

**BERD**  
@NFDI

Forschungsdatenmanagement in der Fachreferatsarbeit (Wirtschafts- und Sozialwissenschaften)

# Forschungsdaten analysieren

Jan Kamlah und Dr. Renat Shigapov (UB Mannheim)

September 2023



Fokus: Wirtschafts- und Sozialwissenschaften

Überblick:

- |            |   |
|------------|---|
| 08.04.2024 | FDM-Grundlagen – Irene Schumm, Jorge Murcia Serra (BERD@NFDI)         |
| 15.04.2024 | Rechtliche Aspekte von Forschungsdaten – Vasilka Stoilova (BERD@NFDI) |
| 22.04.2024 | Forschungsdaten erheben – Renat Shigapov, Jan Kamlah (BERD@NFDI)      |
| 29.04.2024 | Forschungsdaten analysieren – Jan Kamlah, Renat Shigapov (BERD@NFDI)  |
| 06.05.2024 | Forschungsdaten teilen – Anja Perry (GESIS)                           |
| 13.05.2024 | FDM-Infrastrukturen – Irene Schumm, Jorge Murcia Serra (BERD@NFDI)    |

1. Ankommen
2. Forschungsdaten & Lebenszyklus
3. Datenaufbereitung
4. Datenanalyse
5. Wrap-up



## **Online-Dokument**

<https://tinyurl.com/4b9z4bb2>

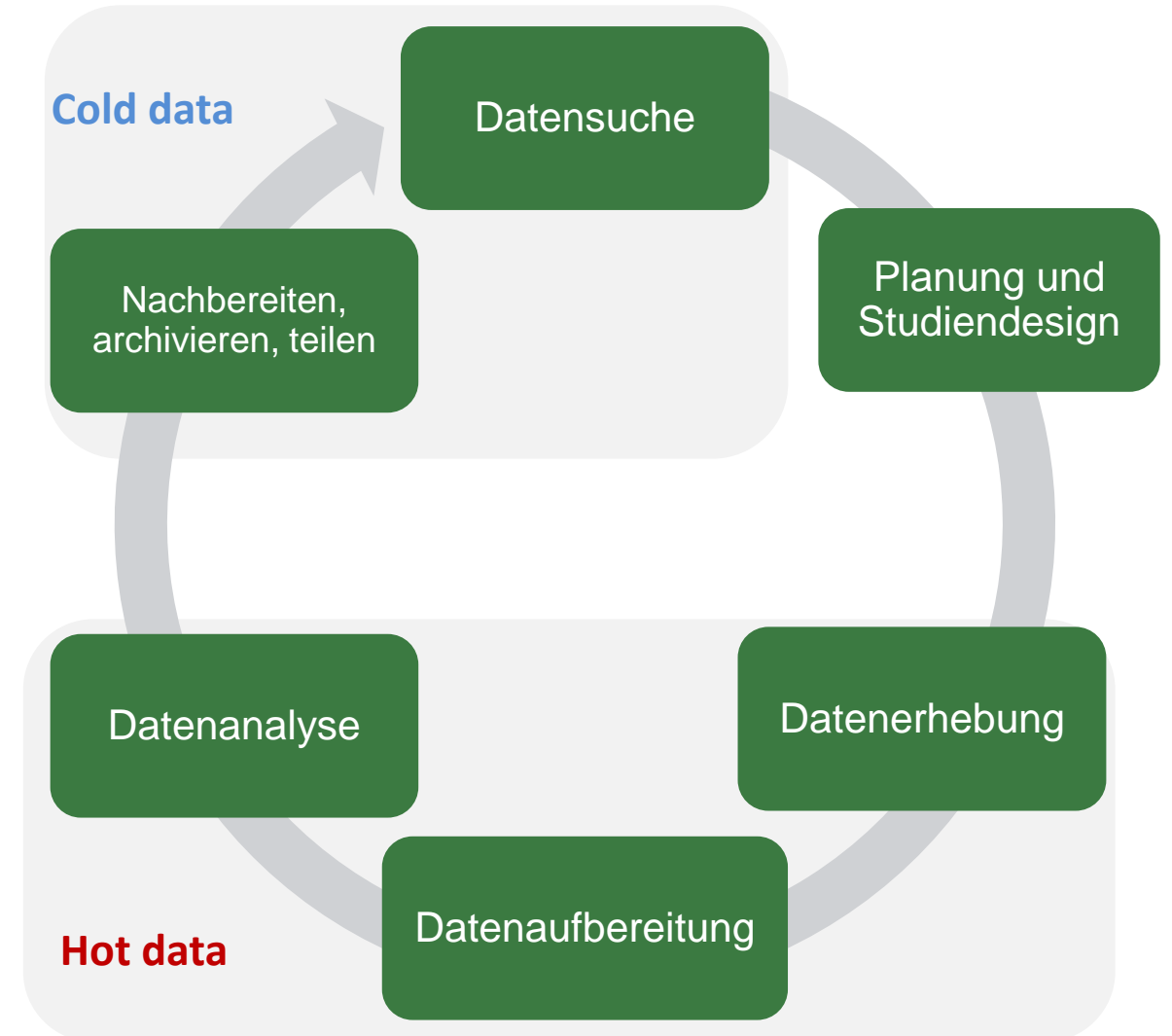
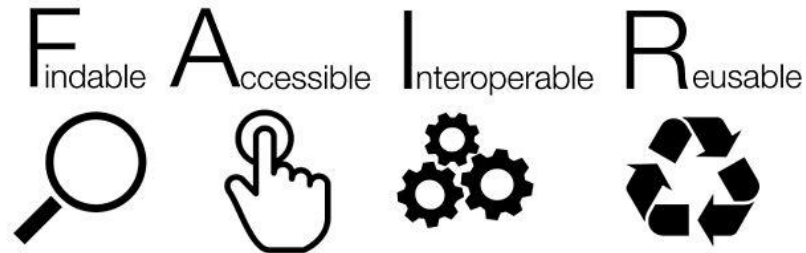
<https://cloud.bib.uni-mannheim.de/s/NCrtqRaA96xx9Fn>



## 2. Forschungsdaten & Lebenszyklus

### Forschungsdaten

Fakten oder Gegebenheiten wie Zahlen, Text, Bild, Audio, Video, die Grundlage einer wissenschaftlichen Auswertung sind bzw. sein können



### Qualitative Daten - unstrukturierte Daten

Q2: Was hat Ihnen an Ihrem Studium am besten gefallen?

Sehr gut fand ich die Wahloptionen bei den Kursen sowie die stets sehr praxisnahen ...



#### Analyse

Interpretationen und Kategorisierungen



### Quantitative Daten - strukturierten Daten

Jahr	BIP	Arbeitslosenquote
2019	300000	3,1
2020	350000	2,9
2021	...	...

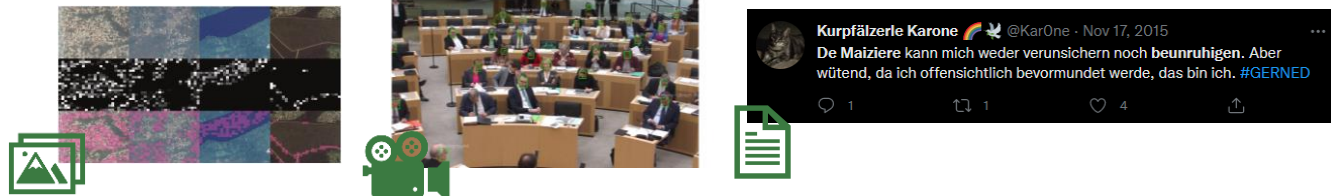


#### Analyse

Statistische Methoden



### Unstrukturierte Datentypen



Soziale Medien	Facebook-Texte, Instagram-Bilder, YouTube-Videos, Twitter-Nachrichten
Allgemeine Medien	Texte, Bilder, Videos, Sprachaufzeichnungen, Musik
Geodaten	GPS-Daten
Log-Daten	Besuch von Webseiten, Verweildauer auf Webseiten, E-Mail-Verhalten
World Wide Web	Webseiten, Nachrichten, Blogs
Allgemeine Dokumente	Texte, PDF-Dateien, eingescannte Dateien
Finanzdaten	Banktransaktionen, Börsendaten
Gesundheitsdaten	Patientenakten, Röntgenbilder, Scanner-Bilder

Quelle: RatSWD Research Report - <https://doi.org/10.17620/02671.73>





## 3. Datenaufbereitung



Welche Prozesse gehören zur Datenaufbereitung?

4 Minuten eintragen in Mentimeter

Homepage: **Menti.com**

Code: **9609 8673**



## Messungen

- **Reliabilität** der Daten – Zuverlässigkeit oder Konsistenz
- **Validität** der Daten – Genauigkeit und Gültigkeit

## Anpassungen

- **Bereinigung** von Daten – Entfernung fehlerhafter Einträge
- **(Vor-)Auswahl** der Daten – Ausschluss nicht benötigter Einträge
- **Anreicherung** von Daten – Ergänzung von fehlenden Einträgen
- **Verknüpfung** von Datensätzen – Zusammenführung verschiedener Datensätze
- **Formatierung** der Daten – Einträge in bestimmte Datentypen umgewandelt bspw. Datum
- **Normalisierung** der Daten – Einträge mit gleicher Bedeutung aber abweichender Bezeichnung werden zusammengeführt
- **Aggregation** von Daten – Gruppierung und/oder Zusammenfassung von Dateneinträgen
- **Kategorisierung** und **Interpretation** von qualitativen Daten

## Export

- **Speichern der Daten in Zielformat** – Benötigtes Format zur Datenanalyse



### OpenRefine - Demo

10 Minuten

# 3. Datenaufbereitung für Audio-, Video- und Bilddaten

---

**Audio-Präprozessierung:** Anwendung von Signalverarbeitungstechniken

**Audiomerkmals-Extraktion:** Extraktion relevanter Merkmale aus Audiosignalen

**Bild-Präprozessierung:** Anwendung von Techniken, um die Bildqualität zu verbessern und das Modelltraining zu erleichtern.

**Video-Framing:** Extraktion einzelner Frames aus Videodateien für Analysen, Merkmalsextraktion oder Objekterkennungsaufgaben.

**Video-Kompression und Kodierung:** Reduzierung der Videodateigröße und Standardisierung von Kodierungsformaten für eine effiziente Speicherung und Verarbeitung.

**Objekterkennung und -verfolgung:** Verwendung von Computer-Vision-Techniken zur Erkennung und Verfolgung von Objekten oder Bewegungen innerhalb von Videobildern.

# 3. Datenaufbereitung für Text- und Social-Media-Daten

---

**Textbereinigung und -vorbereitung:** Reinigung und Vorbereitung von Textdaten aus Social-Media-Beiträgen, einschließlich Behandlung von Hashtags, Erwähnungen, Emojis und Sonderzeichen.

**Sentimentanalyse:** Analyse des Sentiments von Social-Media-Beiträgen, um Benutzermeinungen, Emotionen und Trends zu verstehen.

**Themenmodellierung:** Identifikation und Extraktion von Themen oder Trends aus Social-Media-Gesprächen.

**Tokenisierung:** Aufteilen von Text in einzelne Wörter oder Tokens zur Erleichterung weiterer Analysen.

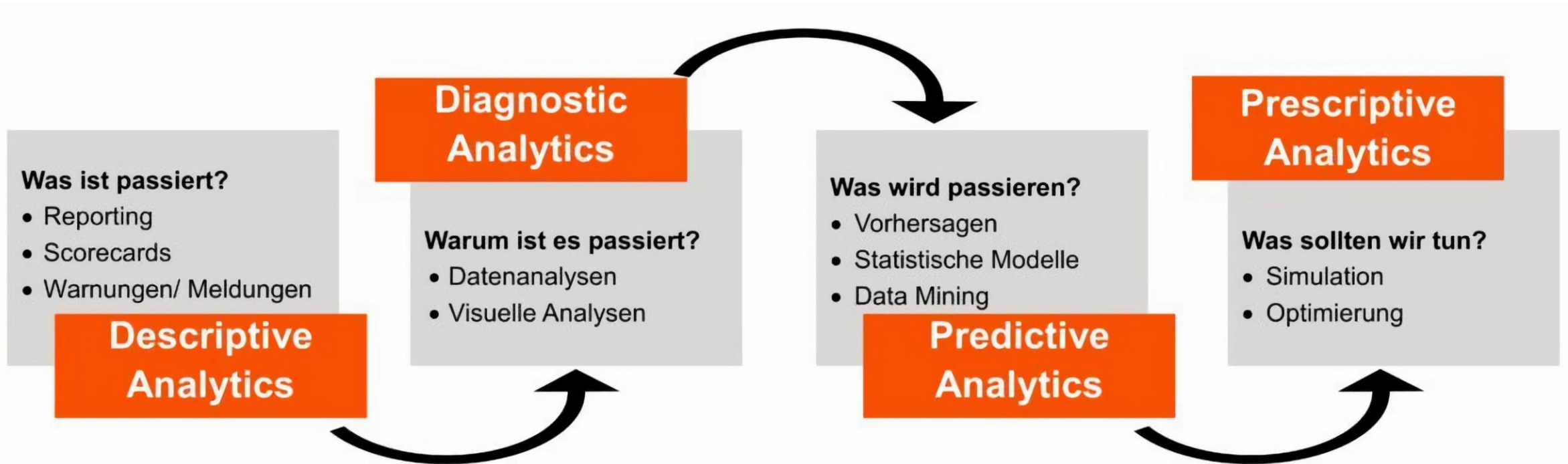
**Stoppwortentfernung:** Entfernen häufig vorkommender Wörter (z. B. "der", "und"), die keine wesentliche Bedeutung haben.

