Explorative Datenanalyse nach der Bereinigung

In [150…

```
#Import Bibliotheken
import pandas as pd
import matplotlib.pyplot as plt
from collections import Counter
```

In [151…

```
#Abruf und Anzeige DataFrame
file_path = r'C:\Users\MarkE\OneDrive\_Career\2021-2024_Master Big Data\5_Semester\_TH
df = pd.read_csv(file_path)
print(df.head())
```

```
  Quelle       Datum                                                Link  \
0    FAZ  11/30/2023  https://www.faz.net/aktuell/wirtschaft/unterne...
1    FAZ  11/30/2023  https://www.faz.net/aktuell/wirtschaft/kuenstl...
2    FAZ  11/29/2023  https://www.faz.net/aktuell/feuilleton/medien/...
3    FAZ  11/28/2023  https://www.faz.net/pro/d-economy/kuenstliche-...
4    FAZ  11/28/2023  https://www.faz.net/pro/d-economy/prompt-der-w...

                                               Titel  \
0  KI: Warum wir nicht mit Roboter-Autos vollauto...
1  Microsoft: Sind Jahrzehnte von einer künstlich...
2  Künstliche Intelligenz: Europa muss von neuen ...
3  Was die Superintelligenz-KI anrichten könnte, ...
4  Künstliche Intelligenz: Wie man sich seine Pro...

                                                Text  Anzahl Woerter Text  \
0  Roboterautos faszinieren viele – die Augen der...                 1963
1  Der Krieg von Mensch gegen Maschine verschiebt...                  310
2  Durch den Streik gegen die Hollywoodbosse habe...                 1285
3  ChatGPT macht Spaß, aber was ist, wenn es erns...                  459
4  Häufig gibt es wiederkehrende Anweisungen an d...                  968

                                     Text_bereinigt  \
0  many fascinating robot car eye world rightly a...
1  war person machine shift accord Microsoft unli...
2  strike Hollywood boss creative United States a...
3  chatgpt fun get serious expert expect superint...
4  often recur instruction artificial intelligenc...

                                   Titel_bereinigt  KI Anteil  \
0                                              NaN   0.042115
1  Microsoft decade remove artificial superintell...   0.058442
2  artificial intelligence Europe learn new rule ...   0.031593
3                  superintelligence AI could today   0.063241
4           artificial intelligence organize prompt   0.046512

   Anzahl KI Wörter  Einmalige KI Wörter  \
0                47                    6
1                 9                    3
2                23                    3
3                16                    4
4                26                    5

                                          KI Wörter
0  ai (33), artificial intelligence (4), robotics...
1   ai (5), artificial intelligence (3), chatgpt (1)
2  ai (17), artificial intelligence (2), chatgpt (4)
3  ai (10), artificial intelligence (3), chatgpt ...
4  ai (14), artificial intelligence (2), chatgpt ...
```

In [152…

```
#Dekriptive Analyse des DataFrame

#Gestaltung (Shape) des DataFrames
df.shape
```

Out[152]:

```
(2049, 12)
```

In [153…

```
#Informationen (Info) über den DataFrame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2049 entries, 0 to 2048
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Quelle               2049 non-null   object
 1   Datum                2049 non-null   object
 2   Link                 2049 non-null   object
 3   Titel                2049 non-null   object
 4   Text                 2049 non-null   object
 5   Anzahl Woerter Text  2049 non-null   int64
 6   Text_bereinigt       2049 non-null   object
 7   Titel_bereinigt      1543 non-null   object
 8   KI Anteil            2049 non-null   float64
 9   Anzahl KI Wörter     2049 non-null   int64
 10  Einmalige KI Wörter  2049 non-null   int64
 11  KI Wörter            2049 non-null   object
dtypes: float64(1), int64(3), object(8)
memory usage: 192.2+ KB
```

