Programming Report

Task1: US Startups Founded per Year

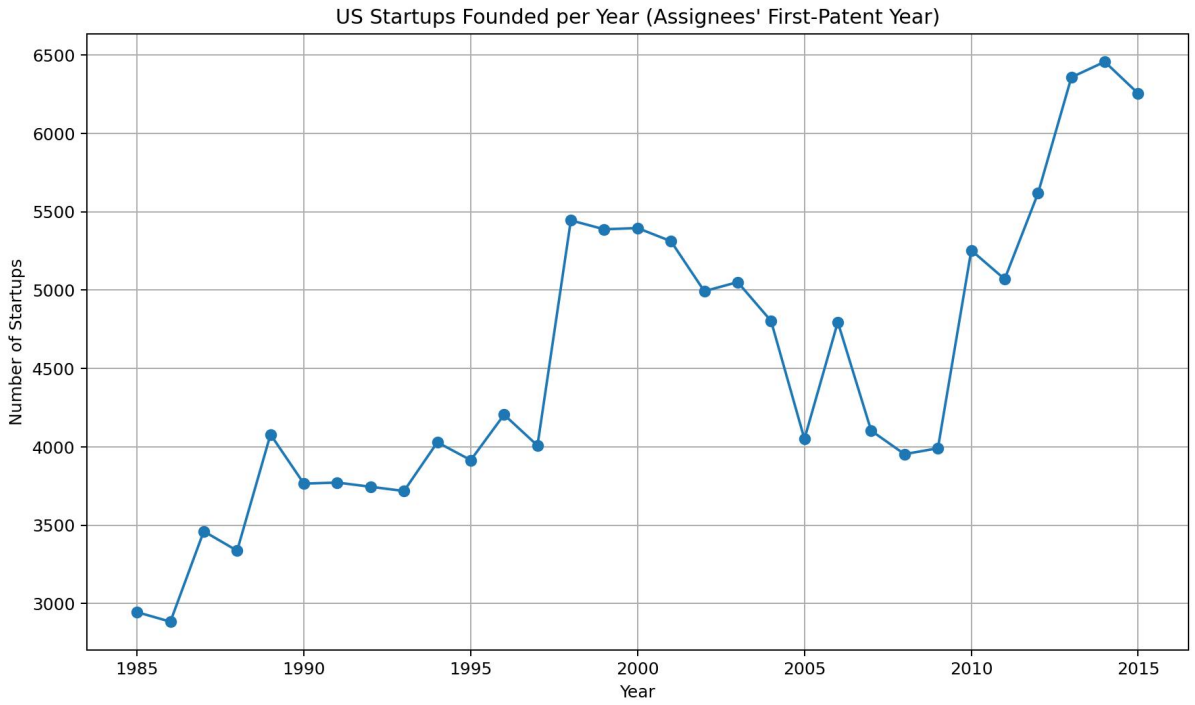
Objective:

To count the number of US startups founded per year between 1985 and 2015, using the first patent grant year as the foundation year of the startup. A startup is defined as an organization with a patent application granted in a specific year.

Main Assumptions:

1. We consider the first patent grant date of each assignee as the foundation date for the startup. This is done by identifying the first patent granted to each assignee in the data set.
2. US startups are defined by their state of incorporation, which we inferred from the state provided in the g_location_disambiguated.csv. Only startups whose state is listed in the US states whitelist (e.g., CA for California, NY for New York, etc.) are considered.
3. Any company listed in the data set can have multiple patents; thus, we aggregate by unique assignees to count the startups accurately.

Results:



/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task1.py

=== US startups (unique assignees) founded per year (1985–2015) ===

	founding_year	num_startups
0	1985	2945
1	1986	2884
2	1987	3461
3	1988	3337
4	1989	4079
5	1990	3765
6	1991	3772
7	1992	3745
8	1993	3718
9	1994	4029
10	1995	3915
11	1996	4205
12	1997	4008
13	1998	5445
14	1999	5388
15	2000	5396
16	2001	5312
17	2002	4993
18	2003	5051
19	2004	4803

20	2005	4050
21	2006	4796
22	2007	4103
23	2008	3953
24	2009	3991
25	2010	5254
26	2011	5070
27	2012	5618
28	2013	6358
29	2014	6457
30	2015	6255

Saved files:

- CSV: /Users/zhangshuhai/Desktop/Programming Test Task1/task1_startup_counts_us_by_assignee_20250923-113745.csv

- TXT: /Users/zhangshuhai/Desktop/Programming Test Task1/task1_console_output_20250923-113745.txt

- FIG: /Users/zhangshuhai/Desktop/Programming Test Task1/task1_startup_counts_us_by_assignee_20250923-113745.png

Process finished with exit code 0

Suggestions:

Many startups may not have filed patents in the early stages of their operations. While the first patent date is used for the analysis, considering startups without patents could offer a more holistic view of startup formation and innovation trends.

While using the first patent date is a reasonable definition for a "startup," further refinement could include distinguishing between different types of startups (e.g., tech vs. non-tech startups).