**Stemming und Lemmatisierung:** Reduzierung von Wörtern auf ihre Basis- oder Stammform zur Normalisierung von Varianten und Verbesserung der Genauigkeit der Textanalyse.

**Benannte-Entitäten-Erkennung (NER):** Identifikation und Klassifizierung von benannten Entitäten wie Personen, Organisationen und Orten in Textdaten.



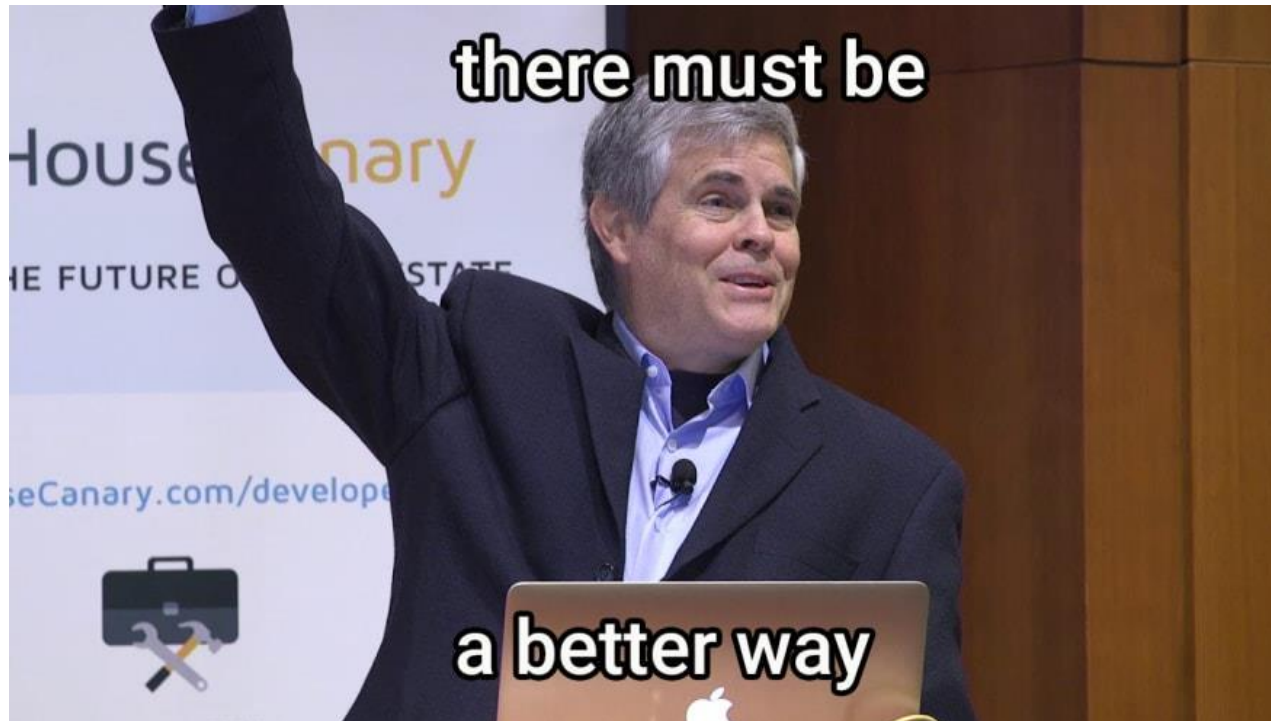
## 3. Datenanalyse





## 4. Datenanalyse – Und was kommt dann..?

1. Können uns die **Daten** also **sagen**, **welche Methoden** benötigt werden um sie auf eine **bestimmte Forschungsfrage** hin zu analysieren?
2. Und kann **moderne Technologie**, diese **Methoden** nicht **direkt anwenden**, die **Ergebnisse visualisieren**, **Vorhersagen treffen** und eventuell sogar **zukünftige Schritte vorschlagen**?

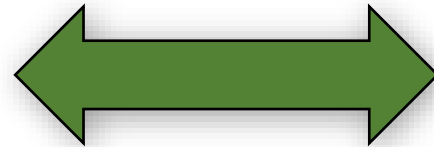


<https://dev.to/artiomkaltovich/constraint-programming-in-python-or-how-to-solve-traveling-salesman-problem-just-describing-it-32k1>

## 4. Moderner Workflow mit virtuellen Assistenten



Researcher




- GWDG (open + OpenAI ChatGPT3.5 / 4): <https://chat-ai.academiccloud.de>
- DFKI (open): <https://opengptx.dfki.de/chat>
- ChatGPT: <https://chat.openai.com>
- Perplexity.ai: <https://www.perplexity.ai>

Virtuelle Assistenten


RE

Daimler-Benz A.G.  
Was ist das?

 Daimler-Benz A.G. war ein deutsches Unternehmen, das durch die Fusion von Benz & Cie. und der Daimler-Motoren-Gesellschaft im Jahr 1926 gegründet wurde. Das Unternehmen ist am besten bekannt für seine Marke Mercedes-Benz und seine hochwertigen Automobile, Lastwagen und Busse. Es war ein Vorläufer der heutigen Daimler AG.

RE

Wie können wir  
die Forschungsdatenanalyse  
in drei Sätzen erklären?

 Forschungsdatenanalyse ist der Prozess der Auswertung von Daten, die durch wissenschaftliche Untersuchungen gesammelt wurden, um Muster, Trends und Schlussfolgerungen zu erkennen. Dabei werden statistische Methoden, Algorithmen und Software-Tools verwendet, um relevante Informationen aus großen Datenmengen zu extrahieren. Dies hilft Forschern, Hypothesen zu bestätigen oder zu widerlegen und neue Erkenntnisse zu gewinnen.

https://www.anaconda.com/anaconda-navigator



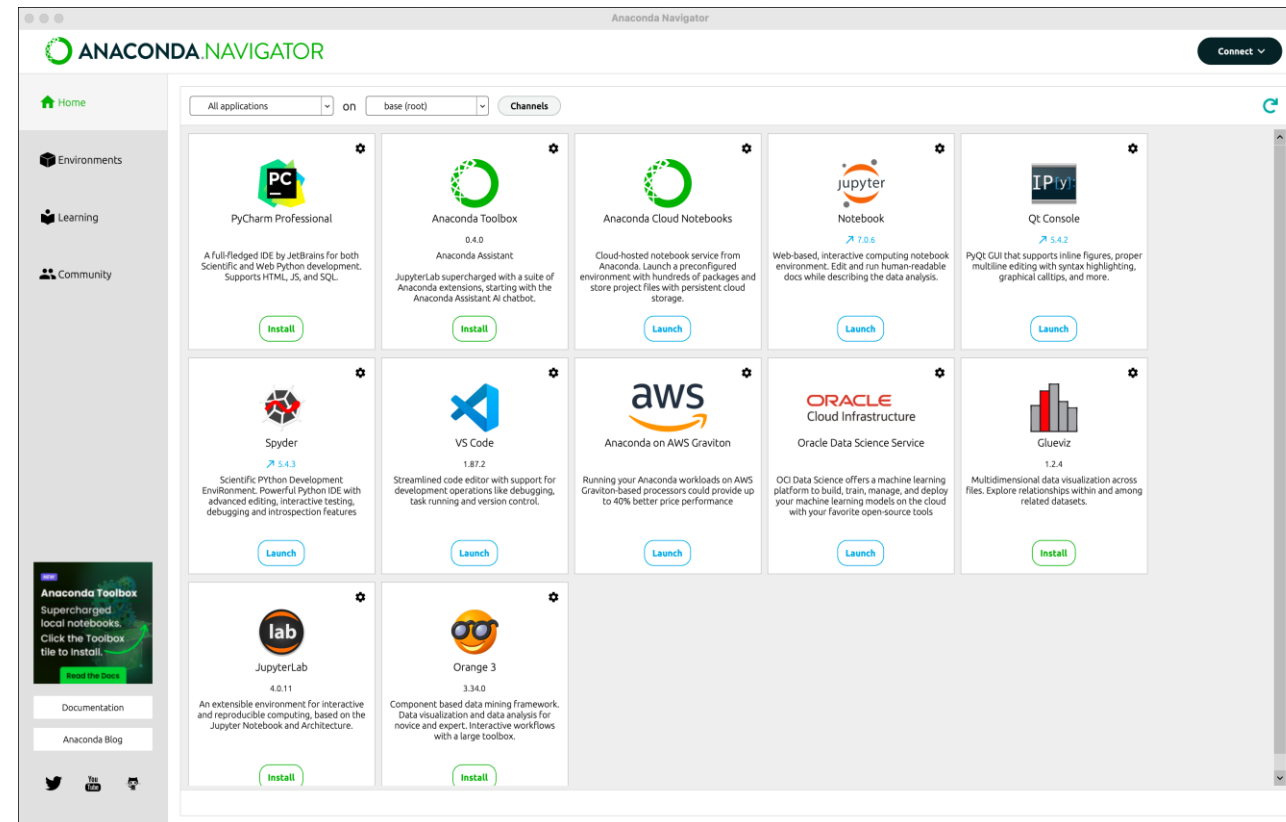
Platform Solutions Pricing Partners Resources Company

Solutions / Anaconda Navigator

## Launch data science applications from your desktop with Anaconda Navigator

The Desktop Portal to Data Science

Download Now >





### 3. Datenanalyse: Descriptive Analytics

## Automatic Descriptive Analysis with Sweetviz

```
python
import sweetviz as sv

# Load the data
data = pd.read_csv('your_data.csv')

# Generate the report
report = sv.analyze(data)
report.show_html('descriptive_analysis.html')
```

## Descriptive Analysis with Pandas Profiling

```
python
import pandas as pd
from pandas_profiling import ProfileReport

# Load the data
data = pd.read_csv('your_data.csv')

# Generate the report
profile = ProfileReport(data, title='Descriptive Analysis Report')
profile.to_file('descriptive_analysis.html')
```

## Automatic Exploratory Data Analysis with Autoviz

```
python
from autoviz.AutoViz_Class import AutoViz_Class

# Load the data
data = pd.read_csv('your_data.csv')

# Generate the report
av = AutoViz_Class()
av.AutoViz(filename='your_data.csv', dfte=data, verbose=0, chart_format='html')
```

## Descriptive Analysis with Dataprep

```
python
from dataprep.eda import create_report

# Load the data
data = pd.read_csv('your_data.csv')

# Generate the report
report = create_report(data)
report.show()
```

## Variables

Select Columns ▾

### Name

Text

MISSING

<b>Distinct</b>	40
<b>Distinct (%)</b>	100.0%
<b>Missing</b>	1
<b>Missing (%)</b>	2.4%
<b>Memory size</b>	460.0 B

Demo

Overview		Words	Characters	Unique		Sample
<b>Length</b>		<b>Characters and Unicode</b>		<b>Unique</b>		
Max length	62	Total characters	632	Unique	40 <span>?</span>	<b>1st row</b> Siemens Energy
Median length	27.5	Distinct characters	54	Unique (%)	100.0%	<b>2nd row</b> Sartorius VzChampionBCDIBCDI Deutschland
Mean length	15.8	Distinct categories	7 <span>?</span>			<b>3rd row</b> Vonovia
Min length	3	Distinct scripts	2 <span>?</span>			<b>4th row</b> RWE
		Distinct blocks	2 <span>?</span>			<b>5th row</b> Bayer

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.



More details

Pandas-Profiling

## 1. Text Analytics: Analyzing Public Policy Documents

Here, we'll use Python's `nltk` library to count the frequency of the term "sustainable development" in a collection of policy documents.