In [154…

```
#Beschreibung (Describe) über den DataFrame
print(df.describe())
```

```
       Anzahl Woerter Text    KI Anteil  Anzahl KI Wörter  Einmalige KI Wörter
count          2049.000000  2049.000000       2049.000000          2049.000000
mean            719.506101     0.037395         12.987311             3.340166
std             463.718056     0.022036         10.328200             1.443247
min             250.000000     0.001439          2.000000             2.000000
25%             398.000000     0.019608          6.000000             2.000000
50%             587.000000     0.035156         11.000000             3.000000
75%             860.000000     0.052533         17.000000             4.000000
max            3955.000000     0.139623         95.000000             9.000000
```

In [155…

```
#Pruefung Duplikate in dem DataFrame
duplicates = df[df.duplicated()]
num_duplicates = len(duplicates)

print("Anzahl der Duplikate:", num_duplicates)
print(duplicates)
```

```
Anzahl der Duplikate: 0
Empty DataFrame
Columns: [Quelle, Datum, Link, Titel, Text, Anzahl Woerter Text, Text_bereinigt, Tite
l_bereinigt, KI Anteil, Anzahl KI Wörter, Einmalige KI Wörter, KI Wörter]
Index: []
```

In [156…

```
#Zaehlen Anzahl der Woerter pro Nachrichtenartikel
df['Anzahl Woerter Titel'] = df['Titel'].apply(lambda x: len(str(x).split()) if pd.not
total_words_titel = df['Anzahl Woerter Titel'].sum()

df['Anzahl Woerter Text'] = df['Text'].apply(lambda x: len(str(x).split()) if pd.notnu
total_words_text = df['Anzahl Woerter Text'].sum()

print("Gesamtanzahl der Woerter im Titel:", total_words_titel)
print("Gesamtanzahl der Woerter im Text:", total_words_text)

#Beschreibung (Describe) über den DataFrame
print(df.describe())

print(df)
```

```
Gesamtanzahl der Woerter im Titel: 21351
Gesamtanzahl der Woerter im Text: 1474268
       Anzahl Woerter Text    KI Anteil  Anzahl KI Wörter  \
count          2049.000000  2049.000000       2049.000000
mean            719.506101     0.037395         12.987311
std             463.718056     0.022036         10.328200
min             250.000000     0.001439          2.000000
25%             398.000000     0.019608          6.000000
50%             587.000000     0.035156         11.000000
75%             860.000000     0.052533         17.000000
max            3955.000000     0.139623         95.000000

       Einmalige KI Wörter  Anzahl Woerter Titel
count          2049.000000           2049.000000
mean              3.340166             10.420205
std               1.443247              2.846256
min               2.000000              2.000000
25%               2.000000              8.000000
50%               3.000000             10.000000
75%               4.000000             12.000000
max               9.000000             27.000000
     Quelle       Datum                                              Link  \
0       FAZ  11/30/2023  https://www.faz.net/aktuell/wirtschaft/unterne...
1       FAZ  11/30/2023  https://www.faz.net/aktuell/wirtschaft/kuenstl...
2       FAZ  11/29/2023  https://www.faz.net/aktuell/feuilleton/medien/...
3       FAZ  11/28/2023  https://www.faz.net/pro/d-economy/kuenstliche-...
4       FAZ  11/28/2023  https://www.faz.net/pro/d-economy/prompt-der-w...
...     ...         ...                                                ...
2044   Zeit   12/6/2023  https://www.zeit.de/news/2023-12/06/google-wil...
2045   Zeit   12/4/2023  https://www.zeit.de/digital/2023-11/ki-gesetz-...
2046   Zeit   12/3/2023  https://www.zeit.de/politik/ausland/2023-12/is...
2047   Zeit   12/2/2023  https://www.zeit.de/2023/51/kuenstliche-intell...
2048   Zeit   12/2/2023  https://www.zeit.de/news/2023-12/02/neuer-poli...

                                                  Titel  \
0     KI: Warum wir nicht mit Roboter-Autos vollauto...
1     Microsoft: Sind Jahrzehnte von einer künstlich...
2     Künstliche Intelligenz: Europa muss von neuen ...
3     Was die Superintelligenz-KI anrichten könnte, ...
4     Künstliche Intelligenz: Wie man sich seine Pro...
...                                                 ...
2044  Sprachmodell Gemini: Google will mit neuem KI-...
