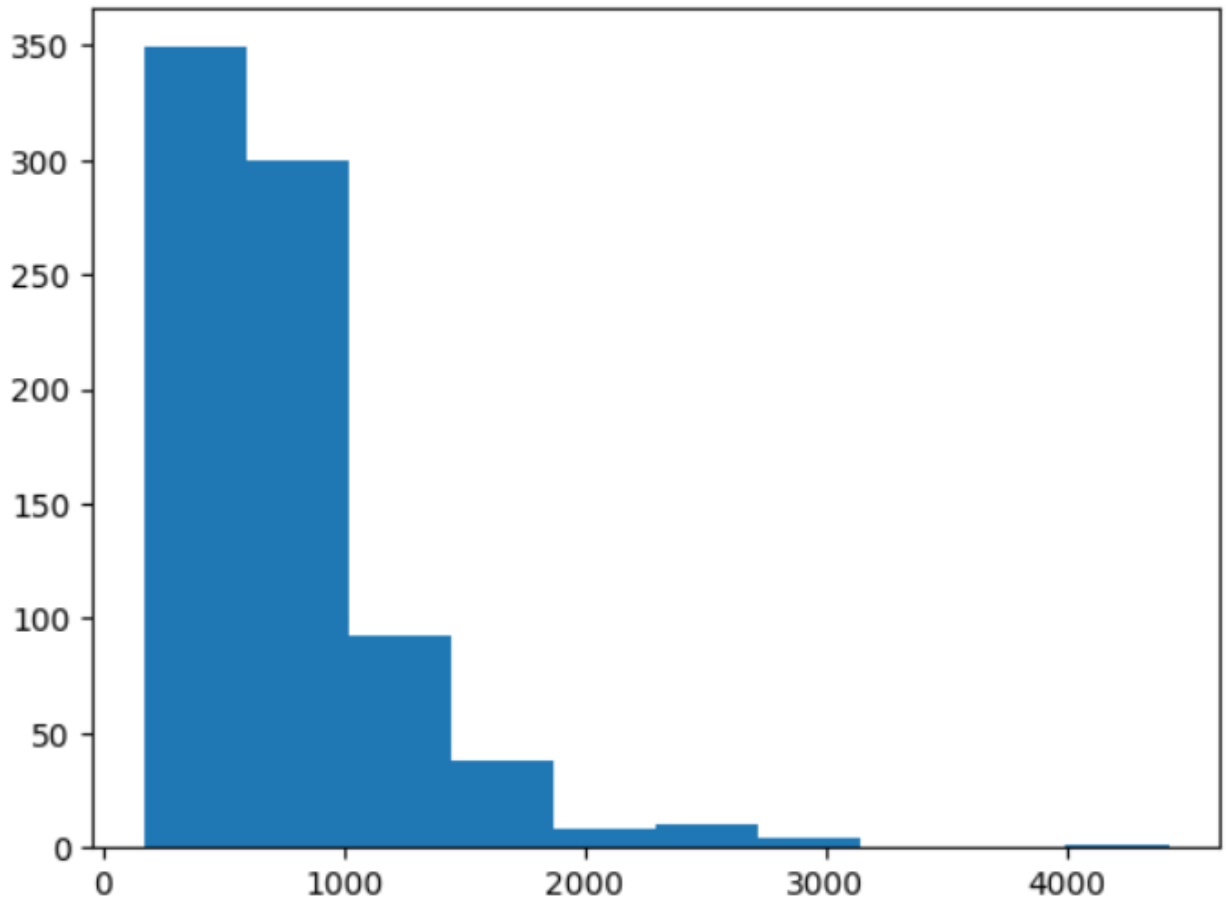
M259 – Laptop-Preise vorhersagen



Inhalt

Datengrundlage....3
Beschreiben der Spalten im Dataset....3
EDA....5
Einige Variablen Analysieren....5
Gruppierungen Analysieren....11
Korrelation von Variablen feststellen....14
Hypothesen....15
Hypothese 1....15
Hypothese 2....16
Hypothese 3....17
Modelle....17
Testdaten....17
Modell 1....17
Modell 2....19
Modell 3....20
Zusammenfassung....20

Datengrundlage

Beschreiben der Spalten im Dataset

```
<class 'pandas.core.frame.DataFrame'>
Index: 802 entries, 0 to 822
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   brand              802 non-null    object
 1   processor_brand    802 non-null    object
 2   processor_name     802 non-null    object
 3   processor_gnrtn    802 non-null    object
 4   ram_gb             802 non-null    object
 5   ram_type           802 non-null    object
 6   ssd                802 non-null    object
 7   hdd                802 non-null    object
 8   os                 802 non-null    object
 9   os_bit             802 non-null    object
 10  graphic_card_gb    802 non-null    object
 11  weight             802 non-null    object
 12  warranty           802 non-null    object
 13  Touchscreen        802 non-null    object
 14  msoffice           802 non-null    object
 15  Price              802 non-null    float64
 16  rating             802 non-null    object
 17  Number of Ratings  802 non-null    int64
 18  Number of Reviews  802 non-null    int64
dtypes: float64(1), int64(2), object(16)
memory usage: 125.3+ KB
```

Abbildung 1 Info des Datensatzes

brand: Brand ist die Marke des Laptops, hier gibt es insgesamt acht verschiedene zur Auswahl. Die Marken sind: Asus, Lenovo, acer, Avita, HP, DELL, MSI und APPLE. In dieser Spalte gibt es keine Null Werte. Der Datentyp ist ein Objekt bzw. ein String.

processor_brand: Hier wird angegeben von welcher Marke der Prozessor des Laptops stammt. Auch hier gibt es keine Null Werte. Es gibt insgesamt drei verschiedene Prozessor Marken. Diese sind: Intel, AMD und M1. Der Datentyp ist ein Objekt.

processor_name: Bei dieser Spalte wird der Name des Prozessors angegeben. Hier gibt es wesentlich mehr unterschiedliche Prozessoren. Insgesamt 11 verschiedene. Der Datentyp ist ein Objekt.

```
df.processor_name.unique()
```

```
array(['Core i3', 'Core i5', 'Celeron Dual', 'Ryzen 5', 'Core i7',
       'Core i9', 'M1', 'Pentium Quad', 'Ryzen 3', 'Ryzen 7', 'Ryzen 9'],
      dtype=object)
```

Abbildung 2 Prozessor Namen

processor_gnrtn: Hier wird angeben aus welcher Generation der Prozessor stammt. Hierbei ist auffällig, dass bei gewissen Prozessoren ein „Not Available“ angegeben ist. Der Datentyp ist ein Objekt.

ram_gb: Es wird angegeben wie viel Ram der Laptop hat. Die verschiedenen Anzahlen sind: 4GB, 8GB, 16GB und 32GB. Der Datentyp ist ein Objekt.

ram_type: Hier wird angegeben, welche Typen von Ram es gibt. Die sechs verschiedenen Arten sind: DDR4, LPDDR4, LPDDR4X, DDR5, DDR3 und LPDDR3. Der Datentyp ist ein Objekt.

ssd: Bei dieser Spalte sieht man wie viel SSD-Speicherplatz der Laptop hat. Es stehen insgesamt sieben Kategorien zur Auswahl. Die verschiedenen Anzahlen an Speicherplatz sind: 0GB, 512GB, 256GB, 128GB, 1024GB, 2048GB und 3072GB. Der Datentyp ist ein Objekt.

