Math 189 - CS Major Enrollement at UC Campuses

Research Question

How do gender enrollment and graduation trends in Computer Science programs at UC schools near major tech hubs (UCB, UCLA, UCSD, UCD, UCSC) compare to those at UCs located farther from these hubs (UCSB, UCR, UCM, UCI) from 2017 to 2022?

Background and Inspiration

The tech industry has long struggled with a gender imbalance. According to the U.S. Bureau of Labor Statistics, there are over 5 million people in the U.S. employed in computer and mathematical occupations but only 26.2% of these occupations are women. These disparities are also shown in education as only 20% of computer science undergraduate degrees go to women. By examining the differences across UC campuses, we can find key insights into the role of industry access and institutional support that goes into shaping diversity in CS programs. Schools located near major tech hubs may benefit from internship pipelines, stronger industry connections, and a more thought out computer science program which could contribute to more diversity compared to campuses which are further from these tech hubs. Understanding these differences can help reform education policies, guiding decisions on resource allocation, and reforming the computer science curriculum to promote greater diversity.

Throughout this project, our findings could help shape policies and improve education strategies. If being near tech hubs plays a major role in CS diversity, this would highlight the importance for non-tech hub campuses to put more resources into building stronger industry connections. On the other hand, if there is no significant difference, it may suggest that university policies have a larger impact, emphasizing the importance of fostering inclusivity in CS education. Our project aims to provide valuable insights to guide decisions on resource distribution and diversity initiatives that will help UCs work towards a more inclusive and equitable CS program.

Hypothesis

For our expected outcomes, we anticipate that UC campuses near major tech hubs (UCB, UCLA, UCSD, UCD, UCSC) will have a higher proportion of female CS graduates compared to campuses further from these hubs(UCSB, UCR, UCM, UCI). We also expect to see an overall increase in the population of female CS majors, but the rate of growth between females and males might be different. From our two-sample t test, we predict that the proportion of female CS graduates will be significantly higher at tech hub campuses. However, if there turns out to be no significant difference, this may mean that university policies and support programs play a larger role than being near a tech hub. From our regression analysis, we expect that location will be a significant predictor in the number of CS graduates. Gender will also likely be a strong factor, with males continuing to be a larger proportion of CS graduates. Additionally, the year variable may also show an overall increase in female CS graduates especially in tech hub campuses if their industry connections have had a strong impact.

Data

Data overview

- Dataset #1
    - Dataset Name: DataUSA Datasets on Graduates' Majors
    - Link to the dataset: https://datausa.io/search/?dimension=University (General Source)
        - We have 18 sub-datasets in total
            - UC Berkeley: https://docs.google.com/spreadsheets/d/1mWQo9-sF1Gv0EzZAiJNDFJ1GXFRtW2ExmL3Xh5oqC94/edit?gid=0#gid=0
            - UC Davis: https://docs.google.com/spreadsheets/d/1mWLnTyUHzo9cedWqUYC-CKvT8ctdReJVNEy-4YPvMpM/edit?gid=0#gid=0
            - UCLA: https://docs.google.com/spreadsheets/d/1TEjgqAb6DKOZlwCFm1TxNvTi7C
              02YMeDB7PXUJE/edit?gid=0#gid=0
            - UC Irvine: https://docs.google.com/spreadsheets/d/1n3frMOAKWhKPswrZlVjLlHyVkG
              _TmfU5dH8/edit?usp=sharing
            - UC San Diego: https://docs.google.com/spreadsheets/d/14HsxrRD9Srs2cLC_pV1BlsBPonx
              gid=275689308#gid=275689308
            - UC Santa Barbra: https://docs.google.com/spreadsheets/d/1AdzX2m0Xb5HcYrh3MVfqP0ruB
              usp=sharing
            - UC Riverside: https://docs.google.com/spreadsheets/d/1WNtmMvjnsUffPmcSkbZhrb8-l2b9qfI8seH94i2Vylk/edit?gid=604871786#gid=604871786

        - UC Santa Cruz: https://docs.google.com/spreadsheets/d/11XbxM0u-ZahWgqercf9cQjkbeqzRiLlpCcMlRJgZMIQ/edit?gid=0#gid=0
        - UC Merced:
          https://docs.google.com/spreadsheets/d/142Jrv9n4m89BffdyU_xgsMTKLM
          OQdI0Vhb5M/edit?usp=sharing
          https://datausa.io/profile/university/university-of-california-merced#graduates
    - Number of observations: 19460 (combined 9 schools and both genders)

    - Number of variables: 5 (Major, Year, University, Gender, Completions)

Important Variables:

- Year (int): The year of graduation, ranging from 2017 to 2022.
- CIP6 (string): The Classification of Instructional Programs (CIP) code that categorizes the major titles. This variable will be used to identify whether a major belongs to "STEM" or "Social Science" based on predefined keywords.
- Completion (int): The number of degrees completed in each major.
- University (string): The name of the university where the students graduated.
- Gender (string): Men and Women

Concepts:

- CIP6 codes serve as proxies for categorizing majors into "STEM" or "Social Science."
- Completion metrics reflect the number of graduates in each major.

Data Integration:

The datasets across the nine schools provide a detailed understanding of each graduating class from 2017 to 2022 by gender. Initially, these datasets contained variables such as the declared major name, number of completed students, university name, and gender. After cleaning, we focus on essential variables: ‘Year’ (int), ‘CIP6’ (string), ‘University’ (string), and ‘Completion’ (int) that are important to our research question. We concatenated the cleaned male and female datasets from all nine UC campuses into one large dataframe. This unified dataset will include all necessary variables for future analysis, enabling a comprehensive examination of gender differences in major choices across UC schools.

DataUSA Datasets on Graduates’ Majors (separated by gender)

In [2]:
```
%pip install pandas
%pip install requests
```

```
import pandas as pd
import requests
```

```
Requirement already satisfied: pandas in c:\users\owent\appdata\local\progra
ms\python\python312\lib\site-packages (2.2.2)Note: you may need to restart t
he kernel to use updated packages.

Requirement already satisfied: numpy>=1.26.0 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\owent\appd
ata\local\programs\python\python312\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\owent\appdata\loca
l\programs\python\python312\lib\site-packages (from pandas) (2024.1)
Requirement already satisfied: six>=1.5 in c:\users\owent\appdata\local\prog
rams\python\python312\lib\site-packages (from python-dateutil>=2.8.2->panda
s) (1.16.0)
```

```
[notice] A new release of pip is available: 24.1.2 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
Requirement already satisfied: requests in c:\users\owent\appdata\local\prog
rams\python\python312\lib\site-packages (2.31.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\owent\ap
pdata\local\programs\python\python312\lib\site-packages (from requests) (3.
3.2)
Requirement already satisfied: idna<4,>=2.5 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from requests) (2.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\owent\appdata
\local\programs\python\python312\lib\site-packages (from requests) (2.1.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\owent\appdata
\local\programs\python\python312\lib\site-packages (from requests) (2023.7.2
2)
Note: you may need to restart the kernel to use updated packages.
```

```
[notice] A new release of pip is available: 24.1.2 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [3]:

```
%pip install seaborn
%pip install patsy
%pip install statsmodels

import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np
import scipy.stats as stats
```

```
Requirement already satisfied: seaborn in c:\users\owent\appdata\local\progr
ams\python\python312\lib\site-packages (0.13.2)Note: you may need to restart
the kernel to use updated packages.

Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\owent\appdat
a\local\programs\python\python312\lib\site-packages (from seaborn) (1.26.4)
Requirement already satisfied: pandas>=1.2 in c:\users\owent\appdata\local\p
rograms\python\python312\lib\site-packages (from seaborn) (2.2.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in c:\users\owent\app
data\local\programs\python\python312\lib\site-packages (from seaborn) (3.8.
4)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\owent\appdata\lo
cal\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=3.
4->seaborn) (1.2.1)
Requirement already satisfied: cycler>=0.10 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=3.4->
seaborn) (0.10.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\owent\appdata\l
ocal\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=
3.4->seaborn) (4.51.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\owent\appdata\l
ocal\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=
3.4->seaborn) (1.4.5)
Requirement already satisfied: packaging>=20.0 in c:\users\owent\appdata\loc
al\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=3.4
->seaborn) (23.2)
Requirement already satisfied: pillow>=8 in c:\users\owent\appdata\local\pro
grams\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seab
orn) (10.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\owent\appdata\lo
cal\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,>=3.
4->seaborn) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\owent\appdat
a\local\programs\python\python312\lib\site-packages (from matplotlib!=3.6.1,
>=3.4->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from pandas>=1.2->seaborn) (20
24.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\owent\appdata\loca
l\programs\python\python312\lib\site-packages (from pandas>=1.2->seaborn) (2
024.1)
Requirement already satisfied: six in c:\users\owent\appdata\local\programs
\python\python312\lib\site-packages (from cycler>=0.10->matplotlib!=3.6.1,>=
3.4->seaborn) (1.16.0)
```

```
[notice] A new release of pip is available: 24.1.2 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
Requirement already satisfied: patsy in c:\users\owent\appdata\local\program
s\python\python312\lib\site-packages (0.5.6)Note: you may need to restart th
e kernel to use updated packages.

Requirement already satisfied: six in c:\users\owent\appdata\local\programs
\python\python312\lib\site-packages (from patsy) (1.16.0)
Requirement already satisfied: numpy>=1.4 in c:\users\owent\appdata\local\pr
ograms\python\python312\lib\site-packages (from patsy) (1.26.4)
```

```
[notice] A new release of pip is available: 24.1.2 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
Requirement already satisfied: statsmodels in c:\users\owent\appdata\local\p
rograms\python\python312\lib\site-packages (0.14.2)
Requirement already satisfied: numpy>=1.22.3 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from statsmodels) (1.26.4)
Requirement already satisfied: scipy!=1.9.2,>=1.8 in c:\users\owent\appdata
\local\programs\python\python312\lib\site-packages (from statsmodels) (1.13.
0)
Requirement already satisfied: pandas!=2.1.0,>=1.4 in c:\users\owent\appdata
\local\programs\python\python312\lib\site-packages (from statsmodels) (2.2.
2)
Requirement already satisfied: patsy>=0.5.6 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from statsmodels) (0.5.6)
Requirement already satisfied: packaging>=21.3 in c:\users\owent\appdata\loc
al\programs\python\python312\lib\site-packages (from statsmodels) (23.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\owent\appd
ata\local\programs\python\python312\lib\site-packages (from pandas!=2.1.0,>=
1.4->statsmodels) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\owent\appdata\local
\programs\python\python312\lib\site-packages (from pandas!=2.1.0,>=1.4->stat
smodels) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\owent\appdata\loca
l\programs\python\python312\lib\site-packages (from pandas!=2.1.0,>=1.4->sta
tsmodels) (2024.1)
Requirement already satisfied: six in c:\users\owent\appdata\local\programs
\python\python312\lib\site-packages (from patsy>=0.5.6->statsmodels) (1.16.
0)
Note: you may need to restart the kernel to use updated packages.
[notice] A new release of pip is available: 24.1.2 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [4]:

```python
# read in female datasets from 9 campuses
response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucb_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucla_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_uci_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucsb_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucsd_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
```

```
df_ucsc_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucd_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=44
data = response.json()
df_ucm_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucr_f = pd.json_normalize(data['data'])

frames = [df_ucb_f, df_ucla_f, df_uci_f, df_ucsb_f, df_ucsd_f, df_ucsc_f, df
result_female = pd.concat(frames)
result_female
```

Out[4]:

	ID CIP6	CIP6	ID Year	Year	ID University	University	ID Gender	Gender	Com
0	260406	Cellular & Molecular Biology	2021	2021	110635	University of California-Berkeley	2	Women	
1	260406	Cellular & Molecular Biology	2022	2022	110635	University of California-Berkeley	2	Women	
2	260406	Cellular & Molecular Biology	2020	2020	110635	University of California-Berkeley	2	Women	
3	260406	Cellular & Molecular Biology	2019	2019	110635	University of California-Berkeley	2	Women	
4	450603	Econometrics & Quantitative Economics	2021	2021	110635	University of California-Berkeley	2	Women	
...	...	...	...	...	...	...	...	...	
862	520301	Accounting	2021	2021	110671	University of California-Riverside	2	Women	
863	520301	Accounting	2022	2022	110671	University of California-Riverside	2	Women	
864	520601	Business & Managerial Economics	2021	2021	110671	University of California-Riverside	2	Women	
865	520601	Business & Managerial Economics	2022	2022	110671	University of California-Riverside	2	Women	
866	540199	Other History	2018	2018	110671	University of California-Riverside	2	Women	

9730 rows × 11 columns

In [5]:

```
# read in male datasets for 9 campuses
response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucb_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucla_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_uci_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucsb_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucsd_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucsc_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucd_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=44
data = response.json()
df_ucm_f = pd.json_normalize(data['data'])

response = requests.get('http://elpaso-app.datausa.io/api/data?University=11
data = response.json()
df_ucr_f = pd.json_normalize(data['data'])

frames = [df_ucb_f, df_ucla_f, df_uci_f, df_ucsb_f, df_ucsd_f, df_ucsc_f, df
result_male = pd.concat(frames)
result_male
```

Out[5]:

	ID CIP6	CIP6	ID Year	Year	ID University	University	ID Gender	Gender	Con
0	110701	Computer Science	2020	2020	110635	University of California-Berkeley	1	Men	
1	141001	Electrical & Electronics Engineering	2022	2022	110635	University of California-Berkeley	1	Men	
2	110701	Computer Science	2022	2022	110635	University of California-Berkeley	1	Men	
3	110701	Computer Science	2021	2021	110635	University of California-Berkeley	1	Men	
4	520201	General Business Administration & Management	2015	2015	110635	University of California-Berkeley	1	Men	
...	...	...	...	...	...	...	...	...	
862	520301	Accounting	2020	2020	110671	University of California-Riverside	1	Men	
863	520301	Accounting	2021	2021	110671	University of California-Riverside	1	Men	
864	520301	Accounting	2022	2022	110671	University of California-Riverside	1	Men	
865	520601	Business & Managerial Economics	2021	2021	110671	University of California-Riverside	1	Men	
866	520601	Business & Managerial Economics	2022	2022	110671	University of California-Riverside	1	Men	

9730 rows × 11 columns

In [6]:

```
#concatenate both gender into one dataset
all_data = [result_female, result_male]
result_all = pd.concat(all_data)
result_all
```

Out[6]:

	ID CIP6	CIP6	ID Year	Year	ID University	University	ID Gender	Gender	Com
0	260406	Cellular & Molecular Biology	2021	2021	110635	University of California-Berkeley	2	Women	
1	260406	Cellular & Molecular Biology	2022	2022	110635	University of California-Berkeley	2	Women	
2	260406	Cellular & Molecular Biology	2020	2020	110635	University of California-Berkeley	2	Women	
3	260406	Cellular & Molecular Biology	2019	2019	110635	University of California-Berkeley	2	Women	
4	450603	Econometrics & Quantitative Economics	2021	2021	110635	University of California-Berkeley	2	Women	
...	...	...	...	...	...	...	...	...	
862	520301	Accounting	2020	2020	110671	University of California-Riverside	1	Men	
863	520301	Accounting	2021	2021	110671	University of California-Riverside	1	Men	
864	520301	Accounting	2022	2022	110671	University of California-Riverside	1	Men	
865	520601	Business & Managerial Economics	2021	2021	110671	University of California-Riverside	1	Men	
866	520601	Business & Managerial Economics	2022	2022	110671	University of California-Riverside	1	Men	

19460 rows × 11 columns

In [7]:

```
#drop unnecessary columns
result_all = result_all.drop(['ID CIP6', 'ID Year', 'ID University', 'ID Ger
result_all
```

Out[7]:

	CIP6	Year	University	Gender	Completions
0	Cellular & Molecular Biology	2021	University of California-Berkeley	Women	555
1	Cellular & Molecular Biology	2022	University of California-Berkeley	Women	535
2	Cellular & Molecular Biology	2020	University of California-Berkeley	Women	474
3	Cellular & Molecular Biology	2019	University of California-Berkeley	Women	424
4	Econometrics & Quantitative Economics	2021	University of California-Berkeley	Women	383
...	...	...	...	...	...
862	Accounting	2020	University of California-Riverside	Men	0
863	Accounting	2021	University of California-Riverside	Men	0
864	Accounting	2022	University of California-Riverside	Men	0
865	Business & Managerial Economics	2021	University of California-Riverside	Men	0
866	Business & Managerial Economics	2022	University of California-Riverside	Men	0

19460 rows × 5 columns

In [8]:

```
#change column name
df=result_all.rename(columns={'CIP6': 'Major'})
df
df.to_csv('output.csv', index=False)
```

In [9]:

```
#convert Year(string) to datatype int
df['Year']=df['Year'].astype(int)
```

In [10]:

```
#select years from 2017 to 2022 and filter out completions that are 0
df = df[(df['Year']>=2017)&(df['Year'] <=2022)&(df['Completions'] > 0)]
df.to_csv('output.csv', index=False)
```

In [11]:

```
df.isna().any()
```

Out[11]:

```
Major          False
Year           False
University     False
Gender         False
Completions    False
dtype: bool
```

Results

Exploratory Data Analysis

Section 1 of EDA - Compare the number of Computer Science majors between gender

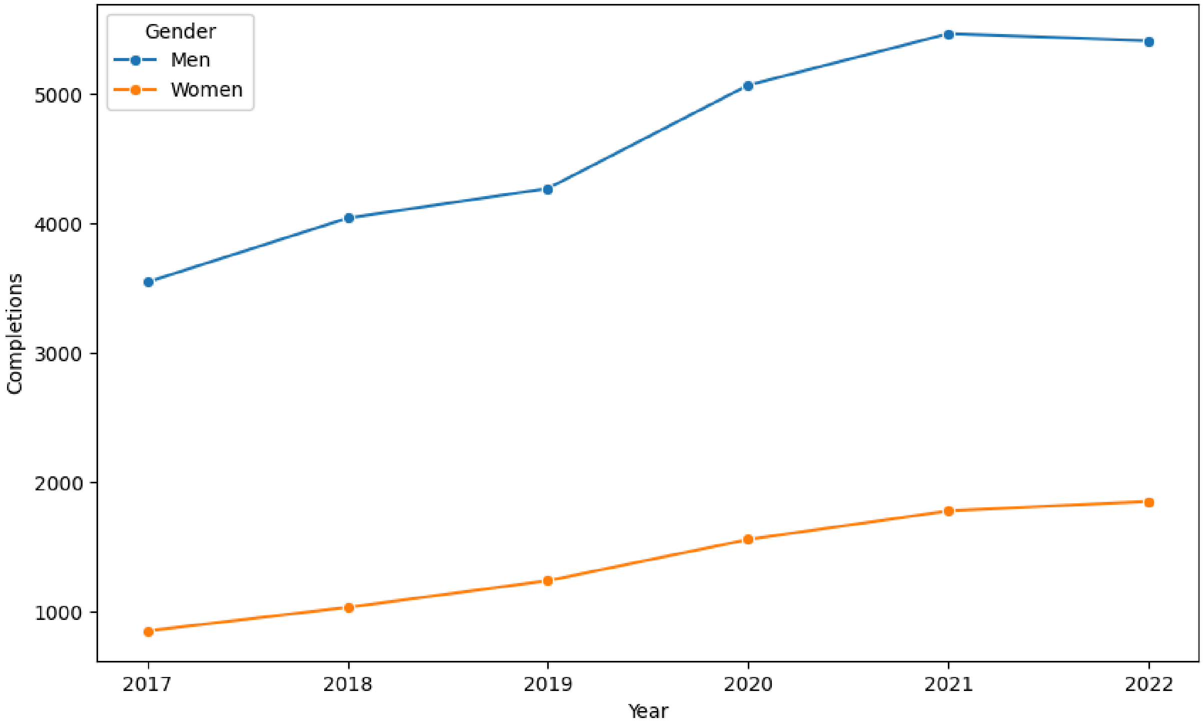
In [12]:

```
df['Major'].unique()
```

```
Out[12]: array(['Cellular & Molecular Biology',
                'Econometrics & Quantitative Economics', 'General Public Health',
                'General Business Administration & Management',
                'Other Computer & Information Sciences',
                'General Political Science & Government', 'General Economics',
                'Other Research & Experimental Psychology', 'General Psychology',
                'Sociology', 'Computer Science', 'Management Science',
                'Programs for Foreign Lawyers', 'Law', 'Cognitive Science',
                'Social Work', 'General English Language & Literature',
                'General Biological Sciences',
                'Electrical & Electronics Engineering', 'Other Social Sciences',
                'General Civil Engineering', 'General Legal Studies',
                'General Statistics', 'Information Science', 'Operations Research',
                'General Applied Mathematics', 'International & Global Studies',
                'General Chemistry', 'Mechanical Engineering',
                'Bioengineering & Biomedical Engineering',
                'General Advanced Legal Studies', 'Anthropology',
                'Other Multidisciplinary Studies', 'General History',
                'General Education', 'Nutrition Sciences',
                'General Public Policy Analysis', 'Optometry',
                'Chemical Engineering', 'Linguistics', 'General Fine Studio Arts',
                'Film, Cinema, & Video Studies', 'Rhetoric & Composition',
                'Peace Studies & Conflict Resolution', 'Sustainability Studies',
                'General Physics', 'Art History, Criticism, & Conservation',
                'General Mathematics', 'Materials Science',
                'General Hispanic & Latin American Languages, Literatures, & Linguis
         tics',
                'Philosophy', 'Teacher Education', 'General Microbiology',
                'General Drama & Theater Arts', 'Financial Mathematics',
                'Urban Studies', 'Comparative Literature', 'General Music',
                'Astrophysics', 'French Language & Literature', 'Geography',
                'General Educational Leadership & Administration',
                'Actuarial Science', 'Nuclear Engineering',
                'Classical, Ancient Mediterranean, & Near Eastern Studies & Archeolo
         gy',
                'Botany & Plant Biology', 'General Geology & Earch Science',
                'Japanese Language & Literature', 'Biostatistics', 'Epidemiology',
                'Chemical & Physical Oceanography', 'German Language & Literature',
                'Other Foreign Languages, Literatures, & Linguistics',
                'Other Business Administration, Management, & Operations',
                'Other Engineering', 'General Engineering', 'Environmental Health',
                'General Slavic Languages, Literatures, & Linguistics',
                'Vision Science & Physiological Optics', 'Geophysics & Seismology',
                'Chinese Language & Literature', 'Italian Language & Literature',
                'Other Middle/Near Eastern & Semetic Languages, Literatures, & Lingu
         istics',
                'Environmental Toxicology', 'Demography & Population Studies',
                'Engineering Science', 'Biochemical Engineering',
                'Other Biological & Biomedical Sciences', 'Health Policy Analysis',
                'General Dance', 'Environmental Health Engineering',
                'General Classical Language, Literature, & Linguistics',
                'Other Legal Professions & Studies', 'Neuroscience',
                'General Special Education & Teaching',
                'Engineering Physics & Applied Physics', 'Endocrinology',
                'Planetary Astronomy & Science', 'Other Education',
                'Scandanavian Languages, Literatures, & Linguistics',
                'Dutch/Flemish Language & Literature', 'Biochemistry',
                'Biophysics', 'Toxicology', 'Judaic Studies',
                'General Atmospheric Sciences & Meteorology',
                'General East Asian Languages, Literatures, & Linguistics',
                'General Romance Languages, Literatures, & Linguistics',
                'Latin Language & Literature', 'Religious Studies',
                'Other Anthropology', 'Agricultural Engineering',
                'Other Electrical, Electronics, & Communications Engineering',
                'Exercise Physiology', 'Computational Biology',
                'Topology & Foundations',
                'Physiological Psychology & Psychobiology',
                'Other Registered Nursing, Nursing Administration, Nursing Research,
         & Clinical Nursing',
                'General Physiology', 'Other Public Health', 'Human Biology',
                'Medicine', 'Business & Managerial Economics',
                'Other Microbiological Science & Immunology',
                'Developmental Economics & International Development',
                'Other Mathematics', 'Dentistry', 'Spanish Language & Literature',
                'Other Linguistic, Comparative, Related Language Studies & Service
         s',
                'General Art Studies', 'Other Geography',
                'Other Design & Applied Arts', 'Other Library Science',
                'Other Biomathematics, Bioinformatics, & Computational Biology',
                'Social & Philosophical Foundations of Education',
                'Applied Economices', 'Labor Studies',
                'General Computer Engineering', 'Materials Engineering',
                'US English Literature', 'Geography And Environmental Studies',
                'Other Management Sciences & Quantitative Methods', 'Ecology',
                'Computational Mathematics', 'Other Fine Arts & Art Studies',
                'Aerospace, Aeronautical, & Astronautical Engineering',
                'Musicology & Ethnomusicology', 'Other Statistics',
                'Nursing Practice', 'Oral Biology & Maxillofacial Pathology',
                'Marine Biology & Biological Oceonography',
                'General Music Performance', 'Korean Language & Literature',
                'Molecular Biochemistry', 'Molecular Biology',
                'Other Atmospheric Sciences & Meteorology',
                'Arabic Language & Literature', 'Molecular Physiology',
                'Bioinformatics', 'Archeology',
                'Music History, Literature, & Theory', 'Genetic Counseling',
                'Community Health & Preventative Medicine',
                'Other East Asian Languages, Literatures, & Linguistics',
                'Mathematics & Statistics',
                'Geographic Information Science & Cartography',
                'Music Teacher Education', 'Jazz & Jazz Studies',
                'Molecular Toxicology', 'Geochemistry',
                'Russian Language & Literature', 'Molecular Pharmacology',
                'Astronomy', 'Music Management', 'Medical Scientist',
                'Other Computer Engineering', 'Human & Medical Genetics',
                'Other Astronomy & Astrophysics', 'Structural Engineering',
                'Manufacturing Engineering',
                'Geological & Geophysical Engineering',
                'General Germanic Languages, Literatures, & Linguistics',
                'Other Liberal Arts & Sciences, General Studies, & Humanities',
                'Pathology & Experimental Pathology',
                'Foreign Language Teacher Education',
                'Portuguese Language & Literature',
                'Other Classical Languages, Literature, & Linguistics',
                'Other Legal Research & Advanced Professional Studies',
                'Islamic Studies', 'Health Services Administration',
                'Social Psychology', 'Criminology', 'Accounting',
                'Pharmaceutical Sciences', 'Informatics', 'Nursing School',
                'Neurobiology & Anatomy', 'Forensic Psychology',
                'Computer Software Engineering', 'Legal Studies & Jurisprudence',
                'Educational Evaluation & Research', 'Human Computer Interaction',
                'General Social Sciences', 'Biochemistry & Molecular Biology',
                'Entrepreneurial Studies', 'Mathematics & Computer Science',
                'Information Resources Management', 'Biotechnology',
                'Cognitive Psychology & Psycholinguistics', 'Other Physics',
                'Public Health Nurse',
                'Modeling, Virtual Environments, & Simulation',
                'Biology Teacher Education', 'Neurobiology & Behavior',
                'Family Practice Nurse', 'Microbiology & Immunology',
                'Ecology & Evolutionary Biology',
                'Health Care Administration & Management', 'Musical Theatre',
                'Pharmacology',
                'Other Ecology, Evolution, Systematics, & Population Biology',
                'General Computer & Information Sciences',
                'Computer Systems Networking & Telecommunications',
                'Creative Writing', 'General Genetics',
                'Other Pharmacy, Pharmaceutical Sciences, & Administration',
                'Cellular Biology & Histology', 'Anatomy',
                'Cultural Studies, Critical Theory, & Analysis',
                'Other Visual & Performing Arts',
                'Medical Microbiology & Bacteriology', 'Other Philosophy',
                'Applied Linguistics', 'Other Applied Mathematics',
                'Systems Science & Theory', 'Other Chemistry',
                'Pharmacology & Toxicology', 'Other Music', 'Geriatric Nurse',
                'Other Economics', 'Other Political Science & Government',
                'Experimental Psychology',
                'Other Cellular Biology & Anatomical Sciences',
                'Other Biochemistry, Biophysics, & Molecular Biology', 'Zoology',
                'Aquatic Biology', 'Other Psychology', 'Other History',
                'Data Science, Other', 'Hydrology & Water Resources Science',
                'Music Theory & Composition', 'Liberal Arts & Sciences',
                'Medieval & Renaissance Studies',
                'Other Communication Disorders Sciences & Services',
                'International Public Health', 'Clinical Psychology',
                'Other Neurobiology & Neurosciences',
                'General Human Development & Family Studies',
                'General Visual & Performing Arts',
                'International Relations & Affairs', 'Information Technology',
                'Other English Language & Literature', 'Pharmacy',
                'Health & Medical Psychology', 'Developmental & Child Psychology',
                'Clinical & Industrial Drug Development', 'Animal Physiology',
                'General Biomedical Sciences', 'Systems Engineering',
                'Biological & Biosystems Engineering', 'US Government & Politics',
                'General Design & Visual Communications',
                'Information Technology Project Management', 'Health Law',
                'Audiology', 'Computational Science',
                'Mathematics Teacher Education', 'Other Physical Sciences',
                'International Economics', 'Technical Theatre Design & Technology',
                'Acting', 'Directing & Theatrical Production',
                'Substance Abuse Counseling',
                'Geotechnical & Geoenvironmental Engineering',
                'Computer Hardware Engineering', 'Playwriting & Screenwriting',
                'Other Dramatic Arts & Stagecraft', 'Behavioral Aspects of Health',
                'Other Literature', 'Chemical Physics',
                'Other Film & Photographic Arts',
                'Other Educational Assessment, Evaluation, & Research',
                'Community Organization & Advocacy',
                'Game & Interactive Media Design',
                'Other Computer & Information Technology Services Management',
                'Physical Sciences', 'Mathematical Statistics & Probability',
                'Documentary Production', 'Artificial Intelligence',
                'Other Computer Software & Media Applications',
                'Computational & Applied Mathematics',
                'General Science Teacher Education', 'Veterinary Medicine',
                'Other Physiology, Pathology, & Related Science',
                'Other Health Professions & Related Clinical Sciences',
                'General Health Services',
                'Other Public Administration & Social Service Professions',
                'Nursing Administration', 'Forensic Science & Technology',
                'Cell Physiology', 'Entomology', 'Maritime Sciences',
                'Forensic Science And Technology', 'General Apparel & Textiles',
                'Health Occupations Teacher Education', 'Human Nutrition',
                'Science, Technology, & Society',
                'Teacher Education & Professional Development',
                'Chemical & Biomolecular Engineering', 'Plant Pathology',
                'Clinical Laboratory Technologist',
                'Agricultural Teacher Education', 'Textile Sciences & Engineering',
                'Veterinary Preventative Medicine, Epidemiology, & Public Health',
                'Transportation & Highway Engineering', 'Immunology',
                'Animal Behavior & Ethology', 'Population Biology',
                'Medical Informatics', 'General Veterinary Sciences',
                'Other Civil Engineering',
                'English & Language Arts Teacher Education',
                'Medicinal & Pharmaceutical Chemistry', 'Humanistic Studies',
                'Behavioral Sciences', 'Cinematography & Film Production',
                'General Finance', 'Public Administration',
                'General Foreign Languages & Literatures', 'Other Dance',
                'Evolutionary Biology',
                'Celtic Languages, Literatures, & Linguistics', 'Buddhist Studies',
                'Other Religious Studies', 'Other Geological & Earth Sciences',
                'Other Germanic Languages, Literatures, & Linguistics',
                'Ancient/Classical Greek Language & Literature',
                'Molecular Genetics', 'Other Materials Sciences',
                'Physical & Biological Anthropology', 'Physics Teacher Education',
                'Physical Education Teaching & Coaching',
                'Mechatronics, Robotics, & Automation Engineering'], dtype=object)
```

