In [2]:
```
# Import Library
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt
```

In [14]:
```
companies = pd.read_csv(r'C:\Users\HP\Desktop\Advance Data Analyst\3. Go beyond Numbers. turning Data into insights\3. Module 3\3. Category to into numeric data\Files\Modified_Unicorn_Companies.csv')
companies.head(10)
```

Out[14]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors
0	Bytedance	180	2017-04-07	Artificial intelligence	Beijing	China	Asia	2012	$8B	Sequoia Capital China, SIG Asia Investments, S...
1	SpaceX	100	2012-12-01	Other	Hawthorne	United States	North America	2002	$7B	Founders Fund, Draper Fisher Jurvetson, Rothen...
2	SHEIN	100	2018-07-03	E-commerce & direct-to-consumer	Shenzhen	China	Asia	2008	$2B	Tiger Global Management, Sequoia Capital China...
3	Stripe	95	2014-01-23	FinTech	San Francisco	United States	North America	2010	$2B	Khosla Ventures, LowercaseCapital, capitalG
4	Klarna	46	2011-12-12	Fintech	Stockholm	Sweden	Europe	2005	$4B	Institutional Venture Partners, Sequoia Capita...
5	Canva	40	2018-01-08	Internet software & services	Surry Hills	Australia	Oceania	2012	$572M	Sequoia Capital China, Blackbird Ventures, Mat...
6	Checkout.com	40	2019-05-02	Fintech	London	United Kingdom	Europe	2012	$2B	Tiger Global Management, Insight Partners, DST...
7	Instacart	39	2014-12-30	Supply chain, logistics, & delivery	San Francisco	United States	North America	2012	$3B	Khosla Ventures, Kleiner Perkins Caufield & By...
8	JUUL Labs	38	2017-12-20	Consumer & retail	San Francisco	United States	North America	2015	$14B	Tiger Global Management
9	Databricks	38	2019-02-05	Data management and analytics	San Francisco	United States	North America	2013	$3B	Andreessen Horowitz, New Enterprise Associates...

In [16]:
```
# Display the data types of the columns.

companies.dtypes
```

Out[16]:
```
Company            object
Valuation           int64
Date Joined        object
Industry           object
City               object
Country/Region     object
Continent          object
Year Founded        int64
Funding            object
Select Investors   object
dtype: object
```

In [18]:
```
# Apply necessary datatype conversions.
companies['Date Joined'] = pd.to_datetime(companies['Date Joined'])
companies.head()
```

Out[18]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors
0	Bytedance	180	2017-04-07	Artificial intelligence	Beijing	China	Asia	2012	$8B	Sequoia Capital China, SIG Asia Investments, S...
1	SpaceX	100	2012-12-01	Other	Hawthorne	United States	North America	2002	$7B	Founders Fund, Draper Fisher Jurvetson, Rothen...
2	SHEIN	100	2018-07-03	E-commerce & direct-to-consumer	Shenzhen	China	Asia	2008	$2B	Tiger Global Management, Sequoia Capital China...
3	Stripe	95	2014-01-23	FinTech	San Francisco	United States	North America	2010	$2B	Khosla Ventures, LowercaseCapital, capitalG
4	Klarna	46	2011-12-12	Fintech	Stockholm	Sweden	Europe	2005	$4B	Institutional Venture Partners, Sequoia Capita...

In [20]:
```
# Create the column Years To Unicorn.

companies['Years To Unicorn'] = companies['Date Joined'].dt.year - companies['Year Founded']
companies.head()
```

Out[20]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors	Years To Unicorn
0	Bytedance	180	2017-04-07	Artificial intelligence	Beijing	China	Asia	2012	$8B	Sequoia Capital China, SIG Asia Investments, S...	5
1	SpaceX	100	2012-12-01	Other	Hawthorne	United States	North America	2002	$7B	Founders Fund, Draper Fisher Jurvetson, Rothen...	10
2	SHEIN	100	2018-07-03	E-commerce & direct-to-consumer	Shenzhen	China	Asia	2008	$2B	Tiger Global Management, Sequoia Capital China...	10
3	Stripe	95	2014-01-23	FinTech	San Francisco	United States	North America	2010	$2B	Khosla Ventures, LowercaseCapital, capitalG	4
4	Klarna	46	2011-12-12	Fintech	Stockholm	Sweden	Europe	2005	$4B	Institutional Venture Partners, Sequoia Capita...	6

In [22]:
```
companies['Years To Unicorn'].describe()
```

Out[22]:
```
count    1074.000000
mean        7.013035
std         5.331842
min        -3.000000
25%         4.000000
50%         6.000000
75%         9.000000
max        98.000000
Name: Years To Unicorn, dtype: float64
```

In [24]:
```
# Isolate any rows where `Years To Unicorn` is negative

companies[companies['Years To Unicorn'] < 0]
```

Out[24]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors	Years To Unicorn
527	InVision	2	2017-11-01	Internet software & services	New York	United States	North America	2020	$349M	FirstMark Capital, Tiger Global Management, IC...	-3

In [26]:
```
# An internet search reveals that InVision was founded in 2011. Replace the value at Year Founded with 2011 for InVision's row.
# Replace InVision's `Year Founded` value with 2011

companies.loc[companies['Company']=='InVision', 'Year Founded'] = 2011

# Verify the change was made properly

companies[companies['Company']=='InVision']
```

Out[26]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors	Years To Unicorn
527	InVision	2	2017-11-01	Internet software & services	New York	United States	North America	2011	$349M	FirstMark Capital, Tiger Global Management, IC...	-3

In [28]:
```
# Recalculate all values in the `Years To Unicorn` column

companies['Years To Unicorn'] = companies['Date Joined'].dt.year - companies['Year Founded']

# Verify that there are no more negative values in the column

companies['Years To Unicorn'].describe()
```

Out[28]:
```
count    1074.000000
mean        7.021415
std         5.323155
min         0.000000
25%         4.000000
50%         6.000000
75%         9.000000
max        98.000000
Name: Years To Unicorn, dtype: float64
```

In [32]:
```
# List provided by the company of the expected industry labels in the data
industry_list = ['Artificial intelligence', 'Other','E-commerce & direct-to-consumer', 'Fintech',\
       'Internet software & services','Supply chain, logistics, & delivery', 'Consumer & retail',\
       'Data management & analytics', 'Edtech', 'Health', 'Hardware','Auto & transportation', \
        'Travel', 'Cybersecurity','Mobile & telecommunications']
set(companies['Industry']) - set(industry_list)
```

Out[32]:
```
{'Artificial Intelligence', 'Data management and analytics', 'FinTech'}
```

In [34]:
```
# 1. Create `replacement_dict`

replacement_dict = {'Artificial Intelligence': 'Artificial intelligence',
                    'Data management and analytics': 'Data management & analytics',
                    'FinTech': 'Fintech'
                    }

# 2. Replace the incorrect values in the `Industry` column

companies['Industry'] = companies['Industry'].replace(replacement_dict)

# 3. Verify that there are no longer any elements in `Industry` that are not in `industry_list`

set(companies['Industry']) - set(industry_list)
```

Out[34]:
```
set()
```

In [36]:
```
# Isolate rows of all companies that have duplicates

companies[companies.duplicated(subset=['Company'], keep=False)]
```

Out[36]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors	Years To Unicorn
385	BrewDog	2	2017-04-10	Consumer & retail	Aberdeen	United Kingdom	Europe	2007	$233M	TSG Consumer Partners, Crowdcube	10
386	BrewDog	2	2017-04-10	Consumer & retail	Aberdeen	UnitedKingdom	Europe	2007	$233M	TSG Consumer Partners	10
510	ZocDoc	2	2015-08-20	Health	New York	United States	North America	2007	$374M	Founders Fund, Khosla Ventures, Goldman Sachs	8
511	ZocDoc	2	2015-08-20	Health	NaN	United States	North America	2007	$374M	Founders Fund	8
1031	SoundHound	1	2018-05-03	Artificial intelligence	Santa Clara	United States	North America	2005	$215M	Tencent Holdings, Walden Venture Capital, Glob...	13
1032	SoundHound	1	2018-05-03	Other	Santa Clara	United States	North America	2005	$215M	Tencent Holdings	13

In [38]:
```
# Drop rows of duplicate companies after their first occurrence

companies = companies.drop_duplicates(subset=['Company'], keep='first')
```