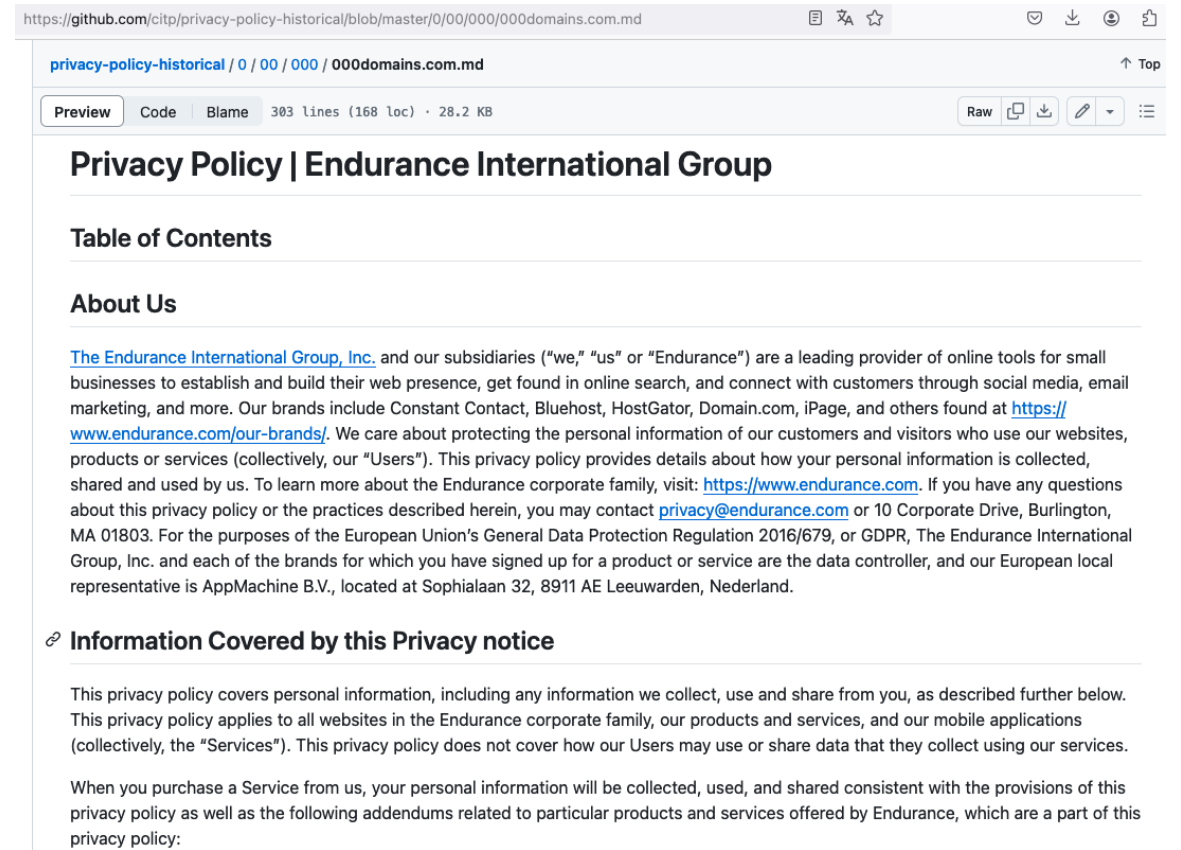
```
python
import nltk
from nltk.corpus import PlaintextCorpusReader
from nltk.probability import FreqDist

# Load your dataset
corpus_root = '/path/to/policy/documents'
wordlists = PlaintextCorpusReader(corpus_root, '.*\.txt') # assuming texts are

# Tokenize the text
words = wordlists.words()
fdist = FreqDist(words)

# Frequency of the term "sustainable development"
term = "sustainable development"
term_freq = sum(1 for phrase in nltk.bigrams(words) if phrase == tuple(term.split(' ')))

print(f"The term '{term}' appears {term_freq} times in the documents.")
```



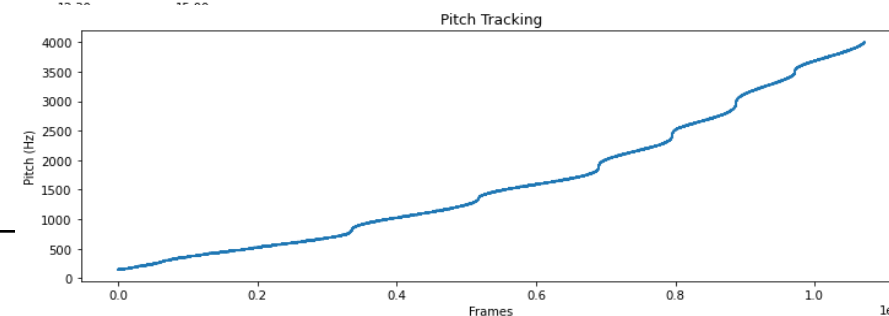
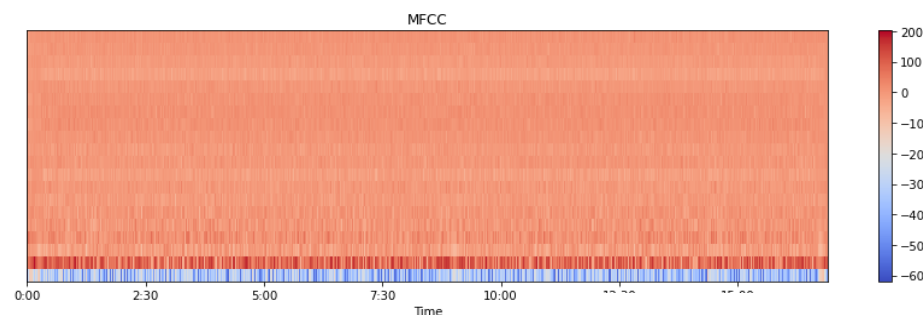
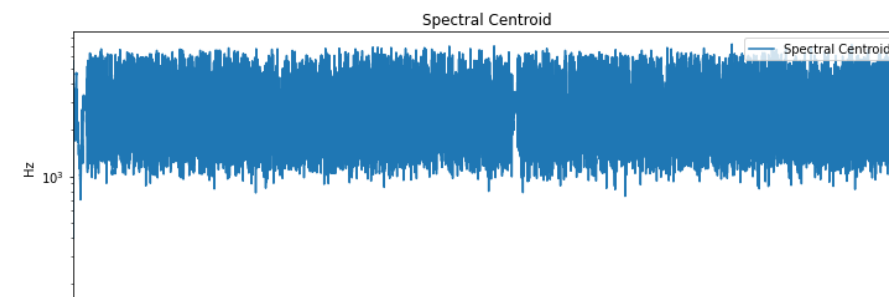
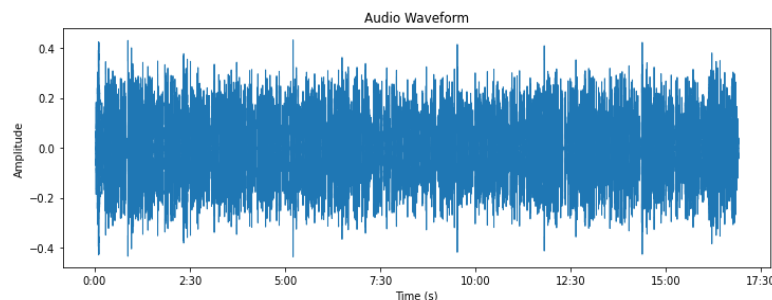
The screenshot shows a GitHub repository page for 'privacy-policy-historical'. The document being viewed is '0 / 00 / 000 / 000domains.com.md'. The document content includes a 'Table of Contents' with sections for 'About Us' and 'Information Covered by this Privacy notice'. The 'About Us' section describes the company and its services. The 'Information Covered by this Privacy notice' section states that the policy covers personal information collected, used, and shared from users.

2024.04.29 'Forschungsdatenanalyse'  
The term 'privacy' appears 41 times in the documents.



# Descriptive Audio Analytics

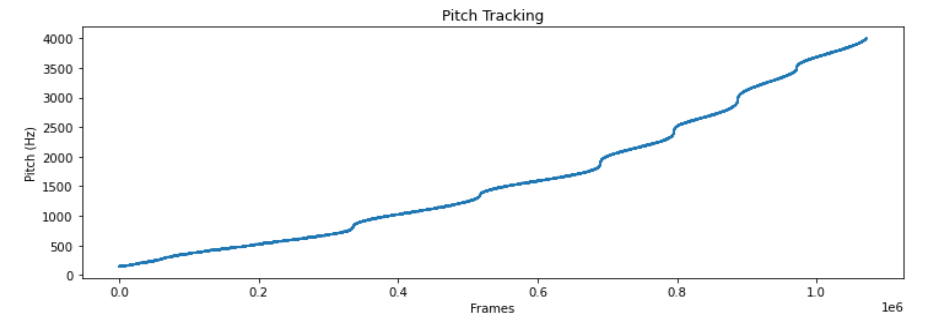
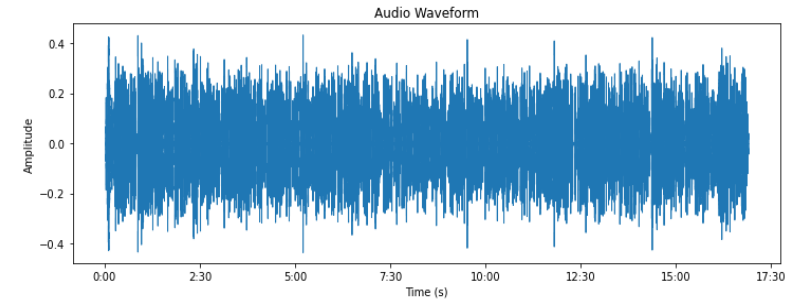
```
python Copy code  
  
import librosa  
import librosa.display  
import matplotlib.pyplot as plt  
import numpy as np  
  
# Load audio file  
audio_path = 'path_to_economic_speech.wav'  
audio_data, sample_rate = librosa.load(audio_path)  
  
# Plot waveform  
plt.figure(figsize=(12, 4))  
librosa.display.waveshow(audio_data, sr=sample_rate)  
plt.title('Audio Waveform')  
plt.xlabel('Time (s)')  
plt.ylabel('Amplitude')  
plt.show()  
  
# Zero Crossing Rate  
zero_crossings = librosa.zero_crossings(audio_data, pad=False)  
print(f"Number of Zero Crossings: {sum(zero_crossings)}")  
  
# Spectral Centroid  
spectral_centroids = librosa.feature.spectral_centroid(audio_data, sr=sample_rate)  
plt.figure(figsize=(12, 4))  
plt.semilogy(spectral_centroids.T, label='Spectral Centroid')  
plt.ylabel('Hz')  
plt.xticks([])  
plt.xlim([0, spectral_centroids.shape[-1]])  
plt.legend(loc='upper right')
```



## 3. Video Analytics: Summarizing Content of Economic Forums

Here, we'll use `moviepy` to extract audio from video and then perform a similar analysis as above.

```
python Copy code  
  
from moviepy.editor import VideoFileClip  
from io import BytesIO  
import speech_recognition as sr  
  
# Load video file  
clip = VideoFileClip('/path/to/economic/forum/video.mp4')  
  
# Extract audio from video  
audio = clip.audio  
audio.write_audiofile('/path/to/output/audio.wav')  
  
# Use the previous speech recognition code here to transcribe and summarize the
```



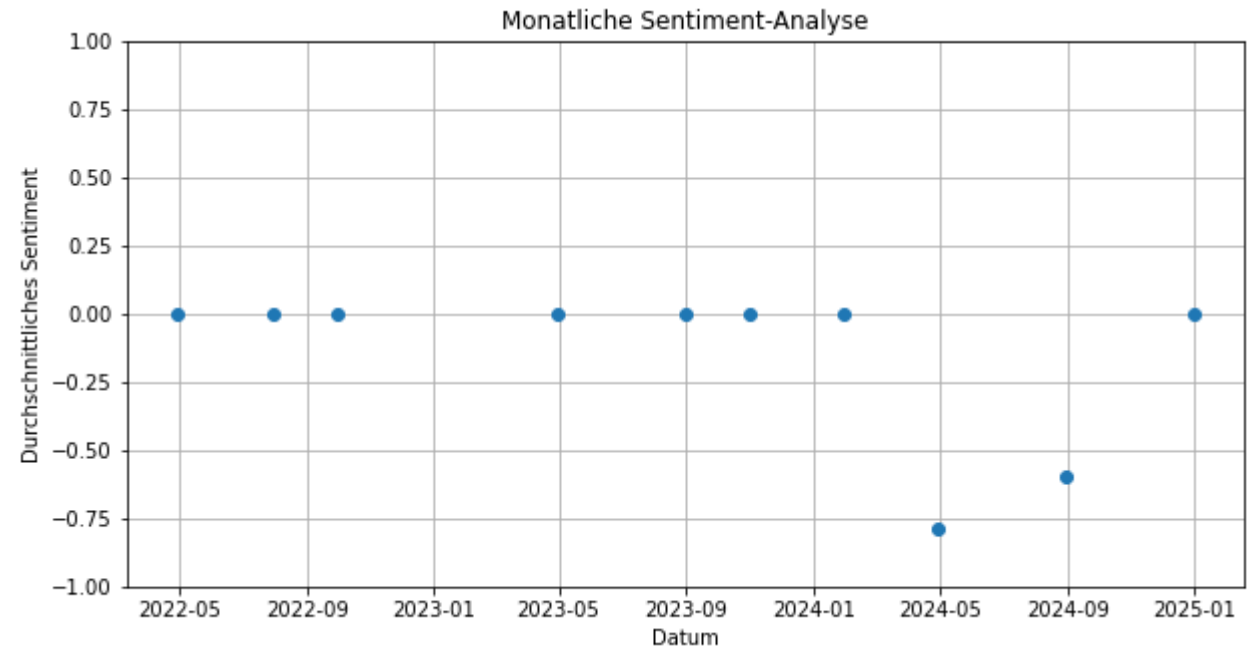


### 3. Datenanalyse: Diagnostic Analytics

**Problemstellung:** Identifizieren der Gründe für plötzliche Änderungen in der Kundenzufriedenheit basierend auf Kundenbewertungen.