In [13]:
```python
df_cs = df[df['Major'].str.contains("Computer")]
```

In [14]:
```python
#general trend between genders
gender_yearly_trend = df_cs.groupby(['Year', 'Gender'])['Completions'].sum()
plt.figure(figsize=(10, 6))
sns.lineplot(data=gender_yearly_trend, x='Year', y='Completions', hue='Gende
Out[14]: <Axes: xlabel='Year', ylabel='Completions'>
```



```
In [15]: tech_hub_universities = ["University of California-Berkeley", "University of
                                   "University of California-Davis", "University of C
                                   "University of California-Los Angeles (110662)", "
         non_tech_hub_universities = ["University of California-Santa Barbara",
                                       "University of California-Merced", "University

         df_cs['Near_tech_Hub'] = df_cs['University'].apply(lambda x: "Yes" if x in t
         grouped_df = df_cs.groupby(['Year', 'Gender', 'Near_tech_Hub'])['Completions
         grouped_df
```

```
C:\Users\owent\AppData\Local\Temp\ipykernel_5920\2721017212.py:7: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_cs['Near_tech_Hub'] = df_cs['University'].apply(lambda x: "Yes" if x in
tech_hub_universities else "No")
```

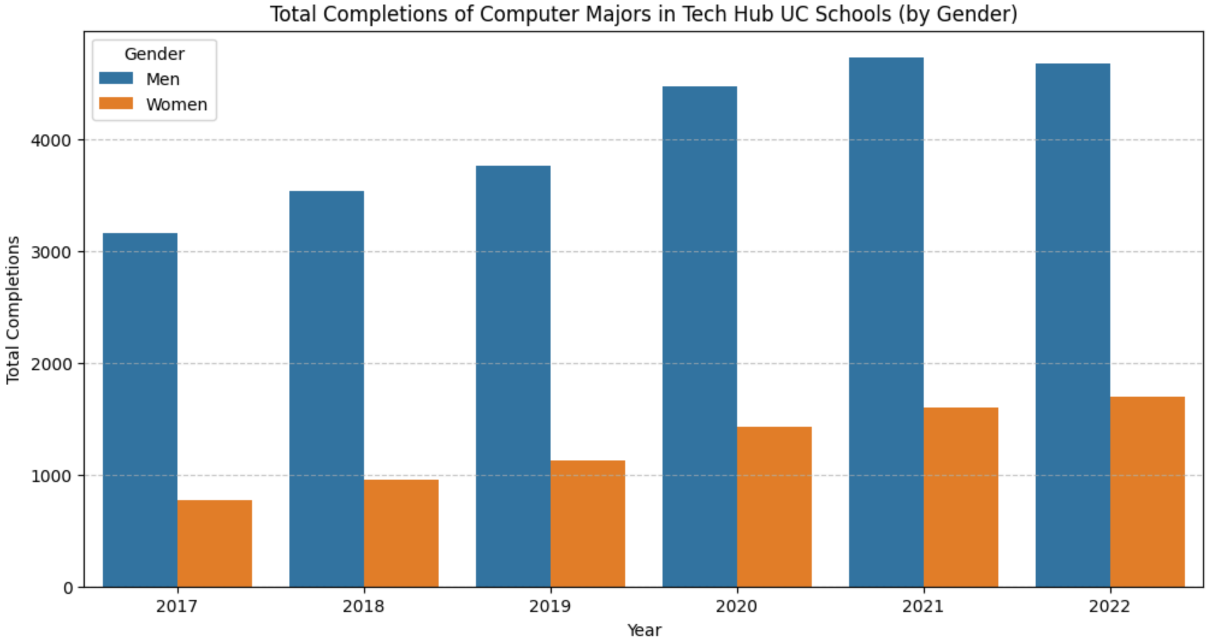
Out[15]:

	Year	Gender	Near_tech_Hub	Completions
0	2017	Men	No	387
1	2017	Men	Yes	3158
2	2017	Women	No	79
3	2017	Women	Yes	769
4	2018	Men	No	498
5	2018	Men	Yes	3542
6	2018	Women	No	73
7	2018	Women	Yes	957
8	2019	Men	No	505
9	2019	Men	Yes	3764
10	2019	Women	No	105
11	2019	Women	Yes	1131
12	2020	Men	No	591
13	2020	Men	Yes	4477
14	2020	Women	No	127
15	2020	Women	Yes	1428
16	2021	Men	No	730
17	2021	Men	Yes	4736
18	2021	Women	No	176
19	2021	Women	Yes	1600
20	2022	Men	No	732
21	2022	Men	Yes	4679
22	2022	Women	No	152
23	2022	Women	Yes	1697

In [16]:

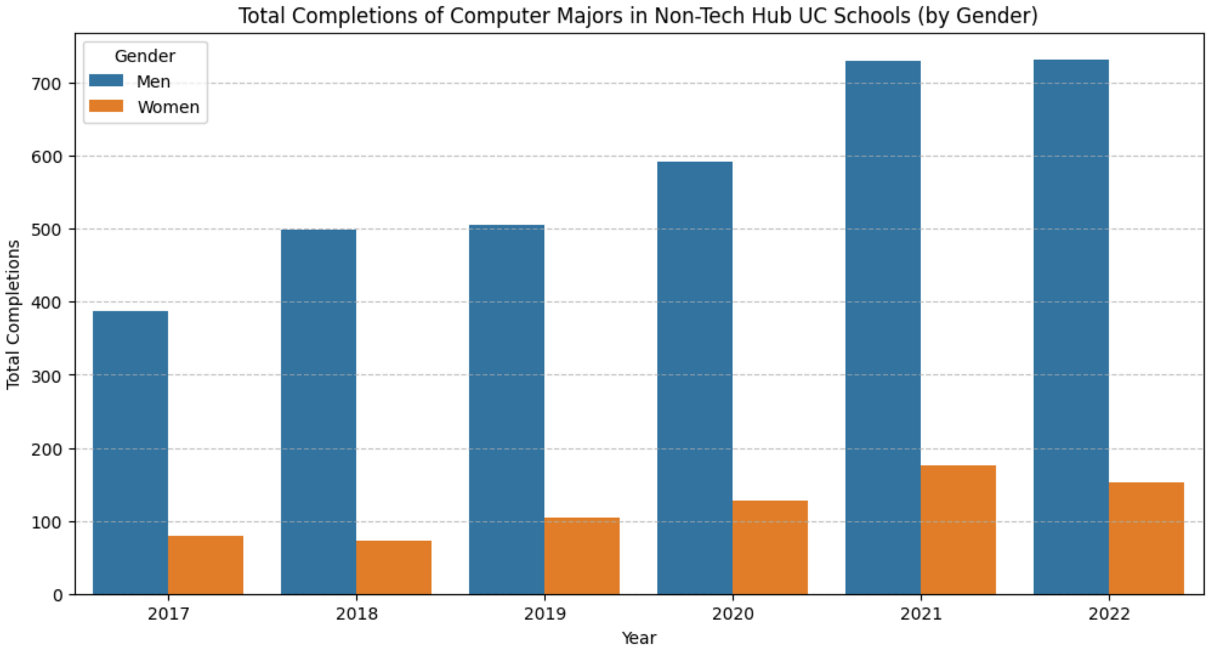
```python
tech_hub_data = grouped_df[grouped_df['Near_tech_Hub'] == "Yes"]

plt.figure(figsize=(12, 6))
sns.barplot(data=tech_hub_data, x='Year', y='Completions', hue='Gender')

# Customize the plot
plt.title("Total Completions of Computer Majors in Tech Hub UC Schools (by G
plt.xlabel("Year")
plt.ylabel("Total Completions")
plt.legend(title="Gender")
plt.grid(axis='y', linestyle="--", alpha=0.7)
```

```
# Show the plot
plt.show()
```



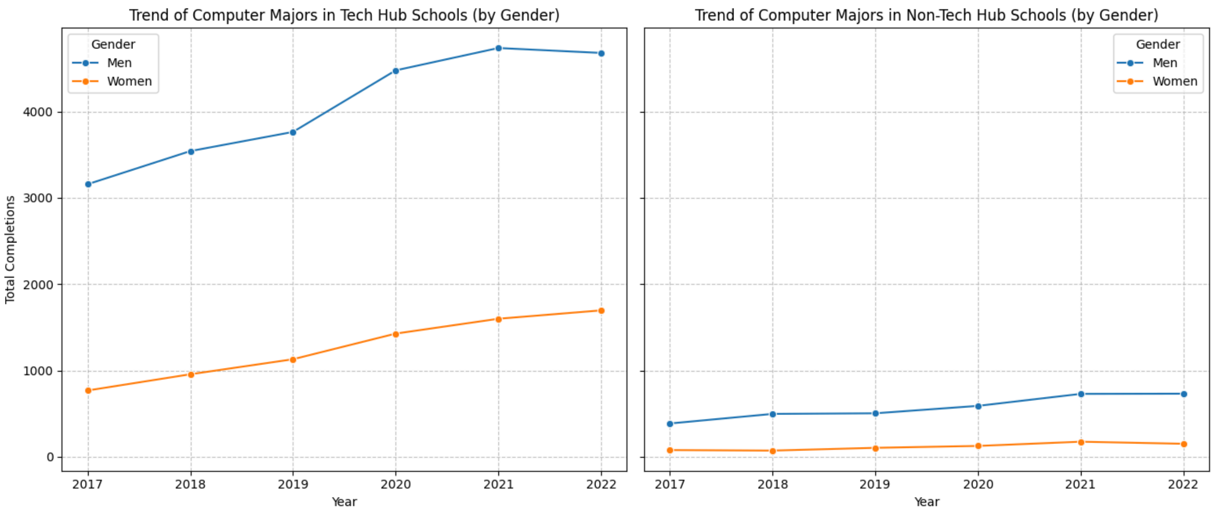
In [17]:

```
non_tech_hub_data = grouped_df[grouped_df['Near_tech_Hub'] == "No"]

plt.figure(figsize=(12, 6))
sns.barplot(data=non_tech_hub_data, x='Year', y='Completions', hue='Gender')

# Customize the plot
plt.title("Total Completions of Computer Majors in Non-Tech Hub UC Schools (
plt.xlabel("Year")
plt.ylabel("Total Completions")
plt.legend(title="Gender")
plt.grid(axis='y', linestyle="--", alpha=0.7)

# Show the plot
plt.show()
```



In [18]:

```
#line graphs comparsion for UC schools near tech hub and not near tech hub

fig, axes = plt.subplots(1, 2, figsize=(14, 6), sharey=True)

# Plot for Tech Hub Schools
sns.lineplot(data=tech_hub_data, x='Year', y='Completions', hue='Gender', ma
axes[0].set_title("Trend of Computer Majors in Tech Hub Schools (by Gender)"
axes[0].set_xlabel("Year")
axes[0].set_ylabel("Total Completions")
axes[0].legend(title="Gender")
axes[0].grid(True, linestyle="--", alpha=0.7)

# Plot for Non-Tech Hub Schools
sns.lineplot(data=non_tech_hub_data, x='Year', y='Completions', hue='Gender'
axes[1].set_title("Trend of Computer Majors in Non-Tech Hub Schools (by Genc
axes[1].set_xlabel("Year")
axes[1].set_ylabel("Total Completions")
axes[1].legend(title="Gender")
axes[1].grid(True, linestyle="--", alpha=0.7)

# Display the plots
plt.tight_layout()
plt.show()
```



Interpretation

Our exploratory data analysis confirms the expected trend that universities near major tech hubs produce a higher number of computer-related degree completions compared to those located outside of these hubs.

For **male graduates**, the total number of completions at **tech hub** campuses increased from **2,525 in 2017 to 3,656 in 2022**, which significantly outpaces the numbers at **non-tech** hub campuses, which ranged from **1,020 to 1,755** over the same period. This suggests that universities located near major technology centers may attract more male students pursuing computer-related degrees, possibly due to stronger industry connections, internship opportunities, or the overall reputation of these institutions in the tech field.

A similar trend is observed for **female graduates**, where completions at **tech hub** universities grew from **638 to 1,376**, whereas **non-tech hub** universities saw lower numbers, ranging from **210 to 473**. This indicates that, for both genders, proximity to tech hubs is associated with higher degree completion rates in computer-related fields.

Overall, **these findings align with expectations**, reinforcing the idea that universities near tech hubs serve as key talent pipelines for the technology industry. The larger enrollment and completion numbers at these institutions may reflect the influence of industry demand, specialized academic programs, and stronger recruitment efforts by tech companies.

Section 2 of EDA - Compare the number of computer completion in UC schools near Tech hub and not near Tech hub within Gender

```
In [19]: df_cs
```

Out[19]:

	Major	Year	University	Gender	Completions	Near_tech_Hub
24	Other Computer & Information Sciences	2022	University of California-Berkeley	Women	327	Yes
44	Other Computer & Information Sciences	2021	University of California-Berkeley	Women	279	Yes
53	Computer Science	2022	University of California-Berkeley	Women	263	Yes
63	Computer Science	2020	University of California-Berkeley	Women	239	Yes
64	Computer Science	2021	University of California-Berkeley	Women	235	Yes
...	...	...	...	...	...	...
179	General Computer Engineering	2018	University of California-Riverside	Men	53	No
180	General Computer Engineering	2020	University of California-Riverside	Men	53	No
212	General Computer Engineering	2017	University of California-Riverside	Men	46	No
213	General Computer Engineering	2022	University of California-Riverside	Men	46	No
221	General Computer Engineering	2019	University of California-Riverside	Men	44	No

340 rows × 6 columns

In [20]:

```
uc_trend = df_cs.groupby(['Year', 'Gender', 'University', 'Near_tech_Hub'])[
uc_trend
```

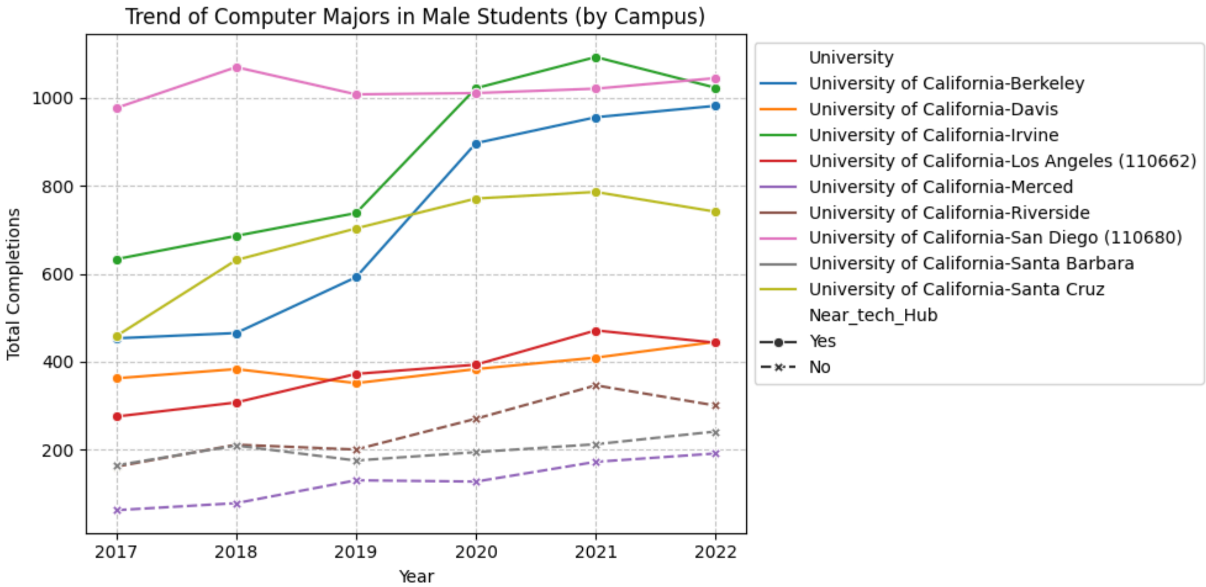
Out[20]:

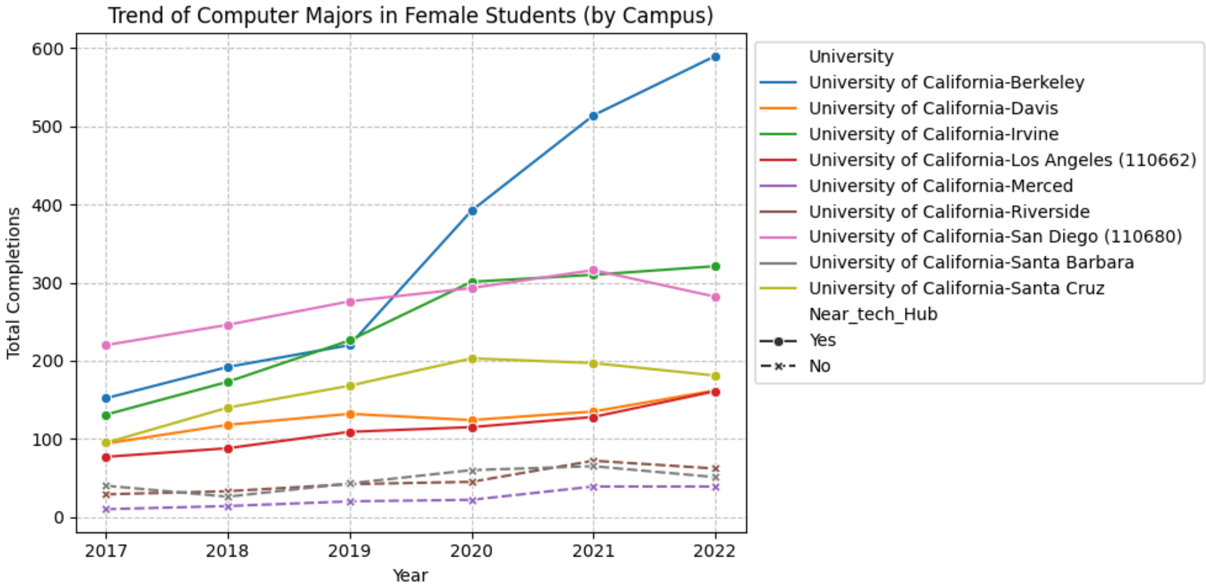
	Year	Gender	University	Near_tech_Hub	Completions
0	2017	Men	University of California-Berkeley	Yes	453
1	2017	Men	University of California-Davis	Yes	362
2	2017	Men	University of California-Irvine	Yes	633
3	2017	Men	University of California-Los Angeles (110662)	Yes	275
4	2017	Men	University of California-Merced	No	62
...	...	...	...	...	...
103	2022	Women	University of California-Merced	No	39
104	2022	Women	University of California-Riverside	No	62
105	2022	Women	University of California-San Diego (110680)	Yes	282
106	2022	Women	University of California-Santa Barbara	No	51
107	2022	Women	University of California-Santa Cruz	Yes	181

108 rows × 5 columns

In [21]:

```
male_data = uc_trend[uc_trend['Gender'] == "Men"]
female_data = uc_trend[uc_trend['Gender'] == "Women"]

# Plot for Male Students
plt.figure(figsize=(10, 5))
sns.lineplot(data=male_data, x='Year', y='Completions', hue='University', st
plt.title("Trend of Computer Majors in Male Students (by Campus)")
plt.xlabel("Year")
plt.ylabel("Total Completions")
plt.grid(True, linestyle="--", alpha=0.7)
plt.legend(loc='upper left', bbox_to_anchor=(1, 1))  # Moves legend outside
plt.tight_layout()
plt.show()

# Plot for Female Students
plt.figure(figsize=(10, 5))
sns.lineplot(data=female_data, x='Year', y='Completions', hue='University',
plt.title("Trend of Computer Majors in Female Students (by Campus)")
plt.xlabel("Year")
plt.ylabel("Total Completions")
plt.grid(True, linestyle="--", alpha=0.7)
plt.legend(loc='upper left', bbox_to_anchor=(1, 1))  # Moves legend outside
plt.tight_layout()
plt.show()
```





Interpretation

For both male and female students, the total number of completions is **consistently higher at tech hub universities** compared to non-tech hub universities. This pattern hints that universities located near major technology centers have a stronger presence in computer-related education.

Additionally, **across all campuses, the number of male graduates surpasses that of female graduates** in computer majors. The gap between male and female completions is present at both tech hub and non-tech hub universities, acting as an evidence saying that the persistent gender disparity in these fields.

One notable exception is UC Irvine, despite being categorized as a non-tech hub university, exhibits higher completion numbers than some tech hub universities. This might be suggesting that UC Irvine may have factors contributing to its strong computer science programs, such as specialized curricula, faculty expertise, or industry partnerships.

Overall, these findings further emphasize the influence of proximity to tech hubs on computer-related degree completions, while also highlighting the gender imbalance in these fields across all institutions.

Section 3 of EDA - Specific Major Counts between Gender

In [22]:

```
# Group data by Major and Gender, then sum the completions
major_gender_counts = df.groupby(['Major', 'Gender'])['Completions'].sum().r

major_gender_counts
```

Out[22]:

	Major	Gender	Completions
0	Accounting	Men	237
1	Accounting	Women	459
2	Acting	Men	25
3	Acting	Women	23
4	Actuarial Science	Men	594
...	...	...	...
683	Veterinary Preventative Medicine, Epidemiology...	Women	11
684	Vision Science & Physiological Optics	Men	15
685	Vision Science & Physiological Optics	Women	28
686	Zoology	Men	115
687	Zoology	Women	451

688 rows × 3 columns

In [23]:

```
# Pivot the table
major_gender_pivot = major_gender_counts.pivot(
    index = 'Major',
    columns = 'Gender',
    values = 'Completions'
).fillna(0)

major_gender_pivot
```

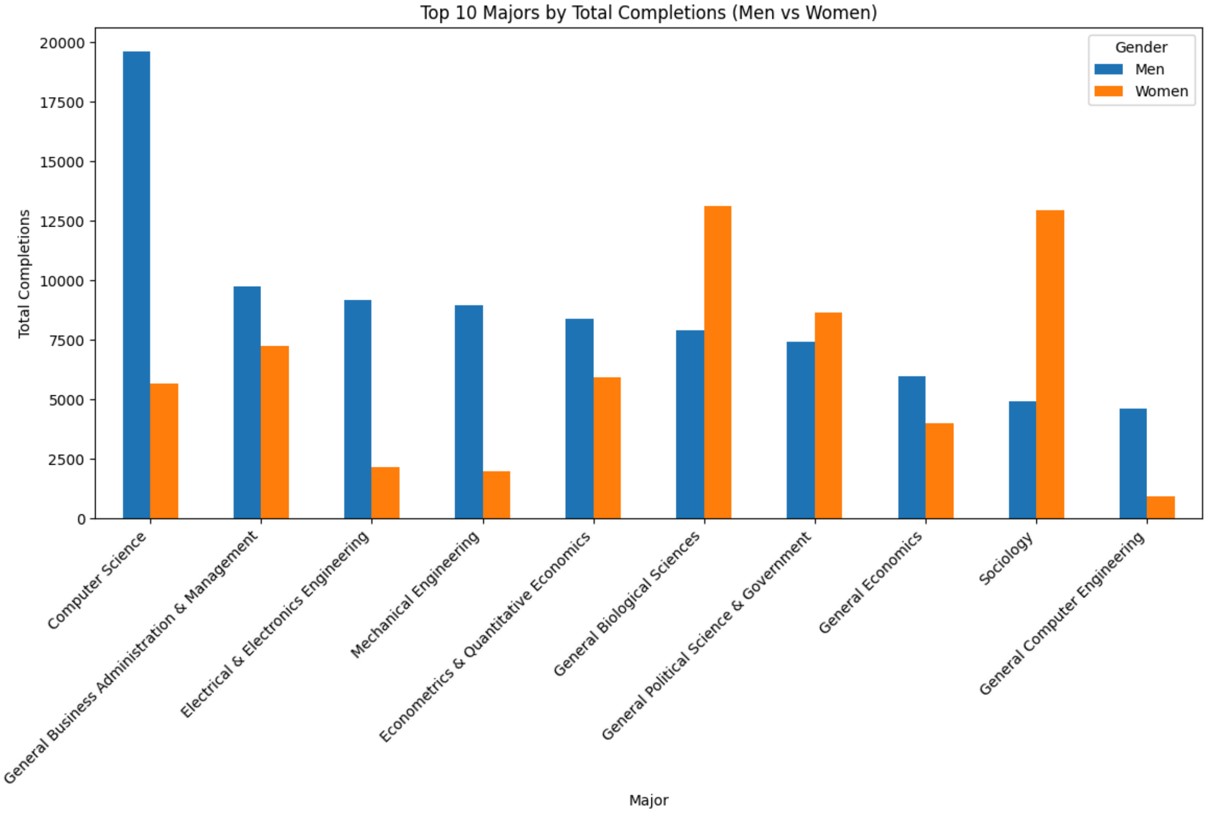
Out[23]:

Gender	Men	Women
Major		
Accounting	237.0	459.0
Acting	25.0	23.0
Actuarial Science	594.0	367.0
Aerospace, Aeronautical, & Astronautical Engineering	1630.0	298.0
Agricultural Engineering	39.0	34.0
...	...	...
Urban Studies	328.0	455.0
Veterinary Medicine	71.0	331.0
Veterinary Preventative Medicine, Epidemiology, & Public Health	8.0	11.0
Vision Science & Physiological Optics	15.0	28.0
Zoology	115.0	451.0

353 rows × 2 columns

In [24]:

```python
major_gender_pivot_sorted = major_gender_pivot.sort_values(by = 'Men', ascer

# Plot a bar chart for the top 10 majors
major_gender_pivot_sorted.head(10).plot(kind = 'bar', figsize = (12,8))
plt.title('Top 10 Majors by Total Completions (Men vs Women)')
plt.xlabel('Major')
plt.ylabel('Total Completions')
plt.xticks(rotation = 45, ha = 'right')
plt.tight_layout()
plt.show()
```

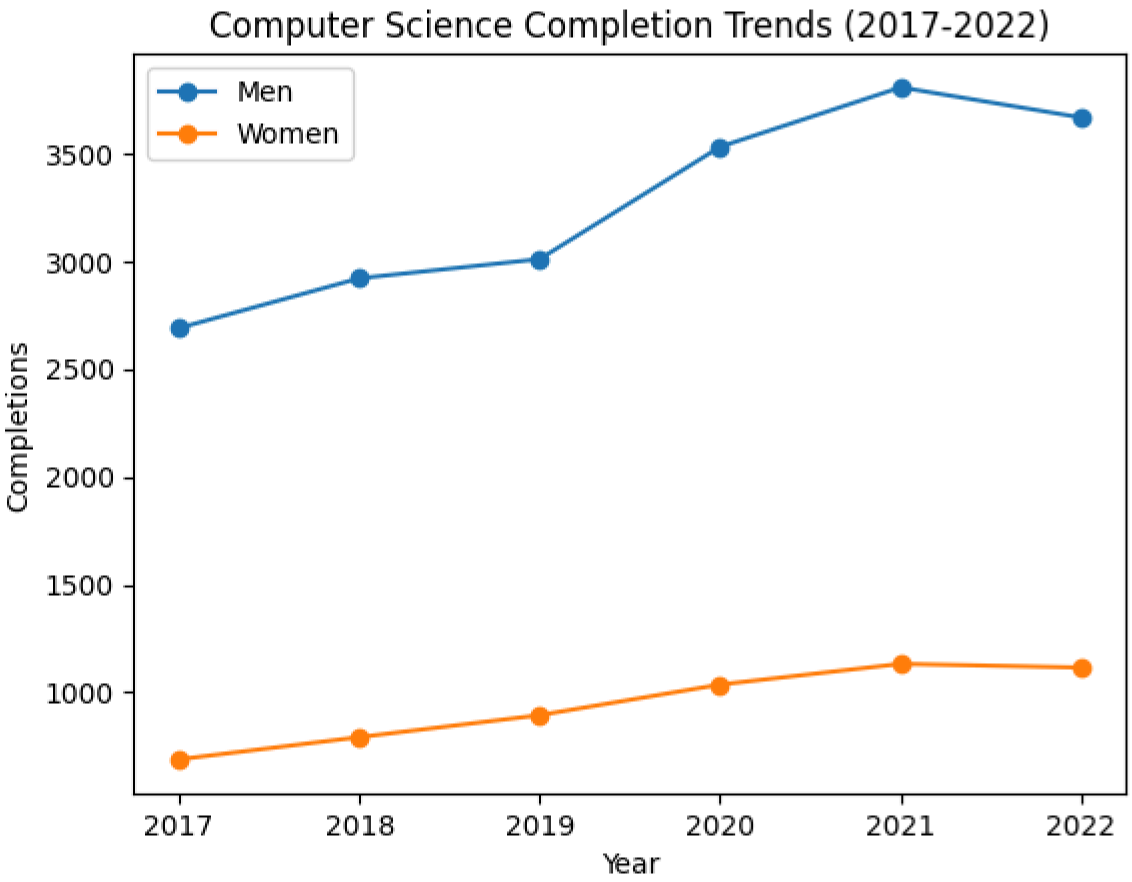


Interpretation

From the tables above, men are more likely to choose **STEM majors** such as Computer Science, Electrical & Electronics Engineering, Mechanical Engineering, and Quantitative Economics, with a particularly significant preference for Computer Science. Additionally, they are more inclined to select Social Science majors like Business Administration, Political Science, and General Economics.