In [40]:
```
# Create new `High Valuation` column

# Use qcut to divide Valuation into 'high' and 'low' Valuation groups
companies['High Valuation'] = pd.qcut(companies['Valuation'], 2, labels = ['low', 'high'])
```

In [42]:
```
# Rank the continents by number of unicorn companies

companies["Continent"].value_counts()
```

Out[42]:
```
Continent
North America    586
Asia             310
Europe           143
South America     21
Oceania            9
Africa             3
Name: count, dtype: int64
```

In [44]:
```
# Create numeric `Continent Number` column

continent_dict = {'North America': 1,
                  'Asia': 2,
                  'Europe': 3,
                  'South America': 4,
                  'Oceania': 5,
                  'Africa': 6
                  }

companies['Continent Number'] = companies['Continent'].replace(continent_dict)
companies.head()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_16820\1487589892.py:10: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
  companies['Continent Number'] = companies['Continent'].replace(continent_dict)
```

Out[44]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors	Years To Unicorn	High Valuation	Continent Number
0	Bytedance	180	2017-04-07	Artificial intelligence	Beijing	China	Asia	2012	$8B	Sequoia Capital China, SIG Asia Investments, S...	5	high	2
1	SpaceX	100	2012-12-01	Other	Hawthorne	United States	North America	2002	$7B	Founders Fund, Draper Fisher Jurvetson, Rothen...	10	high	1
2	SHEIN	100	2018-07-03	E-commerce & direct-to-consumer	Shenzhen	China	Asia	2008	$2B	Tiger Global Management, Sequoia Capital China...	10	high	2
3	Stripe	95	2014-01-23	Fintech	San Francisco	United States	North America	2010	$2B	Khosla Ventures, LowercaseCapital, capitalG	4	high	1
4	Klarna	46	2011-12-12	Fintech	Stockholm	Sweden	Europe	2005	$4B	Institutional Venture Partners, Sequoia Capita...	6	high	3

In [46]:
```
# Create `Country/Region Numeric` column
# Create numeric categories for Country/Region
companies['Country/Region Numeric'] = companies['Country/Region'].astype('category').cat.codes
```

In [48]:
```
# Convert `Industry` to numeric data

# Create dummy variables with Industry values
industry_encoded = pd.get_dummies(companies['Industry'])

# Combine `companies` DataFrame with new dummy Industry columns
companies = pd.concat([companies, industry_encoded], axis=1)
```

In [50]:
```
companies.head()
```

Out[50]:

	Company	Valuation	Date Joined	Industry	City	Country/Region	Continent	Year Founded	Funding	Select Investors	...	E-commerce & direct-to-consumer	Edtech	Fintech	Hardware	Health	Internet software & services	Mobile & telecommunications	Other	Supply chain, logistics, & delivery	Travel
0	Bytedance	180	2017-04-07	Artificial intelligence	Beijing	China	Asia	2012	$8B	Sequoia Capital China, SIG Asia Investments, S...	...	False	False	False	False	False	False	False	False	False	False
1	SpaceX	100	2012-12-01	Other	Hawthorne	United States	North America	2002	$7B	Founders Fund, Draper Fisher Jurvetson, Rothen...	...	False	False	False	False	False	False	False	True	False	False
2	SHEIN	100	2018-07-03	E-commerce & direct-to-consumer	Shenzhen	China	Asia	2008	$2B	Tiger Global Management, Sequoia Capital China...	...	True	False	False	False	False	False	False	False	False	False
3	Stripe	95	2014-01-23	Fintech	San Francisco	United States	North America	2010	$2B	Khosla Ventures, LowercaseCapital, capitalG	...	False	False	True	False	False	False	False	False	False	False
4	Klarna	46	2011-12-12	Fintech	Stockholm	Sweden	Europe	2005	$4B	Institutional Venture Partners, Sequoia Capita...	...	False	False	True	False	False	False	False	False	False	False

5 rows × 29 columns

In [ ]:
```
# Conclusion
# Input validation is essential for ensuring data is high quality and error-free.
# In practice, input validation requires trial and error to identify issues and determine the best way to fix them.
# There are benefits and disadvantages to both label encoding and dummy/one-hot encoding.
# The decision to use label encoding versus dummy/one-hot encoding needs to be made on a case-by-case basis.
```