jupyter 230701091 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [19]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
data=pd.read_csv('books.csv')
print(data.describe())
```

```
         book_id  goodreads_book_id  best_book_id       work_id  books_count  \
count  99.000000       9.900000e+01  9.900000e+01  9.900000e+01    99.000000
mean   50.000000       1.653979e+06  2.020180e+06  4.388198e+06   532.383838
std    28.722813       4.030956e+06  4.681734e+06  6.465936e+06   700.853630
min     1.000000       1.000000e+00  1.000000e+00  5.397000e+03    14.000000
25%    25.500000       3.925000e+03  4.297500e+03  1.681676e+06   172.500000
50%    50.000000       1.813500e+04  1.906300e+04  3.036731e+06   226.000000
75%    74.500000       1.927180e+05  3.135880e+05  3.357144e+06   480.500000
max    99.000000       2.255727e+07  2.255727e+07  4.133543e+07  3455.000000

              isbn13  original_publication_year  average_rating  ratings_count  \
count  9.900000e+01                  99.000000       99.000000   9.900000e+01
mean   9.780473e+12                1943.171717        4.055051   1.263703e+06
std    3.961477e+08                 277.288365        0.245059   7.832598e+05
min    9.780007e+12                -720.000000        3.510000   3.872900e+05
25%    9.780150e+12                1952.500000        3.870000   7.388255e+05
50%    9.780385e+12                1997.000000        4.060000   1.053403e+06
75%    9.780553e+12                2005.000000        4.245000   1.653671e+06
max    9.781612e+12                2015.000000        4.610000   4.780653e+06

       work_ratings_count  work_text_reviews_count      ratings_1  \
count        9.900000e+01                99.000000      99.000000
mean         1.361989e+06             39012.222222   39065.565657
std          7.978995e+05             28693.522227   50901.940472
min          5.493010e+05              4239.000000    4623.000000
25%          8.051315e+05             17789.500000   15337.500000
50%          1.125231e+06             31212.000000   27340.000000
75%          1.729282e+06             48035.000000   45624.000000
max          4.942365e+06            155254.000000  456191.000000