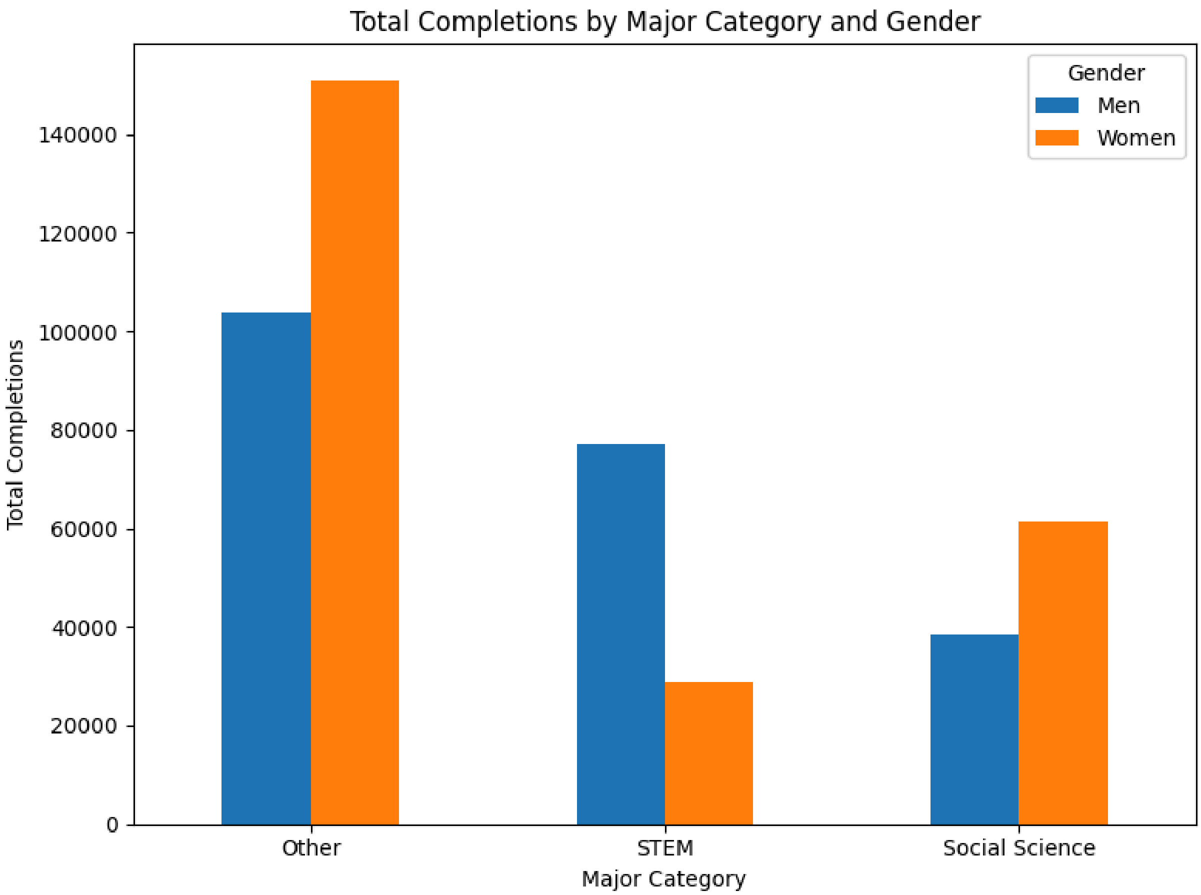
In [25]:

```
major_year_gender = df.groupby(['Major', 'Gender', 'Year'])['Completions'].s

# For instance, filter for Computer Science to visualize its trend
cs_trend = major_year_gender[major_year_gender['Major'] == 'Computer Science

# Plot the trend over time for Computer Science completions by gender
for gender in cs_trend['Gender'].unique():
    subset = cs_trend[cs_trend['Gender'] == gender]
    plt.plot(subset['Year'], subset['Completions'], marker = 'o', label = ge

plt.title('Computer Science Completion Trends (2017-2022)')
plt.xlabel('Year')
plt.ylabel('Completions')
plt.legend()
plt.xticks(cs_trend['Year'].unique())
plt.show()
```



In [26]:

```
def classify_major(major):
    if any(keyword in major for keyword in ['Computer', 'Engineering', 'Math
        return 'STEM'
    elif any(keyword in major for keyword in ['Economics', 'History', 'Socia
        return 'Social Science'
    else:
        return 'Other'

df['Major_Category'] = df['Major'].apply(classify_major)

category_counts = df.groupby(['Major_Category', 'Gender'])['Completions'].su

# Pivot for visualization
category_pivot = category_counts.pivot(
    index = 'Major_Category',
    columns = 'Gender',
    values = 'Completions'
).fillna(0)

# Plot the grouped counts for each category
category_pivot.plot(kind = 'bar', figsize = (8,6))
plt.title('Total Completions by Major Category and Gender')
plt.xlabel('Major Category')
plt.ylabel('Total Completions')
plt.xticks(rotation = 0)
plt.tight_layout()
plt.show()
```



Interpretation

From the tables above, women are more likely to choose **STEM majors** like General Biological Sciences, Econometrics & Quantitative Economics, Computer Science, and Cellular & Molecular Biology. They are also more likely to choose Sociology, General Psychology, and Political Science for **Social Science**.

Hypothesis Testing and Regression Analysis

In this part of our analysis, we are interested in understanding how various variables influence the choice of different fields of study. To do this, we will do regression analysis and hypothesis testing.

We must check our assumptions of normality, independence, homoscadasticity, and linearity are satistified in order to ensure our regression analysis is valid.

We will check every one of these assumptions for each ordinary linear regression analysis using the below functions.

In [27]:

```
# function to check for normality using Shapiro-Wilk test
from scipy.stats import shapiro
def normality(model, alpha = 0.05):
    sw_test = shapiro(model.resid)
    if sw_test.pvalue > alpha:
        print(f'Fail to reject H0, the data is likely normally distributed (
    else:
        print(f'Reject H0, the data is likely not normally distributed (p-va
```

In [28]:

```
# function to check for hetroscedasticity and linearity by plotting residual

def hetro_and_linear(model):
    fitted = model.fittedvalues
    residuals = model.resid
    sns.scatterplot(x=fitted, y=residuals, alpha=0.7)
    plt.axhline(y=0, color='red', linestyle='--', linewidth=1)
    plt.xlabel("Fitted Values")
    plt.ylabel("Residual")
    plt.title("Residuals vs. Fitted Values")
    plt.show()
```

In [29]:

```
# function to check for independence by plotting the ACF fucntion against th
from statsmodels.tsa.stattools import acf
from statsmodels.graphics.tsaplots import plot_acf
def independence(model):
    residuals = model.resid

    plot_acf(residuals, lags=20, alpha=0.05)
    plt.title("Autocorrelation of Residuals (ACF Plot)")
    plt.xlabel("Lags")
    plt.ylabel("Autocorrelation")
    plt.show()
```

Impacts of gender, the year, and the University on Computer Science Major Completions

In [30]:

```
def label_cs(major):
    major_lower = major.lower()
    computer_keywords = ['computer']

    if any(keyword in major_lower for keyword in computer_keywords):
        return 'Computer Science'
    else:
        return 'Other'

df['Field'] = df['Major'].apply(label_cs)

df_stem = df[df['Field'] == 'Computer Science'].copy()

analysis = df_stem.groupby(
    ['University', 'Gender', 'Year'],
    as_index = False
)['Completions'].sum()
```

```
model = smf.ols(formula = 'Completions ~ C(Gender) + Year + C(University)',
```

In [31]:
```
analysis
```

Out[31]:

	University	Gender	Year	Completions
0	University of California-Berkeley	Men	2017	453
1	University of California-Berkeley	Men	2018	465
2	University of California-Berkeley	Men	2019	592
3	University of California-Berkeley	Men	2020	897
4	University of California-Berkeley	Men	2021	956
...	...	...	...	...
103	University of California-Santa Cruz	Women	2018	140
104	University of California-Santa Cruz	Women	2019	168
105	University of California-Santa Cruz	Women	2020	203
106	University of California-Santa Cruz	Women	2021	197
107	University of California-Santa Cruz	Women	2022	181

108 rows × 4 columns

Diagnostics

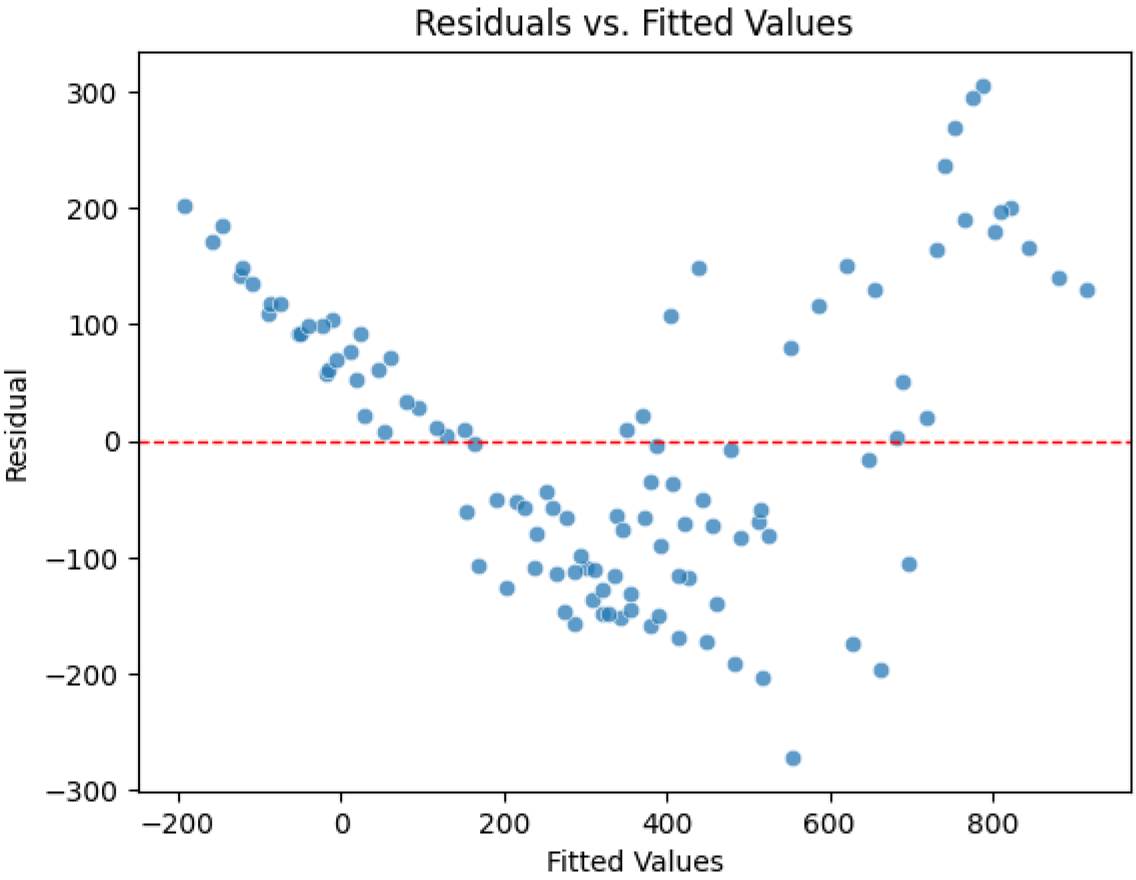
Normality

In [32]:
```
normality(model)
```

```
Reject H0, the data is likely not normally distributed (p-value=0.0182283349
61703452, alpha = 0.05)
```

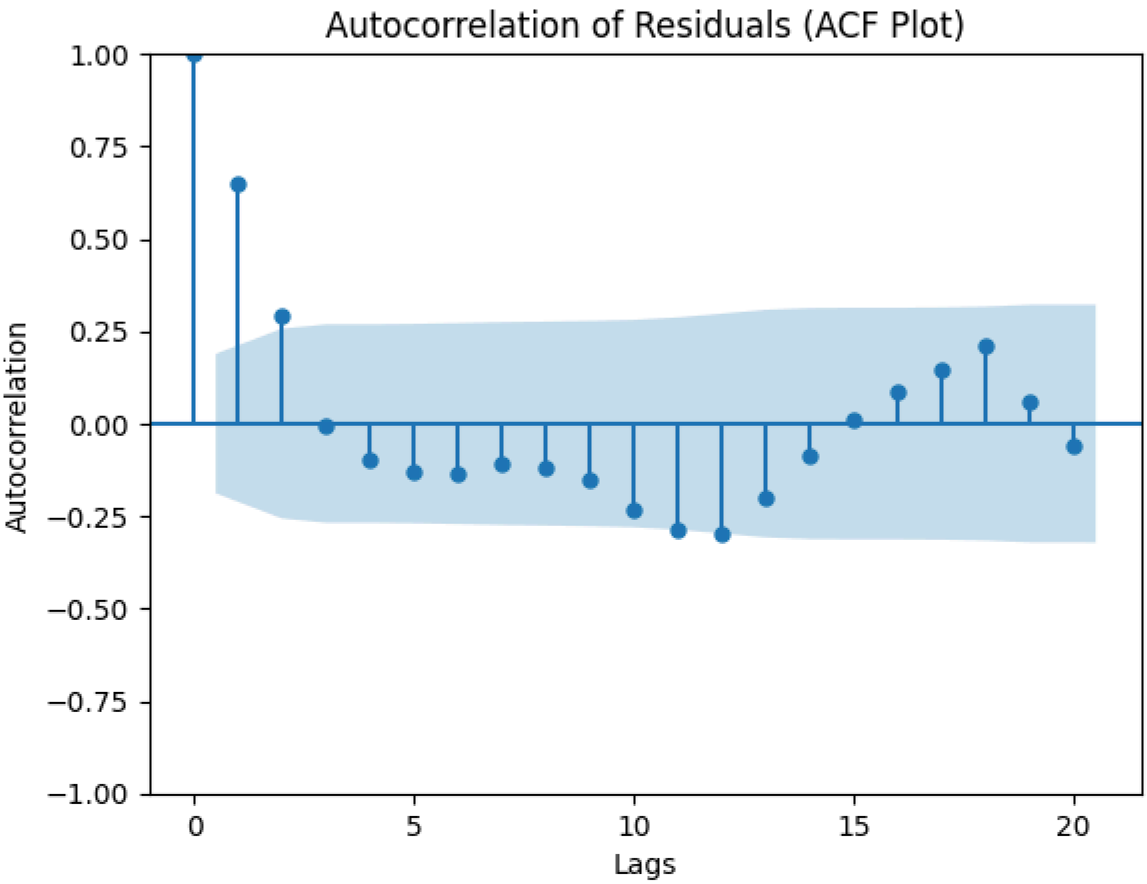
Heteroscedasticity and Linearity

In [33]:
```
hetro_and_linear(model)
```



Independence

```
In [34]: independence(model)
```



Diagnostics Conclusion

```
In [35]: print(analysis.head())
```

```
                          University Gender  Year  Completions
0  University of California-Berkeley    Men  2017          453
1  University of California-Berkeley    Men  2018          465
2  University of California-Berkeley    Men  2019          592
3  University of California-Berkeley    Men  2020          897
4  University of California-Berkeley    Men  2021          956
```

We can see according to the diagnostics above, our regression model fails every diagnostic. This may be because our response variable of majors that include the word "computer" in it has too much variability due to the small amount of data that have a "Computer" major as we see above. The "Computer" major may also have a non-linear relationship to a lot of the covariates. We should use a response variable that is more stable and has more data. What if we tried using every **STEM major** as our response variable?