**Python-Code:**

```
python Copy code  
  
import pandas as pd  
from sklearn.feature_extraction.text import CountVectorizer  
from nltk.sentiment.vader import SentimentIntensityAnalyzer  
  
# Daten laden  
df = pd.read_csv('kundenbewertungen.csv')  
  
# Sentiment Analyse  
sia = SentimentIntensityAnalyzer()  
df['sentiment'] = df['bewertung'].apply(lambda x: sia.polarity_scores(x)['compound'])  
  
# Änderungen in der Sentiment-Bewertung untersuchen  
df['date'] = pd.to_datetime(df['date'])  
df.set_index('date', inplace=True)  
df['sentiment'].resample('M').mean().plot(title='Monatliche Sentiment-Analyse')
```



**Problemstellung:** Ursachenanalyse für Variabilität in der Sprachqualität bei Callcenter-Anrufen.

Python-Code:

python

```
import librosa
import numpy as np

# Audio laden
audio_path = 'callce
data, sr = librosa.l

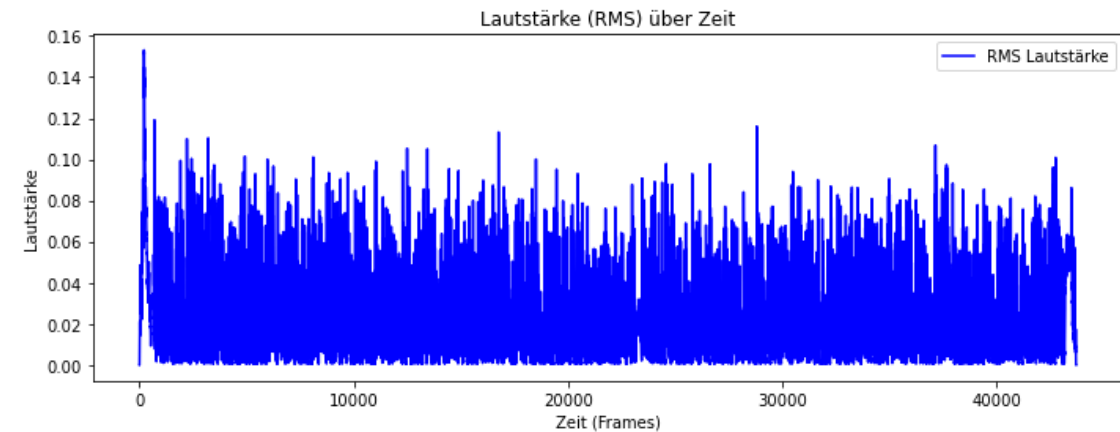
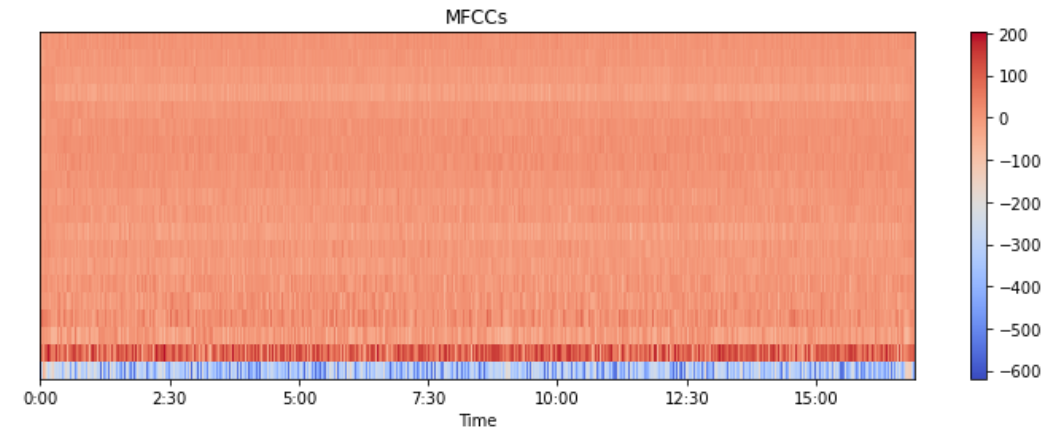
# Feature-Extraktion
mfccs = librosa.feat

# Durchschnittliche
mean_mfccs = np.mean
print("Durchschnittl

# Tonhöhen-Tracking
pitches, magnitudes = librosa.piptrack(y=data, sr=sr)
mean_pitches = np.mean(pitches[pitches > 0]) # Ignoriere Nullwerte für Mittelwe
print("Durchschnittliche Tonhöhe:", mean_pitches)

# Plot 1: MFCCs
plt.figure(figsize=(10, 4))
librosa.display.specshow(mfccs, sr=sr, x_axis='time')
plt.colorbar()
plt.title('MFCCs')
plt.tight_layout()
plt.show()

# Plot 2: Lautstärke
plt.figure(figsize=(10, 4))
plt.plot(rms.T, label='RMS Lautstärke', color='blue')
plt.xlabel('Zeit (Frames)')
plt.ylabel('Lautstärke')
plt.title('Lautstärke (RMS) über Zeit')
plt.legend(loc='upper right')
plt.tight_layout()
plt.show()
```



**Problemstellung:** Bestimmung der Ursachen für Schwankungen in der Markenwahrnehmung auf Basis von Social Media-Daten.

**Python-Code:**

```
python Copy code

import pandas as pd
from textblob import TextBlob
import matplotlib.pyplot as plt

# Daten laden
df = pd.read_csv('social_media_posts.csv')

# Beispiel-Daten anzeigen
print("Beispiel-Daten:")
print(df.head())

# Sentiment-Analyse durchführen
df['sentiment'] = df['post'].apply(lambda x: TextBlob(x).sentiment.polarity)

# Sentiment-Trends analysieren
df['date'] = pd.to_datetime(df['date'])
df.set_index('date', inplace=True)

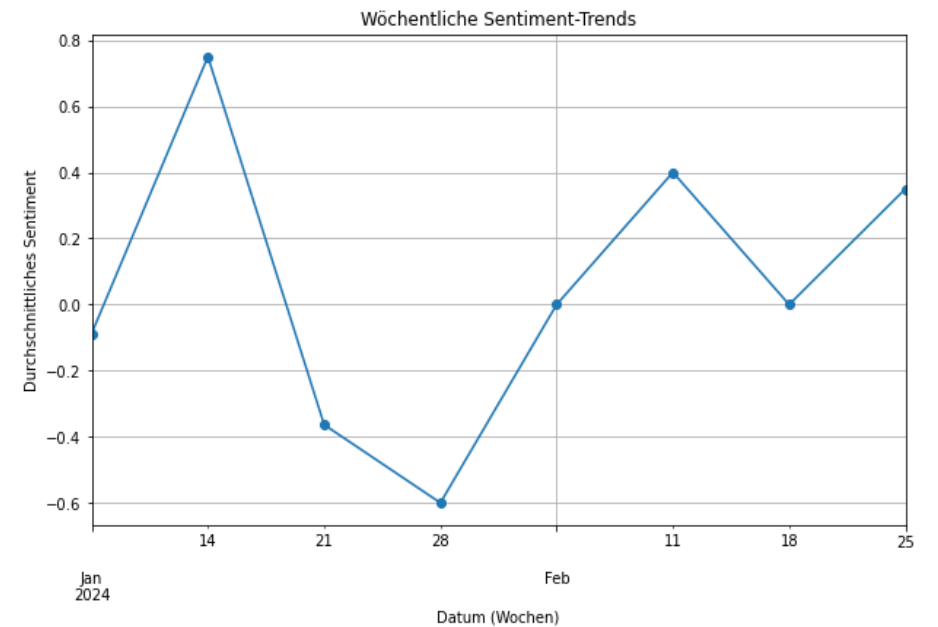
# Durchschnittliche Sentiment-Werte pro Woche berechnen
weekly_sentiment = df['sentiment'].resample('W').mean()

# Visualisierung der wöchentlichen Sentiment-Trends
plt.figure(figsize=(10, 6))
weekly_sentiment.plot(title='Wöchentliche Sentiment-Trends', marker='o', linestyle='solid')
plt.xlabel('Datum (Wochen)')
plt.ylabel('Durchschnittliches Sentiment')
plt.grid(True)
plt.show()
```

**Beispiel für Pseudo-Daten (social\_media\_posts.csv)**

```
csv Copy code

date,post
2024-01-01,Great product! I love it!
2024-01-05,This product is terrible. I'm very disappointed.
2024-01-10,Just received my order. It's amazing!
2024-01-15,The customer service was awful. I'll never buy from this company again.
2024-01-20,Excited to try out this new feature!
2024-01-25,The quality of this product is disappointing.
2024-02-01,Can't wait for the next release!
2024-02-05,The packaging was damaged when I received it.
2024-02-10,Received great support from the customer service team.
2024-02-15,This product exceeded my expectations.
2024-02-20,The user interface is confusing.
2024-02-25,So happy with my purchase!
```





### 3. Datenanalyse: Predictive Analytics

## GWDG LLM Service



Chatbot

Here's an example of text data predictive analytics. This program uses a dataset from scikit-learn (a machine learning library in Python) to classify text into 20 categories:

```
from sklearn.datasets import fetch_20newsgroups
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics

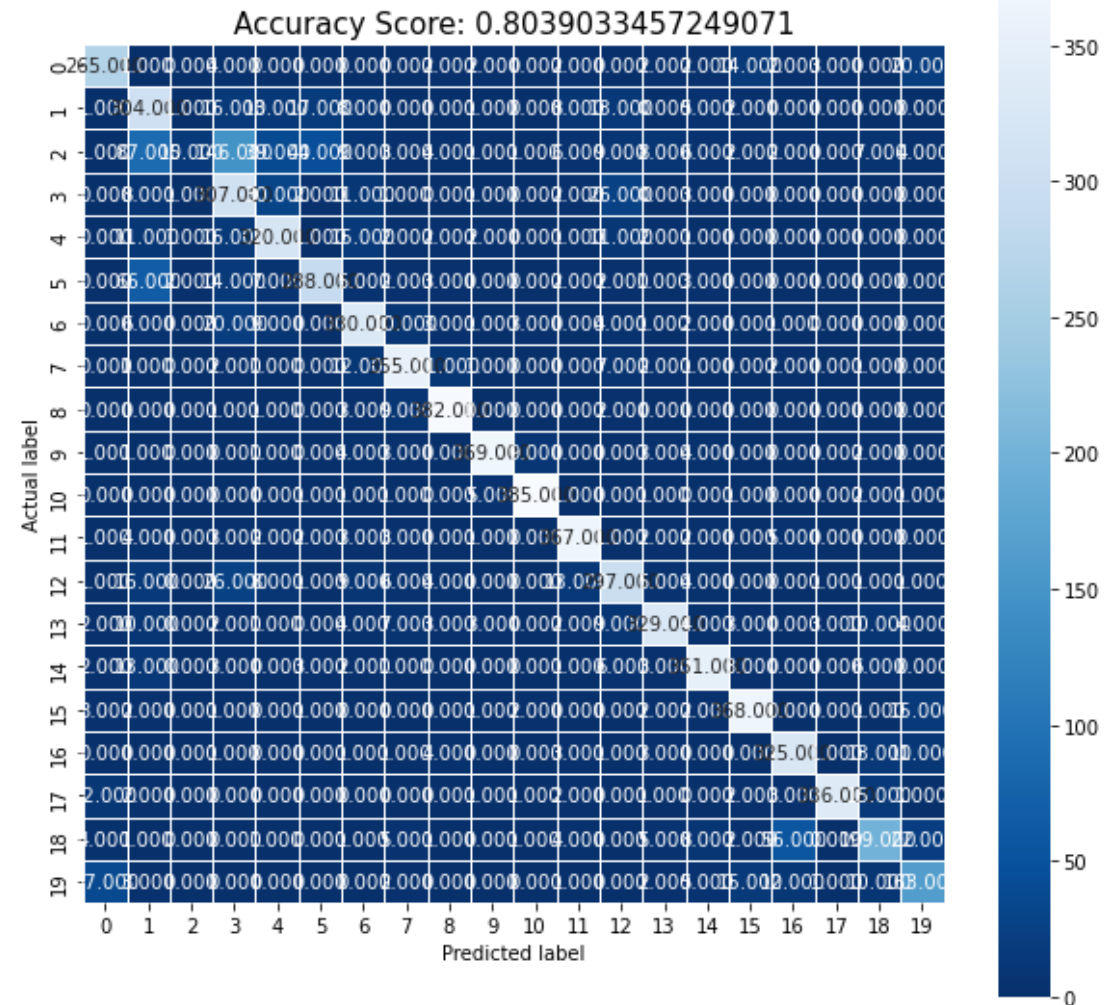
# Load Data
newsgroups_train = fetch_20newsgroups(subset='train')
newsgroups_test = fetch_20newsgroups(subset='test')

# Text Preprocessing
vectorizer = CountVectorizer()
vectors_train = vectorizer.fit_transform(newsgroups_train.data)
vectors_test = vectorizer.transform(newsgroups_test.data)

# Model development and training
clf = MultinomialNB(alpha=.01)
clf.fit(vectors_train, newsgroups_train.target)

# Prediction on testing data
predict = clf.predict(vectors_test)
```