2045  KI-Gesetz der EU: Regulierung oder Innovation?...
2046  Krieg in Gaza: Die "Zielfabrik" der israelisch...
2047  Künstliche Intelligenz: KI kann wissenschaftli...
2048  Polizei: Neuer Polizeipräsident: Bei Verbreche...

                                                  Text  Anzahl Woerter Text  \
0     Roboterautos faszinieren viele – die Augen der...                 1963
1     Der Krieg von Mensch gegen Maschine verschiebt...                  310
2     Durch den Streik gegen die Hollywoodbosse habe...                 1285
3     ChatGPT macht Spaß, aber was ist, wenn es erns...                  459
4     Häufig gibt es wiederkehrende Anweisungen an d...                  968
...                                                 ...                  ...
2044  Im Wettlauf bei Künstlicher Intelligenz will s...                  461
2045  Wenn der Verkehrsminister, der auch Digitalmin...                  457
2046  Seit dem Überfall der Hamas auf Israel fliegt ...                  807
2047  Tina Kretschmer ist Professorin für Erziehungs...                  524
2048  Hamburgs neuer Polizeipräsident Falk Schnabel ...                  363
```

```
                                        Text_bereinigt  \
0     many fascinating robot car eye world rightly a...
1     war person machine shift accord Microsoft unli...
2     strike Hollywood boss creative United States a...
3     chatgpt fun get serious expert expect superint...
4     often recur instruction artificial intelligenc...
...                                                 ...
2044  race artificial intelligence Google want take ...
2045  Minister Transport also digital minister talk ...
2046  since attack Hamas Israel israeli army Israel ...
2047  Tina Kretschmer professor educational sciences...
2048  Hamburg new police chief Falk Schnabel also re...

                                      Titel_bereinigt  KI Anteil  \
0                                                 NaN   0.042115
1     Microsoft decade remove artificial superintell...   0.058442
2     artificial intelligence Europe learn new rule ...   0.031593
3                     superintelligence AI could today   0.063241
4               artificial intelligence organize prompt   0.046512
...                                                 ...        ...
2044  Gemini language model Google want hang new AI ...   0.053191
2045                                              NaN   0.053942
2046   War Gaza target factory israeli army time online   0.014315
2047  artificial intelligence AI replace scientific ...   0.045902
2048             Police new police chief Ki time online   0.014634

      Anzahl KI Wörter  Einmalige KI Wörter  \
0                   47                    6
1                    9                    3
2                   23                    3
3                   16                    4
4                   26                    5
...                ...                  ...
2044                15                    5
2045                13                    5
2046                 7                    2
2047                14                    3
2048                 3                    2

                                          KI Wörter  Anzahl Woerter Titel
0     ai (33), artificial intelligence (4), robotics...                    9
1      ai (5), artificial intelligence (3), chatgpt (1)                    8
2     ai (17), artificial intelligence (2), chatgpt (4)                   10
3     ai (10), artificial intelligence (3), chatgpt ...                   12
4     ai (14), artificial intelligence (2), chatgpt ...                    8
...                                                 ...                  ...
2044  ai (4), artificial intelligence (3), chatbot (...                   12
2045  ai (7), artificial intelligence (3), chatbot (...                   10
2046               ai (2), artificial intelligence (5)                    11
2047   ai (7), artificial intelligence (3), chatgpt (4)                   11
2048               ai (1), artificial intelligence (2)                    10

[2049 rows x 13 columns]
```