Impacts of gender, the year, and the University on STEM Major Completions

In [36]:

```
def label_field(major):
    major_lower = major.lower()
    stem_keywords = ['math', 'biology', 'physics', 'chemistry', 'computer',

    if any(keyword in major_lower for keyword in stem_keywords):
        return 'STEM Major'
    else:
        return 'Other'

df['Field'] = df['Major'].apply(label_field)

df_stem = df[df['Field'] == 'STEM Major'].copy()
analysis = df_stem.groupby(
    ['University', 'Gender', 'Year'],
    as_index = False
)['Completions'].sum()

model = smf.ols(formula='Completions ~ C(University, Treatment("University c
```

Diagnostics

Noramlity

In [37]:

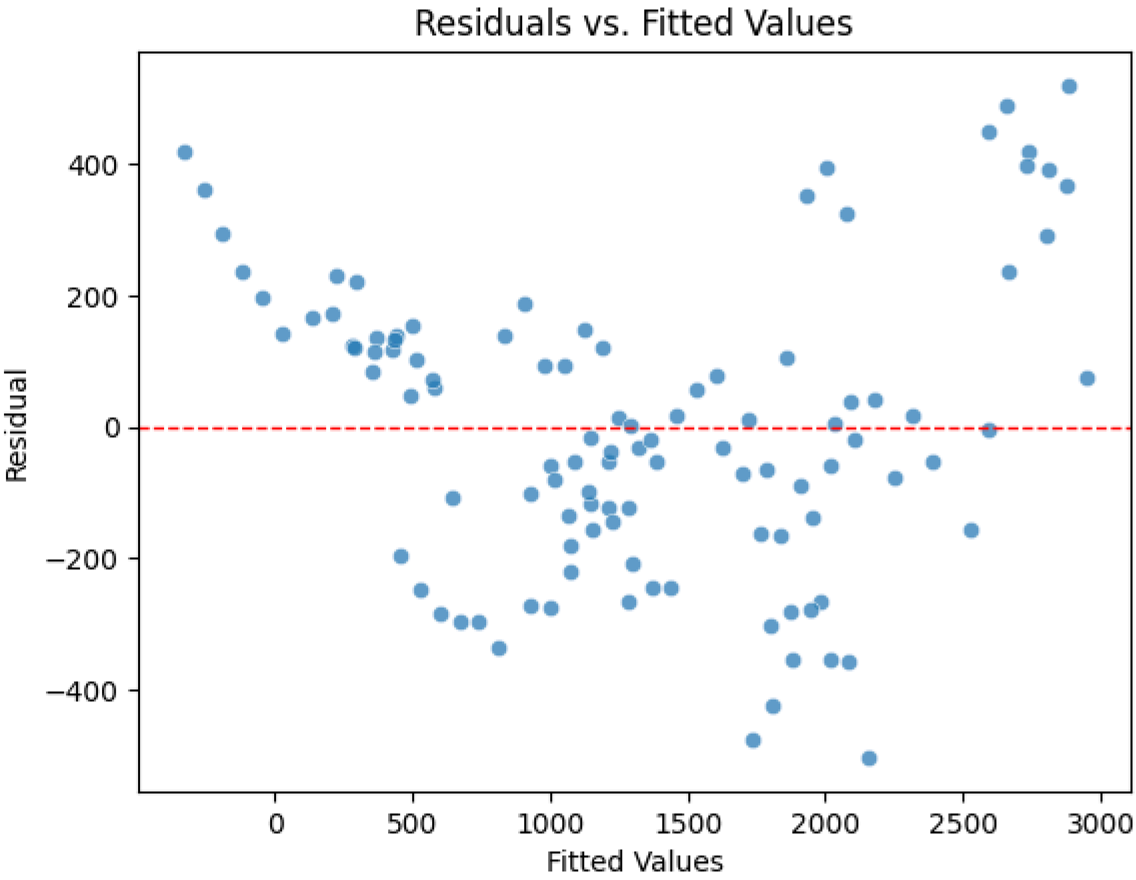
```
normality(model)
```

```
Fail to reject H0, the data is likely normally distributed (p-value=0.438586
2202176716, alpha = 0.05)
```

Heteroscedasticity and Linearity

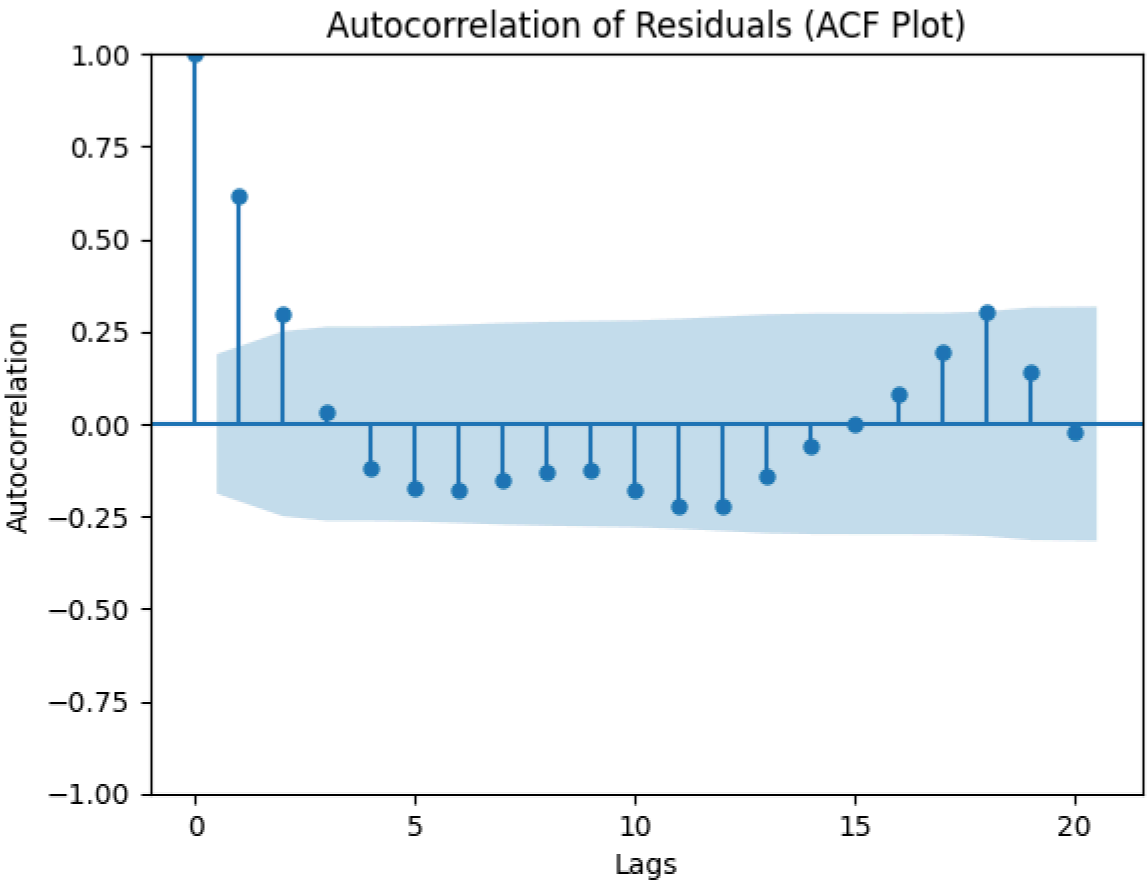
In [38]:

```
hetro_and_linear(model)
```



Independence

```
In [39]: independence(model)
```



Diagnostics Conclusion

According to our diagnostics above, our OLS model satifies every assumption except for independence. For the Shapiro-Wilk test, we obtained a p-value of around 0.43 meaning we fail to reject the null hypothesis according to our selected alpha value of 0.05. This means our residuals are most likely normally distributed. In the residuals vs. fitted values plot, we can see that there is no pattern forming and thus our errors have constant variance and our independent and dependent variables are linear.

We can see that in the ACF plot, we see at some lags, there is a positive correlation.

Further down in the analysis, we can use a different model other than OLS to better model our data. The model we will use is ARIMA since our data is time dependent to satisfy the independence diagnostic that we fail here with the OLS model.

Regression Analysis

In [40]:
```
print(analysis.head())
```

```
                          University Gender  Year  Completions
0  University of California-Berkeley    Men  2017         2368
1  University of California-Berkeley    Men  2018         2592
2  University of California-Berkeley    Men  2019         2903
3  University of California-Berkeley    Men  2020         3159
4  University of California-Berkeley    Men  2021         3202
```

Much more data to work with!

```
In [41]: print(model.summary())
```

```
                            OLS Regression Results
============================================================================
==
Dep. Variable:            Completions   R-squared:                       0.9
29
Model:                            OLS   Adj. R-squared:                  0.9
22
Method:                 Least Squares   F-statistic:                     12
7.4
Date:                Fri, 14 Mar 2025   Prob (F-statistic):           3.40e-
51
Time:                        20:44:06   Log-Likelihood:                -737.
61
No. Observations:                 108   AIC:                             149
7.
Df Residuals:                      97   BIC:                             152
7.
Df Model:                          10
Covariance Type:            nonrobust
============================================================================
============================================================================
====================================

coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------
----------------------------------------------------------------------------
------------------------------------
Intercept
-1.413e+05   2.69e+04     -5.259      0.000   -1.95e+05    -8.8e+04
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Berkeley]                 -65.4167     96.414     -0.67
8      0.499    -256.772     125.938
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Davis]                   -964.7500     96.414    -10.00
6      0.000   -1156.105    -773.395
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Irvine]                  -872.0833     96.414     -9.04
5      0.000   -1063.438    -680.728
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Los Angeles (110662)]  -554.6667     96.414     -5.75
3      0.000    -746.022    -363.312
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Merced]                 -2132.6667     96.414    -22.12
0      0.000   -2324.022   -1941.312
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Riverside]              -1662.2500     96.414    -17.24
1      0.000   -1853.605   -1470.895
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Santa Barbara]          -1575.0000     96.414    -16.33
6      0.000   -1766.355   -1383.645
C(University, Treatment("University of California-San Diego (110680)"))[T.Un
iversity of California-Santa Cruz]             -1512.3333     96.414    -15.68
6      0.000   -1703.688   -1320.978
C(Gender)[T.Women]
-788.3148     45.450    -17.345      0.000    -878.520    -698.109
Year
```

```
71.3540     13.306      5.362      0.000      44.944      97.763
==============================================================================
==
Omnibus:                        1.313   Durbin-Watson:                   0.7
64
Prob(Omnibus):                  0.519   Jarque-Bera (JB):                1.2
98
Skew:                           0.163   Prob(JB):                        0.5
23
Kurtosis:                       2.573   Cond. No.                     2.39e+
06
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The condition number is large, 2.39e+06. This might indicate that there
are
strong multicollinearity or other numerical problems.
```

Every covariate is statstically significant except for UCB as UCB's p-value is greater than 0.05.

Interpretation

Regression Analysis

The regression analysis aimed to explore the relationship between the number of STEM graduates and the predictors: university, gender, and year. The model was defined as follows:

Dependent Variable:

- Completions: Number of STEM degrees awarded.

Independent Variables:

- University: Categorical variable representing the UC campus.
- Gender: Binary variable (Men or Women).
- Year: Graduation year (2017–2022).

Model Diagnostics

1. **Normality:**

   The Shapiro-Wilk test produced a p-value of **0.4386**, suggesting that the residuals most likely follow a normal distribution (fail to reject $H_0$ at $\alpha$ = 0.05). This is an essential assumption for OLS regression, and the result indicates that the model's residuals do not significantly deviate from normality, supporting the validity of the regression results.

2. **Homoscedasticity and Linearity:**

   The residuals against fitted values plot showed no distinct pattern, indicating that the residuals' variance is consistent across the range of fitted values (homoscedasticity). Additionally, the absence of a noticeable trend in the plot suggests a linear relationship between the predictors and the response variable.

3. **Independence:**

   The ACF plot showed some autocorrelation at specific lags, indicating a potential breach of the independence assumption. This suggests that the residuals are not entirely independent, possibly due to the time-dependent nature of the data (e.g., trends over years).

Regression Results

The regression model achieved an R-squared value of **0.929**, indicating that around 92.9% of the variation in STEM completions is accounted for by the predictors. This high R-squared value suggests a strong fit of the model to the data, capturing most of the underlying trends. Key insights include:

1. **University:**

   All universities, except UC Berkeley, exhibited statistically significant differences in STEM completions compared to the reference (UC San Diego). For instance:

   - UC Davis had a coefficient of **-964.75 (p < 0.001)**, indicating significantly fewer STEM completions than UCSD.

   - UC Merced had the largest negative coefficient **(-2132.67, p < 0.001)**, showing the lowest number of STEM completions among the analyzed campuses.

   - UC Riverside and UC Santa Barbara also had significant negative coefficients, indicating fewer STEM completions compared to UCSD.

   UC Berkeley was the only university not statistically significant **(p = 0.499)**, suggesting its STEM completions are not significantly different from UCSD. This could be due to UC Berkeley's strong reputation and resources in STEM fields, which may offset any differences in proximity to tech hubs.

2. **Gender:**

The coefficient for `Gender[T.Women]` was **-788.31 (p < 0.001)**, which means that, on average, women graduate with fewer STEM degrees than men across UC campuses. This aligns with the broader trend of gender disparity in STEM fields. The large magnitude of the coefficient highlights the significant gap in STEM completions between men and women.

3. **Year:**

   The coefficient for `Year` was **71.35 (p < 0.001)**, which suggests a positive trend in STEM completions over time, reflecting an overall increase in STEM graduates across UC campuses from 2017 to 2022. The positive coefficient indicates that, on average, the number of STEM completions increased by approximately **71 students per year**. This trend could be attributed to growing interest in STEM fields, increased enrollment, or institutional efforts to promote STEM education.

4. **Intercept:**

   The intercept term **(-1.413e+05)** represents the predicted number of STEM completions when all predictors are 0. While this value is not directly interpretable in this context, it serves as a baseline for the model.

5. **Multicollinearity:**

   The condition number of **2.39e+06** is large, indicating potential multicollinearity or numerical instability in the model. This might arise from high correlations between predictors (e.g., year and university-specific trends) or the inclusion of too many categorical variables.

   Multicollinearity can increase the variance of the coefficient estimates and make the model less reliable. To address this, future analyses could:

   - Reduce the number of predictors through a method of stepwise selection.

   - Use regularization methods like Ridge Regression to penalize large coefficients and stabilize the model.

Forecasting STEM Major Completion by Gender

In [ ]:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.tsa.arima.model import ARIMA
import matplotlib.pyplot as plt

df_stem = df[df['Field'] == 'STEM Major'].copy()

analysis = df_stem.groupby(['Gender', 'Year'], as_index=False)['Completions'
```

```
pivot_data = analysis.pivot_table(index='Year', columns='Gender', values='Co

men_series = pivot_data['Men']
women_series = pivot_data['Women']

print("Fitting ARIMA for Men...")
model_men = ARIMA(men_series, order=(1,1,1))
results_men = model_men.fit()
print(results_men.summary(), "\n")

print("Fitting ARIMA for Women...")
model_women = ARIMA(women_series, order=(1,1,1))
results_women = model_women.fit()
print(results_women.summary(), "\n")

forecast_steps = 5

forecast_men = results_men.forecast(steps=forecast_steps)
print("Men’s Forecast:")
print(forecast_men, "\n")

forecast_women = results_women.forecast(steps=forecast_steps)
print("Women’s Forecast:")
print(forecast_women, "\n")
```

```
Fitting ARIMA for Men...
```

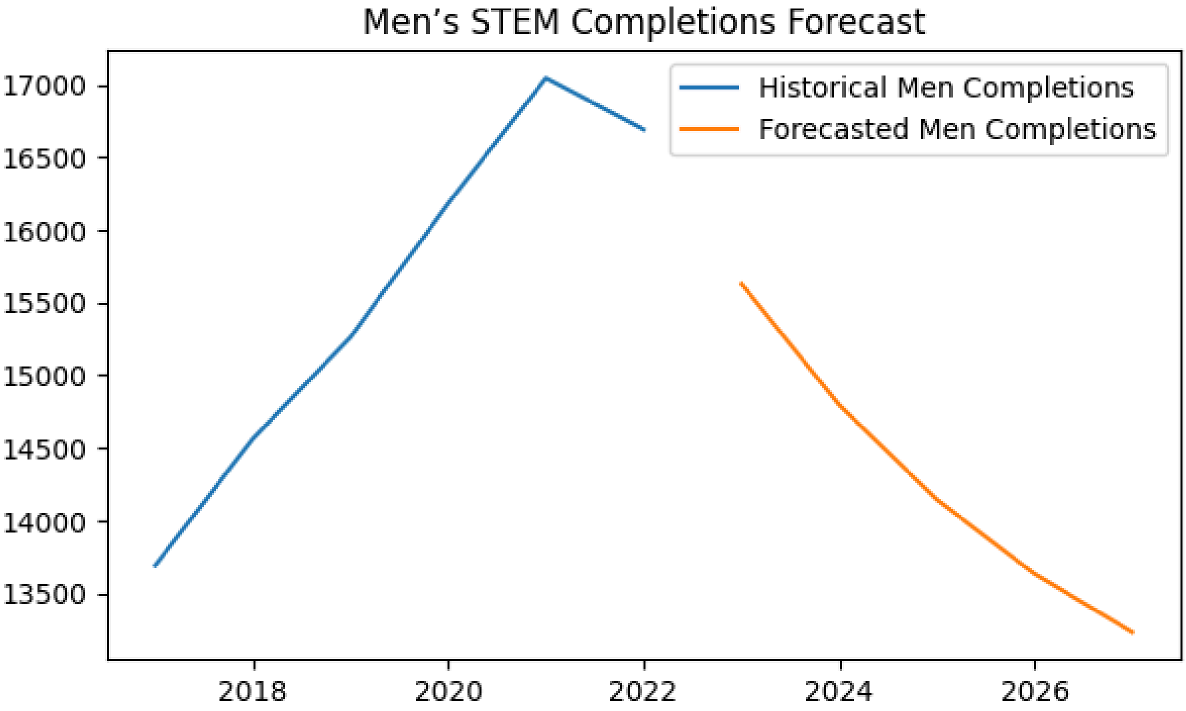
```
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: No frequency information w
as provided, so inferred frequency YS-JAN will be used.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: No frequency information w
as provided, so inferred frequency YS-JAN will be used.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: No frequency information w
as provided, so inferred frequency YS-JAN will be used.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\statespace\sarimax.py:966: UserWarning: Non-stationary starting
autoregressive parameters found. Using zeros as starting parameters.
  warn('Non-stationary starting autoregressive parameters'
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\statespace\sarimax.py:978: UserWarning: Non-invertible starting
MA parameters found. Using zeros as starting parameters.
  warn('Non-invertible starting MA parameters found.'
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\base\model.py:607: ConvergenceWarning: Maximum Likelihood optimizat
ion failed to converge. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: No frequency information w
as provided, so inferred frequency YS-JAN will be used.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: No frequency information w
as provided, so inferred frequency YS-JAN will be used.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: No frequency information w
as provided, so inferred frequency YS-JAN will be used.
  self._init_dates(dates, freq)
```

```
                               SARIMAX Results                                
============================================================================
==
Dep. Variable:                    Men   No. Observations:                   
6
Model:                 ARIMA(1, 1, 1)   Log Likelihood                 -58.0
17
Date:                Fri, 14 Mar 2025   AIC                            122.0
34
Time:                        20:58:24   BIC                            120.8
63
Sample:                    01-01-2017   HQIC                           118.8
90
                         - 01-01-2022                                       
Covariance Type:                  opg                                       
============================================================================
==
                 coef    std err          z      P>|z|      [0.025      0.97
5]
----------------------------------------------------------------------------
--
ar.L1          0.7836      0.025     31.294      0.000       0.735       0.8
33
ma.L1          0.9968      0.048     20.690      0.000       0.902       1.0
91
sigma2      1.674e+04   2.89e-06   5.79e+09      0.000    1.67e+04    1.67e+
04
============================================================================
=======
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):              
0.65
Prob(Q):                              0.83   Prob(JB):                      
0.72
Heteroskedasticity (H):              10.10   Skew:                          
-0.83
Prob(H) (two-sided):                  0.18   Kurtosis:                      
2.42
============================================================================
=======

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (compl
ex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.
73e+24. Standard errors may be unstable.

Fitting ARIMA for Women...
                               SARIMAX Results                                
============================================================================
==
Dep. Variable:                  Women   No. Observations:                   
6
Model:                 ARIMA(1, 1, 1)   Log Likelihood         -1847165085.1
01
Date:                Fri, 14 Mar 2025   AIC                     3694330176.2
02
```

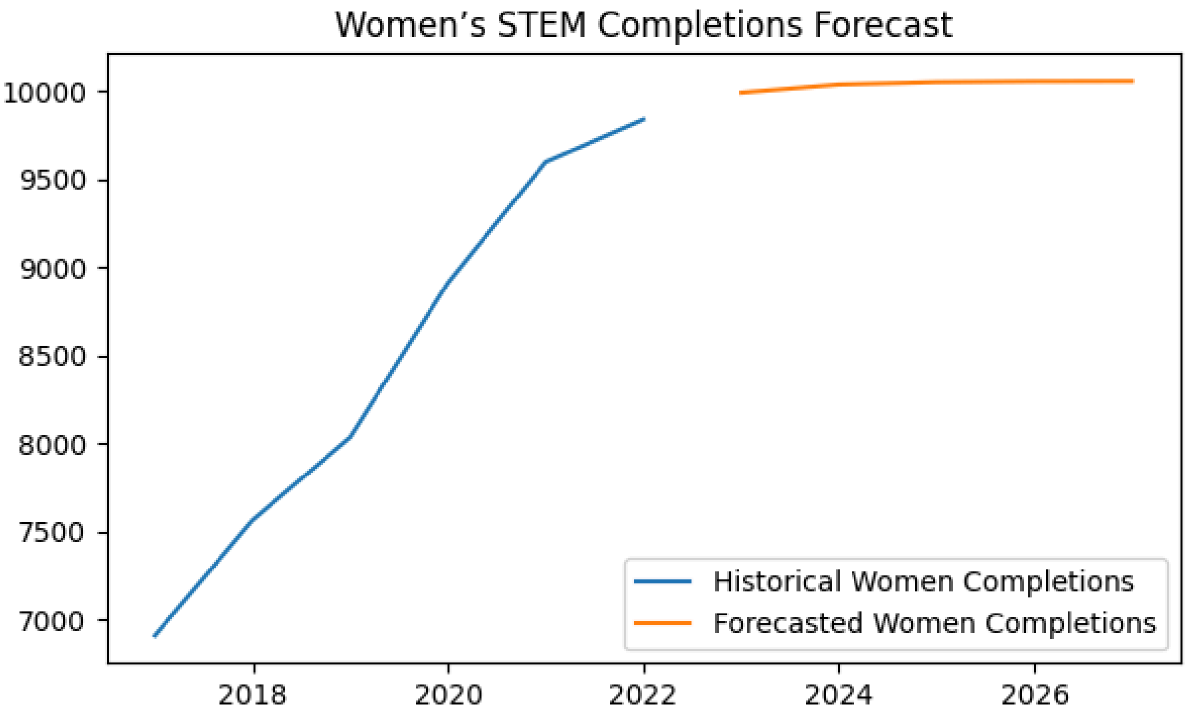
```
Time:                        20:58:25   BIC                   3694330175.0
31
Sample:                    01-01-2017   HQIC                  3694330173.0
58
                         - 01-01-2022
Covariance Type:                  opg
============================================================================
==
                 coef    std err          z      P>|z|      [0.025      0.97
5]
----------------------------------------------------------------------------
--
ar.L1          0.2998   3.14e-09   9.54e+07      0.000       0.300       0.3
00
ma.L1          0.9334   1.68e-08   5.57e+07      0.000       0.933       0.9
33
sigma2         0.0002   2.08e-12   7.28e+07      0.000       0.000       0.0
00
============================================================================
=======
Ljung-Box (L1) (Q):                   2.84   Jarque-Bera (JB):
0.56
Prob(Q):                              0.09   Prob(JB):
0.76
Heteroskedasticity (H):               0.09   Skew:
0.48
Prob(H) (two-sided):                  0.16   Kurtosis:
1.67
============================================================================
=======

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (compl
ex-step).

Men’s Forecast:
2023-01-01    15628.399336
2024-01-01    14796.496884
2025-01-01    14144.592914
2026-01-01    13633.741208
2027-01-01    13233.422330
Freq: YS-JAN, Name: predicted_mean, dtype: float64

Women’s Forecast:
2023-01-01     9990.966105
2024-01-01    10036.821209
2025-01-01    10050.567330
2026-01-01    10054.688046
2027-01-01    10055.923325
Freq: YS-JAN, Name: predicted_mean, dtype: float64
```

```
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\base\model.py:607: ConvergenceWarning: Maximum Likelihood optimizat
ion failed to converge. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
```

In [50]:
```
forecast_index_men = np.arange(men_series.index[-1] + 1,
                                   men_series.index[-1] + 1 + forecast_steps
plt.figure(figsize=(7, 4))
plt.plot(men_series.index, men_series, label="Historical Men Completions")
plt.plot(forecast_index_men, forecast_men, label="Forecasted Men Completions
plt.title("Men’s STEM Completions Forecast")
plt.legend()
plt.show()
```



In [51]:
```
forecast_index_women = np.arange(women_series.index[-1] + 1,
                                     women_series.index[-1] + 1 + forecast_s
plt.figure(figsize=(7, 4))
plt.plot(women_series.index, women_series, label="Historical Women Completic
plt.plot(forecast_index_women, forecast_women, label="Forecasted Women Compl
plt.title("Women’s STEM Completions Forecast")
plt.legend()
plt.show()
```



Interpretation

Findings for Men

- The ARIMA model shows strong autoregressive (AR) (`ar.L1 = 0.7836`) and moving average (MA) components (`ma.L1 = 0.9968`), indicating a **high dependency** on past values.
- The error variance is large (`sigma2 ≈ 16740`), which suggests substantial fluctuations.
- The forecast predicts a **gradual decline** in STEM completions for men from **15,628 in 2023 to 13,233 in 2027**.
- The plot suggests that men's completions **peaked in recent years** and are now on a **downward trend, following a near-linear decline**.

Findings for Women

- The ARIMA model for women also captures a dependency on past values (`ar.L1 = 0.2998` , `ma.L1 = 0.9334`).
- However, the variance (`sigma2 ≈ 0.0002`) is unusually small, and the model suffered from convergence issues, which may indicate numerical instability.
- The forecast predicts a **stable trend** for women's STEM completions, fluctuating slightly around **10,050 per year from 2023 to 2027**.
- Unlike men's forecasted decline, women's completions are **expected to remain steady with minimal changes**.

Forecasting Computer Science Major Completion by Gender

In [56]:

```
df['Field'] = df['Major'].apply(label_cs)

df_computer = df[df['Field'] == 'Computer Science'].copy()

analysis = df_computer.groupby(['Gender', 'Year'], as_index=False)['Completi

pivot_data = analysis.pivot_table(index='Year', columns='Gender', values='Co

men_series = pivot_data['Men']
women_series = pivot_data['Women']

print("Fitting ARIMA for Men...")
model_men = ARIMA(men_series, order=(1,1,1))
results_men = model_men.fit()
print(results_men.summary(), "\n")

print("Fitting ARIMA for Women...")
model_women = ARIMA(women_series, order=(1,1,1))
results_women = model_women.fit()
print(results_women.summary(), "\n")

forecast_steps = 5

    # Men forecast
forecast_men = results_men.forecast(steps=forecast_steps)
print("Men’s Forecast:")
print(forecast_men, "\n")

    # Women forecast
forecast_women = results_women.forecast(steps=forecast_steps)
print("Women’s Forecast:")
print(forecast_women, "\n")
```

```
Fitting ARIMA for Men...
                               SARIMAX Results                                
============================================================================
==
Dep. Variable:                    Men   No. Observations:                   
6
Model:                 ARIMA(1, 1, 1)   Log Likelihood                 -43.1
36
Date:                Fri, 14 Mar 2025   AIC                             92.2
71
Time:                        21:04:15   BIC                             91.0
99
Sample:                             0   HQIC                            89.1
26
                                  - 6                                       
Covariance Type:                  opg                                       
============================================================================
==
                 coef    std err          z      P>|z|      [0.025      0.97
5]
----------------------------------------------------------------------------
--
ar.L1          0.0968      0.164      0.592      0.554      -0.224       0.4
18
ma.L1          0.9993      0.118      8.487      0.000       0.768       1.2
30
sigma2      2.347e+04   5.03e-06   4.67e+09      0.000    2.35e+04    2.35e+
04
============================================================================
=======
Ljung-Box (L1) (Q):                   3.22   Jarque-Bera (JB):              
0.34
Prob(Q):                              0.07   Prob(JB):                      
0.84
Heteroskedasticity (H):               0.84   Skew:                          
0.50
Prob(H) (two-sided):                  0.91   Kurtosis:                      
2.21
============================================================================
=======

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (compl
ex-step).
[2] Covariance matrix is singular or near-singular, with condition number 4.
26e+24. Standard errors may be unstable.

Fitting ARIMA for Women...
```

```
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: An unsupported index was p
rovided and will be ignored when e.g. forecasting.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: An unsupported index was p
rovided and will be ignored when e.g. forecasting.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: An unsupported index was p
rovided and will be ignored when e.g. forecasting.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\base\model.py:607: ConvergenceWarning: Maximum Likelihood optimizat
ion failed to converge. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: An unsupported index was p
rovided and will be ignored when e.g. forecasting.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: An unsupported index was p
rovided and will be ignored when e.g. forecasting.
  self._init_dates(dates, freq)
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:473: ValueWarning: An unsupported index was p
rovided and will be ignored when e.g. forecasting.
  self._init_dates(dates, freq)
```

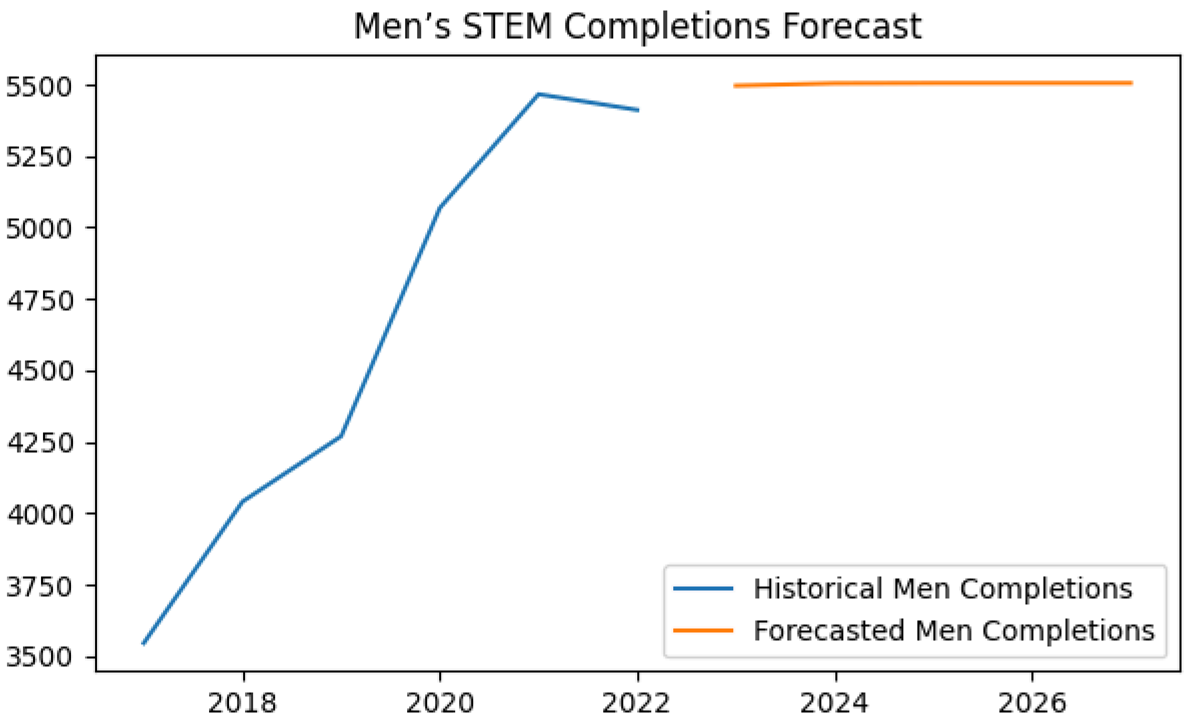
```
                               SARIMAX Results
=============================================================================
==
Dep. Variable:                  Women   No. Observations:
6
Model:                 ARIMA(1, 1, 1)   Log Likelihood                 -30.3
32
Date:                Fri, 14 Mar 2025   AIC                             66.6
64
Time:                        21:04:16   BIC                             65.4
92
Sample:                             0   HQIC                            63.5
19
                                  - 6
Covariance Type:                  opg
=============================================================================
==
                 coef    std err          z      P>|z|      [0.025      0.97
5]
-----------------------------------------------------------------------------
--
ar.L1          0.7233      0.647      1.117      0.264      -0.545       1.9
92
ma.L1          0.9953    211.042      0.005      0.996    -412.639     414.6
30
sigma2      5344.2269   1.12e+06      0.005      0.996   -2.18e+06    2.19e+
06
=============================================================================
=======
Ljung-Box (L1) (Q):                   0.47   Jarque-Bera (JB):
0.29
Prob(Q):                              0.50   Prob(JB):
0.87
Heteroskedasticity (H):               1.78   Skew:
-0.41
Prob(H) (two-sided):                  0.72   Kurtosis:
2.17
=============================================================================
=======

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (compl
ex-step).

