Bibliotheken laden

Daten-Upload

In [2]:
```
import pandas as pd
```

In [3]:
```
EntdeckerbonusKunden = 'raw_data\EntdeckerbonusKunden.csv'
sep = ';'
```

In [4]:
```
# Es werden nur die ersten zwei Zeilen geladen, um die Header-Struktur zu überprüfe
df = pd.read_csv(EntdeckerbonusKunden, sep=sep, nrows=2)
df.head()
```

Out[4]:

	Unnamed: 0	Angebot	KundenNr	
0	1	6640a2057810a93511ae00bf4f66357e	c31d756e5ca8e94b41ff4ebd619b68d9	2021-07-01T1
1	2	10278a9c5032bc7df577045c98e16bda	96d849e1394234ccb29e31e5e7e6722f	2021-07-01T1

In [5]:
```
# Hochladen von Daten ohne die erste Spalte
cols = df.columns
df = pd.read_csv(EntdeckerbonusKunden, sep=sep, usecols=cols[1:])
df.head()
```

Out[5]:

	Angebot	KundenNr	Einloesedat
0	6640a2057810a93511ae00bf4f66357e	c31d756e5ca8e94b41ff4ebd619b68d9	2021-07-01T11:17:11.297+
1	10278a9c5032bc7df577045c98e16bda	96d849e1394234ccb29e31e5e7e6722f	2021-07-01T13:46:29.099+
2	6640a2057810a93511ae00bf4f66357e	777e2512c1a9a28dc4a645e95a5b6afa	2021-07-02T10:52:59.087+
3	a6473e1fccb1f56b594f01ccddc52f8f	702728b8167a14b900f331076db52673	2021-07-03T08:41:21.163+
4	6ce58ed831850a365fbb0c1ee56ae4f0	017d2203c7556fe4deb6d341964717f9	2021-07-04T13:41:18.839+

Basisinformationen zum Datensatz

In [6]:
```
# Datendimensionen
df.shape
```

Out[6]:
```
(931, 3)
```

In [7]:
```
# ob es leere Zellen gibt
df.isna().sum()
```

Out[7]:
```
Angebot          0
KundenNr         0
Einloesedatum    0
dtype: int64
```

In [8]:
```
# Überprüfung des Datentyps
df.dtypes
```

Out[8]:
```
Angebot          object
KundenNr         object
Einloesedatum    object
dtype: object
```

In [9]:
```
df.describe()
```

Out[9]:

	Angebot	KundenNr	Einlo
count	931	931	
unique	75	732	
top	10278a9c5032bc7df577045c98e16bda	c31d756e5ca8e94b41ff4ebd619b68d9	2022-06-17T10:00:00
freq	189	12	

In [10]:
```
# Auf Duplikate prüfen
duplicates = df.duplicated()
duplicates.any()
```

Out[10]:
```
False
```

In [11]:
```
# Konvertieren der Spalte „Einloesedatum“ in einen Datumstyp und Hinzufügen weitere
df['Einloesedatum'] = pd.to_datetime(df['Einloesedatum'], format="%Y-%m-%dT%H:%M:%S

df['Einloesedatum_date'] = df['Einloesedatum'].dt.date
df['Einloesedatum_time'] = df['Einloesedatum'].dt.time
df['Einloesedatum_year'] = df['Einloesedatum'].dt.year
df['Einloesedatum_month'] = df['Einloesedatum'].dt.month
df['Einloesedatum_week_number'] = df['Einloesedatum'].dt.isocalendar().week
```

In [12]:
```
df.head()
```

Out[12]:

	Angebot	KundenNr	Einloesedatum	E
0	6640a2057810a93511ae00bf4f66357e	c31d756e5ca8e94b41ff4ebd619b68d9	2021-07-01 11:17:11.297000+00:00	
1	10278a9c5032bc7df577045c98e16bda	96d849e1394234ccb29e31e5e7e6722f	2021-07-01 13:46:29.099000+00:00	
2	6640a2057810a93511ae00bf4f66357e	777e2512c1a9a28dc4a645e95a5b6afa	2021-07-02 10:52:59.087000+00:00	
3	a6473e1fccb1f56b594f01ccddc52f8f	702728b8167a14b900f331076db52673	2021-07-03 08:41:21.163000+00:00	
4	6ce58ed831850a365fbb0c1ee56ae4f0	017d2203c7556fe4deb6d341964717f9	2021-07-04 13:41:18.839000+00:00	

In [13]:
```
number_of_years = df['Einloesedatum_year'].nunique()
unique_yers = df['Einloesedatum_year'].unique()
number_of_months = df['Einloesedatum_month'].nunique()
unique_months = df['Einloesedatum_month'].unique()

print(f"Anzahl von Jahren: {number_of_years}, Anzahl der Monate: {number_of_months}
print(f"einzigartige Jahre: {sorted(unique_yers)}")
print(f"einzigartige Monate: {sorted(unique_months)}")
```

```
Anzahl von Jahren: 2, Anzahl der Monate: 12
einzigartige Jahre: [2021, 2022]
einzigartige Monate: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
```

Speichern des aktuellen DataFrame in einer CSV-Datei

In [14]:
```
df.to_csv('clean_raw_data\EntdeckerbonusKunden.csv', index=False)
```