In [157…

```python
#Zaehlen der häufigsten Woerter im Titel und im Text

#Woerter in allen Eintraegen der Spalte "Text_bereinigt" aufteilen und in einer Liste
words = df['Text_bereinigt'].str.split(expand=True).stack()

#Zaehle die Woerter mit Counter
```

```
word_counts = Counter(words)

#Abrufen und Sortieren der haeufigsten Woerter
most_common_words = word_counts.most_common()

#Umwandlung der Liste der haeufigsten Woerter in einen DataFrame
df_most_common_words = pd.DataFrame(most_common_words, columns=['Wort', 'Anzahl im Tex

#Anzahl der einzelnen Woerter im Titel

#Woerter in allen Eintraegen der Spalte "Titel_bereinigt" aufteilen und in einer Liste
head_words = df['Titel_bereinigt'].str.split(expand=True).stack()

#Zaehle die Woerter mit Counter
head_word_counts = Counter(head_words)

#Abrufen und Sortieren der haeufigsten Woerter
head_most_common_words = head_word_counts.most_common()

#Umwandlung der Liste der haeufigsten Woerter in einen DataFrame
df_head_most_common_words = pd.DataFrame(head_most_common_words, columns=['Wort', 'Anz

#Full outer Verknuepfung mit der Spalte 'Wort'
df_single_combined = pd.merge(df_most_common_words, df_head_most_common_words, on='Wor

#Fuellen von NaN-Werten mit 0 für die Berechnung in df_single_combined
df_single_combined['Anzahl im Text'] = df_single_combined['Anzahl im Text'].fillna(0)
df_single_combined['Anzahl im Titel'] = df_single_combined['Anzahl im Titel'].fillna(0

# Hinzufuegen einer neuen Spalte 'Anzahl Gesamt' durch Summierung von 'Anzahl im Text'
df_single_combined['Anzahl Gesamt'] = df_single_combined['Anzahl im Text'] + df_single

#Sortieren des df_single_combined DataFrame nach 'Anzahl Gesamt' in absteigender Reihe
df_single_combined = df_single_combined.sort_values(by='Anzahl Gesamt', ascending=Fals

#Zuruecksetzen des Index
df_single_combined = df_single_combined.reset_index(drop=True)

#Speichern des sortierten df_single_combined DataFrame in CSV
df_single_combined.to_csv('2_Einzelwörter.csv', index=False)

#Oberen Teil des sortierten df_single_combined DataFrame anzeigen, um zu überprüfen
print(df_single_combined.head(20))
```

	Wort	Anzahl im Text	Anzahl im Titel	Anzahl Gesamt
0	AI	13566.0	628.0	14194.0
1	also	7807.0	20.0	7827.0
2	use	5276.0	62.0	5338.0
3	company	4745.0	58.0	4803.0
4	intelligence	4315.0	331.0	4646.0
5	say	4499.0	12.0	4511.0
6	artificial	4177.0	330.0	4507.0
7	work	3455.0	43.0	3498.0
8	new	3297.0	81.0	3378.0
9	year	3316.0	19.0	3335.0
10	time	2976.0	209.0	3185.0
11	could	3046.0	24.0	3070.0
12	make	2973.0	40.0	3013.0
13	system	2854.0	9.0	2863.0
14	model	2812.0	15.0	2827.0
15	technology	2773.0	34.0	2807.0
16	example	2793.0	0.0	2793.0
17	date	2639.0	17.0	2656.0
18	people	2621.0	15.0	2636.0
19	one	2609.0	5.0	2614.0

