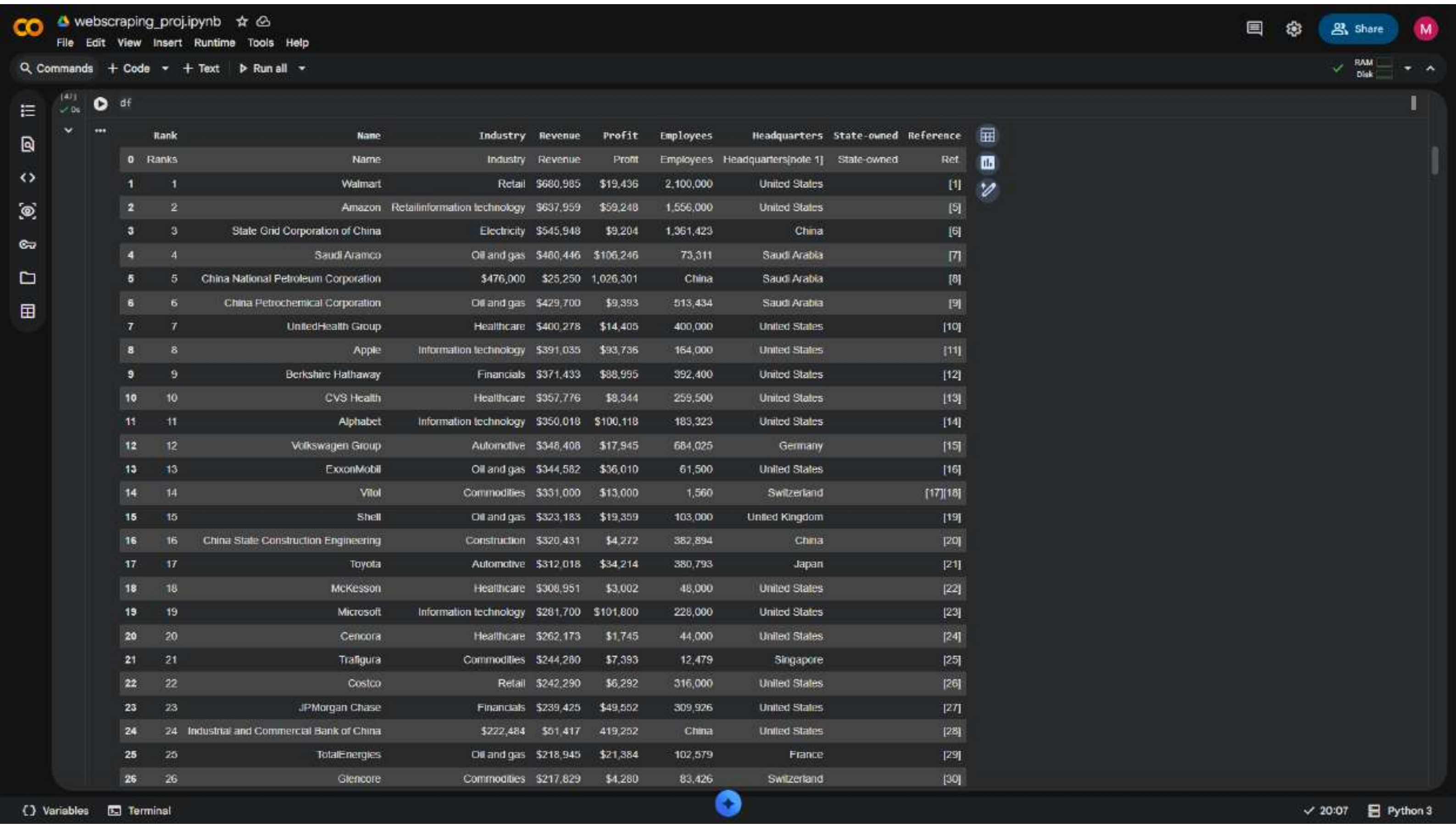
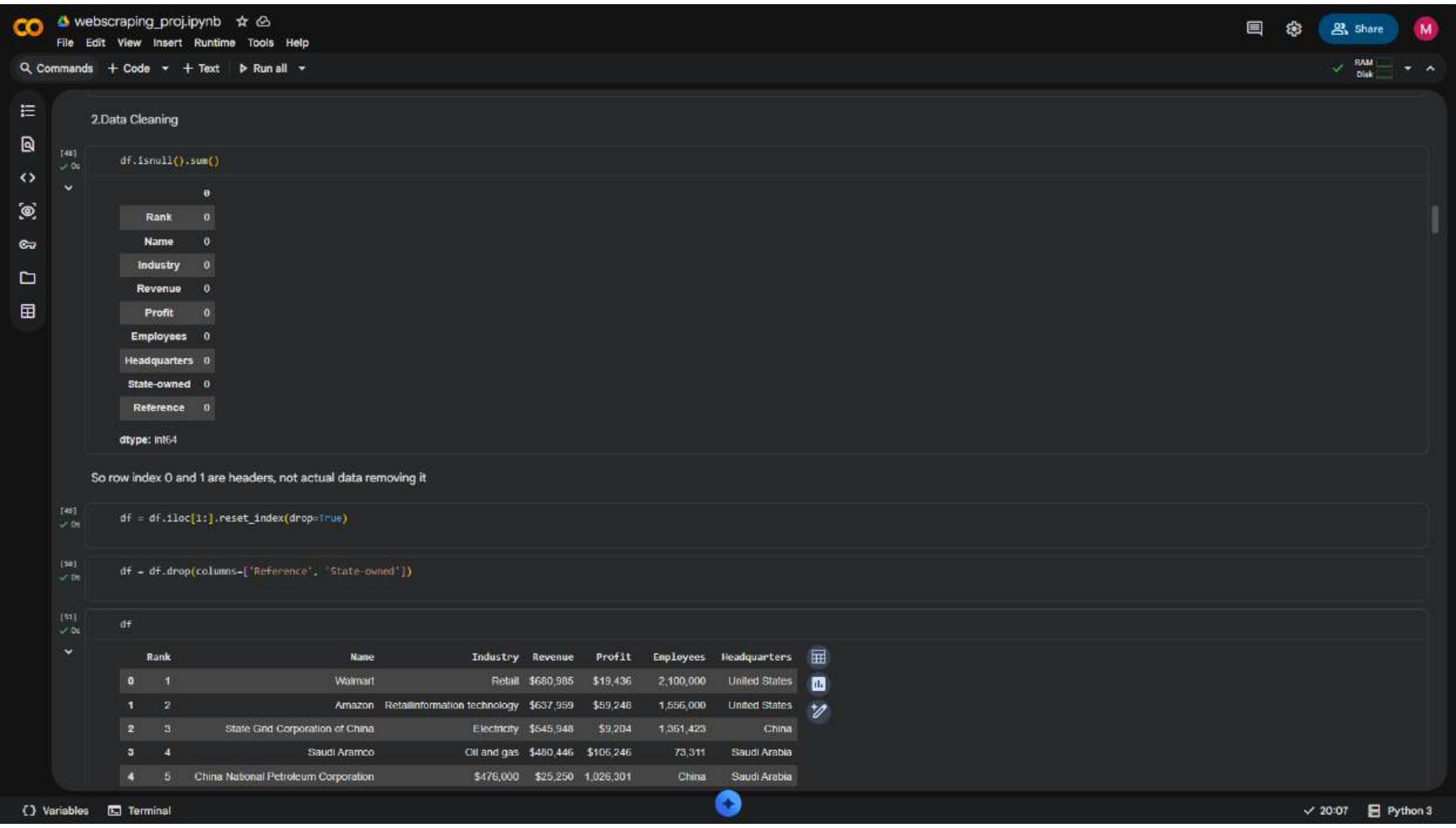


```
df
```

	Rank	Name	Industry	Revenue	Profit	Employees	Headquarters	State-owned	Reference
0	Ranks	Name	Industry	Revenue	Profit	Employees	Headquarters[note 1]	State-owned	Ref.
1	1	Walmart	Retail	$680,985	$19,436	2,100,000	United States		[1]
2	2	Amazon	Retailinformation technology	$637,959	$59,248	1,556,000	United States		[5]
3	3	State Grid Corporation of China	Electricity	$545,948	$9,204	1,361,423	China		[6]
4	4	Saudi Aramco	Oil and gas	$480,446	$106,246	73,311	Saudi Arabia		[7]
5	5	China National Petroleum Corporation	$476,000	$25,250	1,026,301	China	Saudi Arabia		[8]
6	6	China Petrochemical Corporation	Oil and gas	$429,700	$9,393	513,434	Saudi Arabia		[9]
7	7	UnitedHealth Group	Healthcare	$400,278	$14,405	400,000	United States		[10]
8	8	Apple	Information technology	$391,035	$93,736	164,000	United States		[11]
9	9	Berkshire Hathaway	Financials	$371,433	$88,995	392,400	United States		[12]
10	10	CVS Health	Healthcare	$357,776	$8,344	259,500	United States		[13]
11	11	Alphabet	Information technology	$350,018	$100,118	183,323	United States		[14]
12	12	Volkswagen Group	Automotive	$348,408	$17,945	684,025	Germany		[15]
13	13	ExxonMobil	Oil and gas	$344,582	$36,010	61,500	United States		[16]
14	14	Vitol	Commodities	$331,000	$13,000	1,560	Switzerland		[17][18]
15	15	Shell	Oil and gas	$323,183	$19,359	103,000	United Kingdom		[19]
16	16	China State Construction Engineering	Construction	$320,431	$4,272	382,894	China		[20]
17	17	Toyota	Automotive	$312,018	$34,214	380,793	Japan		[21]
18	18	McKesson	Healthcare	$308,951	$3,002	48,000	United States		[22]
19	19	Microsoft	Information technology	$281,700	$101,800	228,000	United States		[23]
20	20	Cencora	Healthcare	$262,173	$1,745	44,000	United States		[24]
21	21	Trafigura	Commodities	$244,280	$7,393	12,479	Singapore		[25]
22	22	Costco	Retail	$242,290	$6,292	316,000	United States		[26]
23	23	JPMorgan Chase	Financials	$239,425	$49,552	309,926	United States		[27]
24	24	Industrial and Commercial Bank of China	$222,484	$51,417	419,252	China	United States		[28]
25	25	TotalEnergies	Oil and gas	$218,945	$21,384	102,579	France		[29]
26	26	Glencore	Commodities	$217,829	$4,280	83,426	Switzerland		[30]



2.Data Cleaning

[48]

```
df.isnull().sum()
```

```
              0
Rank          0
Name          0
Industry      0
Revenue       0
Profit        0
Employees     0
Headquarters  0
State-owned   0
Reference     0

dtype: int64
```

So row index 0 and 1 are headers, not actual data removing it

[49]

```
df = df.iloc[1:].reset_index(drop=True)
```

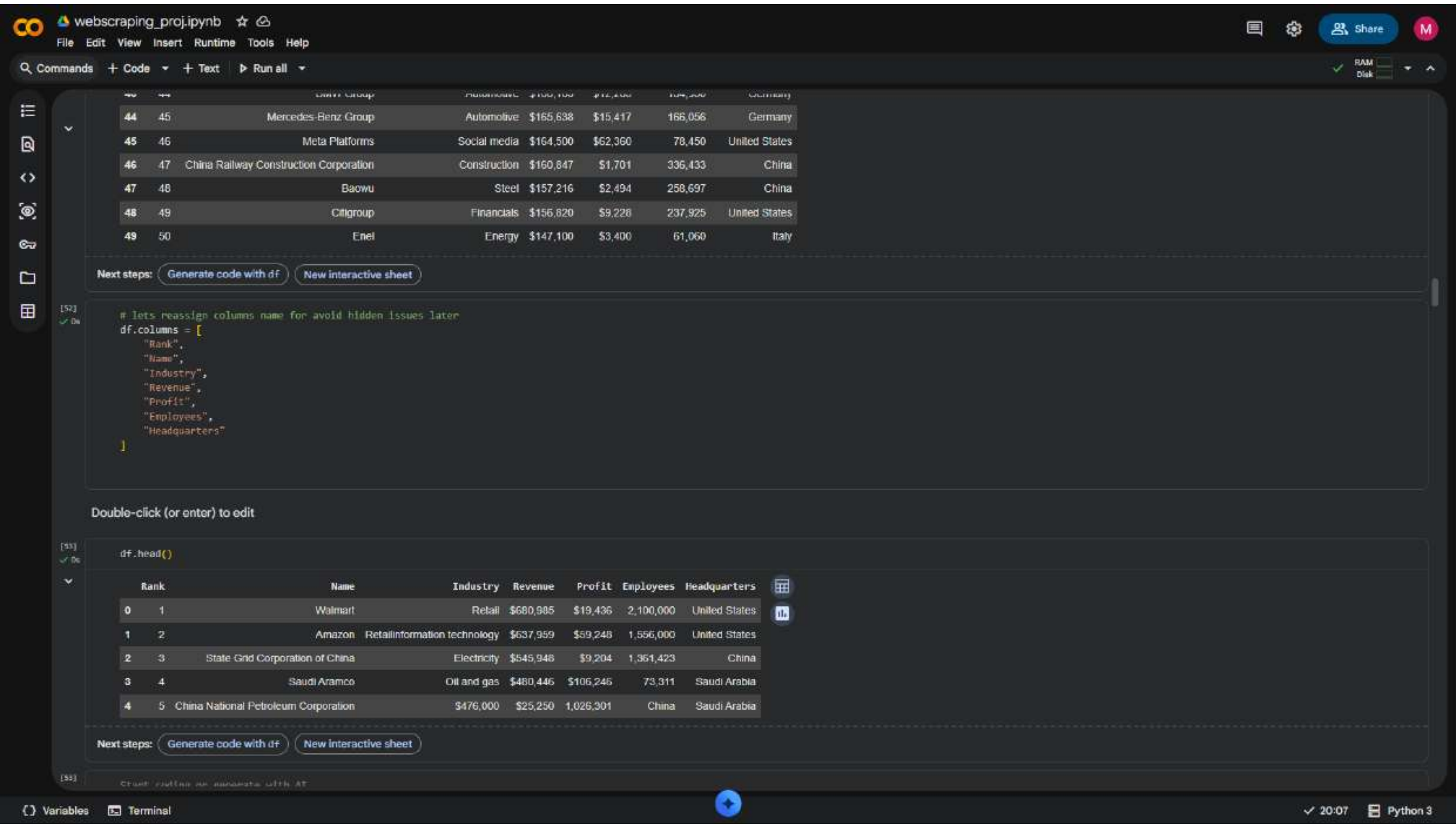
[50]

```
df = df.drop(columns=['Reference', 'State-owned'])
```

[51]

```
df
```

```
   Rank                                Name                       Industry   Revenue    Profit  Employees   Headquarters
0     1                             Walmart                         Retail  $680,985   $19,436  2,100,000  United States
1     2                              Amazon  Retailinformation technology  $637,959   $59,248  1,556,000  United States
2     3      State Grid Corporation of China                    Electricity  $545,948    $9,204  1,361,423          China
3     4                        Saudi Aramco                    Oil and gas  $480,446  $106,246     73,311   Saudi Arabia
4     5  China National Petroleum Corporation                   $476,000   $25,250  1,026,301      China   Saudi Arabia
```



43	44	BMW Group	Automotive	[illegible]	[illegible]	[illegible]	Germany
44	45	Mercedes-Benz Group	Automotive	$165,638	$15,417	166,056	Germany
45	46	Meta Platforms	Social media	$164,500	$62,360	78,450	United States
46	47	China Railway Construction Corporation	Construction	$160,847	$1,701	336,433	China
47	48	Baowu	Steel	$157,216	$2,494	258,697	China
48	49	Citigroup	Financials	$156,820	$9,228	237,925	United States
49	50	Enel	Energy	$147,100	$3,400	61,060	Italy

Next steps: Generate code with df New interactive sheet

[52]

```
# lets reassign columns name for avoid hidden issues later
df.columns = [
    "Rank",
    "Name",
    "Industry",
    "Revenue",
    "Profit",
    "Employees",
    "Headquarters"
]
```

Double-click (or enter) to edit

[53]

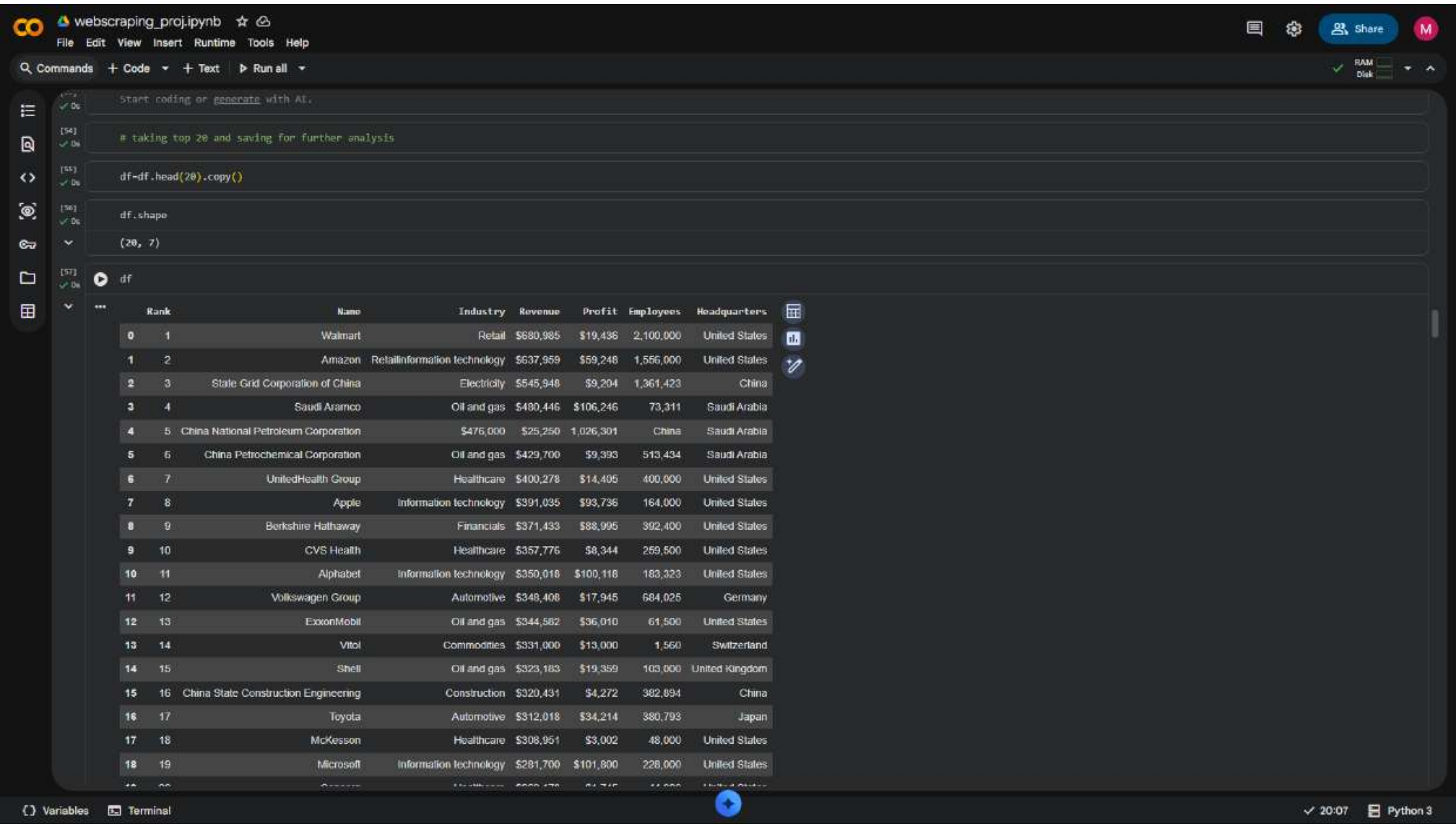
```
df.head()
```

	Rank	Name	Industry	Revenue	Profit	Employees	Headquarters
0	1	Walmart	Retail	$680,985	$19,436	2,100,000	United States
1	2	Amazon	Retailinformation technology	$637,959	$59,248	1,556,000	United States
2	3	State Grid Corporation of China	Electricity	$545,948	$9,204	1,361,423	China
3	4	Saudi Aramco	Oil and gas	$480,446	$106,246	73,311	Saudi Arabia
4	5	China National Petroleum Corporation	$476,000	$25,250	1,026,301	China	Saudi Arabia

Next steps: Generate code with df New interactive sheet

[53]

```
Start coding or generate with AI
```


webscraping_proj.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
Share
RAM
Disk
Variables
Terminal
20:07
Python 3

```
Start coding or generate with AI.
```

[54]

```
# taking top 20 and saving for further analysis
```

[55]

```
df=df.head(20).copy()
```

[56]

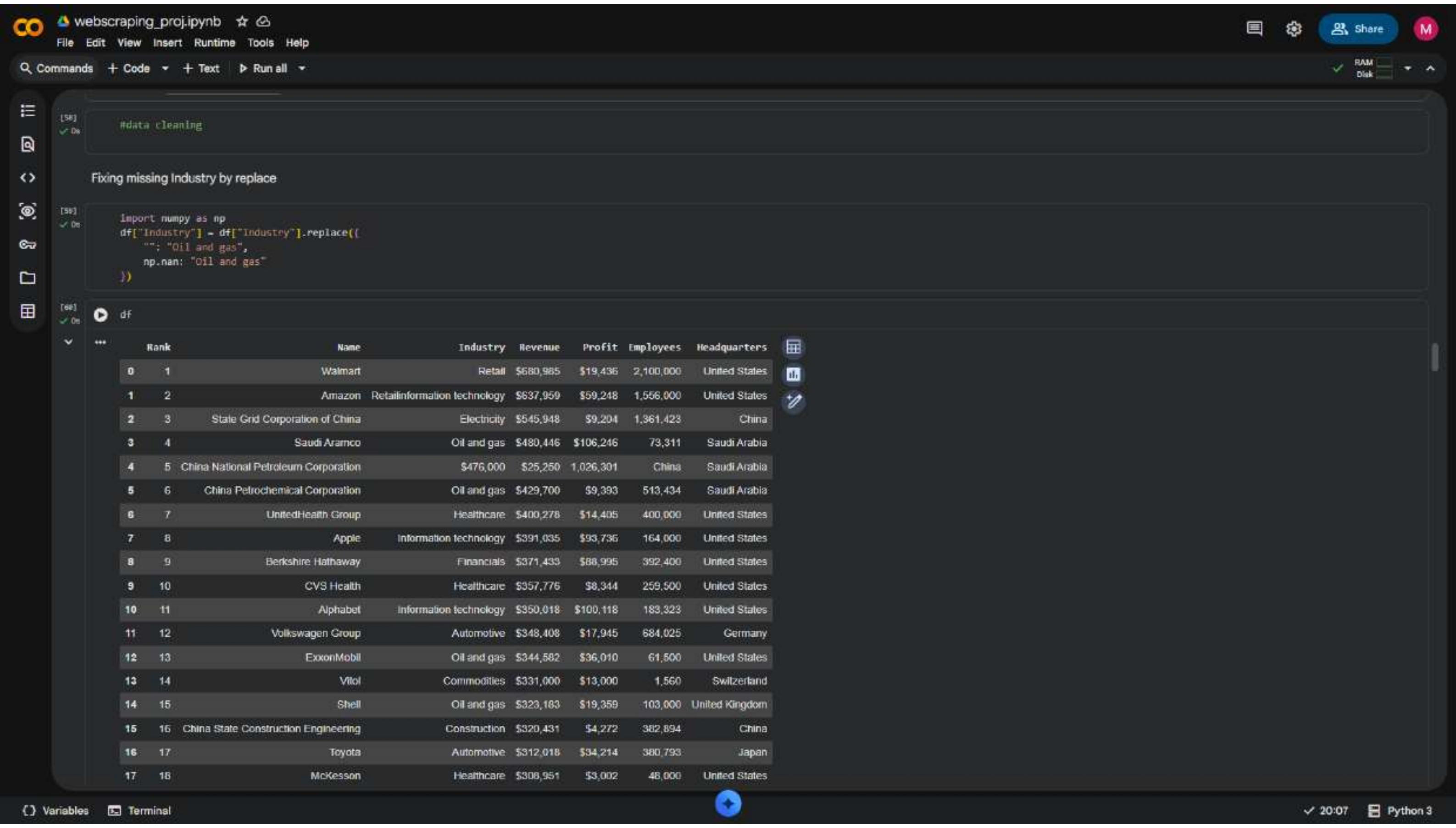
```
df.shape
```

```
(20, 7)
```

[57]

```
df
```

	Rank	Name	Industry	Revenue	Profit	Employees	Headquarters
0	1	Walmart	Retail	$680,985	$19,436	2,100,000	United States
1	2	Amazon	Retailinformation technology	$637,959	$59,248	1,556,000	United States
2	3	State Grid Corporation of China	Electricity	$545,948	$9,204	1,361,423	China
3	4	Saudi Aramco	Oil and gas	$480,446	$106,246	73,311	Saudi Arabia
4	5	China National Petroleum Corporation	$476,000	$25,250	1,026,301	China	Saudi Arabia
5	6	China Petrochemical Corporation	Oil and gas	$429,700	$9,393	513,434	Saudi Arabia
6	7	UnitedHealth Group	Healthcare	$400,278	$14,405	400,000	United States
7	8	Apple	Information technology	$391,035	$93,736	164,000	United States
8	9	Berkshire Hathaway	Financials	$371,433	$88,995	392,400	United States
9	10	CVS Health	Healthcare	$357,776	$8,344	259,500	United States
10	11	Alphabet	Information technology	$350,018	$100,118	183,323	United States
11	12	Volkswagen Group	Automotive	$348,408	$17,945	684,025	Germany
12	13	ExxonMobil	Oil and gas	$344,582	$36,010	61,500	United States
13	14	Vitol	Commodities	$331,000	$13,000	1,560	Switzerland
14	15	Shell	Oil and gas	$323,183	$19,359	103,000	United Kingdom
15	16	China State Construction Engineering	Construction	$320,431	$4,272	382,894	China
16	17	Toyota	Automotive	$312,018	$34,214	380,793	Japan
17	18	McKesson	Healthcare	$308,951	$3,002	48,000	United States
18	19	Microsoft	Information technology	$281,700	$101,800	228,000	United States



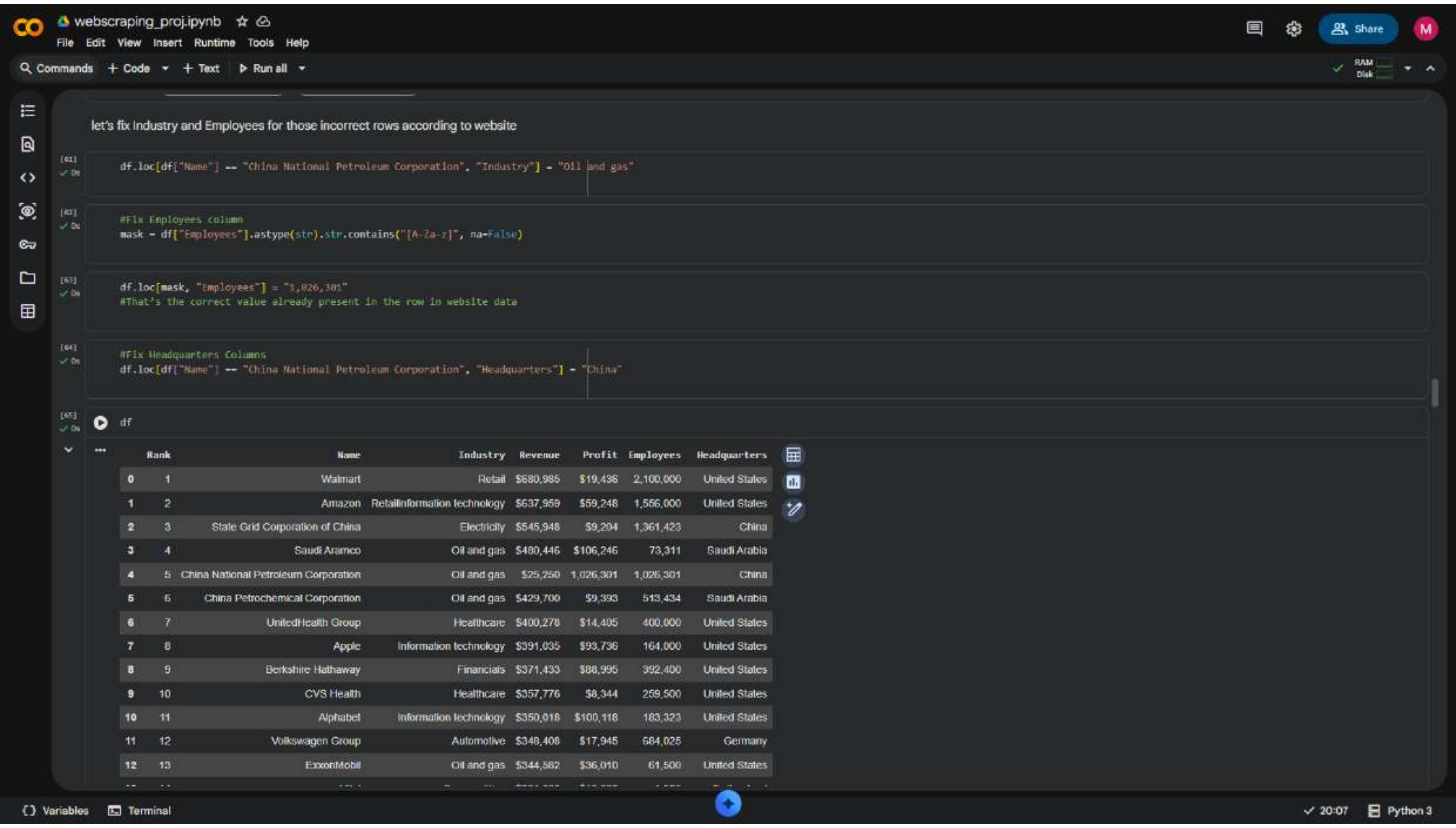
```
#data cleaning
```

Fixing missing Industry by replace

```
import numpy as np
df["Industry"] = df["Industry"].replace({
    "": "Oil and gas",
    np.nan: "Oil and gas"
})
```

```
df
```

	Rank	Name	Industry	Revenue	Profit	Employees	Headquarters
0	1	Walmart	Retail	$680,985	$19,436	2,100,000	United States
1	2	Amazon	Retailinformation technology	$637,959	$59,248	1,556,000	United States
2	3	State Grid Corporation of China	Electricity	$545,948	$9,204	1,361,423	China
3	4	Saudi Aramco	Oil and gas	$480,446	$106,246	73,311	Saudi Arabia
4	5	China National Petroleum Corporation	$476,000	$25,250	1,026,301	China	Saudi Arabia
5	6	China Petrochemical Corporation	Oil and gas	$429,700	$9,393	513,434	Saudi Arabia
6	7	UnitedHealth Group	Healthcare	$400,278	$14,405	400,000	United States
7	8	Apple	Information technology	$391,035	$93,736	164,000	United States
8	9	Berkshire Hathaway	Financials	$371,433	$88,995	392,400	United States
9	10	CVS Health	Healthcare	$357,776	$8,344	259,500	United States
10	11	Alphabet	Information technology	$350,018	$100,118	183,323	United States
11	12	Volkswagen Group	Automotive	$348,408	$17,945	684,025	Germany
12	13	ExxonMobil	Oil and gas	$344,582	$36,010	61,500	United States
13	14	Vitol	Commodities	$331,000	$13,000	1,560	Switzerland
14	15	Shell	Oil and gas	$323,183	$19,359	103,000	United Kingdom
15	16	China State Construction Engineering	Construction	$320,431	$4,272	382,894	China
16	17	Toyota	Automotive	$312,018	$34,214	380,793	Japan
17	18	McKesson	Healthcare	$308,951	$3,002	48,000	United States



let's fix Industry and Employees for those incorrect rows according to website

[61]
```
df.loc[df["Name"] == "China National Petroleum Corporation", "Industry"] = "Oil and gas"
```

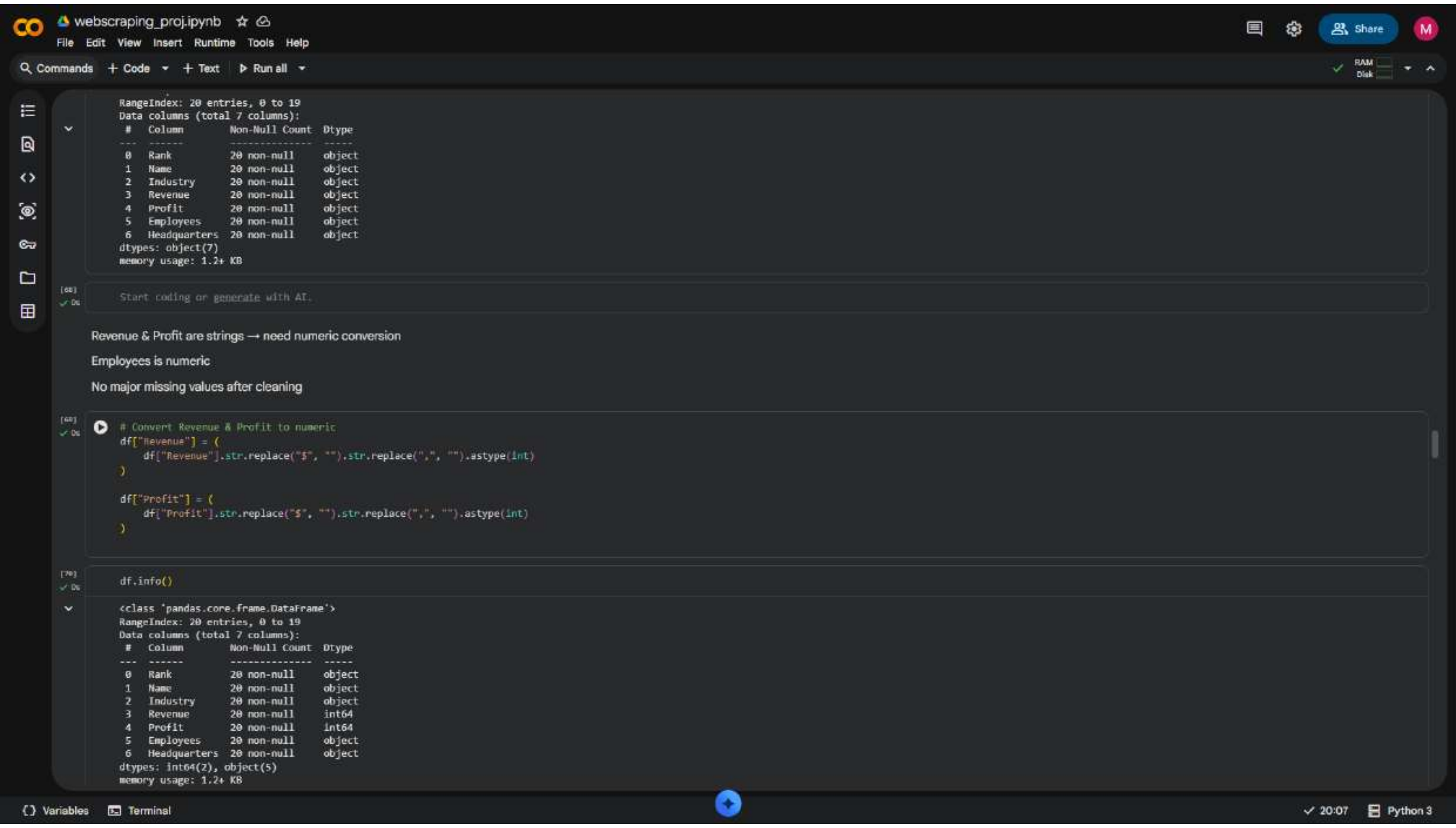
[62]
```
#Fix Employees column
mask = df["Employees"].astype(str).str.contains("[A-Za-z]", na=False)
```

[63]
```
df.loc[mask, "Employees"] = "1,026,301"
#That's the correct value already present in the row in website data
```

[64]
```
#Fix Headquarters Columns
df.loc[df["Name"] == "China National Petroleum Corporation", "Headquarters"] = "China"
```

[65]
```
df
```

	Rank	Name	Industry	Revenue	Profit	Employees	Headquarters
0	1	Walmart	Retail	$680,985	$19,436	2,100,000	United States
1	2	Amazon	Retailinformation technology	$637,959	$59,248	1,556,000	United States
2	3	State Grid Corporation of China	Electricity	$545,948	$9,204	1,361,423	China
3	4	Saudi Aramco	Oil and gas	$480,446	$106,246	73,311	Saudi Arabia
4	5	China National Petroleum Corporation	Oil and gas	$25,250	1,026,301	1,026,301	China
5	6	China Petrochemical Corporation	Oil and gas	$429,700	$9,393	513,434	Saudi Arabia
6	7	UnitedHealth Group	Healthcare	$400,278	$14,405	400,000	United States
7	8	Apple	Information technology	$391,035	$93,736	164,000	United States
8	9	Berkshire Hathaway	Financials	$371,433	$88,995	392,400	United States
9	10	CVS Health	Healthcare	$357,776	$8,344	259,500	United States
10	11	Alphabet	Information technology	$350,018	$100,118	183,323	United States
11	12	Volkswagen Group	Automotive	$348,408	$17,945	684,025	Germany
12	13	ExxonMobil	Oil and gas	$344,582	$36,010	61,500	United States


webscraping_proj.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
Share
RAM Disk

```
RangeIndex: 20 entries, 0 to 19
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Rank          20 non-null     object
 1   Name          20 non-null     object
 2   Industry      20 non-null     object
 3   Revenue       20 non-null     object
 4   Profit        20 non-null     object
 5   Employees     20 non-null     object
 6   Headquarters  20 non-null     object
dtypes: object(7)
memory usage: 1.2+ KB
```

[68]

Start coding or generate with AI.

Revenue & Profit are strings → need numeric conversion

Employees is numeric

No major missing values after cleaning

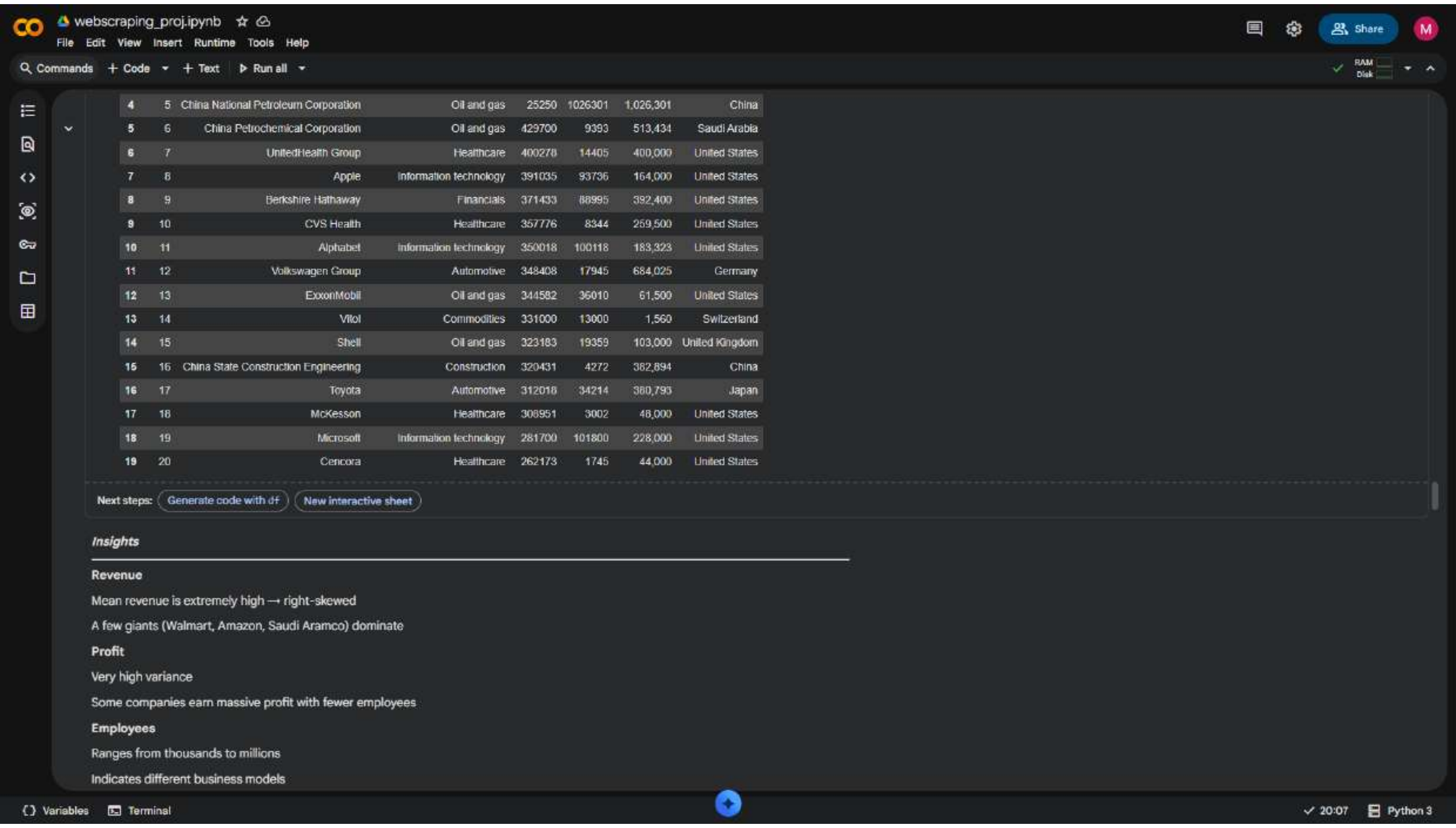
[69]

```
# Convert Revenue & Profit to numeric
df["Revenue"] = (
    df["Revenue"].str.replace("$", "").str.replace(",", "").astype(int)
)

df["Profit"] = (
    df["Profit"].str.replace("$", "").str.replace(",", "").astype(int)
)
```

[70]

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Rank          20 non-null     object
 1   Name          20 non-null     object
 2   Industry      20 non-null     object
 3   Revenue       20 non-null     int64 
 4   Profit        20 non-null     int64 
 5   Employees     20 non-null     object
 6   Headquarters  20 non-null     object
dtypes: int64(2), object(5)
memory usage: 1.2+ KB
```


webscraping_proj.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text ▷ Run all
Share
M
RAM
Disk

4	5	China National Petroleum Corporation	Oil and gas	25250	1026301	1,026,301	China
5	6	China Petrochemical Corporation	Oil and gas	429700	9393	513,434	Saudi Arabia
6	7	UnitedHealth Group	Healthcare	400278	14405	400,000	United States
7	8	Apple	Information technology	391035	93736	164,000	United States
8	9	Berkshire Hathaway	Financials	371433	88995	392,400	United States
9	10	CVS Health	Healthcare	357776	8344	259,500	United States
10	11	Alphabet	Information technology	350018	100118	183,323	United States
11	12	Volkswagen Group	Automotive	348408	17945	684,025	Germany
12	13	ExxonMobil	Oil and gas	344582	36010	61,500	United States
13	14	Vitol	Commodities	331000	13000	1,560	Switzerland
14	15	Shell	Oil and gas	323183	19359	103,000	United Kingdom
15	16	China State Construction Engineering	Construction	320431	4272	382,894	China
16	17	Toyota	Automotive	312018	34214	380,793	Japan
17	18	McKesson	Healthcare	308951	3002	48,000	United States
18	19	Microsoft	Information technology	281700	101800	228,000	United States
19	20	Cencora	Healthcare	262173	1745	44,000	United States

Insights

Revenue

Mean revenue is extremely high → right-skewed

A few giants (Walmart, Amazon, Saudi Aramco) dominate

Profit

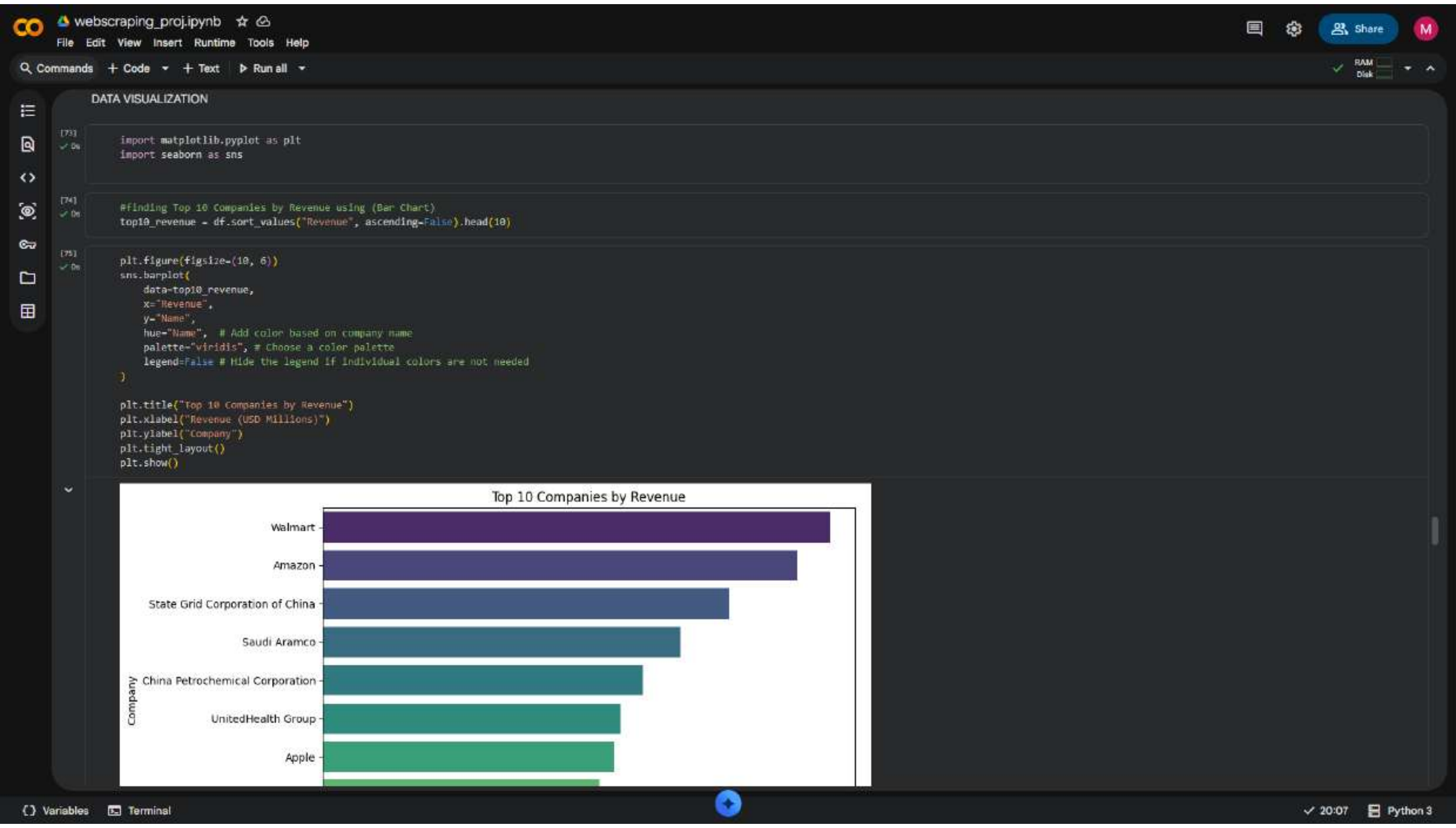
Very high variance

Some companies earn massive profit with fewer employees

Employees

Ranges from thousands to millions

Indicates different business models



DATA VISUALIZATION

[73]

```
import matplotlib.pyplot as plt
import seaborn as sns
```

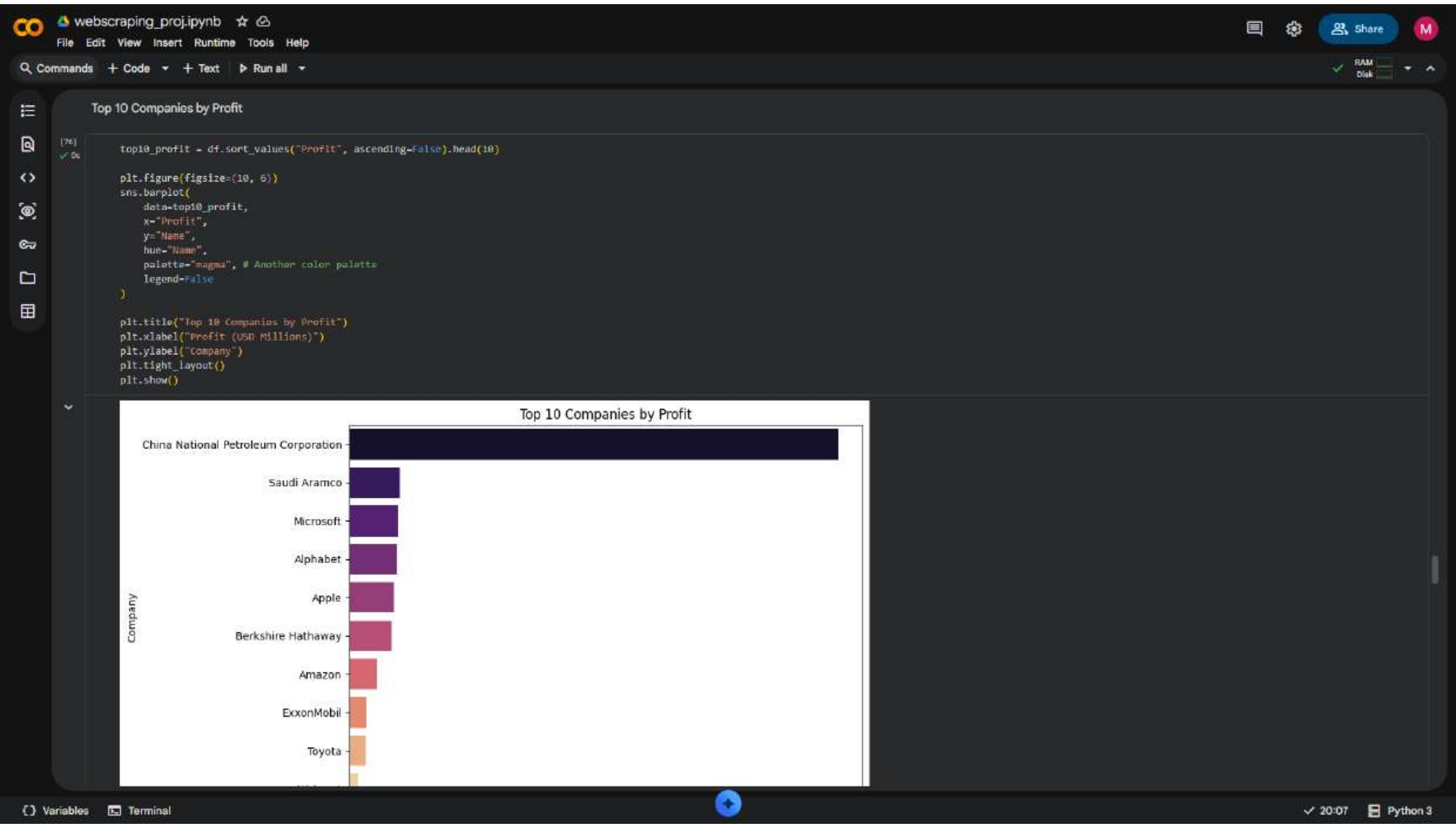
[74]

```
#Finding Top 10 Companies by Revenue using (Bar Chart)
top10_revenue = df.sort_values("Revenue", ascending=False).head(10)
```

[75]

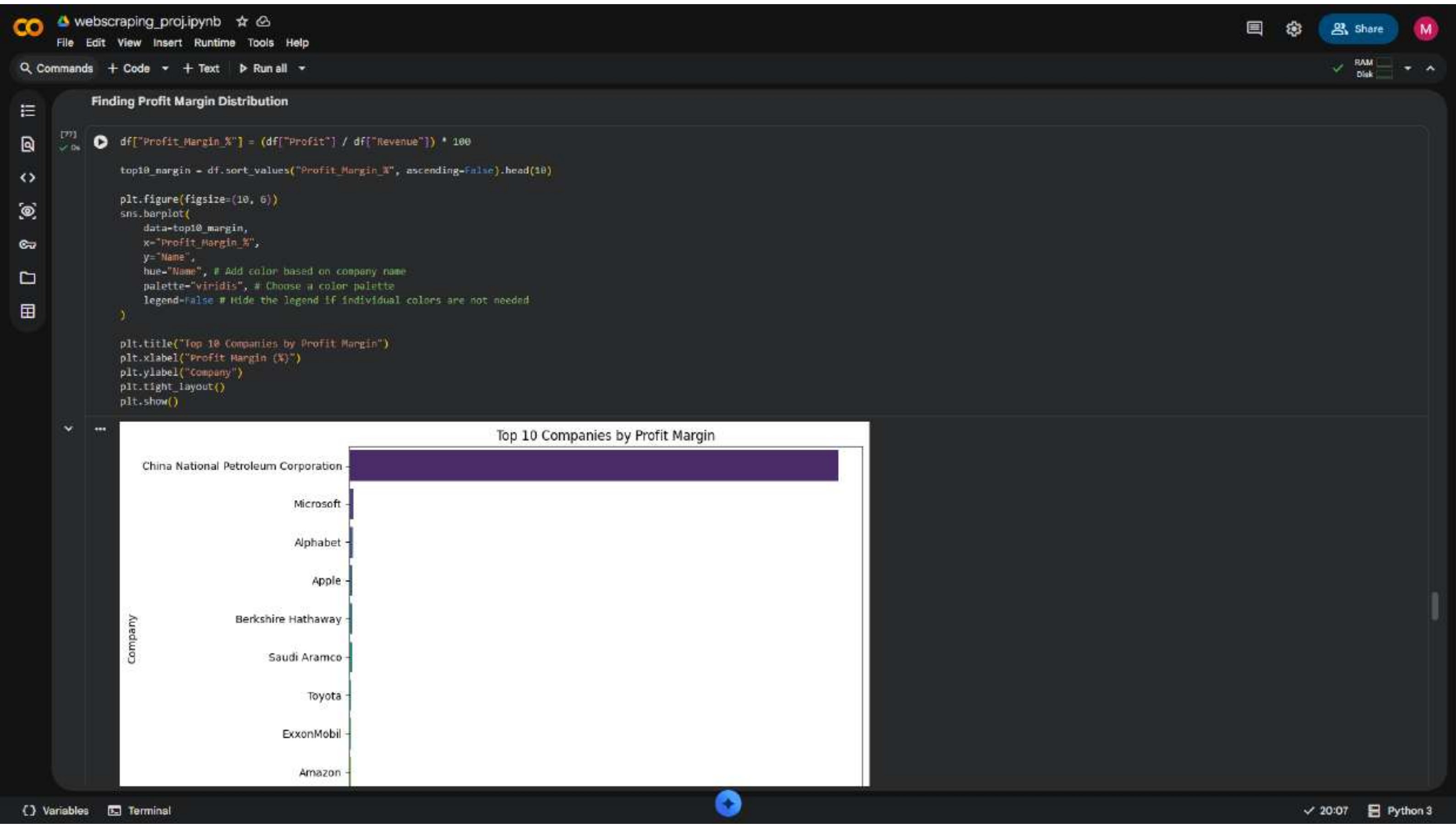
```
plt.figure(figsize=(10, 6))
sns.barplot(
    data=top10_revenue,
    x="Revenue",
    y="Name",
    hue="Name",  # Add color based on company name
    palette="viridis", # Choose a color palette
    legend=False # Hide the legend if individual colors are not needed
)

plt.title("Top 10 Companies by Revenue")
plt.xlabel("Revenue (USD Millions)")
plt.ylabel("Company")
plt.tight_layout()
plt.show()
```

Top 10 Companies by Revenue
Company
Walmart
Amazon
State Grid Corporation of China
Saudi Aramco
China Petrochemical Corporation
UnitedHealth Group
Apple



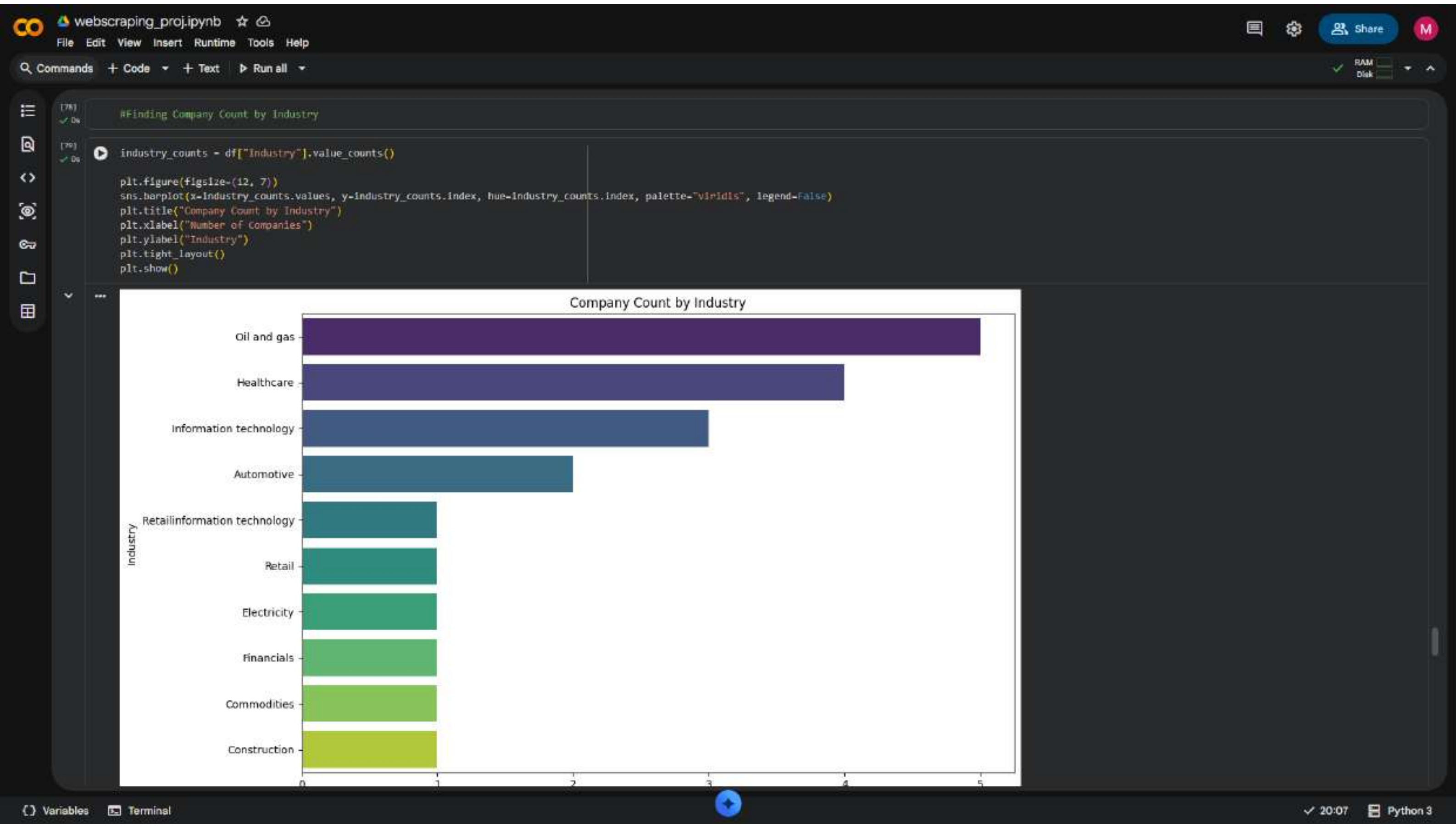
Top 10 Companies by Profit

```
top10_profit = df.sort_values("Profit", ascending=False).head(10)

plt.figure(figsize=(10, 6))
sns.barplot(
    data=top10_profit,
    x="Profit",
    y="Name",
    hue="Name",
    palette="magma", # Another color palette
    legend=False
)

plt.title("Top 10 Companies by Profit")
plt.xlabel("Profit (USD Millions)")
plt.ylabel("Company")
plt.tight_layout()
plt.show()
```



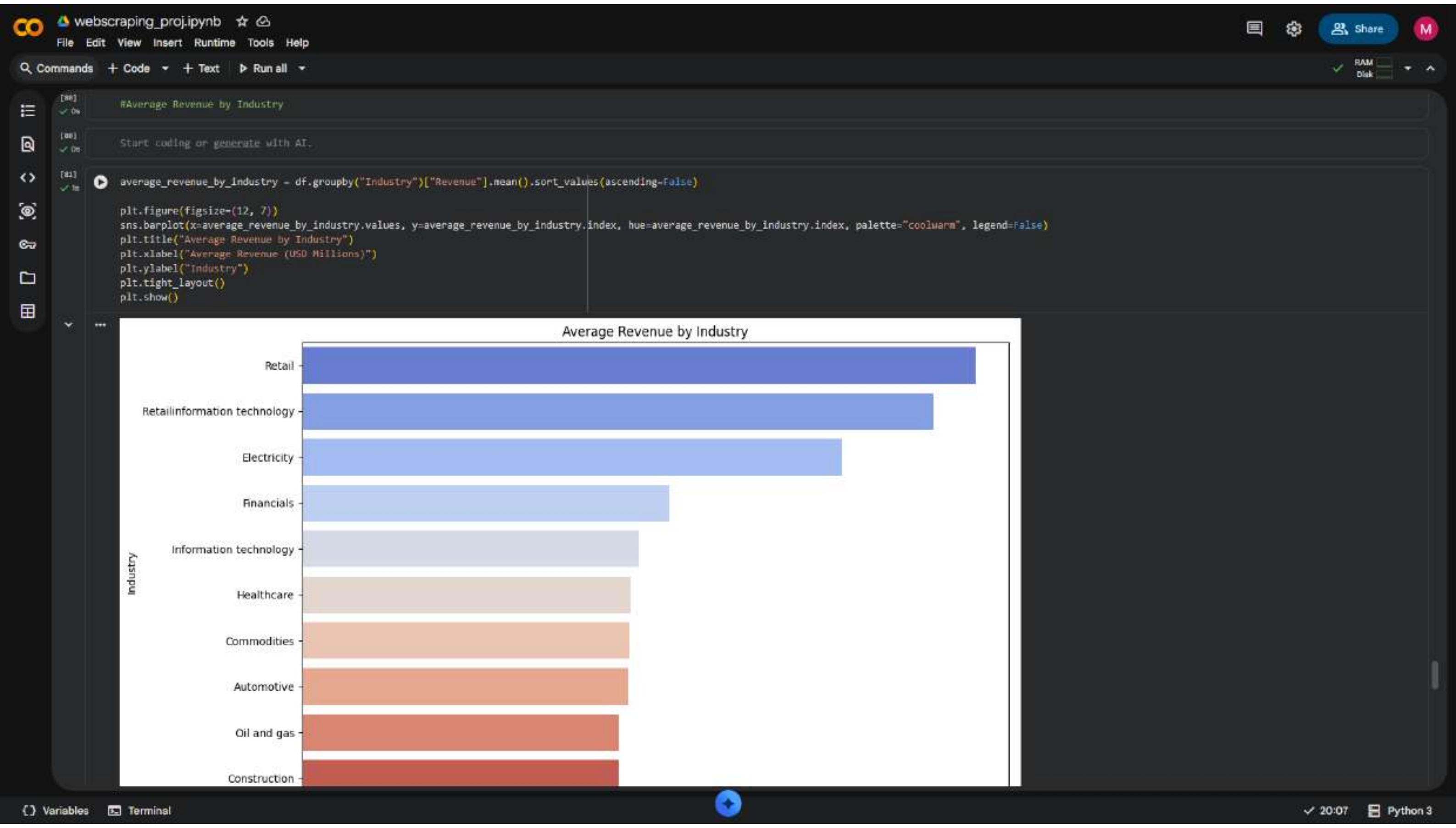
Finding Profit Margin Distribution

```
df["Profit_Margin_%"] = (df["Profit"] / df["Revenue"]) * 100

top10_margin = df.sort_values("Profit_Margin_%", ascending=False).head(10)

plt.figure(figsize=(10, 6))
sns.barplot(
    data=top10_margin,
    x="Profit_Margin_%",
    y="Name",
    hue="Name", # Add color based on company name
    palette="viridis", # Choose a color palette
    legend=False # Hide the legend if individual colors are not needed
)

plt.title("Top 10 Companies by Profit Margin")
plt.xlabel("Profit Margin (%)")
plt.ylabel("Company")
plt.tight_layout()
plt.show()
```



```
#Finding Company Count by Industry
```

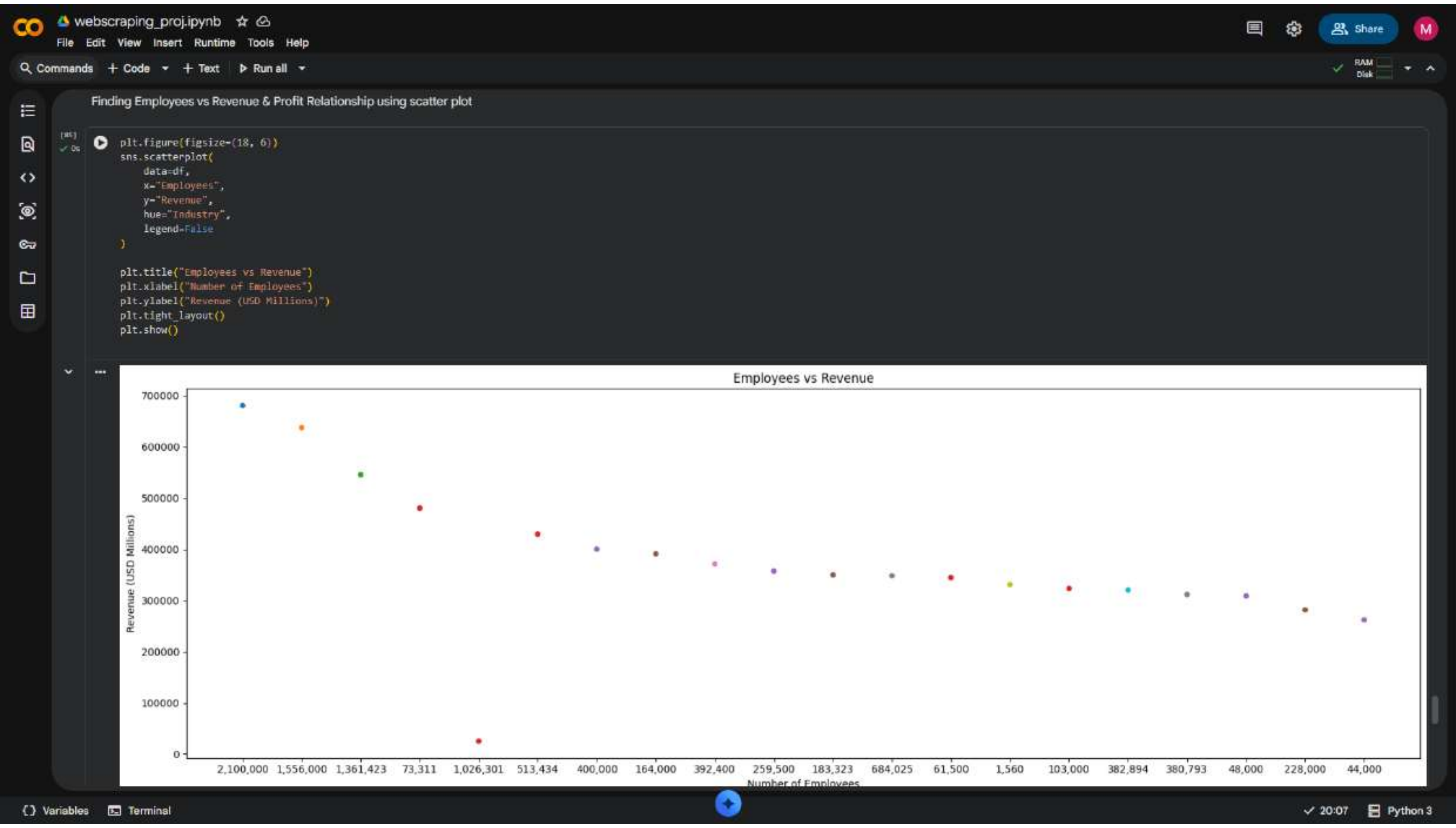
```
industry_counts = df["Industry"].value_counts()

plt.figure(figsize=(12, 7))
sns.barplot(x=industry_counts.values, y=industry_counts.index, hue=industry_counts.index, palette="viridis", legend=False)
plt.title("Company Count by Industry")
plt.xlabel("Number of Companies")
plt.ylabel("Industry")
plt.tight_layout()
plt.show()
```



```
#Average Revenue by Industry
```

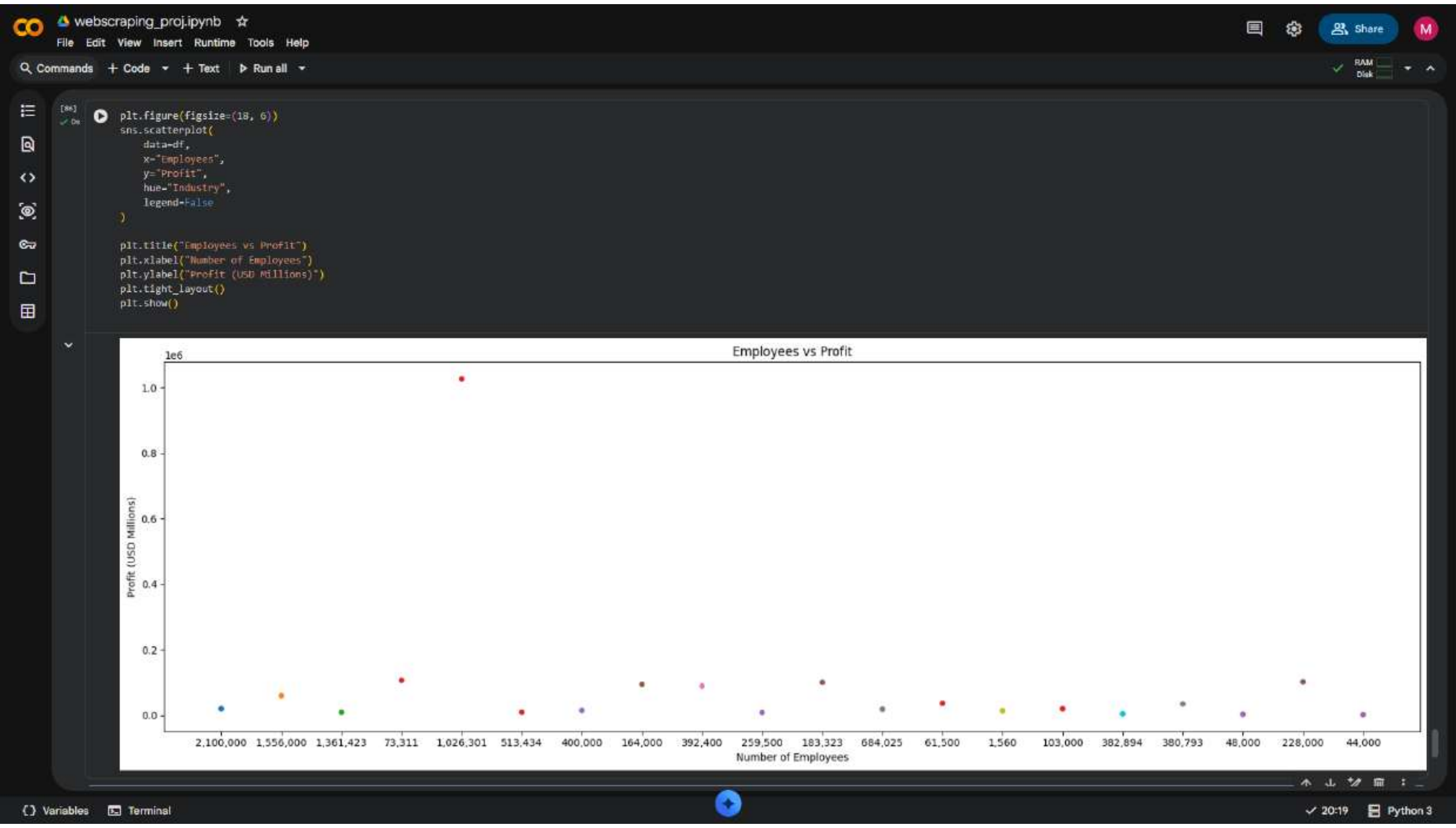
```
Start coding or generate with AI.
```

```
average_revenue_by_industry = df.groupby("Industry")["Revenue"].mean().sort_values(ascending=False)

plt.figure(figsize=(12, 7))
sns.barplot(x=average_revenue_by_industry.values, y=average_revenue_by_industry.index, hue=average_revenue_by_industry.index, palette="coolwarm", legend=False)
plt.title("Average Revenue by Industry")
plt.xlabel("Average Revenue (USD Millions)")
plt.ylabel("Industry")
plt.tight_layout()
plt.show()
```

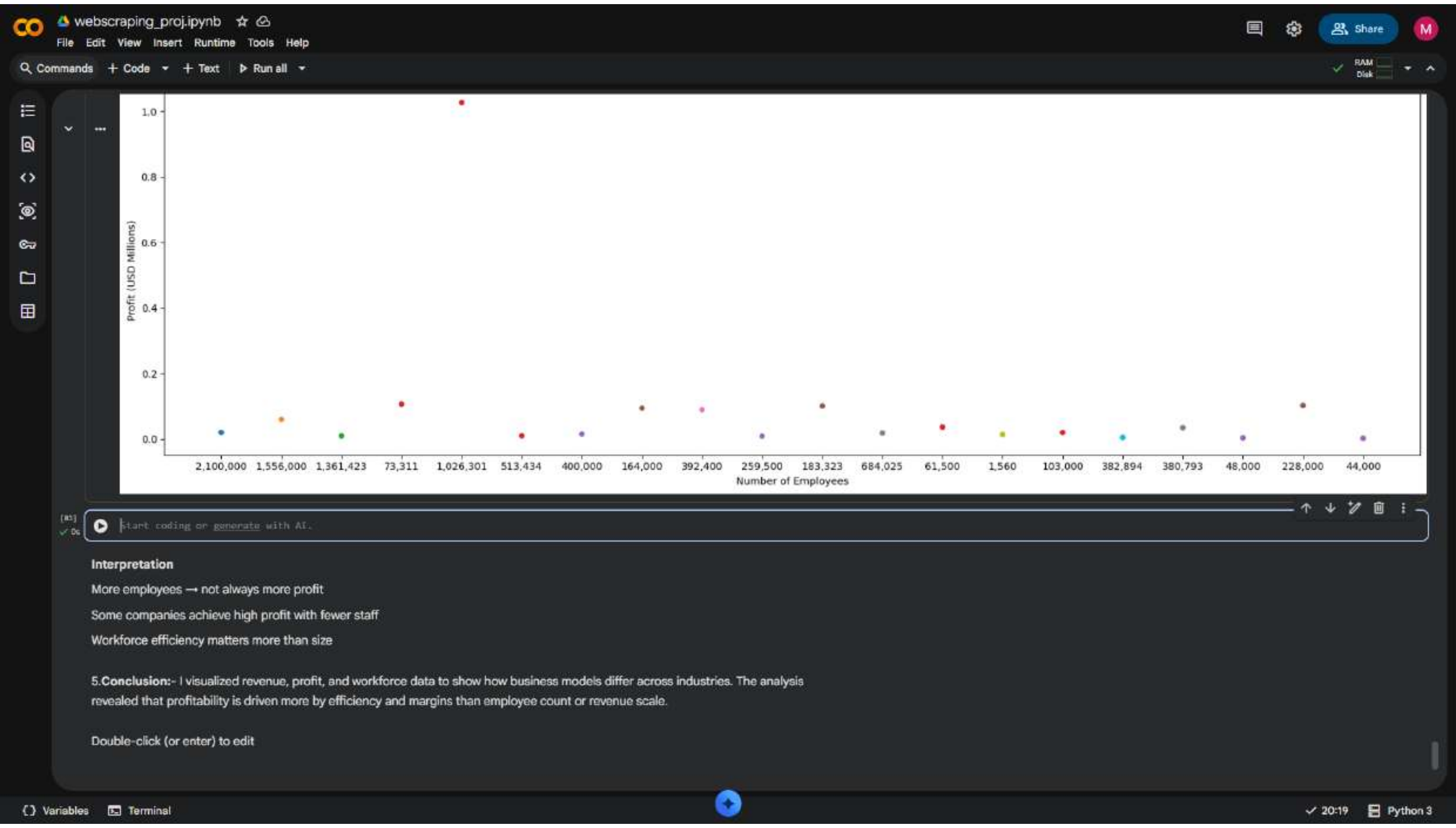


Finding Employees vs Revenue & Profit Relationship using scatter plot

```
plt.figure(figsize=(18, 6))
sns.scatterplot(
    data=df,
    x="Employees",
    y="Revenue",
    hue="Industry",
    legend=False
)

plt.title("Employees vs Revenue")
plt.xlabel("Number of Employees")
plt.ylabel("Revenue (USD Millions)")
plt.tight_layout()
plt.show()
```



```
plt.figure(figsize=(18, 6))
sns.scatterplot(
    data=df,
    x="Employees",
    y="Profit",
    hue="Industry",
    legend=False
)

plt.title("Employees vs Profit")
plt.xlabel("Number of Employees")
plt.ylabel("Profit (USD Millions)")
plt.tight_layout()
plt.show()
```



Interpretation

More employees → not always more profit

Some companies achieve high profit with fewer staff

Workforce efficiency matters more than size

5.**Conclusion:**- I visualized revenue, profit, and workforce data to show how business models differ across industries. The analysis revealed that profitability is driven more by efficiency and margins than employee count or revenue scale.

Double-click (or enter) to edit