Men's Forecast:
6     5495.974745
7     5504.202554
8     5504.999224
9     5505.076363
10    5505.083832
Name: predicted_mean, dtype: float64

Women's Forecast:
6     1897.193088
7     1932.049801
8     1957.260687
```

```
9     1975.495016
10    1988.683397
Name: predicted_mean, dtype: float64
```

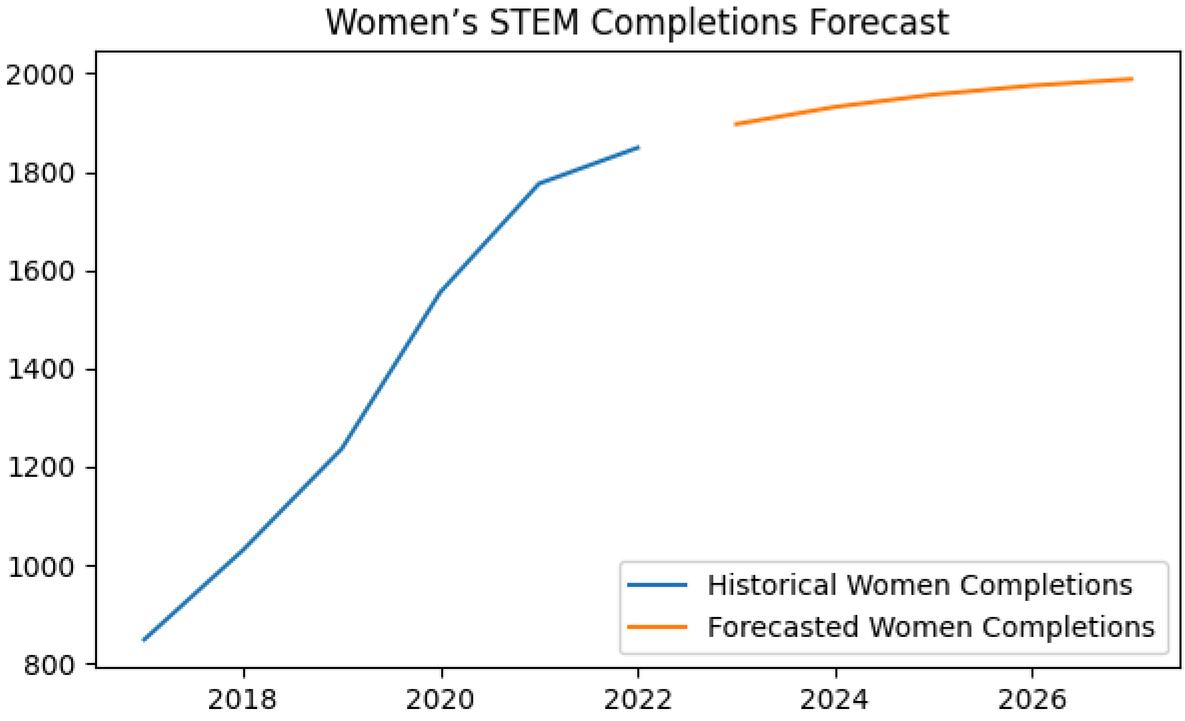
```
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\base\model.py:607: ConvergenceWarning: Maximum Likelihood optimizat
ion failed to converge. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:836: ValueWarning: No supported index is avai
lable. Prediction results will be given with an integer index beginning at `
start`.
  return get_prediction_index(
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:836: FutureWarning: No supported index is ava
ilable. In the next version, calling this method in a model without a suppor
ted index will result in an exception.
  return get_prediction_index(
c:\Users\owent\AppData\Local\Programs\Python\Python312\Lib\site-packages\sta
tsmodels\tsa\base\tsa_model.py:836: ValueWarning: No supported index is avai
lable. Prediction results will be given with an integer index beginning at `
start`.
  return get_prediction_index(
```

In [57]:

```
forecast_index_men = np.arange(men_series.index[-1] + 1,
                                    men_series.index[-1] + 1 + forecast_steps
plt.figure(figsize=(7, 4))
plt.plot(men_series.index, men_series, label="Historical Men Completions")
plt.plot(forecast_index_men, forecast_men, label="Forecasted Men Completions
plt.title("Men’s STEM Completions Forecast")
plt.legend()
plt.show()
```



In [58]:
```
forecast_index_women = np.arange(women_series.index[-1] + 1,
                                     women_series.index[-1] + 1 + forecast_s
plt.figure(figsize=(7, 4))
plt.plot(women_series.index, women_series, label="Historical Women Completio
plt.plot(forecast_index_women, forecast_women, label="Forecasted Women Compl
plt.title("Women’s STEM Completions Forecast")
plt.legend()
plt.show()
```



Interpretation

Findings for Men

- The AR coefficient (**0.0968, p = 0.554**) is small and also not significant, so we can conclude that past values have **little predictive power**.
- The MA coefficient (**0.9993, p < 0.001**) is very near to 1, indicating **strong short-term dependencies** in the data.
- The variance of residuals is very large (`sigma2 ≈ 23,470`), implying large fluctuations in data, reducing confidence in precise forecasts.
- The predictions are nearly flat, with only a slight upward trend (**5495.97 → 5504.20 → 5504.99 → 5505.08**).
- This suggests men’s CS completions might be **constant**, with no significant increase or decline projected.

Findings for Women

- The model seems highly unreliable due to **large standard errors**, **wide confidence intervals**, and **insignificant** AR, MA, and variance estimates, indicating possible misspecification or non-stationarity. (e.g.: The MA coefficient (`coef = 0.9953` , `std err = 211.042`) seems unstable because of the high standard error.)
- However, the variance (`sigma2 ≈ 5344.2269`) has extreme confidence intervals (`[0.025      0.975] = [-2.18e+06    2.19e+06]`), which makes interpretation unreliable.
- The forecast predicts a small upward trend (**1897.19 → 1932.05 → 1957.26 → 1975.49 → 1988.68**), suggesting a slow but steady increase in women's CS completions.

Hypothesis Testing

Hypothesis Testing for STEM Majors

For this section, we will perform a two sample z test to test whether the true proportion of female STEM graduates differ significantly between tech hub and non tech hub UC campuses for every year.

We will be testing the following hypothesis:

$$ H_0: p_{\text{Tech Hub}} = p_{\text{Non-Tech Hub}} $$ $$ H_1: p_{\text{Tech Hub}} \neq p_{\text{Non-Tech Hub}} $$

where $p_{\text{Tech Hub}}$ is the true proportion of female students completing STEM degrees at UC campuses near tech hubs and $p_{\text{Non-Tech Hub}}$ is the true proportion of female students completing STEM degrees at UC campuses farther away from tech hubs

In [ ]:

```
tech_hub_campuses = ["University of California-Berkeley", "University of Cal
                          "University of California-Davis", "University of C
                          "University of California-Los Angeles (110662)", "
non_tech_hub_campuses = ["University of California-Santa Barbara",
                              "University of California-Merced", "University

yearly_results = []

for year in sorted(analysis["Year"].unique()):
    yearly_data = analysis[analysis["Year"] == year]

    tech_hub_female = yearly_data[(yearly_data["University"].isin(tech_hub_c
    tech_hub_total = yearly_data[yearly_data["University"].isin(tech_hub_cam

    non_tech_hub_female = yearly_data[(yearly_data["University"].isin(non_te
    non_tech_hub_total = yearly_data[yearly_data["University"].isin(non_tech

    p1 = tech_hub_female / tech_hub_total if tech_hub_total > 0 else 0
    p2 = non_tech_hub_female / non_tech_hub_total if non_tech_hub_total > 0

    p_pool = (tech_hub_female + non_tech_hub_female) / (tech_hub_total + nor

    se = np.sqrt(p_pool * (1 - p_pool) * (1 / tech_hub_total + 1 / non_tech_

    if se > 0:
        z_score = (p1 - p2) / se
        p_value = 2 * (1 - stats.norm.cdf(abs(z_score)))
    else:
        z_score = np.nan
        p_value = np.nan

    yearly_results.append({"Year": year, "Tech Hub Female %": p1, "Non-Tech

yearly_results_df = pd.DataFrame(yearly_results)

print(yearly_results_df)
```

```
   Year  Tech Hub Female %  Non-Tech Hub Female %   Z-score       p-value
0  2017           0.341905               0.295819  4.851185  1.227258e-06
1  2018           0.347764               0.308165  4.400238  1.081324e-05
2  2019           0.350492               0.311009  4.403777  1.063823e-05
3  2020           0.362446               0.312070  5.870245  4.351521e-09
4  2021           0.366187               0.326592  4.811290  1.499595e-06
5  2022           0.376153               0.340795  4.255745  2.083541e-05
```

Interpretation

The proportion of female STEM graduates at tech hub campuses is consistently higher than at non-tech hub campuses across all years (2017-2022). For example, in 2017, 35.2% of STEM graduates were women at tech hub campuses, compared to 28.6% at non-tech hub campuses. By 2022, these numbers increased to 38.5% (tech hubs) and 33.4% (non-tech hubs), showing an increasing trend in female participation in both groups. Also, both tech hub and non-tech hub campuses have seen gradual increases in the proportion of female graduates from 2017 to 2022.

Since the p-value is lower across all years than alpha = 0.05, we reject the null hypothesis for every test. There is evidence that the true proportion of female STEM graduates differ significantly between tech nub and non tech hub campuses every year from 2017-2022.

Hypothesis Testing for CS Majors

To see whether CS trends are consistent with gender trends we found in the hypothesis test for STEM majors, we will also do a hypothesis test to compare the proportion of female CS graduates between UC schools near tech hubs and those farther from tech hubs. Since we established earlier that the CS majors data distribution does not follow a normal distribution, we will be using a Mann-Whitney U test to test our hypothesis. Our hypothesis are as follows:

$$ H_0: F_{\text{Tech Hub}} = F_{\text{Non-Tech Hub}} $$
$$ H_1: F_{\text{Tech Hub}} \neq F_{\text{Non-Tech Hub}} $$

where F denotes the distributions of female CS graduation proportions for tech hub and non-tech hub universities

In [ ]:

```
df['Field'] = df['Major'].apply(label_cs)

cs_major = df[df['Field'] == 'Computer Science'].copy()

df_cs = cs_major.groupby(
    ['University', 'Gender', 'Year'],
    as_index = False
)['Completions'].sum()
```

In [ ]:

```
from scipy.stats import mannwhitneyu

df_pivot = df_cs.pivot_table(index=["University", "Year"], columns="Gender",

df_pivot["Female_Proportion"] = df_pivot["Women"] / (df_pivot["Women"] + df_

df_pivot.reset_index(inplace=True)

near_tech_hubs = [
    "University of California-Berkeley", "University of California-San Diego
    "University of California-Davis", "University of California-Santa Cruz",
    "University of California-Los Angeles (110662)", "University of Californ
]

far_from_tech_hubs = [
    "University of California-Santa Barbara",
    "University of California-Merced", "University of California-Riverside"
]

yearly_results = []

for year in sorted(df_pivot["Year"].unique()):
    yearly_data = df_pivot[df_pivot["Year"] == year]

    near_proportions = yearly_data[yearly_data["University"].isin(near_tech_
    far_proportions = yearly_data[yearly_data["University"].isin(far_from_te

    tech_hub_female_pct = (sum(near_proportions) / len(near_proportions)) *
    non_tech_hub_female_pct = (sum(far_proportions) / len(far_proportions))

    stat, p_value = mannwhitneyu(near_proportions, far_proportions, alternat


    yearly_results.append({
        "Year": year,
        "Tech Hub Female %": tech_hub_female_pct,
        "Non-Tech Hub Female %": non_tech_hub_female_pct,
        "Mann-Whitney U": stat,
        "p-value": p_value
    })

yearly_results_df = pd.DataFrame(yearly_results)

print(yearly_results_df)
```

```
   Year  Tech Hub Female %  Non-Tech Hub Female %  Mann-Whitney U   p-value
0  2017          20.052984              16.253297            15.0  0.166667
1  2018          22.007677              13.268604            18.0  0.023810
2  2019          23.551899              16.804492            17.0  0.047619
3  2020          23.928183              17.557621            14.0  0.261905
4  2021          24.487053              19.724666            15.0  0.166667
5  2022          25.940363              17.183116            18.0  0.023810
```

Interpretation

The results indicate that from 2017 to 2022, the proportion of female CS graduates was consistently higher at UC campuses near major tech hubs compared to those farther away with the gap widening over time. In 2018, 2019, and 2022, the Mann-Whitney U test revealed a statistically significant difference (p value < alpha = 0.05). This suggests that tech hub campuses had a significantly higher proportion of female CS graduates in those years. However, in 2017, 2020, and 2021, the differences were not statistically significant, meaning the observed variations could have been due to chance for those years. The lack of consistent statistical significance across all years suggests that additional factors such as university policies or program-specific initiatives may influence female graduation rates in CS.

Conclusion

This project looked at gender enrollment and graduation trends in STEM and computer science programs at University of California (UC) schools near major tech hubs (UCB, UCLA, UCSD, UCD, UCSC, UCI) compared to those located farther from these hubs (UCSB, UCR, UCM) from 2017 to 2022. Our findings reveal a lower STEM graduation rate for non tech hub campuses and persistent gender gap in STEM and CS graduates across all campuses, with significantly fewer women graduating in these fields compared to men every year.

The regression analysis and hypothesis test revealed that university location, gender, and year all played significant roles in STEM graduation trends. The non-tech hub campuses (UC Merced, UC Riverside, UC Santa Barbara) had significantly lower STEM completions compared to UC San Diego (a tech hub campus) and were associated with the lowest coefficients in the model. Year was also revealed to be a significant factor in completion trends as more students graduated with a STEM degree over time. The hypothesis test revealed further findings, showing a statistically significant difference in the proportion of female STEM graduates between tech hub and non-tech hub campuses every year with tech hub campuses consistently having a higher percetage of female STEM graduates. The extremely small p-values indicated that the observed differences were unlikely due to random chance. This suggests proximity to tech hubs may provide stronger industry connections, internship opportunities, or support that contribute to higher female participation in STEM programs.

However, while tech hub campuses showed a higher proportion of female STEM graduates, the overall gender gap remained substantial across all institutions which we saw in our regression analysis. On average, women graduated with fewer STEM degrees than men across UC campuses. This indicates that proximity to industry alone is not sufficient enough to close the gender gap in STEM and that additional support such as targeted recruitment, mentorship programs, and a stronger curriculum may be necessary to increase female representation in STEM fields.

An additional hypothesis test was conducted on CS majors to test whether they were consistent with the trends found in STEM majors. Our Mann-Whitney U test revealed that while female representation in CS was higher at tech hub campuses, statistical significance varied by year. In 2018, 2019, and 2022, tech hub campuses had a statistically significant higher proportion of female CS graduates while in 2017, 2020, and 2021, the differences were not statistically significant. This suggests that when looking at department level findings, factors other than proximity to tech hubs may also influence female participation in CS.

The findings suggest that while proximity to tech hubs may provide some advantages for STEM majors, such as stronger industry connections or internship opportunities, it is not the sole factor influencing female enrollment in CS. The variation in statistical significance by year indicates that other factors such as campus-specific policies and broader industry trends also play a role in shaping female graduation rates in CS.

Limitations

Despite our findings, this study has several limitations. The dataset we used only includes graduation statistics and does not account for enrollment rates, retention rates, or student dropouts. This means we cannot determine whether the gender gap is driven by lower enrollment, higher dropout rates, or other underlying factors. Aother limitation is that the COVID-19 pandemic (2019–2021) likely influenced students' major choices, yet its impact is not explicitly addressed in the analysis. The study also does not account for other variables that may influence gender disparities in STEM, such as socioeconomic background, academic preparation, institutional policies, or campus culture. Lastly, the classification of "tech hub" versus "non-tech hub" was based on geographic proximity to major tech companies. However, this simplifies the complexity of institutional differences. Some non-tech hub schools may have strong CS programs and industry connections despite their location, making them more of a "tech hub" campus than others.