In [158…

```python
#Zaehlen der haeufigsten "2 zusammenhängenden Woerter" im Titel und im Text

#Funktion zum Erzeugen von Bigrammen aus einer Zeichenkette
def create_bigrams(text):
    if isinstance(text, str):  # Überprüfe, ob der Text ein String ist
        # Aufteilung des Textes in Wörter
        words = text.split()
        # Erstellen von Bigrammen, indem aufeinanderfolgende Wörter gepaart werden
        bigrams = [' '.join(pair) for pair in zip(words[:-1], words[1:])]
        return bigrams
    else:
        return []  # Wenn der Text kein String ist, gib eine leere Liste zurück

#Anwendung der Funktion auf jeden Eintrag in der Spalte "Text_bereinigt" und Sammlung
bigrams_list = df['Text_bereinigt'].apply(create_bigrams).sum()
#Zaehlen der Bigramme mit Counter
bigram_counts = Counter(bigrams_list)

#Abrufen und Sortieren der haeufigsten Bigramme
most_common_bigrams = bigram_counts.most_common()

#Umwandlung der Liste der haeufigsten Bigramme in einen DataFrame
df_most_common_bigrams = pd.DataFrame(most_common_bigrams, columns=['Wörter', 'Anzahl

#Anzahl mit zwei Woertern in der Ueberschrift

#Funktion zum Erzeugen von Bigrammen aus einer Zeichenkette
def create_bigrams(text):
    if pd.isna(text):
        return []
    words = text.split()
    bigrams = [' '.join(pair) for pair in zip(words[:-1], words[1:])]
    return bigrams

#Anwendung der Funktion auf jeden Eintrag in der Spalte "Titel_bereinigt" und Sammlung
bigrams_list = df['Titel_bereinigt'].apply(create_bigrams).sum()

#Zaehlen der Bigramme mit Counter
```

```
bigram_counts = Counter(bigrams_list)

#Abrufen und Sortieren der haeufigsten Bigramme
most_common_bigrams = bigram_counts.most_common()

#Konvertieren der Liste haeufig vorkommender Bigramme in einen DataFrame
df_head_most_common_bigrams = pd.DataFrame(most_common_bigrams, columns=['Wörter', 'Ar

#Verknuefung mit der Spalte 'Wort'
df_bigrams_combined = pd.merge(df_most_common_bigrams, df_head_most_common_bigrams, or

#Fuellen von NaN-Werten mit 0 für die Berechnung in df_single_combined
df_bigrams_combined['Anzahl im Text'] = df_bigrams_combined['Anzahl im Text'].fillna(0
df_bigrams_combined['Anzahl im Titel'] = df_bigrams_combined['Anzahl im Titel'].fillna

#Hinzufuegen einer neuen Spalte 'Anzahl Gesamt' durch Summierung von 'Anzahl im Text'
df_bigrams_combined['Anzahl Gesamt'] = df_bigrams_combined['Anzahl im Text'] + df_bigr

#Sortieren des df_single_combined DataFrame nach 'Anzahl Gesamt' in absteigender Reihe
df_bigrams_combined = df_bigrams_combined.sort_values(by='Anzahl Gesamt', ascending=Fa

#Zuruecksetzen des Index
df_bigrams_combined = df_bigrams_combined.reset_index(drop=True)

#Speichern des sortierten df_single_combined DataFrame in CSV
df_bigrams_combined.to_csv('3_Zwei_Wörter.csv', index=False)

#Anzeige des oberen Teils des sortierten df_bigrams_combined DataFrame, um zu ueberpru
print(df_bigrams_combined.head(20))
```

	Wörter	Anzahl im Text	Anzahl im Titel	Anzahl Gesamt
0	artificial intelligence	3934.0	326.0	4260.0
1	intelligence AI	806.0	7.0	813.0
2	AI system	757.0	2.0	759.0
3	use AI	682.0	17.0	699.0
4	generative AI	466.0	4.0	470.0
5	voice model	465.0	3.0	468.0
6	Open Ai	385.0	15.0	400.0
7	language model	384.0	4.0	388.0
8	year ago	377.0	0.0	377.0
9	United States	366.0	3.0	369.0
10	among thing	350.0	0.0	350.0
11	AI model	346.0	3.0	349.0
12	GPT 4	338.0	7.0	345.0
13	AI application	322.0	2.0	324.0
14	use artificial	304.0	9.0	313.0
15	year old	287.0	1.0	288.0
16	Elon Musk	257.0	25.0	282.0
17	Sam Altman	249.0	32.0	281.0
18	last year	280.0	0.0	280.0
19	Open AI	264.0	15.0	279.0

In [ ]: