DSC550-T301

Chitramoy Mukherjee

Week-4

Date: 12/20/2023

In [41]:
```
import warnings
warnings.filterwarnings('ignore')

# Required python basic libraries

import numpy as np
import pandas as pd
import textblob
from textblob import TextBlob
import seaborn as sns
import matplotlib.pyplot as plt




from os.path import basename, exists


def download(url):
    filename = basename(url)
    if not exists(filename):
        from urllib.request import urlretrieve

        local, _ = urlretrieve(url, filename)
        print("Downloaded " + local)

### Reading the auto-mpg.csv file into DataFrame
auto_df = pd.read_csv("C:\\Users\\14024\\OneDrive\\Desktop\\MS-DSC\\DSC-550\Week-4\\au

# Display the first few rows of the DataFrame to ensure it's loaded properly
print(auto_df.head())
```

```
    mpg  cylinders  displacement horsepower  weight  acceleration  model year  \
0  18.0          8         307.0        130    3504          12.0          70
1  15.0          8         350.0        165    3693          11.5          70
2  18.0          8         318.0        150    3436          11.0          70
3  16.0          8         304.0        150    3433          12.0          70
4  17.0          8         302.0        140    3449          10.5          70

   origin                   car name
0       1  chevrolet chevelle malibu
1       1          buick skylark 320
2       1         plymouth satellite
3       1              amc rebel sst
4       1                ford torino
```

In [3]:
```
# Remove the 'car name' column
auto_df = auto_df.drop('car name', axis=1)

# Display the DataFrame after removing the 'car name' column
```

```
print("\nDataFrame after removing 'car name' column:")
print(auto_df.head())
```

```
DataFrame after removing 'car name' column:
    mpg  cylinders  displacement horsepower  weight  acceleration  model year  \
0  18.0          8         307.0        130    3504          12.0          70
1  15.0          8         350.0        165    3693          11.5          70
2  18.0          8         318.0        150    3436          11.0          70
3  16.0          8         304.0        150    3433          12.0          70
4  17.0          8         302.0        140    3449          10.5          70

   origin
0       1
1       1
2       1
3       1
4       1
```

In [4]:
```
# Create dummy variables for the 'origin' column
df = pd.get_dummies(auto_df, columns=['origin'], drop_first=True)

# Display the modified DataFrame
print("\nDataFrame after preprocessing:")
print(auto_df.head())
```