Retry Undo Clear





```
import spacy

nlp = spacy.load("en_core_web_sm")
text = "Apple is looking to buy a UK startup for $1 billion"

doc = nlp(text)

for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```



```
Apple 0 5 ORG
UK 26 28 GPE
$1 billion 41 51 MONEY
```

Opposition stimmt dagegen

## Bundestag beschließt Reform des Klimaschutzgesetzes

Stand: 26.04.2024 13:32 Uhr

Die Verabschiedung stand auf der Kippe. Doch nun hat der Bundestag Änderungen des Klimaschutzgesetzes beschlossen. Die Opposition stimmte dagegen. Die CDU sprach von einem Rückschritt für den Klimaschutz.

Der Bundestag hat die umstrittene Reform des Klimaschutzgesetzes beschlossen. Für die Gesetzesänderung stimmten die Abgeordneten der Ampel-Fraktionen. Die Opposition votierte dagegen. Die Reform des Gesetzes kam vor allem auf Verlangen der FDP zustande. Das Gesetz muss noch den Bundesrat passieren.

Die Reform des Klimaschutzgesetzes sieht grundlegende Änderungen vor. Bisher gilt: Wenn einzelne Sektoren wie der Verkehrs- oder Gebäudebereich gesetzliche Vorgaben zum Kohlendioxidausstoß verfehlen, müssen die zuständigen Ministerien im nachfolgenden Jahr Sofortprogramme vorlegen.

Im vergangenen Jahr verfehlten der Verkehrs- sowie der Gebäudebereich die Vorgaben. Verkehrsminister Volker Wissing (FDP) hatte mit drastischen Maßnahmen bis hin zu Fahrverboten am Wochenende gedroht, sollte der Bundestag die Reform des Klimaschutzgesetzes nicht bis Sommer beschließen - dann hätte Wissing ein Sofortprogramm vorlegen müssen, damit der Verkehrssektor die Klimaziele einhält.



### ChatGPT

The Bundestag has passed changes to the Climate Protection Act, despite opposition from the CDU, who view it as a setback for climate protection. The reform, pushed primarily by the FDP, shifts the focus from retroactive sector-specific controls to forward-looking, multi-year, cross-sectoral targets to ensure overall climate goals are met. If the government falls off track for its 2030 climate target for two consecutive years, adjustments must be made. Critics argue this weakens the law's effectiveness, with the CDU calling it a hollowing out of climate protection, while the FDP defends it as a move away from overly bureaucratic planning. The law sets ambitious targets for reducing CO2 emissions by 2030, aiming for a 65% reduction compared to 1990 levels, with further reductions planned by 2040 and 2045 towards achieving carbon neutrality.



### 3. Datenanalyse: Prescriptive Analytics

```
python Copy code

from textblob import TextBlob

def analyze_sentiment(text):
    blob = TextBlob(text)
    sentiment = blob.sentiment.polarity
    if sentiment > 0:
        return "Positive"
    elif sentiment < 0:
        return "Negative"
    else:
        return "Neutral"

def policy_recommendation(sentiment):
    if sentiment == "Positive":
        return "Recommendation: Proceed with implementing the new economic policy"
    elif sentiment == "Negative":
        return "Recommendation: Review and address concerns before proceeding with"
    else:
        return "Recommendation: Further analysis may be needed to determine the"

text = "The new economic policy received positive feedback from experts."
sentiment = analyze_sentiment(text)
print("Sentiment Analysis Result:", sentiment)
print(policy_recommendation(sentiment))
```

## Sentiment analysis and Decision Making

```
sql Copy code

+-----+
|           Text Sentiment Analysis           |
+-----+
| Text: "The new economic policy received positive |
| feedback from experts."                       |
|                                               |
| Sentiment Analysis Result: Positive           |
|                                               |
| Recommendation:                               |
| Proceed with implementing the new economic    |
| policy.                                       |
+-----+
```

## Audio Data – Speech Emotion Recognition

```
import librosa
import numpy as np
from sklearn.preprocessing import StandardScaler
from tensorflow.keras.models import load_model

def extract_features(audio_file):
    y, sr = librosa.load(audio_file, sr=None)
    mfccs = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13)
    return np.mean(mfccs.T, axis=0)

def predict_emotion(audio_file, num_jokes=0):
    # Incorporating the number of jokes parameter into the feature extraction process
    # For simplicity, we'll just add it as an additional feature
    scaler = StandardScaler()
    model = load_model("emotion_detection_model.h5")
    features = extract_features(audio_file)
    # Appending the number of jokes as an additional feature
    features_with_jokes = np.append(features, num_jokes)
    features_scaled = scaler.fit_transform(features_with_jokes.reshape(1, -1))
    predicted_class = model.predict_classes(features_scaled)[0]
    emotions = ["Neutral", "Happy", "Sad", "Angry"]
    return emotions[predicted_class]

def improve_emotion_recognition():
    print("Recommendation:")
    print("Consider incorporating contextual factors such as the number of jokes in the audio")
    print("This can provide the model with more information to better understand the speaker's emotional state.")

audio_file = "sample_audio.wav"
num_jokes = 2 # Example parameter: number of jokes in the audio
predicted_emotion = predict_emotion(audio_file, num_jokes)
print("Predicted Emotion:", predicted_emotion)
if predicted_emotion == "Sad":
    improve_emotion_recognition()
```

```
vbnet Copy code
+-----+
|           Audio Data - Emotion Analysis           |
+-----+
| Audio: "sample_audio.wav"                         |
|                                                    |
| Predicted Emotion: Sad                            |
|                                                    |
| Recommendation:                                   |
| Consider incorporating contextual factors such    |
| as the number of jokes in a lecture as          |
| additional features. This can provide the model |
| with more information to better understand the  |
| speaker's emotional state.                        |
+-----+
```

## Video Data – Object Detection and Recommendation

```
python Copy code  
  
import cv2  
  
def detect_objects(video_file):  
    cap = cv2.VideoCapture(video_file)  
    while cap.isOpened():  
        ret, frame = cap.read()  
        if not ret:  
            break  
        # Object detection code using OpenCV or other libraries  
        # Here, we'll assume we've detected a car and a person  
        # Based on the detected objects, we'll suggest actions  
        if "car" in detected_objects:  
            print("Action: Implement traffic monitoring and control measures.")  
        if "person" in detected_objects:  
            print("Action: Enhance security measures and crowd management.")  
    cap.release()  
  
video_file = "sample_video.mp4"  
detect_objects(video_file)
```

```
css Copy code  
  
+-----+  
|           Video Object Detection           |  
+-----+  
| Video: "sample_video.mp4"                 |  
|                                           |  
| Detected Objects:                         |  
| - Car                                     |  
| - Person                                 |  
|                                           |  
| Actions:                                  |  
| - Implement traffic monitoring and control |  
|   measures.                              |  
| - Enhance security measures and crowd    |  
|   management.                            |  
+-----+
```

## 4. Image Data - Image Classification and Action Recommendation:

```
python Copy code  
  
import tensorflow as tf  
from tensorflow.keras.preprocessing import image  
import numpy as np  
  
def predict_image_class(image_path):  
    model = tf.keras.models.load_model("image_classification_model.h5")  
    img = image.load_img(image_path, target_size=(224, 224))  
    img_array = image.img_to_array(img)  
    img_array = np.expand_dims(img_array, axis=0)  
    img_array = tf.keras.applications.resnet.preprocess_input(img_array)  
    predictions = model.predict(img_array)  
    predicted_class = np.argmax(predictions, axis=1)[0]  
    classes = ["Class1", "Class2", "Class3"] # Define your classes  
    return classes[predicted_class]  
  
image_path = "sample_image.jpg"  
predicted_class = predict_image_class(image_path)  
print("Predicted Image Class:", predicted_class)
```

```
lua Copy code  
  
+-----+  
|           Image Data - Classification           |  
+-----+  
| Image: "sample_image.jpg"                       |  
|                                                    |  
| Predicted Image Class: Class2                    |  
|                                                    |  
| Recommendation:                                   |  
| Further investigate features contributing to     |  
| misclassification. Consider collecting more     |  
| diverse training data and fine-tuning the model |  
| architecture for improved classification        |  
| accuracy.                                        |  
+-----+
```

## 5. Social Media Data - Opinion Mining and Decision Support:

```
python Copy code  
  
import tweepy  
  
def get_twitter_sentiment(query, count=100):  
    consumer_key = 'YOUR_CONSUMER_KEY'  
    consumer_secret = 'YOUR_CONSUMER_SECRET'  
    access_token = 'YOUR_ACCESS_TOKEN'  
    access_token_secret = 'YOUR_ACCESS_TOKEN_SECRET'  
  
    auth = tweepy.OAuth1UserHandler(consumer_key, consumer_secret, access_token,  
    api = tweepy.API(auth)  
  
    tweets = api.search(q=query, count=count)  
    positive_count = sum(1 for tweet in tweets if TextBlob(tweet.text).sentiment  
    negative_count = sum(1 for tweet in tweets if TextBlob(tweet.text).sentiment  
    if positive_count > negative_count:  
        return "Positive"  
    elif positive_count < negative_count:  
        return "Negative"  
    else:  
        return "Neutral"  
  
query = "economic policy"  
sentiment = get_twitter_sentiment(query)  
print("Twitter Sentiment on", query, ":", sentiment)
```

