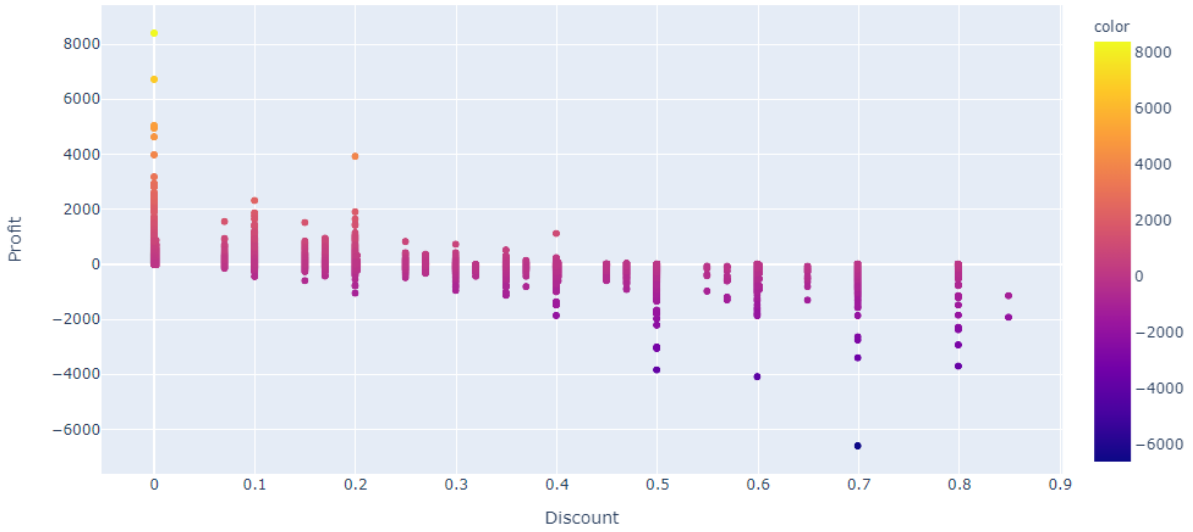
Wert aus Heatmap: 0,01

```
calculateAnovaScore("Order Priority","Profit")
```

```
P-Wert für Anova ist:  0.22283640166560595
F Score für Anova ist:  1.461514297034312
```

Data Exploration – Interpretation der Ergebnisse: Beispiel Discount

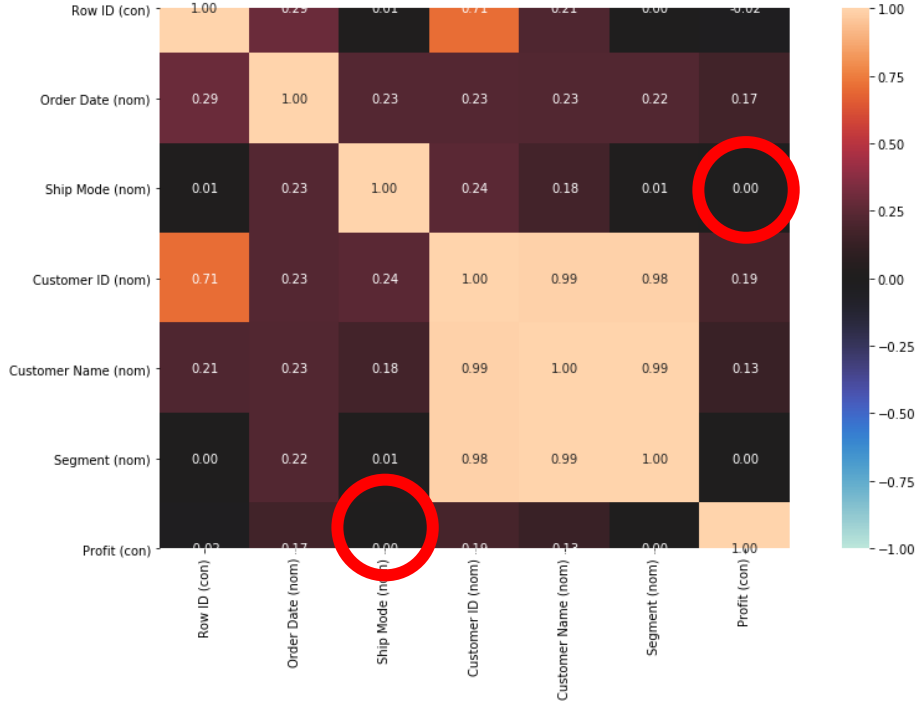


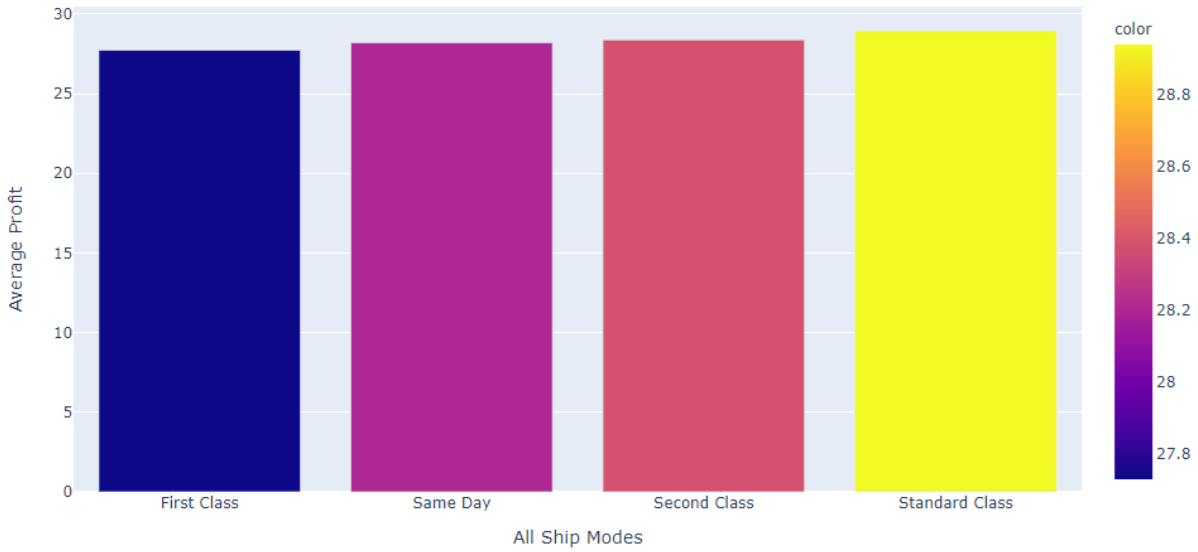


4.4 Fazit

Discount sollte definitiv für das finale Modell genutzt werden. Denn es gibt eindeutig eine Korrelation zwischen Discount und Profit. Diese Korrelation ist negativ. Im Sachzusammenhang heißt das, dass je höher der Discount ist, desto niedriger ist der Wert für Profit. Dies zeigt sich sowohl im Korrelationswert,als auch vor Allem im Streudiagramm.

Data Exploration – Interpretation der Ergebnisse: Beispiel Ship Mode





Wert aus Heatmap: 0

```
calculateAnovaScore("Ship Mode","Profit")
```

```
P-Wert für Anova ist:  0.9535491202134 93
F Score für Anova ist:  0.11126944003460387
```

Data Exploration – Löschen der unabhängigen Attribute

```
del df["State"]
```

```
del df["City"]
```

```
del df["Postal Code"]
```

```
del df["Order Priority"]
```

```
del df["Order ID"]
```

	Row ID	Country	Market	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Shipping Cost
0	32298	United States	US	East	Technology	Accessories	2309.650	7	0.0	762.1845	933.57
1	26341	Australia	APAC	Oceania	Furniture	Chairs	3709.395	9	0.1	-288.7650	923.63
2	25330	Australia	APAC	Oceania	Technology	Phones	5175.171	9	0.1	919.9710	915.49
3	13524	Germany	EU	Central	Technology	Phones	2892.510	5	0.1	-96.5400	910.16
4	47221	Senegal	Africa	Africa	Technology	Copiers	2832.960	8	0.0	311.5200	903.04
...	...	...	...	...	...	...	...	...	...	...	...
51285	29002	Japan	APAC	North Asia	Office Supplies	Fasteners	65.100	5	0.0	4.5000	0.01
51286	35398	United States	US	Central	Office Supplies	Appliances	0.444	1	0.8	-1.1100	0.01
51287	40470	United States	US	West	Office Supplies	Envelopes	22.920	3	0.0	11.2308	0.01
51288	9596	Brazil	LATAM	South	Office Supplies	Binders	13.440	2	0.0	2.4000	0.00
51289	6147	Nicaragua	LATAM	Central	Office Supplies	Paper	61.380	3	0.0	1.8000	0.00

51290 rows × 11 columns

Data Exploration – Modell aufsetzen

```
df['Category'] =df['Category'].astype('category').cat.codes
df['Sub-Category'] =df['Sub-Category'].astype('category').cat.codes
df['Country'] =df['Country'].astype('category').cat.codes
df['Market'] =df['Market'].astype('category').cat.codes
df['Region'] =df['Region'].astype('category').cat.codes
```

```
df
```

	Row ID	Country	Market	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Shipping Cost
0	32298	139	6	6	2	0	2309.650	7	0.0	762.1845	933.57
1	26341	6	0	9	0	5	3709.395	9	0.1	-288.7650	923.63
2	25330	6	0	9	2	13	5175.171	9	0.1	919.9710	915.49
3	13524	47	4	3	2	13	2892.510	5	0.1	-96.5400	910.16
4	47221	110	1	0	2	6	2832.960	8	0.0	311.5200	903.04
...	...	...	...	...	...	...	...	...	...	...	...
51285	29002	65	0	8	1	8	65.100	5	0.0	4.5000	0.01
51286	35398	139	6	3	1	1	0.444	1	0.8	-1.1100	0.01
51287	40470	139	6	12	1	7	22.920	3	0.0	11.2308	0.01
51288	9596	17	5	10	1	3	13.440	2	0.0	2.4000	0.00
51289	6147	92	5	3	1	12	61.380	3	0.0	1.8000	0.00

4. Ridge Regression

```
reg = linear_model.Ridge (alpha = .55)
reg.fit(X_train, y_train)
print('Score: ', reg.score(X_test, y_test))
print('Weights: ', reg.coef_)

plt.plot(reg.predict(X_test))
plt.plot(y_test)
plt.show()
```

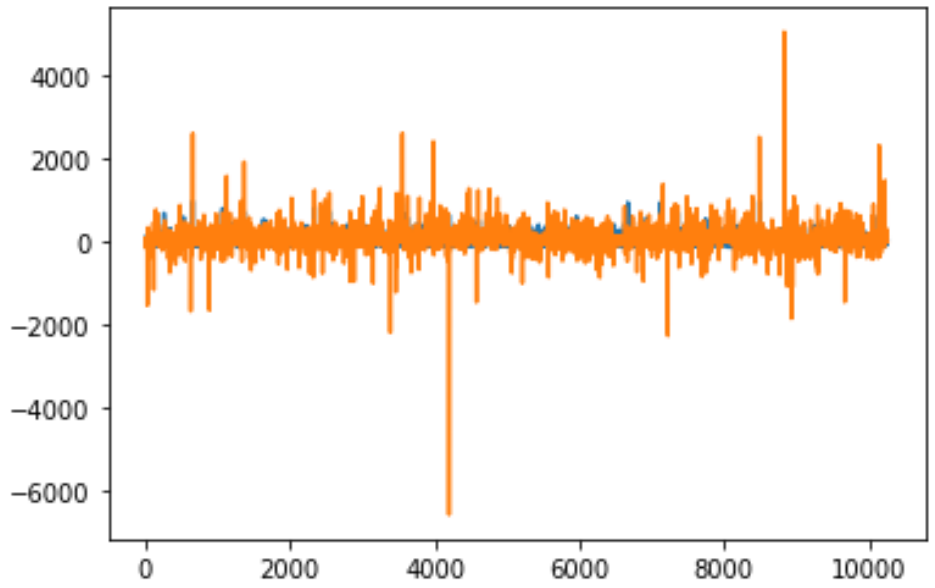
```
Score:  0.2851107659420663
Weights:  [ 9.75163222e-02 -6.78488133e-01 -2.36060740e-01  9.00804676e+00
 -1.75012557e+00  1.78819909e-01 -3.34300358e+00 -2.29024277e+02
 -9.36847813e-02]
```

```
X = np.asarray(df[['Country','Market','Region','Category','Sub-Category','Sales','Quantity','Discount','Shipping Cost']])
Y = np.asarray(df['Profit'])

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, shuffle= True)
```

```
ScoreLinearRegression(50)
ScoreRidgeRegression(50)
```

```
Linear Regression
0.31351977115824076
[ 9.62680903e-02 -5.18547620e-01 -1.66981826e-01  1.00533668e+01
 -1.70204605e+00  1.81090347e-01 -3.55344448e+00 -2.29934252e+02
 -1.37444936e-01]

Ridge Regression
0.29477039286070505
[ 9.86004724e-02 -4.91344459e-01 -1.79858040e-01  1.00986354e+01
 -1.71819788e+00  1.85668790e-01 -3.67016459e+00 -2.29366512e+02
 -1.57310611e-01]
```