hdd: Hier wird angegeben, welche Anzahl and HDD-Speicherplatz der Laptop hat. Die vier verschiedenen Anzahlen sind: 1024GB, 0GB, 512GB und 2048GB. Der Datentyp ist ein Objekt.

os: In dieser Spalte wird angegeben welches Betriebssystem auf dem Laptop installiert ist. Dabei stehen drei zur Auswahl. Die Betriebssysteme sind: Windows, DOS und Mac. Der Datentyp ist ein Objekt.

os_bit: Hier wird angegeben welche Bitrate das Betriebssystem hat. Es stehen nur 32Bit oder 64Bit zur Auswahl. Der Datentyp ist ein Objekt.

graphic_card_gb: Es gibt die Leistung der Grafikkarte an. Hierbei gibt es insgesamt fünf verschiedene. Diese sind: 0GB, 2GB, 4GB, 6GB und 8GB. Der Datentyp ist ein Objekt.

weight: In dieser Spalte wird das Gewicht des Laptops in verschiedene Kategorien unterteilt. Die drei Kategorien heissen: Casual, ThinNlight und Gaming. Der Datentyp ist natürlich auch ein Objekt.

warranty: Hier wird die Garantie länge des Laptops angegeben. Gewisse Laptops haben keine Garantie also „No warranty“. Bei den restlichen Laptops gibt es eine Auswahl zwischen einem Jahr, zwei Jahren oder drei Jahren Garantie. Der Datentyp ist ein Objekt.

Touchscreen: Hier wird angegeben, ob der Bildschirm des Laptops ein Touchscreen ist oder nicht. In dieser Spalte gibt es insgesamt nur zwei Werte, entweder No oder Yes. Der Datentyp ist ein Objekt/Boolean.

msoffice: In dieser Spalte wird angegeben, ob der Laptop ein Microsoft Office vorinstalliert hat oder nicht. Die zwei Auswahlen sind: No und Yes. Der Datentyp ist ein Objekt/Boolean.

Price: In dieser Spalte wird der Preis des Laptops in Rupee angegeben. Ich habe den Preis jedoch in CHF umgewandelt. Der Datentyp ist ein Int64 (Integer). Ich habe diesen jedoch in einen float umgewandelt.

rating: Hier wird die Bewertung des Laptops in Sternen angegeben, es gibt eine Auswahl von 1-5 Sterne. Der Datentyp ist ein Objekt.

Number of Ratings: In dieser Spalte wird die Anzahl von Ratings angegeben. Der Datentyp ist ein Int64(Integer).

Number of Reviews: In dieser Spalte wird die Anzahl an Reviews über einen Laptop angegeben. Der Datentyp ist ein Int64(Integer).

EDA

Einige Variablen Analysieren

Anzahl Datensätze: 823 Datensätze

Anzahl NaN: 823 bei je 19 Spalten

Grössenordnung: Der günstigste Laptop kostet: 169.90 CHF, der teuerste kostet: 4419.90 CHF, Im Durchschnitt kostet ein Laptop 766.25 CHF. Im Durchschnitt hat ein Laptop 300 Ratings und 36 Reviews. Die höchste Anzahl Ratings für einen Laptop sind 15279 und die kleinste Anzahl Ratings sind 0. Die höchste Anzahl von Reviews an einem Laptop sind 1947 Reviews und die niedrigste Anzahl ist 0.

	Price	Number of Ratings	Number of Reviews
count	802.000000	802.00000	802.000000
mean	766.255436	299.84414	36.089776
std	452.329844	1001.78442	118.313553
min	169.900000	0.00000	0.000000
25%	459.900000	0.00000	0.000000
50%	639.900000	17.00000	2.000000
75%	895.250000	140.25000	18.000000
max	4419.900000	15279.00000	1947.000000

Abbildung 3 Describe des Datensatzes

Messeinheiten: Bei denn Messeinheiten gibt es nur die Garantie in Jahren und die verschiedenen Speicherplätze in GB. Das Rating wird als Sterne angegeben. Bei den Os-Bit werden die Daten als Bit angegeben.

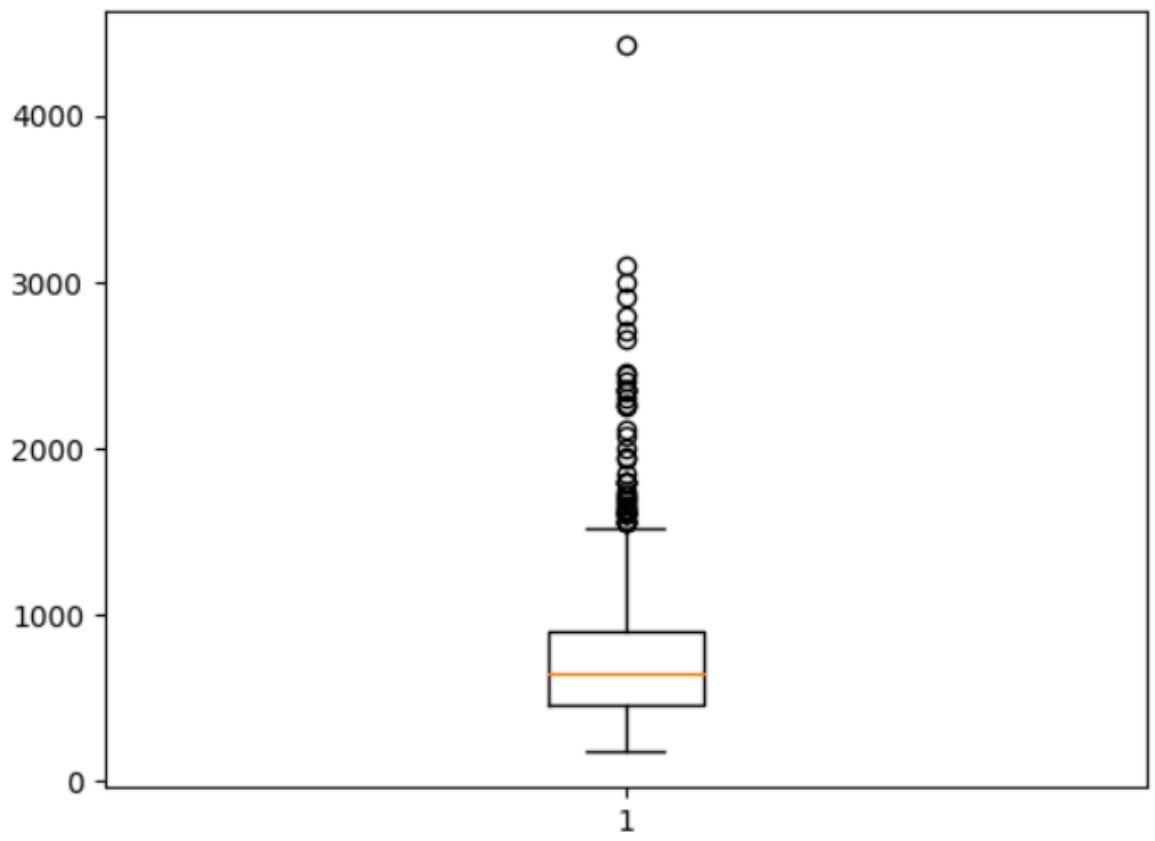


Abbildung 4 BoxPlot des Preises in CHF

Hier ist ein BoxPlot der Laptop Preise zu sehen. Wie man gut sehen kosten die meisten Laptops zwischen 459.90 und 895.25 CHF. Dabei gibt es auch einige Ausreisser. Man erkennt sehr deutlich der Laptop der 4419.90 CHF kostet und da mit Abstand am teuersten ist.

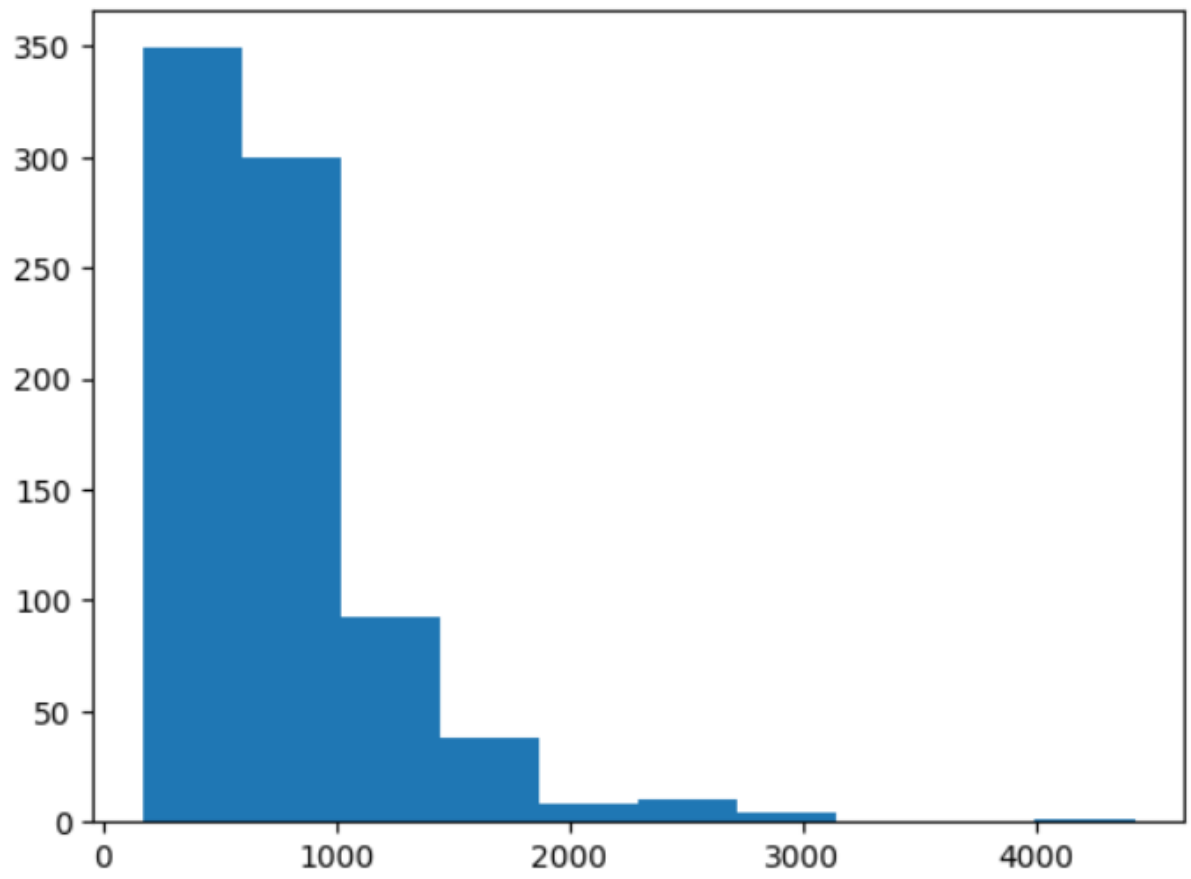


Abbildung 5 Histogramm der Laptop Preise

In dieser Grafik erkannt man gut, dass die meisten Laptops zwischen 169.90 und 594.90 CHF liegen. Und das mit einer Anzahl von 349 Laptops. Man kann gut sehen, dass es von den teureren Laptops nur wenige gibt.

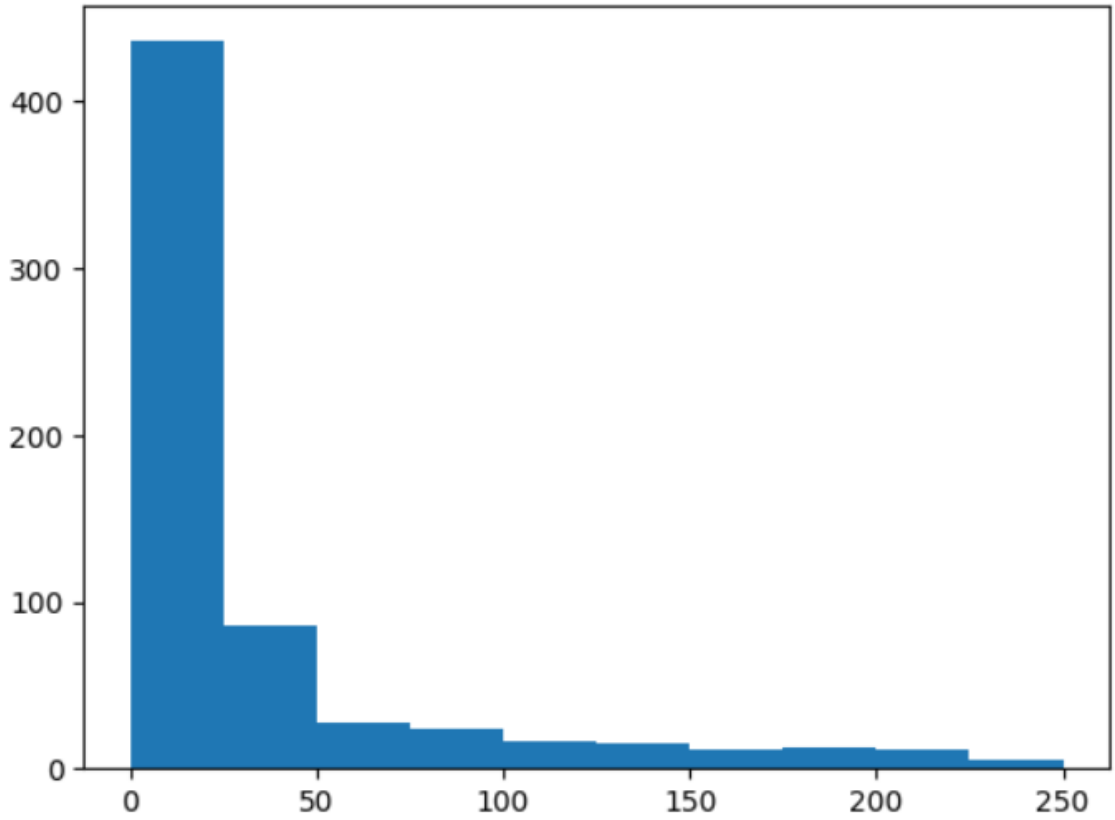


Abbildung 6 Histogramm der Anzahl Ratings bis 250

Gut bei diesem Histogramm zu sehen ist, dass die meisten Laptops gar keine Ratings bis sehr wenige Ratings haben. Nur wenige Laptops haben eine grössere Anzahl Ratings.

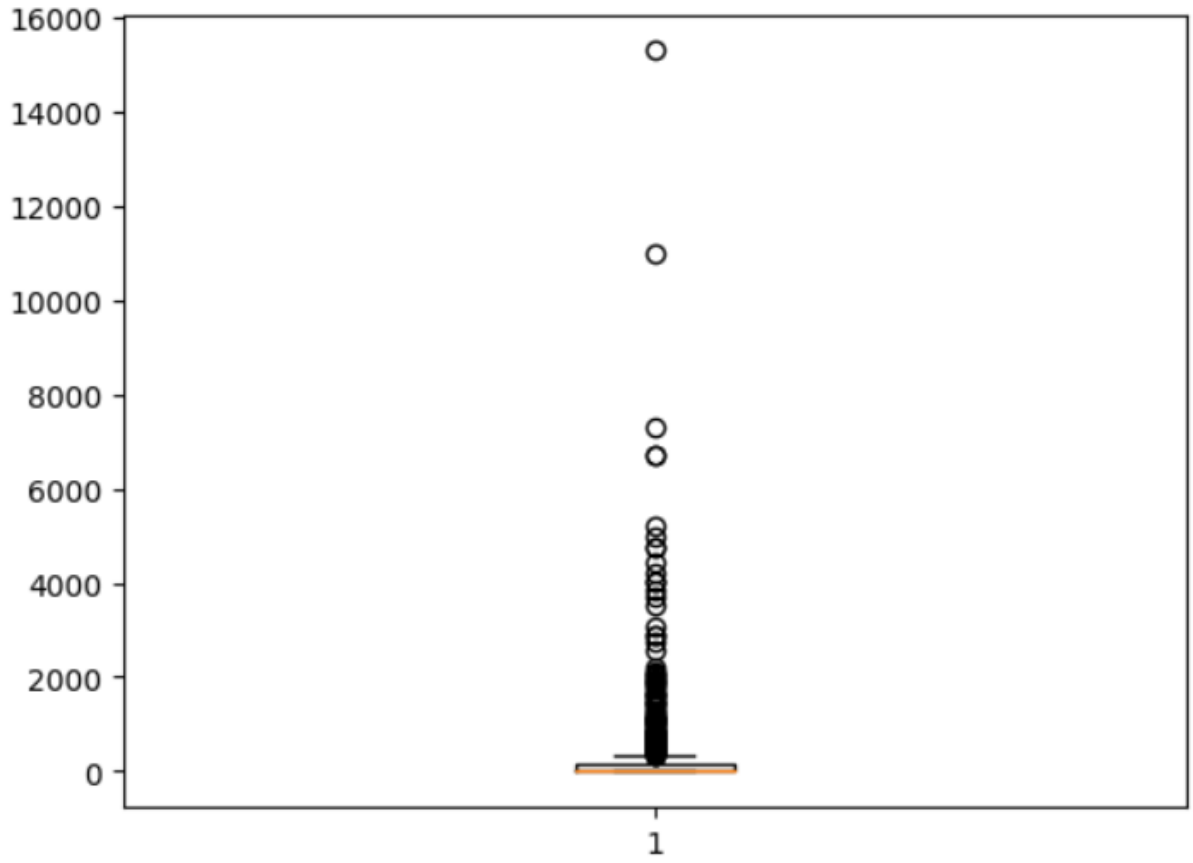


Abbildung 7 BoxPlot der Anzahl Reviews

Gut zu erkennen ist die weite Streuung der Anzahl Ratings. Die Range geht von 0 bis 15279. Der BoxPlot ist sehr unübersichtlich. Doch man erkennt gut, dass die meisten Laptops keine bis wenig Ratings haben. Gut zu sehen sind, die vielen Ausreisser.

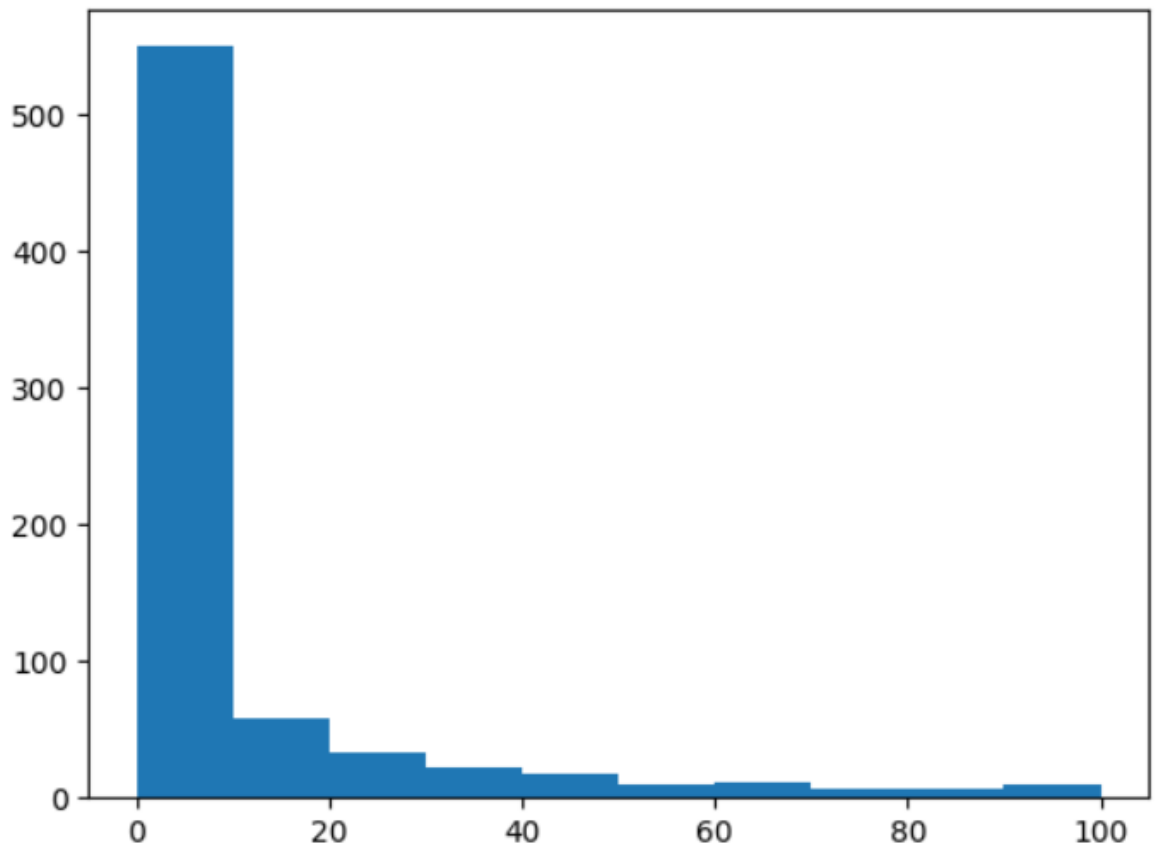


Abbildung 8 Histogramm der Anzahl Reviews bis 100

Bei diesem Histogramm erkennt man gut, dass die meisten Laptops keine bis wenig Reviews haben. Es gibt insgesamt 550 Laptops, die zwischen 0 und 10 Reviews haben.

Gruppierungen Analysieren

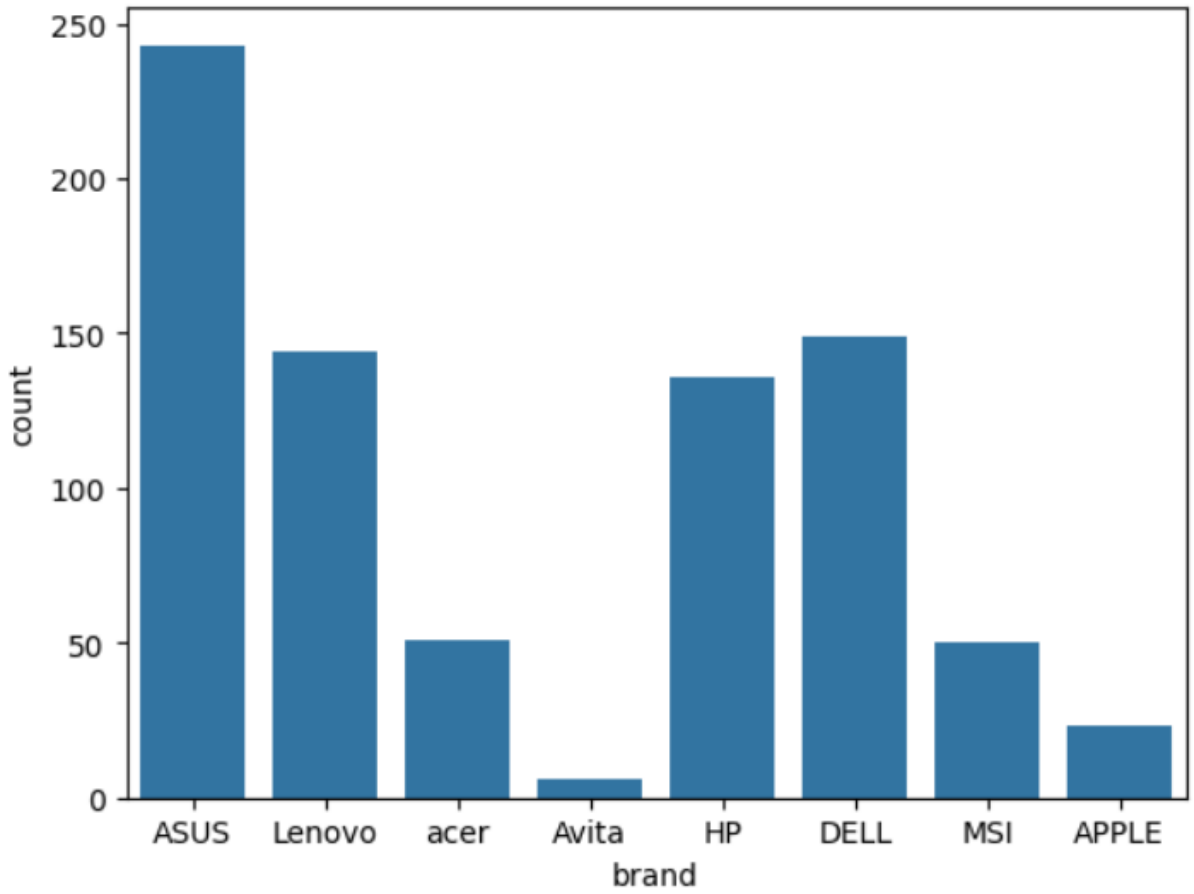


Abbildung 9 Countplot der Laptop-Marken

Von den 823 Laptops sind insgesamt:

APPLE: 23

Asus: 243

Avita: 6

DELL: 149

HP: 136

Lenovo: 144

MIS: 50

Acer: 51

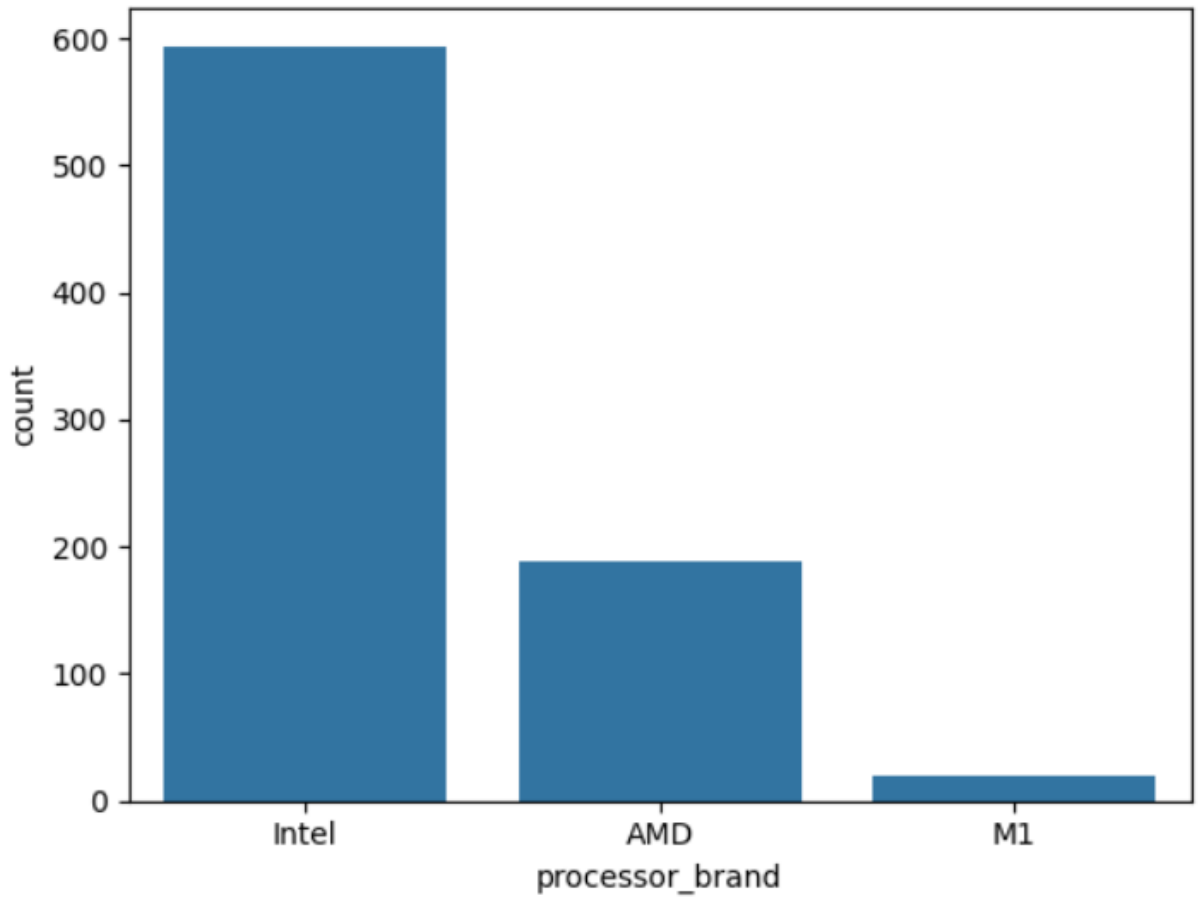


Abbildung 10 Countplot der Prozessor-Marken

Von den 823 Prozessoren sind insgesamt:

AMD: 189

Intel: 594

M1: 19

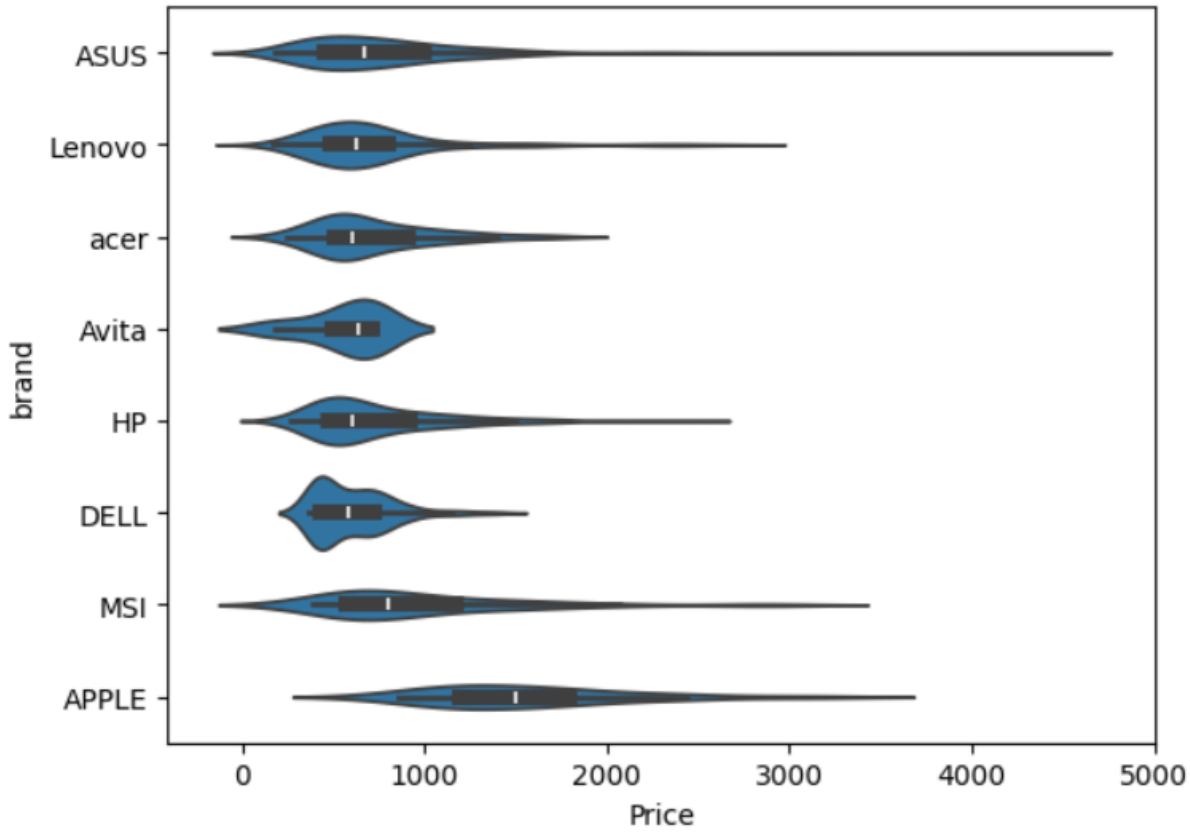


Abbildung 11 Violinplot der Brand Preise

Gut auf dem Violinplot zu erkennen ist, dass Asus Laptop, die Günstig bis sehr teuer sind. Apple hingegen hat im durchschnitt die teureren Laptops und dass, mit einem deutlichen Unterschied. Dell hat mehrheitlich Günstigere Laptops. Avitas teuerster Laptop ist knapp über 1000 CHF.

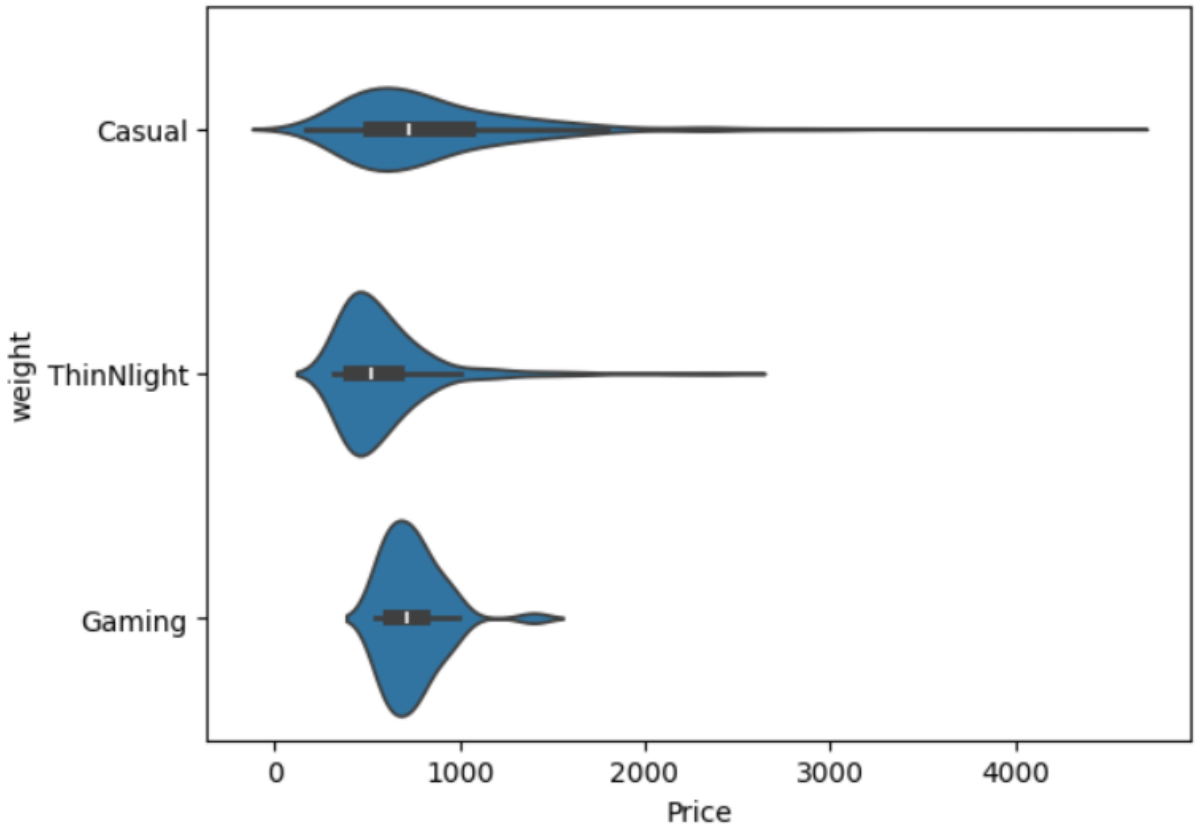


Abbildung 12 Violinplot des Preisverhältnises der verschiedenen Gewichts-Klassen

Auf dem Plot kann man gut erkennen, dass die Gaming Laptops eine kleine Preis-Range haben im Vergleich zu den anderen. Die Casual Laptops haben aber auch durch, denn einen Ausreisser eine sehr grosse Preis-Range.

Korrelation von Variablen feststellen

	Price	Number of Ratings	Number of Reviews
Price	1.000000	-0.152553	-0.156791
Number of Ratings	-0.152553	1.000000	0.991062
Number of Reviews	-0.156791	0.991062	1.000000

Abbildung 13 Korrelation der Zahlen

Die Korrelation zwischen den «Number of Ratings» und «Number of Reviews» ist sehr hoch. Dies bedeutet somit, dass man sagen könnte, dass die Laptops mit mehr Ratings auch gleichzeitig mehr Reviews haben. Beim Preis ist es sehr schwierig etwas einzuschätzen, da die Korrelationen sehr niedrig sind.

Hypothesen

Hypothese 1

Aussage: Ein Laptop der in der Regel mehr Ratings hat, hat auch gleichzeitig mehr Reviews.

	Price	Number of Ratings	Number of Reviews
Price	1.000000	-0.152553	-0.156791
Number of Ratings	-0.152553	1.000000	0.991062
Number of Reviews	-0.156791	0.991062	1.000000

Abbildung 14 Korrelation zwischen Ratings und Reviews

Antwort: Diese Aussage stimmt. Die Korrelation zwischen Ratings und Reviews ist mit 0.99 eine der höchsten Korrelationen, die auftreten können. Somit ist mit grosser Wahrscheinlichkeit die Anzahl der Ratings von der Anzahl der Reviews abhängig. Durch diese hohe Korrelation hat sich die Hypothese als wahr herausgestellt. Auf dem Scaterplot ist auch eine sehr lineare Funktion zu sehen.

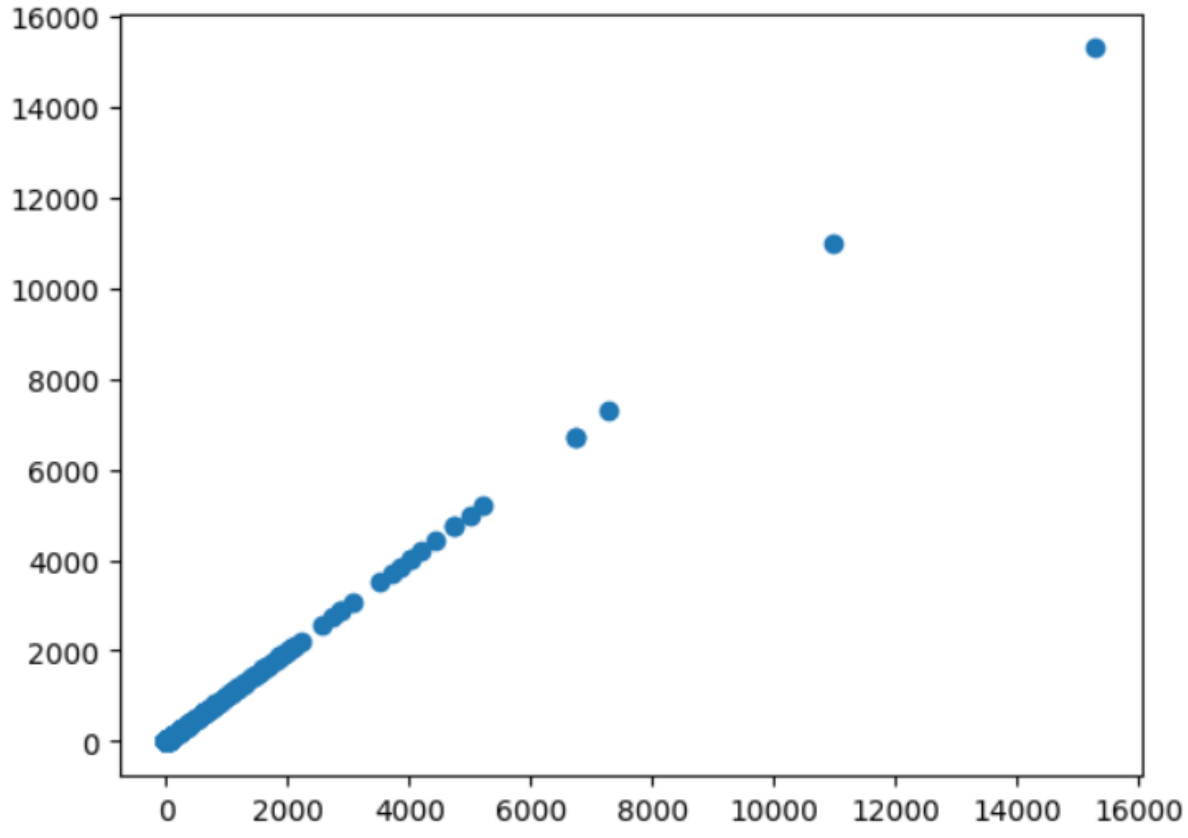


Abbildung 15 ScaterPlot der Anzahl Ratings und Reviews

Hypothese 2

Aussage: Die Apple Laptops sind im Durchschnitt am teuersten.

brand	
APPLE	1570.813043
ASUS	794.260206
Avita	563.465000
DELL	607.889530
HP	736.402721
Lenovo	729.202083
MSI	987.130200
acer	724.200392

Abbildung 16 Preisdurchschnitt der Marken

Antwort: Auf der Berechnung ist klar zu erkennen, dass die Apple Laptops im Durchschnitt um einiges teurer sind als die der anderen Marken. Obwohl Asus einen starken Ausreisser hat, ist der Durchschnitt um einiges tiefer als der von Apple. Wichtig noch zu berücksichtigen ist, dass Apple weniger Laptops als andere Marken auf dem Markt haben. Somit kann man mit grosser Wahrscheinlichkeit sagen, dass Apple Laptops im Durchschnitt teurer sind. Auf der untenstehend Grafik ist auch sehr gut zu sehen das der Median sowie das 1. Quartal und 3. Quartal deutlich im Preis höher sind als bei den anderen Marken.

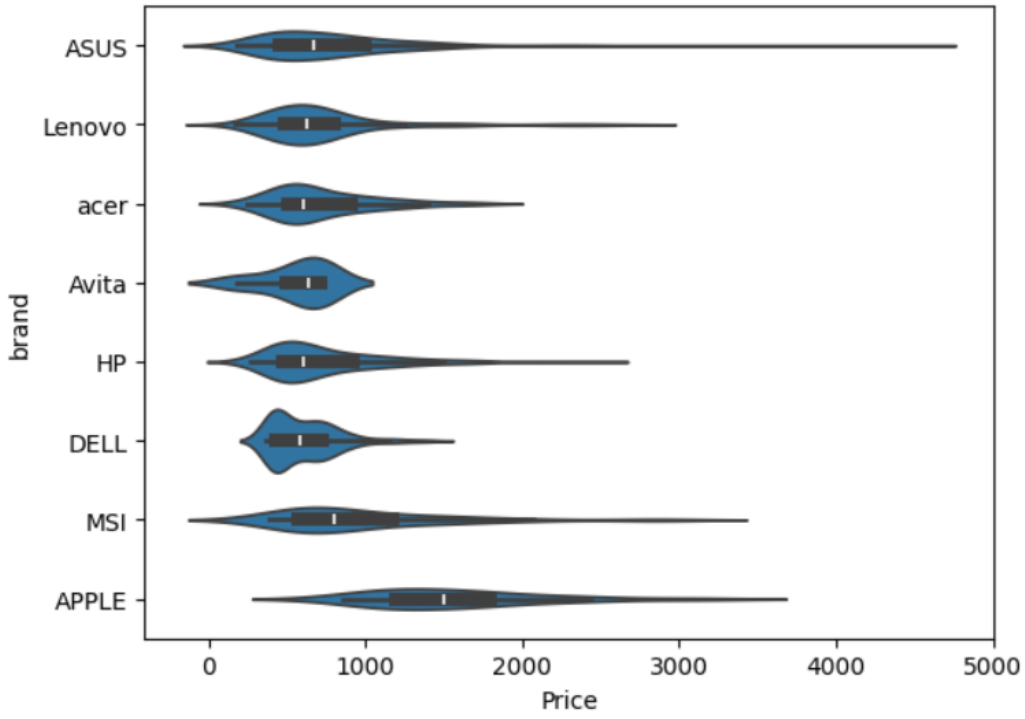


Abbildung 17 ViolinPlot des Preises zu den einzelnen Laptop Marken

Hypothese 3

Aussage: Laptops die mehr Ratings haben sind auch günstiger.

	Price	Number of Ratings	Number of Reviews
Price	1.000000	-0.152553	-0.156791
Number of Ratings	-0.152553	1.000000	0.991062
Number of Reviews	-0.156791	0.991062	1.000000

Abbildung 18 Korrelation zwischen dem Preis und der Anzahl Ratings

Antwort: Die Aussage stimmt nicht. Die Korrelation ist viel zu tief und man kann nicht erwarten das wenn ein Laptop günstiger ist, er auch mehr Ratings hat. Die Aussage trifft eventuell bei wenigen Laptops zu. Im grossen und ganzen könnte man begründen, dass es zwar zu einem kleinen Grad stimmt, doch dafür hat es viel zu wenig Ratings und wenn jemand einen Laptop kauft, muss ein Rating nicht unbedingt abgegeben werden. Somit ist die Hypothese unmöglich zu hinterlegen.

Modelle

Testdaten

```
X_price = df_ex_onehot.drop(['Price', 'processor_gnrtn', 'ram_type', 'hdd', 'os', 'os_bit', 'warranty', 'Touchscreen', 'msoffice'], axis=1)
y_price = df_ex_onehot['Price']

X_price_train, X_price_test, y_price_train, y_price_test = train_test_split(X_price, y_price, test_size=0.4, shuffle=True)
```

Abbildung 19 X und Y Achse bestimmen

Für die X-Achse habe ich die Daten, die ich nicht für wichtig empfinde, aussortiert und die restlichen Daten, dann in den „x_price“ eingefügt.

Bei der Y-Achse habe ich mich für den Preis entschieden, da ich laut der Aufgabenstellung den Preis der Laptops mithilfe bekannter Merkmale möglichst genau vorherzusagen.

Modell 1

Bei meinem ersten Modell handelt es sich um eine lineare Regression.

Der MSE-Wert war 73336.20 und der R2-Score 0.65. Die Standartabweichung ist damit bei ca. 270.Es ist ein guter Wert für die Preiseinschätzung, doch dieser Wert ist mit grosser Wahrscheinlichkeit durch die Ausreisser beeinflusst worden und ist aus diesem Grund nicht niedriger. Der R2.Score von 0.65 ist auch ein guter Wert. Im Gegensatz zum dritten Modell ist dieses mehr geeignet, doch das zweite ist noch leicht besser. Auf dem DataFrame sind links die geschätzten Daten und rechts die wirklichen Daten.

	0	1
0	1434.129024	1510.98
1	338.904951	389.40
2	513.425264	419.90
3	543.863556	364.90
4	401.358822	359.90
5	772.865031	849.90
6	787.394629	729.00
7	1420.824465	1289.90
8	391.407519	429.90
9	1224.310209	1229.90

Abbildung 20 DataFrame der linearen Regression

Modell 2

Beim zweiten Modell handelt es sich um ein Ridge.

Der MSE-Wert war 64075.75 und der R2-Score 0.70. Die beiden Werte sind somit leicht besser als die Werte der linearen Regression. Die Standardabweichung liegt bei ca. 253. Somit ist die Einschätzung noch etwas genauer. Ich denke dieses Modell ist für diese spezifische Aufgabe am besten geeignet. Zum einen, weil es etwas besser vor Ausreisser geschützt ist und zum anderen hat es eine sehr gute Einschätzung gemacht. Auf dem DataFrame sind links die geschätzten Daten und rechts die wirklichen Daten.

0	1495.519597	1510.98
1	349.706254	389.40
2	509.376051	419.90
3	510.385999	364.90
4	420.008665	359.90
5	776.252812	849.90
6	775.824899	729.00
7	1440.155453	1289.90
8	408.963949	429.90
9	1246.916333	1229.90

Abbildung 21 DataFrame der Ridge

Modell 3

Mein drittes Modell ist ein DescisionTreeRegressor.

Der MSE-Wert war 133044.52 und der R2-Score 0.37. Die Standardabweichung liegt somit bei ca. 365 und ist somit deutlich das schlechteste Modell. Zumindest für diese Aufgabe. Der R2-Score ist nur halb so gross wie der, der anderen Modell. Auf dem DataFrame sind links die geschätzten Daten und rechts die wirklichen Daten.

0	1148.135288	1510.98
1	602.818324	389.40
2	602.818324	419.90
3	602.818324	364.90
4	602.818324	359.90
5	602.818324	849.90
6	602.818324	729.00
7	1148.135288	1289.90
8	602.818324	429.90
9	1148.135288	1229.90

Abbildung 22 DataFrame des DescisionTree

Zusammenfassung

Meine Aufgabe war es mit der Hilfe von bekannten Merkmalen eine Model zu entwickeln, dass den Preis eines Laptops möglichst genau vorhersagen kann. Ich konnte diese durch ein Ridge gut lösen und konnte ein R2-Score von 0.70 erhalten.

Ich denke da es sehr schwierig ist einen Laptop anhand von Merkmalen einzuschätzen ist mir dies gut gelungen. Ich denke man kann dieses Model eher schlecht einsetzt, da die Laptop Preise ständig steigen und man müsste denn Datensatz fast jede Woche oder Monat überarbeiten. Doch wenn die Laptoppreise stabil sein würden, dann würde das Model eine gute Arbeit leisten, da es sich bei den Laptoppreise um werte die über 4000 gehen können, ist eine Standardabweichung von ca. 250 gut bis sehr gut.

Ich denke durch noch mehr Merkmale und einen grösseren Datensatz könnte das Model eine noch höhere Performance erhalten. Am Model selbst könnte man eventuell noch mehr der verfügbaren Merkmalen benutzen auch wenn sie nicht direkt mit den Preisen in Verbindung stehen.