```
vbnet Copy code  
  
+-----+  
|          Social Media Sentiment Analysis          |  
+-----+  
| Query: "economic policy"                          |  
|                                                    |  
| Twitter Sentiment: Positive                        |  
|                                                    |  
| Recommendation:                                    |  
| Continue monitoring sentiment trends on social    |  
| media platforms. Consider engaging with users,   |  
| addressing concerns, and highlighting positive   |  
| feedback to reinforce public perception of the   |  
| economic policy.                                  |  
+-----+
```



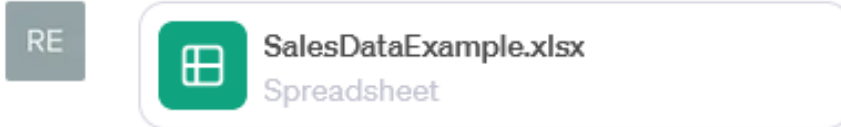
### 3. Datenanalyse mit ChatGPT Enterprise



# ChatGPT Enterprise: Advanced Data Analysis

	Date	Region	Product Category	Product	Customer Name	Sales	Cost	Profit	City	State	Country
1	31 Mai, 2017	Central	Grocery	Fruits and Vegetables	Mary Elizabeth	\$530	\$23	\$507	Woodbury	Minnesota	United States
2	#####	Central	Grocery	Meat	Vincent Herbert	\$1230	\$24	\$1205	Woodbury	Minnesota	United States
3	23 Jul, 2018	Central	Stationery	Copy Paper	Rozario Diego	\$76	\$25	\$51	Woodbury	Minnesota	United States
4	13 Okt, 2016	Central	Grocery	Fruits and Vegetables	Catherine Rosa	\$1523	\$135	\$1408	Wichita	Kansas	United States
5	13 Okt, 2018	Central	Grocery	Fruits and Vegetables	Cath						
6	12 Feb, 2018	Central	Grocery	Fruits and Vegetables	Benj						
7	#####	Central	Grocery	Meat	Jim						
8	08 Nov, 2017	West	Grocery	Fruits and Vegetables	Rhin						
9	27 Jan, 2017	West	Stationery	Copy Paper	Stev						
10	28 Jan, 2017	West	Stationery	Surge Protectors	Man						
11	13 Feb, 2017	West	Grocery	Fruits and Vegetables	And						
12	22 Okt, 2017	West	Stationery	Computer Paper	Bar						
13	30 Mrz, 2017	West	Stationery	Binder Clips	Jose						
14	23 Mrz, 2019	West	Grocery	Fruits and Vegetables	Rick						
15	09 Okt, 2015	Central	Stationery	Copy Paper	Harc						
16	#####	Central	Grocery	Fruits and Vegetables	Stev						
17	18 Jan, 2019	Central	Grocery	Fruits and Vegetables	And						
18	20 Apr, 2017	Central	Stationery	Round Ring Binders	Dea						
19	06 Okt, 2015	Central	Stationery	Art Supplies	Hall						
20	13 Feb, 2018	Central	Stationery	Copy Paper	Mike						
21	29 Jul, 2018	Central	Grocery	Fruits and Vegetables	Shui						
22	24 Apr, 2016	Central	Furniture	Office Chairs	Shui						
23	#####	West	Grocery	Fruits and Vegetables	Patr						
24	30 Jan, 2019	West	Stationery	Heavy Duty Binders	Hele						
25	26 Jun, 2016	West	Grocery	Fruits and Vegetables	Stev						
26	29 Apr, 2016	West	Stationery	Copy Paper	Paul						
27	25 Apr, 2019	West	Stationery	Computer Paper	Willi						
28	01 Jul, 2015	West	Stationery	Computer Paper	Paul						
29	03 Mai, 2015	West	Stationery	Art Supplies	Hilar						
30	07 Jul, 2015	West	Stationery	Portable Storage	Benj						
31	#####	West	Grocery	Fruits and Vegetables	Johr						
32	14 Feb, 2019	West	Stationery	Art Supplies	Max						
33	06 Jan, 2017	West	Grocery	Fruits and Vegetables	Steff						
34	11 Jun, 2018	West	Stationery	Art Supplies	Carl						
35	#####	West	Stationery	Air Purifiers	Steff						
36	15 Mrz, 2016	West	Grocery	Fruits and Vegetables	Rhin						
37	16 Feb, 2019	Central	Stationery	File Labels	And						
38											

Advanced Data Analysis



Führen Sie eine detaillierte grafische Analyse der Umsatzrentabilität in wichtigen Städten durch und visualisieren Sie diese.



Zuerst werde ich das bereitgestellte Excel-Dokument öffnen, um den Inhalt zu überprüfen und zu verstehen, welche Daten es enthält.

Finished working Show work

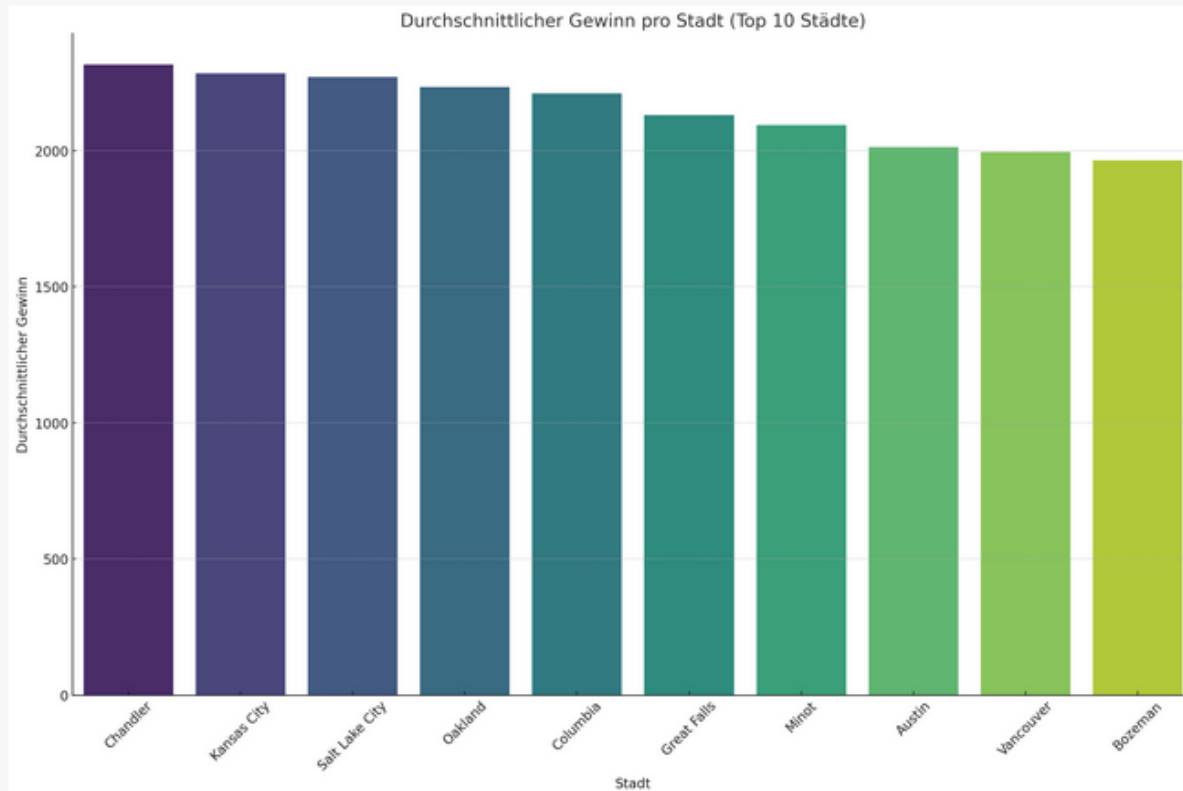
Die Daten enthalten Informationen über Verkaufsdaten in verschiedenen Regionen. Jeder Datensatz enthält das Datum des Verkaufs, die Region, die Produktkategorie, das Produkt selbst, den Kundennamen, den Umsatz, die Kosten, den Gewinn, die Stadt, den Bundesstaat und das Land

Daten

Um die Umsatzrentabilität in wichtigen Städten zu analysieren, werde ich die Daten nach Städten gruppieren und den durchschnittlichen Gewinn (Profit) berechnen. Anschließend werde ich die Ergebnisse grafisch darstellen.

Finished working

Show work ▾



Hier ist ein Balkendiagramm, das den durchschnittlichen Gewinn in den Top-10-Städten zeigt. Dies gibt einen Überblick darüber, welche Städte in Bezug auf den durchschnittlichen Gewinn am profitabelsten sind.

Wenn Sie weitere spezifische Analysen oder Visualisierungen wünschen, lassen Sie es mich bitte wissen!

## EDA = explorative Datenanalyse

RE

Ja, mach bitte EDA und andere Visualisierungen



Natürlich! Eine explorative Datenanalyse (EDA) hilft uns, einen tieferen Einblick in die Daten zu bekommen. Ich werde die folgenden Schritte durchführen:

1. **Deskriptive Statistik** - Grundlegende statistische Maße für die numerischen Spalten.
2. **Verteilung des Gewinns** - Ein Histogramm, um die Verteilung des Gewinns zu zeigen.
3. **Umsatz und Gewinn nach Region** - Balkendiagramme, um den Umsatz und Gewinn nach Region zu zeigen.
4. **Umsatz und Gewinn nach Produktkategorie** - Balkendiagramme, um den Umsatz und Gewinn nach Produktkategorie zu zeigen.
5. **Trend von Umsatz und Gewinn über die Zeit** - Ein Zeitreihendiagramm, um den Umsatz- und Gewinntrend über die Zeit zu zeigen.


	Sales	Cost	Profit
count	756.0	756.0	756.0
mean	1726.9647883597881	636.9062037037037	1090.0585846560846
std	2209.3478148980503	923.3881294228048	1403.7992188206867
min	3.44	0.08	1.13
25%	144.5025	38.42	99.74000000000001
50%	766.28	222.08	482.60999999999996
75%	2519.805	892.2950000000001	1621.1425000000002
max	11751.55	6365.29	9766.390000000001

Hier ist die deskriptive Statistik für die numerischen Spalten "Sales", "Cost" und "Profit":

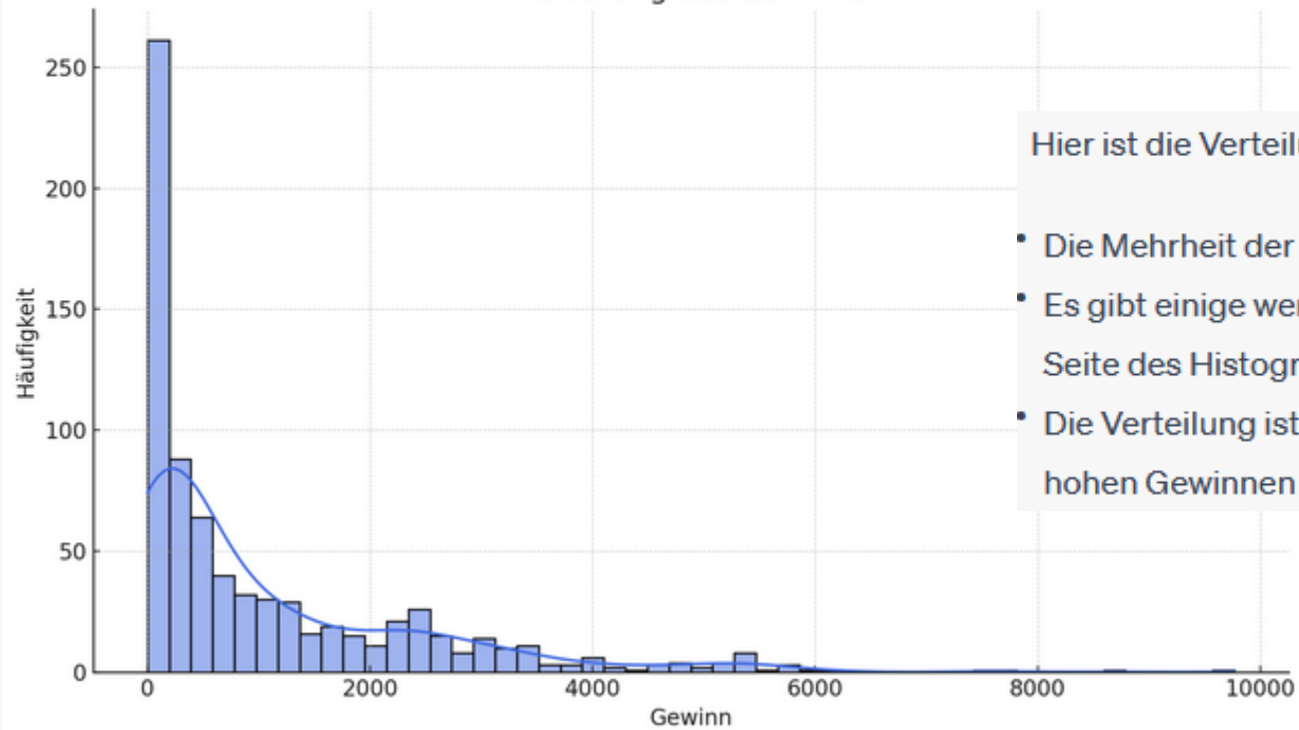