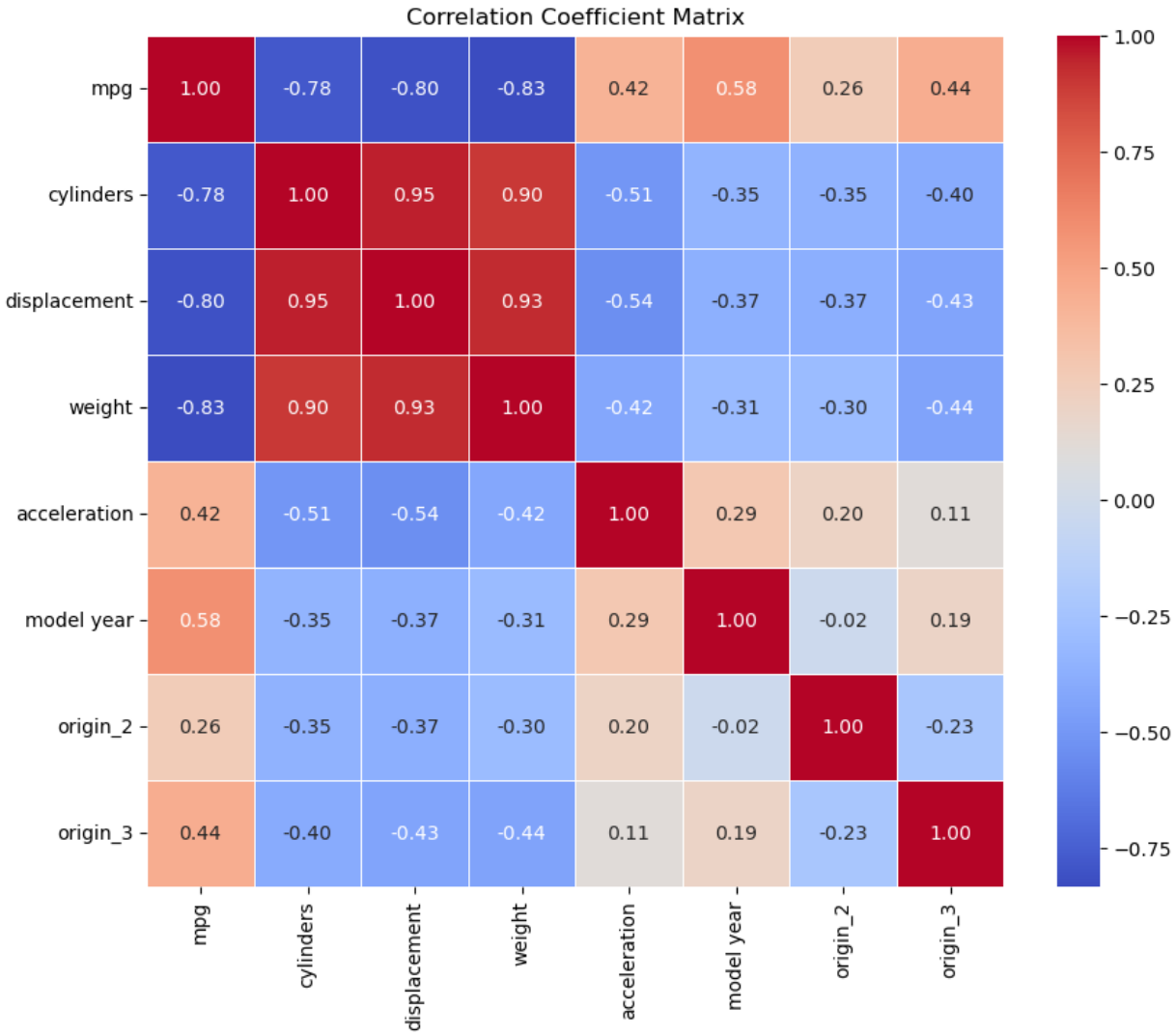
```
DataFrame after preprocessing:
    mpg  cylinders  displacement horsepower  weight  acceleration  model year  \
0  18.0          8         307.0        130    3504          12.0          70
1  15.0          8         350.0        165    3693          11.5          70
2  18.0          8         318.0        150    3436          11.0          70
3  16.0          8         304.0        150    3433          12.0          70
4  17.0          8         302.0        140    3449          10.5          70

   origin
0       1
1       1
2       1
3       1
4       1
```

In [5]:
```
# Create a correlation coefficient matrix
correlation_matrix = df.corr()

# Create a heatmap for visualization
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5
plt.title("Correlation Coefficient Matrix")
plt.show()

# Display features highly correlated with 'mpg'
mpg_correlations = correlation_matrix['mpg'].sort_values(ascending=False)
print("Features highly correlated with 'mpg':\n", mpg_correlations)
```




```
Features highly correlated with 'mpg':
 mpg             1.000000
model year       0.579267
origin_3         0.442174
acceleration     0.420289
origin_2         0.259022
cylinders       -0.775396
displacement    -0.804203
weight          -0.831741
Name: mpg, dtype: float64
```