           ratings_2      ratings_3     ratings_4     ratings_5
count      99.000000      99.000000  9.900000e+01  9.900000e+01
mean    70538.828283  228209.626263  4.204259e+05  6.037495e+05
std     55694.613546  124117.176087  2.231151e+05  4.628464e+05
min     15781.000000   76071.000000  1.403040e+05  1.760720e+05
```

jupyter 230701091 Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

75%	86238.000000	277074.000000	5.184895e+05	7.283425e+05
max	436802.000000	793319.000000	1.481305e+06	3.011543e+06

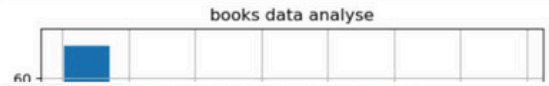
In [8]:
```
data.head(10)
```

Out[8]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...	ratings_
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	...	47
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	...	46
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	...	36
3	4	2657	2657	3275794	487	61120081	9.780061e+12	Harper Lee	1960.0	To Kill a Mockingbird	...	34
4	5	4671	4671	245494	1356	743273567	9.780743e+12	F. Scott Fitzgerald	1925.0	The Great Gatsby	...	26
5	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	2012.0	The Fault in Our Stars	...	23
6	7	5907	5907	1540236	969	618260307	9.780618e+12	J.R.R. Tolkien	1937.0	The Hobbit or There and Back Again	...	20
7	8	5107	5107	3036731	360	316769177	9.780317e+12	J.D. Salinger	1951.0	The Catcher in the Rye	...	20
8	9	960	960	3338963	311	1416524797	9.781417e+12	Dan Brown	2000.0	Angels & Demons	...	20
9	10	1885	1885	3060926	3455	679783261	9.780680e+12	Jane Austen	1813.0	Pride and Prejudice	...	20

10 rows × 23 columns

In [20]:
```
data['books_count'].hist(bins=10)
plt.title("books data analyse")
plt.show()
```



jupyter 230701091 Last Checkpoint: an hour ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

75%	86238.000000	277074.000000	5.184895e+05	7.283425e+05
max	436802.000000	793319.000000	1.481305e+06	3.011543e+06

In [41]:
```
data.head()
data.info()
```

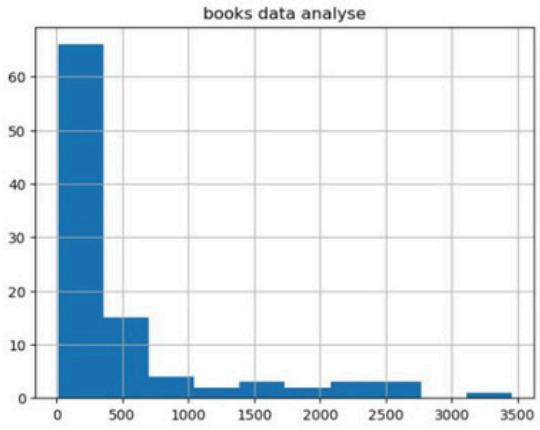
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 23 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   book_id                    99 non-null     int64
 1   goodreads_book_id          99 non-null     int64
 2   best_book_id               99 non-null     int64
 3   work_id                    99 non-null     int64
 4   books_count                99 non-null     int64
 5   isbn                       99 non-null     object
 6   isbn13                     99 non-null     float64
 7   authors                    99 non-null     object
 8   original_publication_year  99 non-null     float64
 9   original_title             98 non-null     object
 10  title                      99 non-null     object
 11  language_code              98 non-null     object
 12  average_rating             99 non-null     float64
 13  ratings_count              99 non-null     int64
 14  work_ratings_count         99 non-null     int64
 15  work_text_reviews_count    99 non-null     int64
 16  ratings_1                  99 non-null     int64
 17  ratings_2                  99 non-null     int64
 18  ratings_3                  99 non-null     int64
 19  ratings_4                  99 non-null     int64
 20  ratings_5                  99 non-null     int64
 21  image_url                  99 non-null     object
 22  small_image_url            99 non-null     object
dtypes: float64(3), int64(13), object(7)
memory usage: 17.9+ KB
```

jupyte 230701091 Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [20]:
```
data['books_count'].hist(bins=10)
plt.title("books data analyse")
plt.show()
```



jupyter 230701091 Last Checkpoint: an hour ago (autosaved)

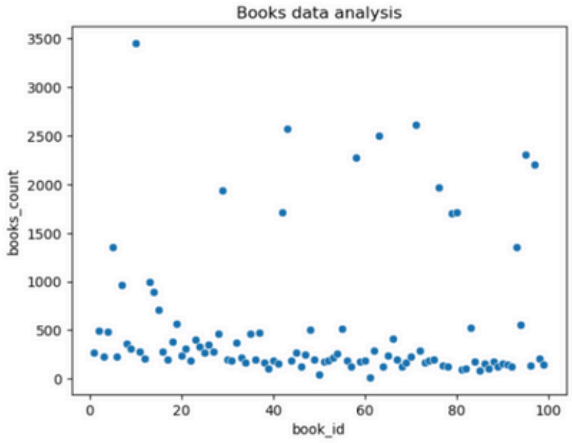
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

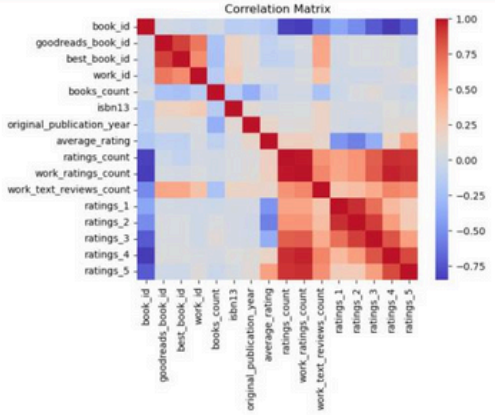
In [24]:
```
sns.scatterplot(x='book_id', y='books_count', data=data)
plt.title('Books data analysis')
plt.show()
```



localhost:8888/nbconvert/html/230701121%584%5D.ipynb?download=false

In [33]:
```
sns.heatmap(data.corr(), annot=False, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_1864\1628057676.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or spe
value of numeric_only to silence this warning.
  sns.heatmap(data.corr(), annot=False, cmap='coolwarm')
```



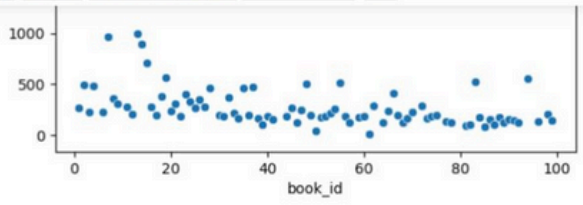
jupyter 230701091 Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

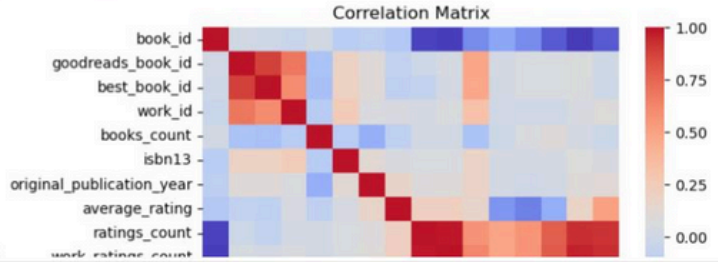
Trusted | Python 3 (ipykernel)

Run Code



In [33]:
```
sns.heatmap(data.corr(), annot=False, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_1864\1628057676.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  sns.heatmap(data.corr(), annot=False, cmap='coolwarm')
```



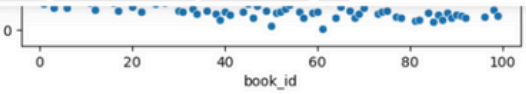
jupyter 230701091 Last Checkpoint: an hour ago (autosaved)

Logout

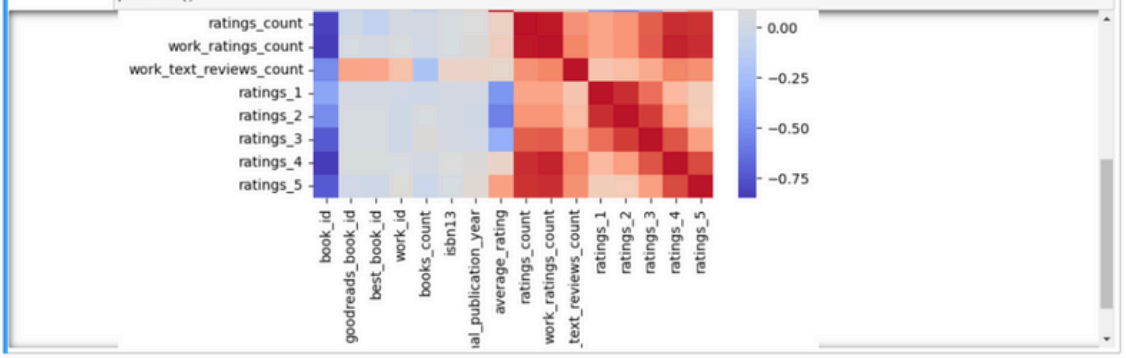
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code



In [33]:
```
sns.heatmap(data.corr(), annot=False, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



jupyter 230701091 Last Checkpoint: an hour ago (autosaved)

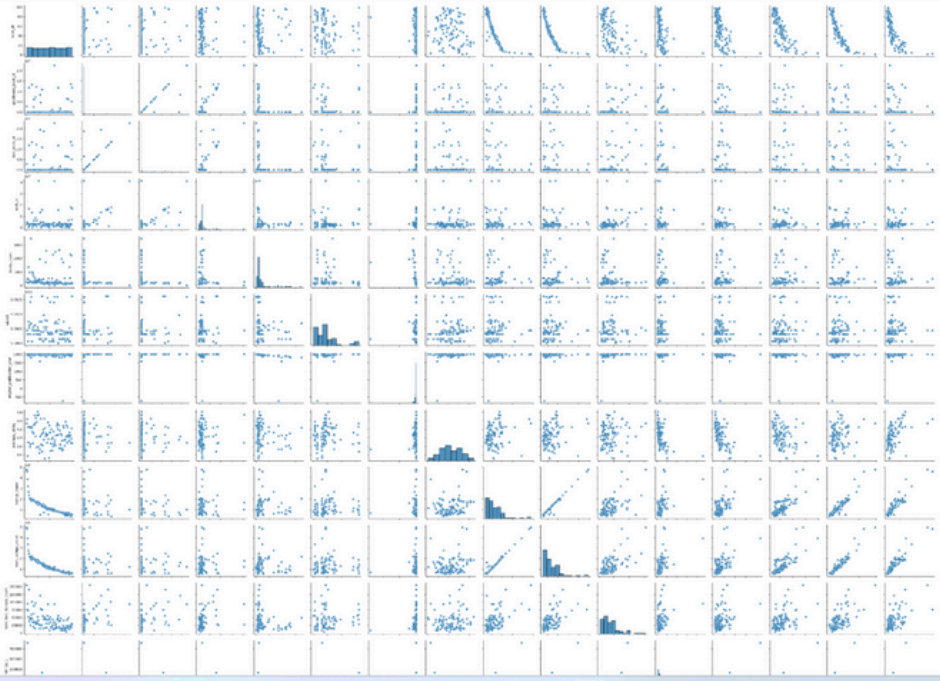
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [39]:
```
sns.pairplot(data)
plt.show()
```



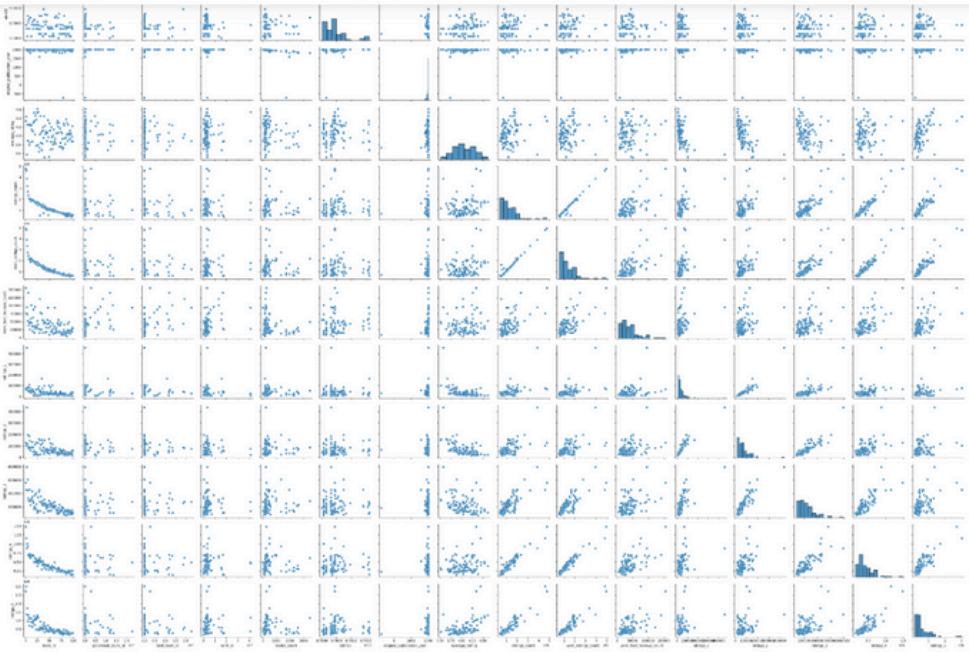
jupyter 230701091 Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code



jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved)

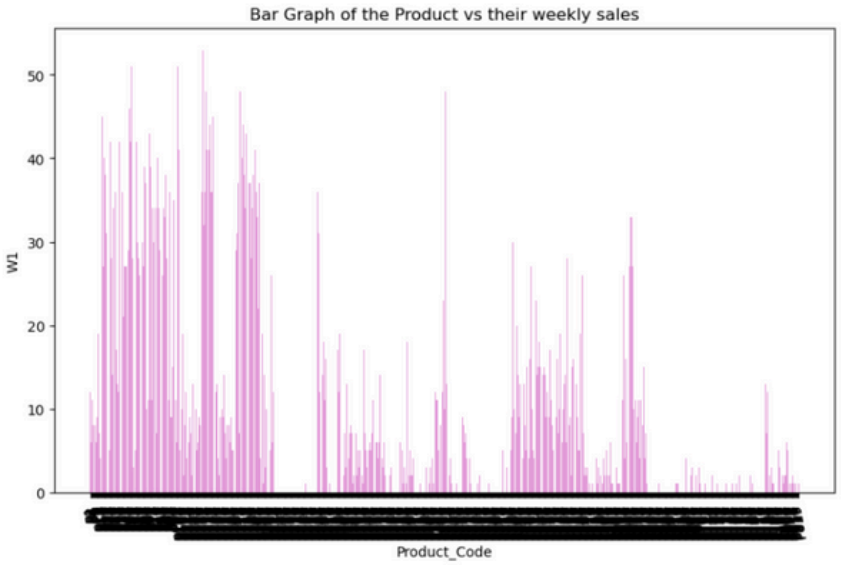
Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Markdown

In [34]:
```
plt.figure(figsize=(10, 6))
plt.bar(df['Product_Code'],df['W1'],color='plum')
plt.xlabel('Product_Code')
plt.ylabel('W1')
plt.title('Bar Graph of the Product vs their weekly sales')
plt.xticks(rotation=100)
plt.show()
```



jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Markdown

```
structured__data=pd.Dataframed({ 'ID':[1,2,3], 'Name':['Ram','John','Geeta'], 'Age':[25,30,35] })
```

In [5]:
```
import pandas as pd
structured_data=pd.DataFrame({ 'ID':[1,2,3], 'Name':['Ram','John','Geeta'], 'Age':[25,30,35] })
print("Structured Data: \n",structured_data)
```

```
Structured Data:
    ID   Name  Age
0   1    Ram   25
1   2   John   30
2   3  Geeta   35
```

In [7]:
```
import pandas as pd
unstructured_data="This is an example of unstructured data"
print("unstructured data: \n", unstructured_data)
```

```
unstructured data:
 This is an example of unstructured data
```

In [8]:
```
import pandas as pd
semistructured_data={ 'ID':[1,2,3], 'Name': ['Ram','John','Geeta'],'Age':[25,30,35]}
print("semistructured data: \n", semistructured_data)
```

```
semistructured data:
 {'ID': [1, 2, 3], 'Name': ['Ram', 'John', 'Geeta'], 'Age': [25, 30, 35]}
```

jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved)

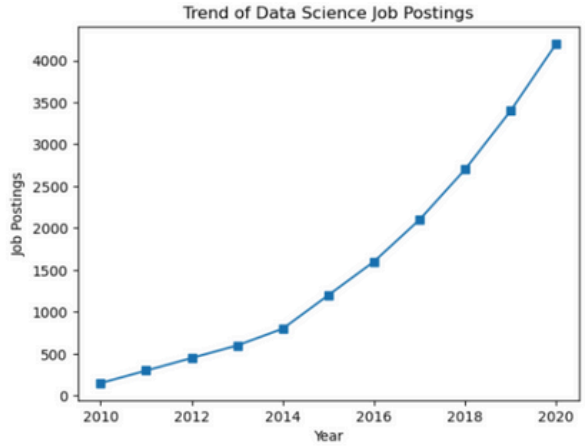
Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Markdown

In [22]:
```
import pandas as pd
import matplotlib.pyplot as plt
data = {'Year':list(range(2010,2021)),'Job Postings':[150,300,450,600,800,1200,1600,2100,2700,3400,4200]}
df=pd.DataFrame(data)
plt.plot(df['Year'],df['Job Postings'], marker='s')
plt.title('Trend of Data Science Job Postings')
plt.xlabel('Year')
plt.ylabel('Job Postings')
plt.show()
```



jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved)

Logout

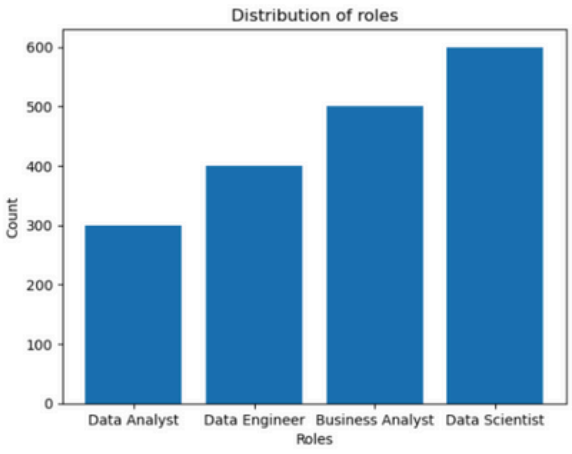
File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Markdown

Year

In [20]:
```
import pandas as pd
import matplotlib.pyplot as plt
roles=['Data Analyst','Data Engineer','Business Analyst','Data Scientist']
count=[300,400,500,600]
plt.bar(roles,count)
plt.title('Distribution of roles')
plt.xlabel('Roles')
plt.ylabel('Count')
plt.show()
```



jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved)

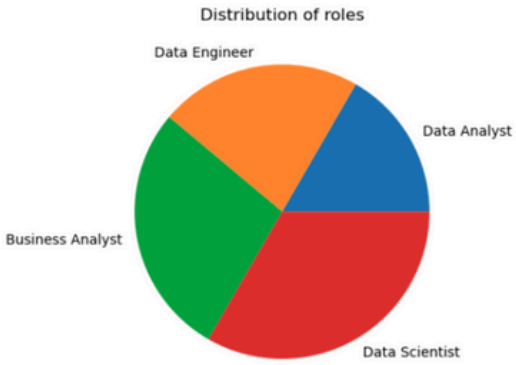
Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Markdown

In [21]:
```
import pandas as pd
import matplotlib.pyplot
roles=['Data Analyst','Data Engineer','Business Analyst','Data Scientist']
count=[300,400,500,600]
plt.pie(count,labels=roles)
plt.title('Distribution of roles')
plt.show()
```



jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Markdown

In [6]:
```
import pandas as pd
import matplotlib.pyplot as plt
pf=pd.read_csv("Downloads/Sales_Transactions_Dataset_Weekly.csv")
pf.head()
```

Out[6]:

	Product_Code	W0	W1	W2	W3	W4	W5	W6	W7	W8	...	Normalized 42	Normalized 43	Normalized 44	Normalized 45	Normalized 46	Normalized 47	Normalized 48	Normal
0	P1	11	12	10	8	13	12	14	21	6	...	0.06	0.22	0.28	0.39	0.50	0.00	0.22	
1	P2	7	6	3	2	7	1	6	3	3	...	0.20	0.40	0.50	0.10	0.10	0.40	0.50	
2	P3	7	11	8	9	10	8	7	13	12	...	0.27	1.00	0.18	0.18	0.36	0.45	1.00	
3	P4	12	8	13	5	9	6	9	13	13	...	0.41	0.47	0.06	0.12	0.24	0.35	0.71	
4	P5	8	5	13	11	6	7	9	14	9	...	0.27	0.53	0.27	0.60	0.20	0.20	0.13	

5 rows × 107 columns

jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Markdown

In [16]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot
#Load the data into a pandas Dataframe
file_path='Downloads/Sales_Transactions_Dataset_Weekly.csv'
df=pd.read_csv(file_path)
#Display the first few rows of the dataframe
print(df.head())
#check for the missing values
print(df.isnull().sum())
#fill or drop missing values if necessary
df['Sales'].fillna(df['Sales'].mean(),inplace=True)
df.dropna(subset=['Product', 'Quantity', 'Region'], inplace=True)
#summary statistics
print(df.describe())
product_summary=df.groupby['Product'].agg({
    'Sales':'sum',
    'Quantity':'sum'
}).reset_index()
print(product_summary)
plt.figure(figsize=(10,6))
plt.bar(product_summary['Product'], product_summary['Sales'])
plt.xlabel('Product')
plt.ylabel('Sales')
plt.show()
```

```
  Product_Code  W0  W1  W2  W3  W4  W5  W6  W7  W8  ...  Normalized 42  \
0           P1  11  12  10   8  13  12  14  21   6  ...           0.06
1           P2   7   6   3   2   7   1   6   3   3  ...           0.20
2           P3   7  11   8   9  10   8   7  13  12  ...           0.27
3           P4  12   8  13   5   9   6   9  13  13  ...           0.41
4           P5   8   5  13  11   6   7   9  14   9  ...           0.27

   Normalized 43  Normalized 44  Normalized 45  Normalized 46  Normalized 47  \
0           0.22           0.28           0.39           0.50           0.00
1           0.40           0.50           0.10           0.10           0.40
2           1.00           0.18           0.18           0.36           0.45
3           0.47           0.06           0.12           0.24           0.35
4           0.53           0.27           0.60           0.20           0.20

   Normalized 48  Normalized 49  Normalized 50  Normalized 51
0           0.22           0.17           0.11           0.39
1           0.50           0.10           0.60           0.00
2           1.00           0.45           0.45           0.36
3           0.71           0.35           0.29           0.35
```

jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Markdown

```
1           P2   7   6   3   2   7   1   6   3   3  ...           0.20
2           P3   7  11   8   9  10   8   7  13  12  ...           0.27
3           P4  12   8  13   5   9   6   9  13  13  ...           0.41
4           P5   8   5  13  11   6   7   9  14   9  ...           0.27

   Normalized 43  Normalized 44  Normalized 45  Normalized 46  Normalized 47  \
0           0.22           0.28           0.39           0.50           0.00
1           0.40           0.50           0.10           0.10           0.40
2           1.00           0.18           0.18           0.36           0.45
3           0.47           0.06           0.12           0.24           0.35
4           0.53           0.27           0.60           0.20           0.20

   Normalized 48  Normalized 49  Normalized 50  Normalized 51
0           0.22           0.17           0.11           0.39
1           0.50           0.10           0.60           0.00
2           1.00           0.45           0.45           0.36
3           0.71           0.35           0.29           0.35
4           0.13           0.53           0.33           0.40

[5 rows x 107 columns]
Product_Code     0
W0               0
W1               0
W2               0
W3               0
                ..
Normalized 47    0
Normalized 48    0
Normalized 49    0
Normalized 50    0
Normalized 51    0
Length: 107, dtype: int64
```

jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Markdown

In [13]:
```
import numpy as np
import matplotlib.pyplot
file_path='Downloads/Sales_Transactions_Dataset_Weekly.csv'
df.isnull().sum()
```

Out[13]:
```
Product_Code     0
W0               0
W1               0
W2               0
W3               0
                ..
Normalized 47    0
Normalized 48    0
Normalized 49    0
Normalized 50    0
Normalized 51    0
Length: 107, dtype: int64
```

jupyter 230701091 Last Checkpoint: 08/22/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Markdown

```
<Figure size 1000x600 with 0 Axes>
```

In [20]:
```
import pandas as pd
import matplotlib.pyplot as plt
pf=pd.read_csv("Downloads/Sales_Transactions_Dataset_Weekly.csv")
pf.head()
```

Out[20]:

	Product_Code	W0	W1	W2	W3	W4	W5	W6	W7	W8	...	Normalized 42	Normalized 43	Normalized 44	Normalized 45	Normalized 46	Normalized 47	Normalized 48	Normal
0	P1	11	12	10	8	13	12	14	21	6	...	0.06	0.22	0.28	0.39	0.50	0.00	0.22	
1	P2	7	6	3	2	7	1	6	3	3	...	0.20	0.40	0.50	0.10	0.10	0.40	0.50	
2	P3	7	11	8	9	10	8	7	13	12	...	0.27	1.00	0.18	0.18	0.36	0.45	1.00	
3	P4	12	8	13	5	9	6	9	13	13	...	0.41	0.47	0.06	0.12	0.24	0.35	0.71	
4	P5	8	5	13	11	6	7	9	14	9	...	0.27	0.53	0.27	0.60	0.20	0.20	0.13	

5 rows × 107 columns