- **Count:** Es gibt 756 Datensätze.
- **Mean (Durchschnitt):**
  - Der durchschnittliche Umsatz beträgt etwa 1726.96 Einheiten.
  - Die durchschnittlichen Kosten betragen etwa 636.91 Einheiten.
  - Der durchschnittliche Gewinn beträgt etwa 1090.06 Einheiten.
- **Std (Standardabweichung):**
  - Der Umsatz hat eine Standardabweichung von etwa 2209.35, was auf eine breite Streuung der Daten hinweist.
  - Die Kosten haben eine Standardabweichung von etwa 923.39.
  - Der Gewinn hat eine Standardabweichung von etwa 1403.80.
- **Min und Max:**
  - Der Umsatz reicht von 3.44 bis 11751.55 Einheiten.
  - Die Kosten reichen von 0.08 bis 6365.29 Einheiten.
  - Der Gewinn reicht von 1.13 bis 9766.39 Einheiten.

Als nächstes werde ich die Verteilung des Gewinns visualisieren.

Finished working

Show work 

Verteilung des Gewinns



Hier ist die Verteilung des Gewinns:

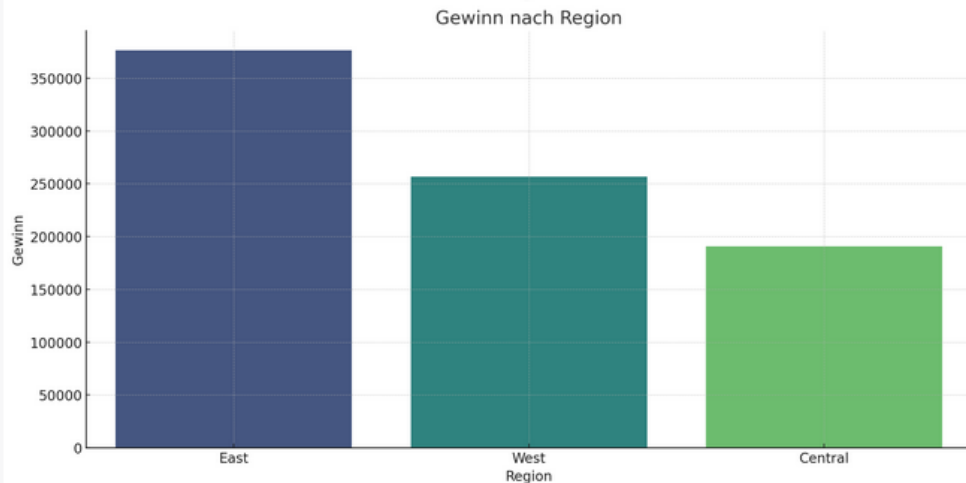
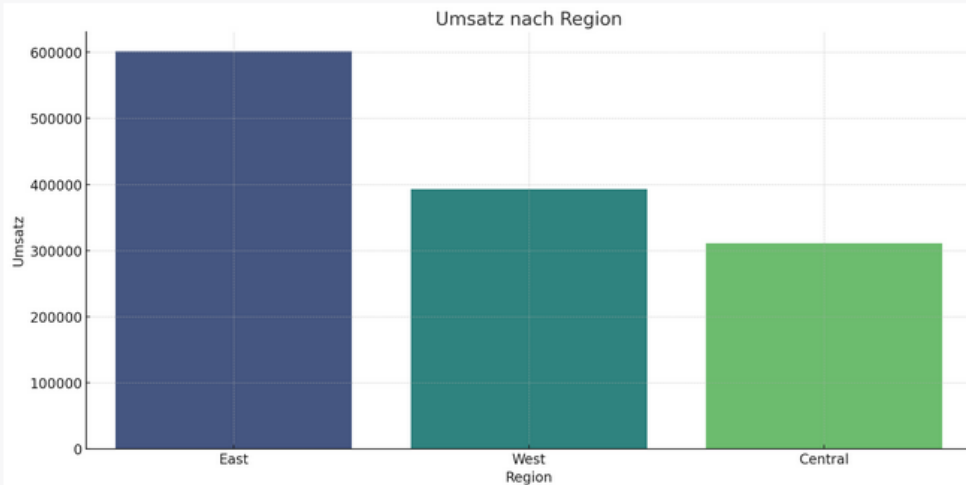
- Die Mehrheit der Gewinne liegt im Bereich von 0 bis 2.500 Einheiten.
- Es gibt einige wenige Transaktionen mit sehr hohen Gewinnen, die sich auf der rechten Seite des Histogramms zeigen.
- Die Verteilung ist rechtsschief, was bedeutet, dass es einige wenige Transaktionen mit sehr hohen Gewinnen gibt, während die meisten Transaktionen moderate Gewinne aufweisen.

# Gewinn nach Region

Als nächstes werde ich den Umsatz und den Gewinn nach Region visualisieren.

Finished working

Show work ▾



Hier sind die Balkendiagramme für Umsatz und Gewinn nach Region:

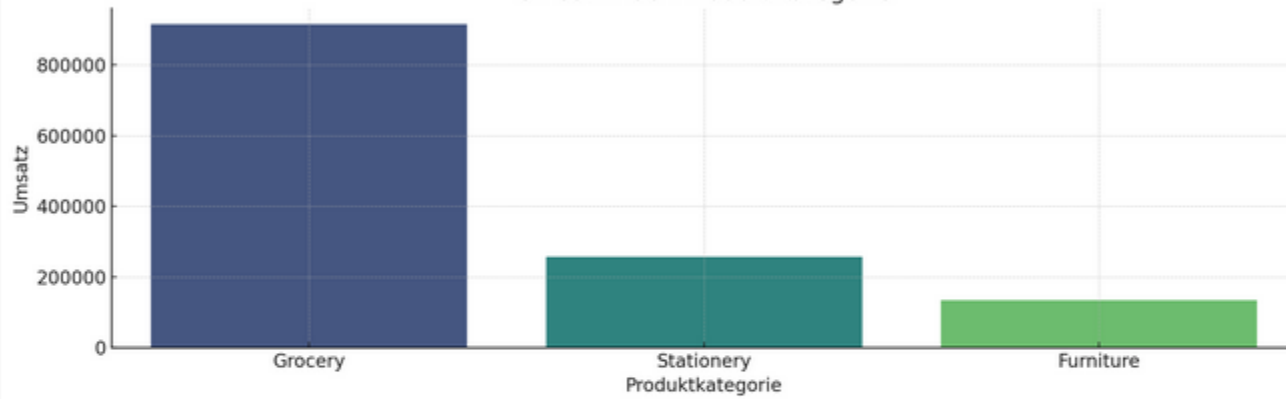
- **Umsatz nach Region:** Die Region "West" hat den höchsten Umsatz, gefolgt von "East", "Central" und "South".
- **Gewinn nach Region:** Auch hier führt die Region "West" in Bezug auf den Gesamtgewinn, gefolgt von "East", "Central" und "South".

# Gewinn nach Produktkategorie

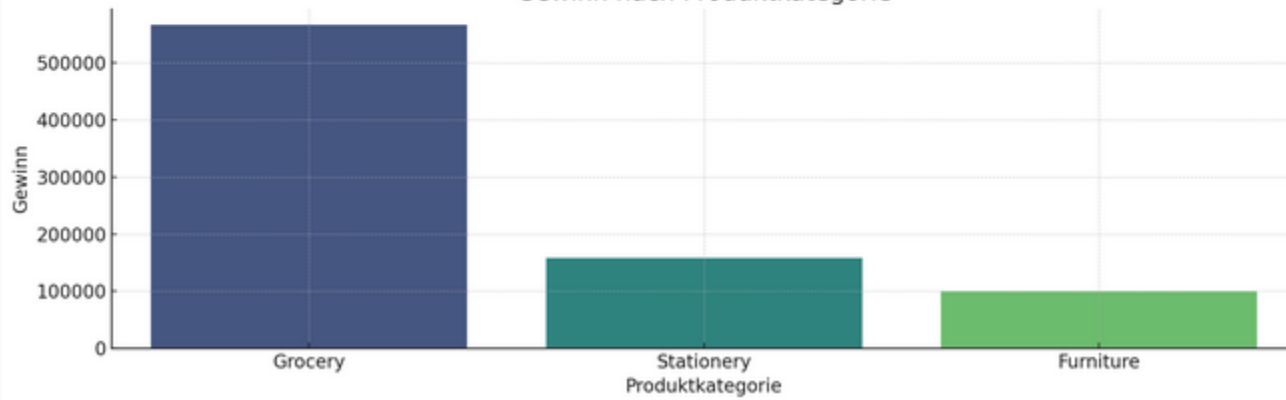
Als nächstes werde ich den Umsatz und den Gewinn nach Produktkategorie visualisieren.

Finished working Show work ▾

Umsatz nach Produktkategorie



Gewinn nach Produktkategorie



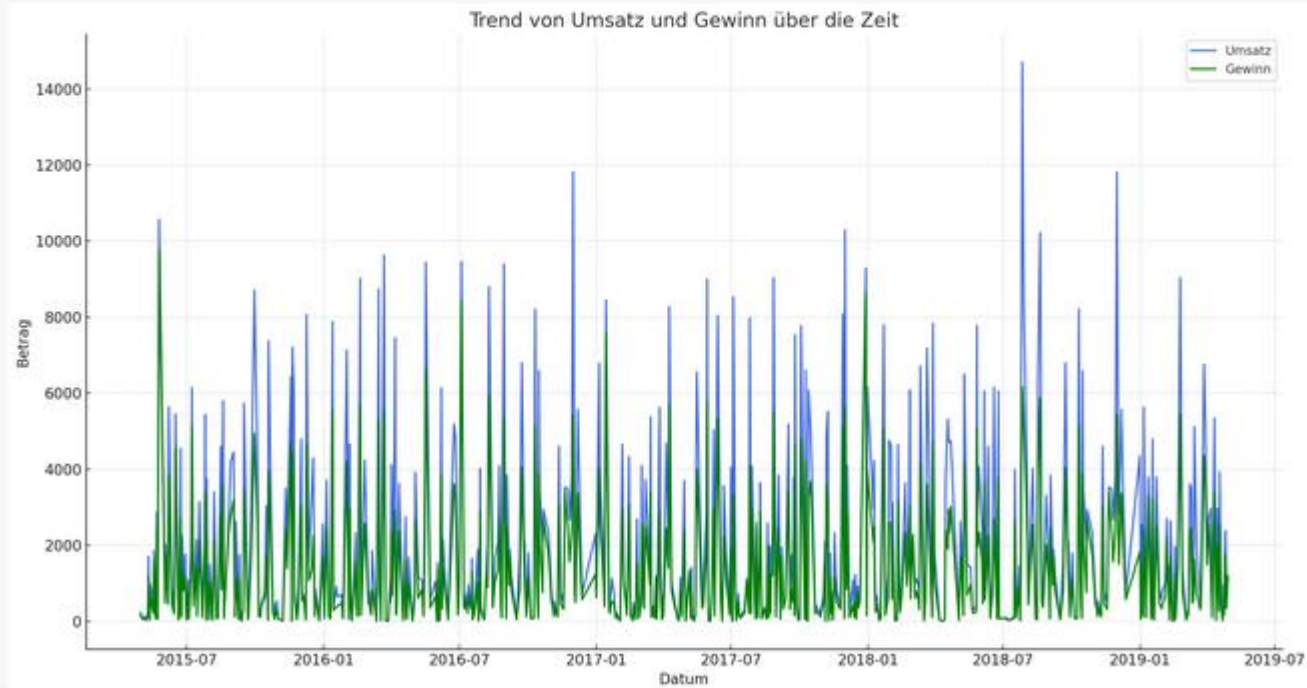
Hier sind die Balkendiagramme für Umsatz und Gewinn nach Produktkategorie:

- **Umsatz nach Produktkategorie:** Die Kategorie "Technology" hat den höchsten Umsatz, gefolgt von "Furniture" und "Office Supplies".
- **Gewinn nach Produktkategorie:** Die Kategorie "Technology" hat auch den höchsten Gewinn, gefolgt von "Furniture" und "Office Supplies".

Als nächstes werde ich den Trend von Umsatz und Gewinn über die Zeit visualisieren.

Finished working

Show work ▼



Hier ist das Zeitreihendiagramm, das den Trend von Umsatz und Gewinn über die Zeit zeigt:

- Der Umsatz und der Gewinn zeigen im Laufe der Zeit Schwankungen.
- Es gibt bestimmte Zeiträume, in denen sowohl der Umsatz als auch der Gewinn Spitzenwerte erreichen.
- Insgesamt scheint der Trend für Umsatz und Gewinn positiv zu sein, mit einem allgemeinen Anstieg über den dargestellten Zeitraum.

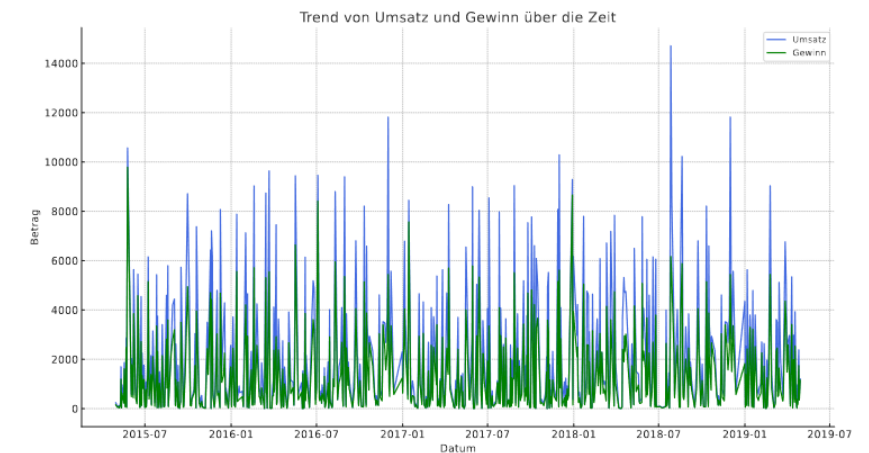
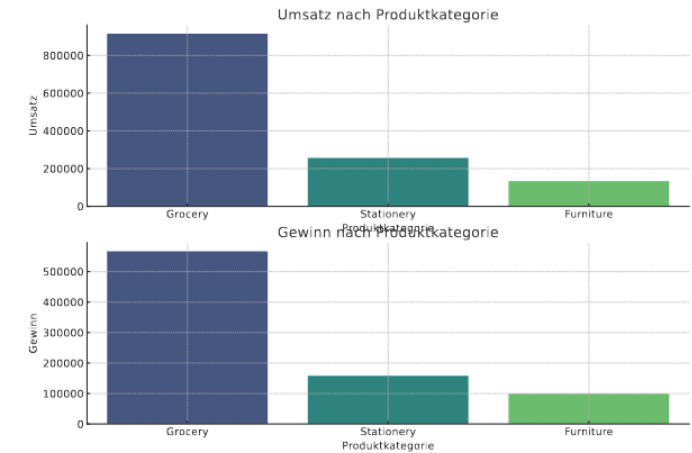
# Als PDF-File speichern? Kein Problem

Die Analyse und Visualisierungen wurden in ein PDF-Dokument exportiert. Sie können das PDF-Dokument über den folgenden Link herunterladen:

[SalesDataAnalysis.pdf herunterladen](#)

/Downloads/SalesDataAnalysis.pdf

60%





RE

Gibt es Ausreißer in den Daten?




Um festzustellen, ob es Ausreißer in den Daten gibt, können wir verschiedene Methoden verwenden:

1. **Boxplots:** Diese Plots zeigen die Verteilung der Daten und kennzeichnen potenzielle Ausreißer visuell.
2. **IQR (Interquartilsabstand):** Ein häufig verwendeter Ansatz zur Identifizierung von Ausreißern. Datenpunkte, die außerhalb des Bereichs von  $Q1 - 1.5 \times IQR$  und  $Q3 + 1.5 \times IQR$  liegen, gelten als Ausreißer.
3. **Z-Score:** Der Z-Score gibt an, wie viele Standardabweichungen ein Datenpunkt vom Mittelwert entfernt ist. Datenpunkte mit einem Z-Score von mehr als 3 oder weniger als -3 können als Ausreißer betrachtet werden.

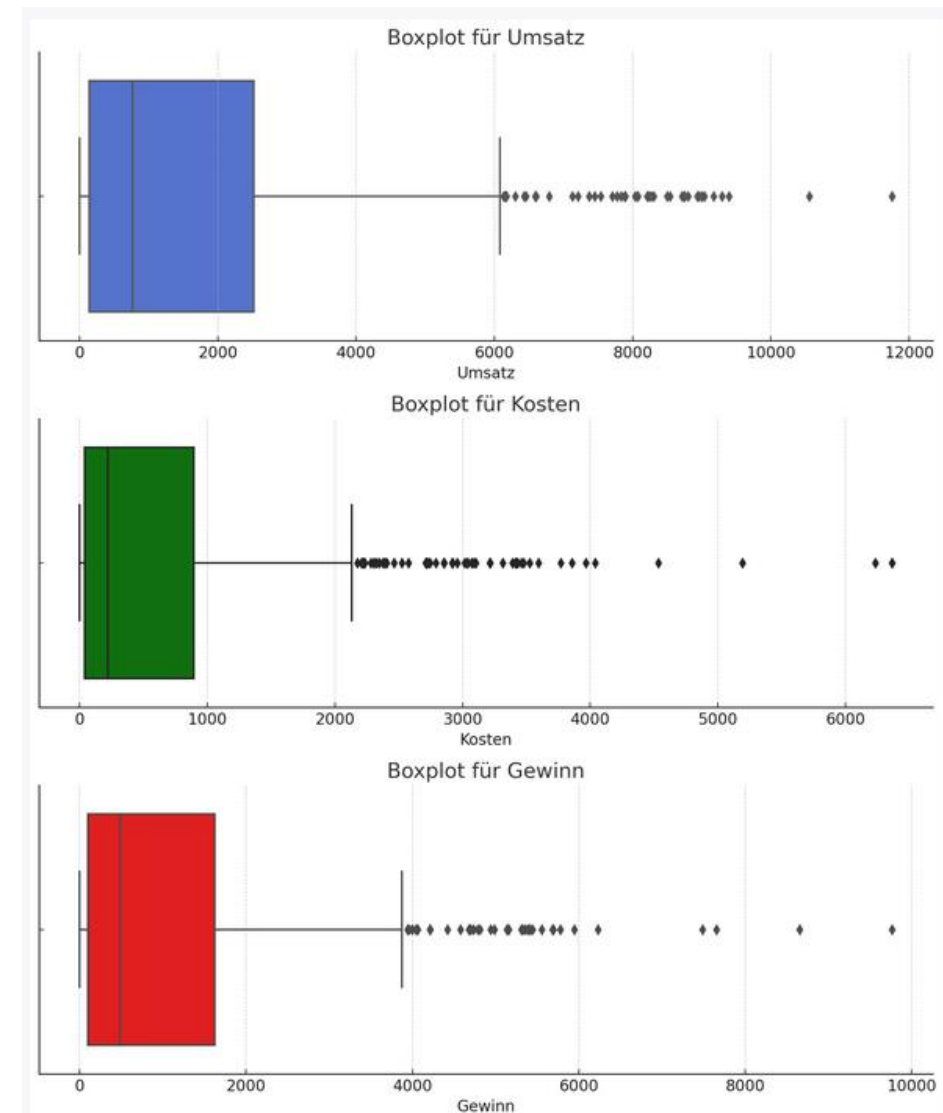
Ich werde zuerst Boxplots für die Spalten "Sales", "Cost" und "Profit" erstellen, um visuell nach Ausreißern zu suchen.

Finished working

Show work 

Basierend auf der IQR-Methode haben wir die folgenden Ausreißer in den Daten:

- **Umsatz (Sales):** 47 Ausreißer
- **Kosten (Cost):** 61 Ausreißer
- **Gewinn (Profit):** 40 Ausreißer





### 3. Datenanalyse mit ChatGPT Plus und GPTs







## GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Top Picks DALL-E Writing Productivity Research & Analysis Programming Education Lifestyle

### Featured

Curated top picks from this week

 <p><b>Website Generator</b> A GPT for website creation, design, copywriting, and code. Integrated with DALL-E 3. Powered by B12. By <a href="https://www.websitegenerator.b12.io">websitegenerator.b12.io</a></p>	 <p><b>PDF Ai PDF</b> Securely store and chat with ALL your PDFs for FREE, no matter the size. Free Chrome extension to sa... By <a href="https://myaidrive.com">myaidrive.com</a></p>
 <p><b>Diagrams: Show Me</b> Diagram creation: flowcharts, mindmaps, UML, chart, PlotUML, workflow, sequence, ERD, databas... By <a href="https://helpful.dev">helpful.dev</a></p>	 <p><b>Automated Writer</b> I craft well written, optimized, and researched, content in your unique voice. By <a href="https://octaneai.com">octaneai.com</a></p>
 <p><b>Universal Primer</b> The fastest way to learn anything. By <a href="https://siqi.chen">Siqi Chen</a></p>	 <p><b>Travel Guide</b> Expert on global travel destinations, trip planning, budget building, and exploring the world!... By <a href="https://capchair.com">capchair.com</a></p>

<https://chat.openai.com/g/g-BkMR28wIV-fair>

FAIR ▾



**FAIR**

By Renat Shigapov

Helps to make data FAIR

Create documentation for my data

Review my metadata

My dataset is uploaded. Make it FAIR

Write a DMP for my metadata

 Message FAIR...




<https://chat.openai.com/g/g-BkMR28wIV-fair>

# GPT: SigTech Finance

https://chat.openai.com/g/g-3yf7GNkby-sigtech-finance

SigTech Finance



**SigTech Finance**

By SIG Technologies US LLC

Used by top hedge funds to get cross-asset market data, test investment/trading portfolios & get finance charts/graphs

Compare the performance of Apple with an S&P ...

How would a 60% 40% basket of rolling S&P & ...

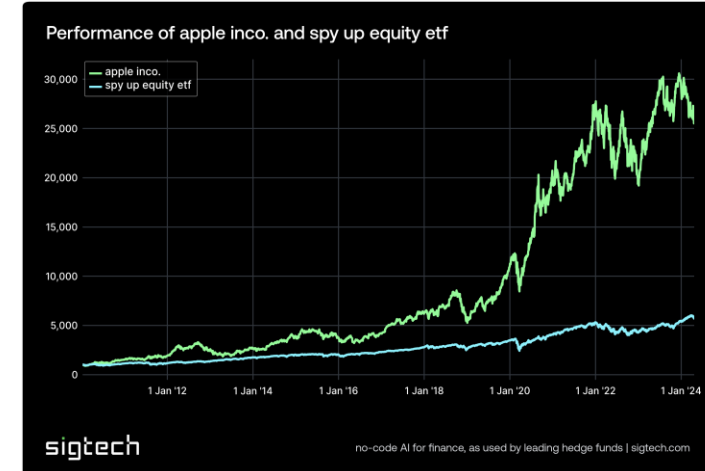
Can you chart EURUSD versus US GDP data fr...

Please analyze an uploaded chart and provide ...

Message SigTech Finance...

**You**  
Compare the performance of Apple with an S&P ETF since 2010?

- SigTech Finance**
- ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com
  - ✓ Talked to api.sigtech.com



The performance comparison of Apple Inc. (AAPL) and the SPDR S&P 500 ETF Trust (SPY) since 2010 yields the following insights based on their cumulative returns and various financial metrics:

### Performance Summary (2010 to 2024)

Metric	Apple Inc. (AAPL)	SPDR S&P 500 ETF Trust (SPY)
Annualized Excess Return (%)	24.32%	11.96%
Annualized Volatility (%)	28.06%	17.21%
Sharpe Ratio	0.92	0.74
Sortino Ratio	1.53	1.15
Max Drawdown (%)	-43.78%	-33.69%
Total Return %	2518.95%	485.25%
Max Drawdown Start Date	2012-09-19	2020-02-19
Max Drawdown End Date	2013-07-19	2020-03-23

# GPT: Stock Analyst by DAIZY

https://chat.openai.com/g/g-mhjpgUbn8g-stock-analyst-by-d

## Stock Analyst by Daizy



## Stock Analyst by Daizy

By community builder

DAIZY is the most advanced AI companion for all your US stock and ETF research. Powered by institutional-grade data and a real-time analytics engine..

Should I buy META?

How risky is Amazon?

How is NVIDIA doing?

Tesla sentiment analysis

**You**  
Should I buy META?

**Stock Analyst by Daizy**  
Talked to plugin.daizy.com

Regarding whether to buy Meta Platforms, Inc. (META), let's look at what analysts are currently saying. The analyst ratings for META are quite positive:

- **Target Price:** \$341.68
- **Average Analyst Rating:** 4.49 (on a scale where 5.0 is a strong buy)
- **Recommendations Breakdown:**
  - **Strong Buy:** 41 analysts
  - **Buy:** 11 analysts
  - **Hold:** 7 analysts
  - **Sell:** 2 analysts
  - **Strong Sell:** 0 analysts

This information indicates that a significant majority of analysts have a favorable outlook on META, suggesting it could be a good investment based on current analysis. Here are some visual insights:

## Meta Platforms

META  
Apr 26 2024  
(close)

What do the analysts think?



Source: IEX Cloud, EOD Historical  
© 2024 Vesti.AI



## Meta Platforms

META  
Apr 26 2024  
(close)

What do the analysts think?



Source: IEX Cloud, EOD Historical  
© 2024 Vesti.AI



https://chat.openai.com/g/g-CMdGQ6ENA-uk-companies-house-explorer

## UK Companies House Explorer



### UK Companies House Explorer

By Noland Smith

Talk Directly with the UK Companies House Database to find information on businesses, officers, and documentation

Tell me how this works

You

provide me a list of of ceos from the three UK companies with the highest revenue



UK Companies House Explorer

To identify and provide information about the CEOs from the three UK companies with the highest revenue, I'll use a common reference for business rankings like the latest "FTSE 100" list or similar financial summaries to determine which companies top the revenue charts. Here are the companies traditionally known for high revenue figures:

1. **BP plc (British Petroleum)**
2. **Royal Dutch Shell plc**
3. **Tesco plc**

Please note, these companies are often at the top due to their scale and global operations, particularly in sectors like oil and energy, and retail. Next, I'll find the current CEOs of these companies. Let's start with BP plc.

The current CEO of BP plc, one of the UK companies with the highest revenue, is Murray Auchincloss. He officially assumed the role in January 2024 after serving as the interim CEO since September 2023. Prior to this, he held the position of Chief Financial Officer at BP (BP).

Next, I'll look up the current CEOs for Royal Dutch Shell plc and Tesco plc to complete the list. Let's proceed with Royal Dutch Shell plc.

The current CEOs of the two UK companies with the highest revenue are:

1. **BP plc:** Murray Auchincloss is the CEO. He was appointed in January 2024 after serving as the interim CEO since September 2023 (BP).
2. **Royal Dutch Shell plc:** Wael Sawan has been the CEO since January 2023. He succeeded Ben van Beurden, who had been in the role since 2014 (Shell Global) (Wikipedia) (Royal Dutch Shell Plc .com).

Next, I'll look up the current CEO of Tesco plc to complete your request.

The current Group Chief Executive Officer of Tesco plc is Ken Murphy. He has held this position since October 2020, after succeeding Dave Lewis .

To summarize, here are the CEOs of the three UK companies with the highest revenue:

1. **BP plc:** Murray Auchincloss
2. **Royal Dutch Shell plc:** Wael Sawan
3. **Tesco plc:** Ken Murphy

If you need more detailed profiles of these CEOs or further assistance on other companies, feel free to ask!



## Datenanalysepraktikum mit ChatGPT-3.5

10 Minuten Gruppenarbeit

<https://chat.openai.com>

<https://chat-ai.academiccloud.de>

Aufgabe: Führen Sie eine möglichst umfassende NLP-Analyse zum folgenden Text durch

Daten: unstrukturierter Text



### Offene Diskussionsrunde



**Mo., 06.05.2024**

Anja Perry

**5.Forschungsdaten  
teilen (5/6)**



Keeping  
research  
data



Sharing  
research  
data