In [44]:
```
pip install flair
```

```
Collecting flairNote: you may need to restart the kernel to use updated packages.

  Downloading flair-0.13.0-py3-none-any.whl (387 kB)
     ------------------------------------ 387.2/387.2 kB 4.0 MB/s eta 0:00:00
Collecting segtok>=1.5.11
  Downloading segtok-1.5.11-py3-none-any.whl (24 kB)
Collecting gensim>=4.2.0
  Downloading gensim-4.3.2-cp39-cp39-win_amd64.whl (24.0 MB)
     ---------------------------------------- 24.0/24.0 MB 10.7 MB/s eta 0:00:00
Collecting langdetect>=1.0.9
  Downloading langdetect-1.0.9.tar.gz (981 kB)
     ------------------------------------ 981.5/981.5 kB 4.4 MB/s eta 0:00:00
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Collecting wikipedia-api>=0.5.7
  Downloading Wikipedia_API-0.6.0-py3-none-any.whl (14 kB)
Collecting gdown>=4.4.0
  Downloading gdown-4.7.1-py3-none-any.whl (15 kB)
Collecting conllu>=4.0
  Downloading conllu-4.5.3-py2.py3-none-any.whl (16 kB)
Collecting ftfy>=6.1.0
  Downloading ftfy-6.1.3-py3-none-any.whl (53 kB)
     ---------------------------------------- 53.4/53.4 kB 2.7 MB/s eta 0:00:00
Requirement already satisfied: lxml>=4.8.0 in c:\users\14024\anaconda3\lib\site-packa
ges (from flair) (4.9.1)
Collecting deprecated>=1.2.13
  Downloading Deprecated-1.2.14-py2.py3-none-any.whl (9.6 kB)
Collecting transformer-smaller-training-vocab>=0.2.3
  Downloading transformer_smaller_training_vocab-0.3.3-py3-none-any.whl (14 kB)
Collecting mpld3>=0.3
  Downloading mpld3-0.5.9-py3-none-any.whl (201 kB)
     ------------------------------------ 201.2/201.2 kB 1.4 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.63.0 in c:\users\14024\anaconda3\lib\site-pack
ages (from flair) (4.64.1)
Collecting semver<4.0.0,>=3.0.0
  Downloading semver-3.0.2-py3-none-any.whl (17 kB)
Collecting sqlitedict>=2.0.0
  Downloading sqlitedict-2.1.0.tar.gz (21 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\14024\anaconda3\lib
\site-packages (from flair) (2.8.2)
Collecting huggingface-hub>=0.10.0
  Downloading huggingface_hub-0.19.4-py3-none-any.whl (311 kB)
     ------------------------------------ 311.7/311.7 kB 9.7 MB/s eta 0:00:00
Collecting torch!=1.8,>=1.5.0
  Downloading torch-2.1.2-cp39-cp39-win_amd64.whl (192.2 MB)
     ---------------------------------------- 192.2/192.2 MB 4.1 MB/s eta 0:00:00
Collecting bpemb>=0.3.2
  Downloading bpemb-0.3.4-py3-none-any.whl (19 kB)
Requirement already satisfied: tabulate>=0.8.10 in c:\users\14024\anaconda3\lib\site-
packages (from flair) (0.8.10)
Requirement already satisfied: boto3>=1.20.27 in c:\users\14024\anaconda3\lib\site-pa
ckages (from flair) (1.24.28)
Collecting more-itertools>=8.13.0
  Downloading more_itertools-10.1.0-py3-none-any.whl (55 kB)
     ---------------------------------------- 55.8/55.8 kB 3.0 MB/s eta 0:00:00
Collecting pptree>=3.1
  Downloading pptree-3.1.tar.gz (3.0 kB)
  Preparing metadata (setup.py): started
```

```
  Preparing metadata (setup.py): finished with status 'done'
Requirement already satisfied: regex>=2022.1.18 in c:\users\14024\anaconda3\lib\site-
packages (from flair) (2022.7.9)
Collecting transformers[sentencepiece]<5.0.0,>=4.18.0
  Downloading transformers-4.36.1-py3-none-any.whl (8.3 MB)
     ---------------------------------------- 8.3/8.3 MB 3.1 MB/s eta 0:00:00
Collecting pytorch-revgrad>=0.2.0
  Downloading pytorch_revgrad-0.2.0-py3-none-any.whl (4.6 kB)
Requirement already satisfied: scikit-learn>=1.0.2 in c:\users\14024\anaconda3\lib\si
te-packages (from flair) (1.0.2)
Requirement already satisfied: urllib3<2.0.0,>=1.0.0 in c:\users\14024\anaconda3\lib
\site-packages (from flair) (1.26.11)
Requirement already satisfied: matplotlib>=2.2.3 in c:\users\14024\anaconda3\lib\site
-packages (from flair) (3.5.2)
Collecting janome>=0.4.2
  Downloading Janome-0.5.0-py2.py3-none-any.whl (19.7 MB)
     ---------------------------------------- 19.7/19.7 MB 4.5 MB/s eta 0:00:00
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in c:\users\14024\anaconda3\lib
\site-packages (from boto3>=1.20.27->flair) (0.10.0)
Requirement already satisfied: s3transfer<0.7.0,>=0.6.0 in c:\users\14024\anaconda3\l
ib\site-packages (from boto3>=1.20.27->flair) (0.6.0)
Requirement already satisfied: botocore<1.28.0,>=1.27.28 in c:\users\14024\anaconda3
\lib\site-packages (from boto3>=1.20.27->flair) (1.27.28)
Requirement already satisfied: numpy in c:\users\14024\anaconda3\lib\site-packages (f
rom bpemb>=0.3.2->flair) (1.21.5)
Collecting sentencepiece
  Downloading sentencepiece-0.1.99-cp39-cp39-win_amd64.whl (977 kB)
     ---------------------------------------- 977.6/977.6 kB 3.9 MB/s eta 0:00:00
Requirement already satisfied: requests in c:\users\14024\anaconda3\lib\site-packages
(from bpemb>=0.3.2->flair) (2.28.1)
Requirement already satisfied: wrapt<2,>=1.10 in c:\users\14024\anaconda3\lib\site-pa
ckages (from deprecated>=1.2.13->flair) (1.14.1)
Collecting wcwidth<0.3.0,>=0.2.12
  Downloading wcwidth-0.2.12-py2.py3-none-any.whl (34 kB)
Requirement already satisfied: beautifulsoup4 in c:\users\14024\anaconda3\lib\site-pa
ckages (from gdown>=4.4.0->flair) (4.11.1)
Requirement already satisfied: filelock in c:\users\14024\anaconda3\lib\site-packages
(from gdown>=4.4.0->flair) (3.6.0)
Requirement already satisfied: six in c:\users\14024\anaconda3\lib\site-packages (fro
m gdown>=4.4.0->flair) (1.16.0)
Requirement already satisfied: scipy>=1.7.0 in c:\users\14024\anaconda3\lib\site-pack
ages (from gensim>=4.2.0->flair) (1.9.1)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\14024\anaconda3\lib\site
-packages (from gensim>=4.2.0->flair) (5.2.1)
Requirement already satisfied: packaging>=20.9 in c:\users\14024\anaconda3\lib\site-p
ackages (from huggingface-hub>=0.10.0->flair) (21.3)
Collecting fsspec>=2023.5.0
  Downloading fsspec-2023.12.2-py3-none-any.whl (168 kB)
     ---------------------------------------- 169.0/169.0 kB 9.9 MB/s eta 0:00:00
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\14024\anaconda3
\lib\site-packages (from huggingface-hub>=0.10.0->flair) (4.3.0)
Requirement already satisfied: pyyaml>=5.1 in c:\users\14024\anaconda3\lib\site-packa
ges (from huggingface-hub>=0.10.0->flair) (6.0)
Requirement already satisfied: pillow>=6.2.0 in c:\users\14024\anaconda3\lib\site-pac
kages (from matplotlib>=2.2.3->flair) (9.2.0)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\14024\anaconda3\lib\site-
packages (from matplotlib>=2.2.3->flair) (3.0.9)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\14024\anaconda3\lib\site
-packages (from matplotlib>=2.2.3->flair) (4.25.0)
Requirement already satisfied: cycler>=0.10 in c:\users\14024\anaconda3\lib\site-pack
```