Data Exploration – Modell aufsetzen




```
tabelleErzeugen("Market")
```

	Name	Zahl
0	APAC	0
1	Africa	1
2	Canada	2
3	EMEA	3
4	EU	4
5	LATAM	5
6	US	6

reg.predict([['Country','Market','Region','Category','Sub-Category','Sales','Quantity','Discount','Shipping Cost']])

```
reg.predict([[139, 6, 6, 2, 13, 1000, 1, 0.1, 400]])
```

```
reg.predict([[110, 1, 0, 2, 13, 1000, 1, 0.1, 1200]])
```

Ergebnisse

- Prüfung der Korrelationen
- Filterung der relevanten Attribute
- Aufgesetztes Modell
- Ridge Regression
- Methodenaufruf zur Vorhersage des Profits
- Parameter nehmen Einfluss auf Profit

Ergebnis des Beispiels

1. Variante:

Country: 139 (United States), Market: 6 (US), Region: 7 (North), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 700

```
getCountryNumber("United States")
```

	Name	Zahl
139	United States	139

```
reg.predict([[139, 6, 7, 2, 13, 1000, 1, 0.1, 400]])
```

```
array([149.83844811])
```

2. Variante:

Country: 139 (United States), Market: 6 (US), Region: 6 (East), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 700

```
reg.predict([[139, 6, 6, 2, 13, 1000, 1, 0.1, 400]])
```

```
array([150.07450885])
```

3. Variante:

Country: 139 (United States), Market: 6 (US), Region: 12 (West), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 700

```
reg.predict([[139, 6, 12, 2, 13, 1000, 1, 0.1, 400]])
```

```
array([148.65814441])
```

4. Variante:

Country: 17 (Brazil), Market: 5 (LATAM), Region: 10 (South), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 700

```
getCountryNumber("Brazil")
```

	Name	Zahl
17	Brazil	17

```
reg.predict([[17, 5, 10, 2, 13, 1000, 1, 0.1, 700]])
```

```
array([109.80632833])
```

5. Variante:

Country: 47 (Germany), Market: 4 (EU), Region: 3 (Central), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 650

```
reg.predict([[47, 4, 3, 2, 13, 1000, 1, 0.1, 650]])
```

```
array([119.74697037])
```

Ergebnis des Beispiels

Ergebnis des Beispiels

6. Variante:

Country: 110 (Senegal), Market: 1 (Africa), Region: 0 (Africa), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 1200

```
getCountryNumber("Senegal")
```

	Name	Zahl
110	Senegal	110

```
reg.predict([[110, 1, 0, 2, 13, 1000, 1, 0.1, 1200]])
```

```
array([77.10751558])
```

7. Variante:

Country: 26 (China), Market: 0 (APAC), Region: 8 (North Asia), Category: 2 (Technology), Sub-Category: 13 (Phones), Quantity: 1, Discount: 0.1, Shipping Cost: 950

```
getCountryNumber("China")
```

	Name	Zahl
26	China	26

```
reg.predict([[26, 0, 8, 2, 13, 1000, 1, 0.1, 950]])
```

```
array([91.12734206])
```

Fazit / kritische Reflexion

Positiv

- Sehr detaillierte Datenexploration
- Entscheidungen über Attributwahl valide und repräsentativ
- Modell liefert passende Ergebnisse
- Ziel der Profitvorhersage erreicht
- Betriebswirtschaftlicher Nutzen anwendbar

Negativ

- Modell nicht zu Ende trainiert
- Parameter noch nicht perfekt zur Abbildung der Daten
- Größe der Datenmenge
- Ggf. Noch andere Modelle als Regressionen möglich


Danke für die Aufmerksamkeit
Gibt es Fragen?