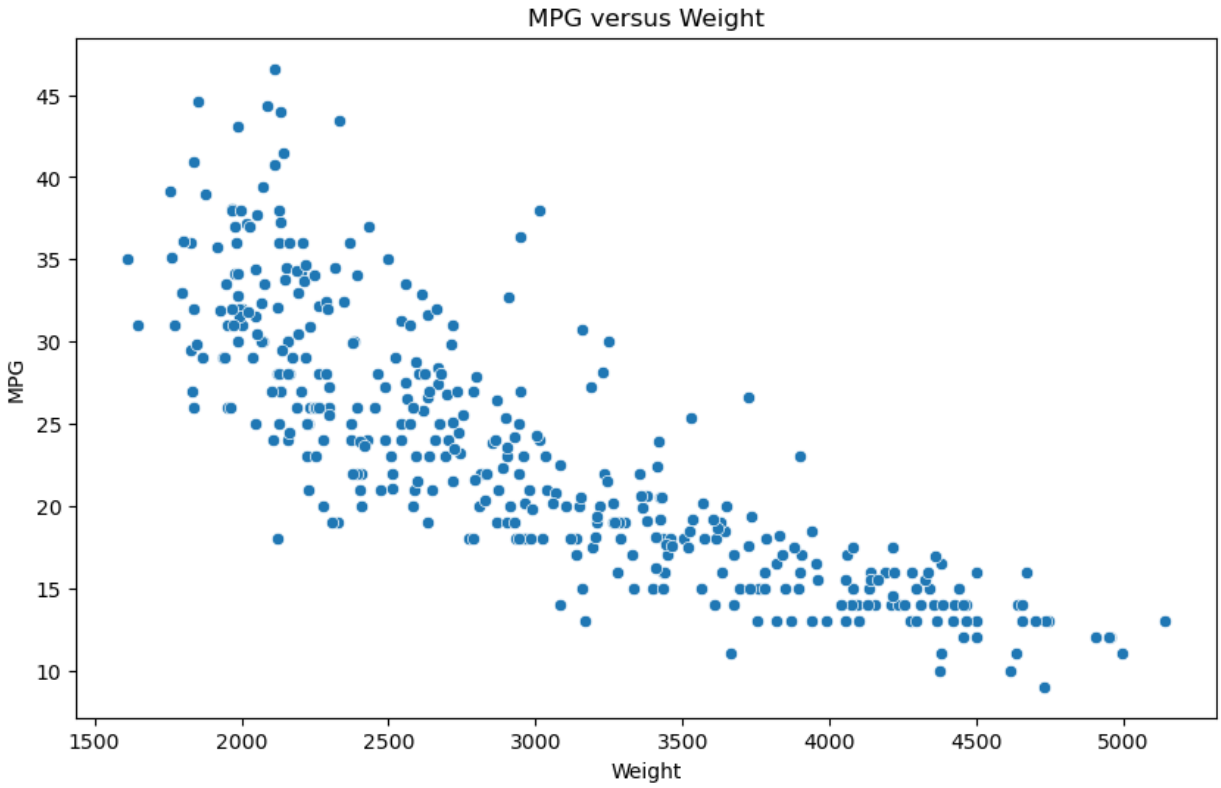
```
ages (from matplotlib>=2.2.3->flair) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\14024\anaconda3\lib\site
-packages (from matplotlib>=2.2.3->flair) (1.4.2)
Requirement already satisfied: jinja2 in c:\users\14024\anaconda3\lib\site-packages
(from mpld3>=0.3->flair) (2.11.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\14024\anaconda3\lib\s
ite-packages (from scikit-learn>=1.0.2->flair) (2.2.0)
Requirement already satisfied: joblib>=0.11 in c:\users\14024\anaconda3\lib\site-pack
ages (from scikit-learn>=1.0.2->flair) (1.1.0)
Requirement already satisfied: sympy in c:\users\14024\anaconda3\lib\site-packages (f
rom torch!=1.8,>=1.5.0->flair) (1.10.1)
Requirement already satisfied: networkx in c:\users\14024\anaconda3\lib\site-packages
(from torch!=1.8,>=1.5.0->flair) (2.8.4)
Requirement already satisfied: colorama in c:\users\14024\anaconda3\lib\site-packages
(from tqdm>=4.63.0->flair) (0.4.5)
Collecting tokenizers<0.19,>=0.14
  Downloading tokenizers-0.15.0-cp39-none-win_amd64.whl (2.2 MB)
     ---------------------------------------- 2.2/2.2 MB 3.2 MB/s eta 0:00:00
Collecting safetensors>=0.3.1
  Downloading safetensors-0.4.1-cp39-none-win_amd64.whl (277 kB)
     ---------------------------------------- 277.8/277.8 kB 4.3 MB/s eta 0:00:00
Collecting protobuf
  Downloading protobuf-4.25.1-cp39-cp39-win_amd64.whl (413 kB)
     ---------------------------------------- 413.4/413.4 kB 5.2 MB/s eta 0:00:00
Collecting accelerate>=0.21.0
  Downloading accelerate-0.25.0-py3-none-any.whl (265 kB)
     ---------------------------------------- 265.7/265.7 kB 5.4 MB/s eta 0:00:00
Requirement already satisfied: soupsieve>1.2 in c:\users\14024\anaconda3\lib\site-pac
kages (from beautifulsoup4->gdown>=4.4.0->flair) (2.3.1)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\14024\anaconda3\lib\site-
packages (from jinja2->mpld3>=0.3->flair) (2.0.1)
Requirement already satisfied: idna<4,>=2.5 in c:\users\14024\anaconda3\lib\site-pack
ages (from requests->bpemb>=0.3.2->flair) (3.3)
Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\14024\anaconda3\l
ib\site-packages (from requests->bpemb>=0.3.2->flair) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\14024\anaconda3\lib\sit
e-packages (from requests->bpemb>=0.3.2->flair) (2022.9.14)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in c:\users\14024\anaconda3\lib
\site-packages (from requests->bpemb>=0.3.2->flair) (1.7.1)
Requirement already satisfied: mpmath>=0.19 in c:\users\14024\anaconda3\lib\site-pack
ages (from sympy->torch!=1.8,>=1.5.0->flair) (1.2.1)
Requirement already satisfied: psutil in c:\users\14024\anaconda3\lib\site-packages
(from accelerate>=0.21.0->transformers[sentencepiece]<5.0.0,>=4.18.0->flair) (5.9.0)
Building wheels for collected packages: langdetect, pptree, sqlitedict
  Building wheel for langdetect (setup.py): started
  Building wheel for langdetect (setup.py): finished with status 'done'
  Created wheel for langdetect: filename=langdetect-1.0.9-py3-none-any.whl size=99322
5 sha256=a3cf388b0018e2c805c2bff2f118cf0dfa21e61845ebc3aa5941a682da4c3eda
  Stored in directory: c:\users\14024\appdata\local\pip\cache\wheels\d1\c1\d9\7e068de
779d863bc8f8fc9467d85e25cfe47fa5051fff1a1bb
  Building wheel for pptree (setup.py): started
  Building wheel for pptree (setup.py): finished with status 'done'
  Created wheel for pptree: filename=pptree-3.1-py3-none-any.whl size=4609 sha256=e38
3d8cf4a836ff0d4f3af95a8db4e98a2f494a56052a6475385445f626720f0
  Stored in directory: c:\users\14024\appdata\local\pip\cache\wheels\52\0e\51\514e690
004ea9713bc3fdb678d5e2768fcc597d0c3b6a3abd2
  Building wheel for sqlitedict (setup.py): started
  Building wheel for sqlitedict (setup.py): finished with status 'done'
  Created wheel for sqlitedict: filename=sqlitedict-2.1.0-py3-none-any.whl size=16864
sha256=3ff2797b84a62ba7be9bc9982b1675bc06fb76f749b7d6aedd62b8ab06b8a17c
```

```
  Stored in directory: c:\users\14024\appdata\local\pip\cache\wheels\f6\48\c4\942f7a1
d556fddd2348cb9ac262f251873dfd8a39afec5678e
Successfully built langdetect pptree sqlitedict
Installing collected packages: wcwidth, sqlitedict, sentencepiece, pptree, janome, se
mver, segtok, safetensors, protobuf, more-itertools, langdetect, ftfy, fsspec, deprec
ated, conllu, wikipedia-api, torch, huggingface-hub, gensim, tokenizers, pytorch-revg
rad, mpld3, gdown, bpemb, accelerate, transformers, transformer-smaller-training-voca
b, flair
  Attempting uninstall: wcwidth
    Found existing installation: wcwidth 0.2.5
    Uninstalling wcwidth-0.2.5:
      Successfully uninstalled wcwidth-0.2.5
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2022.7.1
    Uninstalling fsspec-2022.7.1:
      Successfully uninstalled fsspec-2022.7.1
  Attempting uninstall: gensim
    Found existing installation: gensim 4.1.2
    Uninstalling gensim-4.1.2:
      Successfully uninstalled gensim-4.1.2
Successfully installed accelerate-0.25.0 bpemb-0.3.4 conllu-4.5.3 deprecated-1.2.14 f
lair-0.13.0 fsspec-2023.12.2 ftfy-6.1.3 gdown-4.7.1 gensim-4.3.2 huggingface-hub-0.1
9.4 janome-0.5.0 langdetect-1.0.9 more-itertools-10.1.0 mpld3-0.5.9 pptree-3.1 protob
uf-4.25.1 pytorch-revgrad-0.2.0 safetensors-0.4.1 segtok-1.5.11 semver-3.0.2 sentence
piece-0.1.99 sqlitedict-2.1.0 tokenizers-0.15.0 torch-2.1.2 transformer-smaller-train
ing-vocab-0.3.3 transformers-4.36.1 wcwidth-0.2.12 wikipedia-api-0.6.0
```

In [12]:

```
# Plot mpg versus weight
# Plot mpg versus weight
plt.figure(figsize=(10, 6))
sns.scatterplot(x='weight', y='mpg', data=df)
plt.title('MPG versus Weight')
plt.xlabel('Weight')
plt.ylabel('MPG')
plt.show()

# Explain the relationship between MPG and Weight
print("Analyzing the relationship between MPG and Weight:")
print("The scatterplot shows a negative correlation between MPG and Weight.")
print("As weight increases, MPG tends to decrease.")
print("This is consistent with the negative correlation coefficient observed in the cc
```



Analyzing the relationship between MPG and Weight:
The scatterplot shows a negative correlation between MPG and Weight.
As weight increases, MPG tends to decrease.
This is consistent with the negative correlation coefficient observed in the correlat
ion matrix.

In [32]:
```
# Randomly split the data into 80% training data and 20% test data
X = df.drop('mpg', axis=1)
y = df['mpg']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=

# Check the shape of the training and test sets
print("Training set shape:", X_train.shape, y_train.shape)
print("Test set shape:", X_test.shape, y_test.shape)
```

```
Training set shape: (318, 8) (318,)
Test set shape: (80, 8) (80,)
```

In [38]:
```
# Train an ordinary linear regression model
linear_reg_model = LinearRegression()
linear_reg_model.fit(X_train, y_train)

# Predict on the training data
y_train_pred = linear_reg_model.predict(X_train)

# Evaluate the model using training data
r2_train = r2_score(y_train, y_train_pred)
rmse_train = mean_squared_error(y_train, y_train_pred, squared=False)
mae_train = mean_absolute_error(y_train, y_train_pred)

# Print the results of R^2, RMSE and MAE
print("Results on the Training Set:")
print(f'R^2: {r2_train:.4f}')
print(f'RMSE: {rmse_train:.4f}')
print(f'MAE: {mae_train:.4f}')
```

```
rmse_train = mean_squared_error(y_train, y_train_pred, squared=False)
mae_train = mean_absolute_error(y_train, y_train_pred)

# Interpret the results
print(f"R2 (Training): {r2_train}")
print(f"RMSE (Training): {rmse_train}")
print(f"MAE (Training): {mae_train}")

# R-squared is a measure of how well the model explains the variability of the target
# R^2 of 0.8188 means that approximately 81.88% of the variability in the target varia
# A lower RMSE indicates better model performance. In this case, an RMSE of 3.3703 mea
# Lower MAE indicates better model performance. In this case, an MAE of 2.6055 means t
```

```
Results on the Training Set:
R^2: 0.8188
RMSE: 3.3703
MAE: 2.6055
R2 (Training): 0.8188288951042786
RMSE (Training): 3.3702735639389054
MAE (Training): 2.6054846937710354
```

In [40]:

```
from sklearn.ensemble import RandomForestRegressor

# Train a Random Forest Regression model
rf_model = RandomForestRegressor(random_state=42)
rf_model.fit(X_train, y_train)

# Predict on training and test sets
y_train_pred_rf = rf_model.predict(X_train)

# Calculate R2, RMSE, and MAE on both training and test sets for Random Forest model
r2_train_rf = r2_score(y_train, y_train_pred_rf)

rmse_train_rf = mean_squared_error(y_train, y_train_pred_rf, squared=False)

mae_train_rf = mean_absolute_error(y_train, y_train_pred_rf)

# Interpret the results for Random Forest model
print("Random Forest Regression Results:")
print(f"R2 (Training): {r2_train_rf}")
print(f"RMSE (Training): {rmse_train_rf}")
print(f"MAE (Training): {mae_train_rf}")

# R-squared is a measure of how well the model explains the variability of the target
# R^2 of 0.9810 means that approximately 98.10% of the variability in the target varia
# A lower RMSE indicates better model performance. In this case, an RMSE of 1.0908 mea
# Lower MAE indicates better model performance. In this case, an MAE of 0.7477 means t
```

```
Random Forest Regression Results:
R2 (Training): 0.9810189898945959
RMSE (Training): 1.0908884599607205
MAE (Training): 0.7477955974842765
```