Task2: Share of Startups' First Patents Among All Patents (1985–2015)

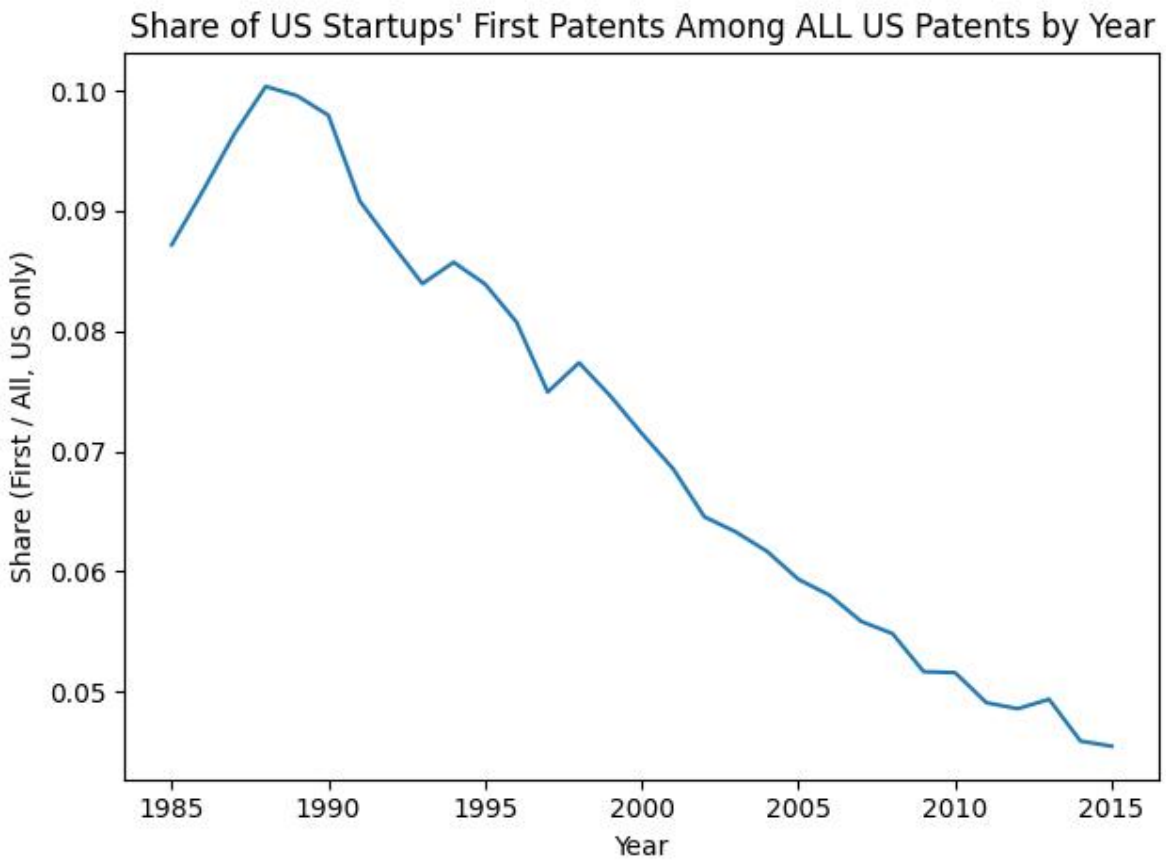
Objective:

We need to compute, for each calendar year from 1985 to 2015, the share of patents that are first patents of U.S. startups.

Main Assumptions:

1. The task focuses on U.S. startups. We operationalize "U.S." via the assignee location on the relevant patent
2. A startup is founded in the grant year of its first patent. We identify each assignee's earliest (min) grant and call that patent its first patent; the grant year is the founding year.
3. A patent is treated as U.S. if at least one assignee location on that patent is in the United States
4. A U.S. startup is an assignee for which its first patent has at least one U.S. assignee location.

Results:



/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task2.py

Computing per-year stats (US only): 100%|[redacted]| 47/47 [00:00<00:00, 137.04it/s]

Share of US startups' first patents among ALL US patents (1985–2015):

year	all_patents_us	first_patents_us	share_first
1985	32924	2870	0.087170
1986	31107	2852	0.091684
1987	35424	3415	0.096404
1988	32814	3293	0.100354
1989	40301	4013	0.099576
1990	37906	3713	0.097953
1991	41123	3735	0.090825
1992	42222	3687	0.087324
1993	43720	3670	0.083943
1994	46054	3948	0.085725
1995	45847	3847	0.083910
1996	50606	4087	0.080761
1997	52116	3905	0.074929
1998	68376	5290	0.077366
1999	71718	5350	0.074598
2000	73193	5232	0.071482
2001	76719	5259	0.068549
2002	76287	4924	0.064546

2003	77418	4900	0.063293
2004	75074	4630	0.061672
2005	66712	3959	0.059345
2006	80633	4678	0.058016
2007	72030	4023	0.055852
2008	71143	3901	0.054833
2009	76154	3934	0.051658
2010	99314	5123	0.051584
2011	101199	4967	0.049082
2012	113660	5522	0.048583
2013	126782	6259	0.049368
2014	138596	6361	0.045896
2015	135614	6165	0.045460

Saved files:

- Console output: /Users/zhangshuhai/Desktop/Programming Test Task2/task2_console_output_20250923-114009.txt

- Figure: /Users/zhangshuhai/Desktop/Programming Test Task2/task2_figure_20250923-114009.png

Process finished with exit code 0

Suggestions:

Any residual disambiguation error in assignee_id (splits/merges) can bias the counts. If time permits, spot-check large assignees for ID stability.

We accepted all state values when country == "US". A stricter filter could whitelist only valid U.S. state/territory codes to drop rare miscodings.

Multiple first patents on same day: Extremely rare ties (same-day grants) could be broken deterministically (e.g., min patent_id) to ensure exactly one first patent per startup.

Task3: Top U.S. States by Startup Count (1985–2015)

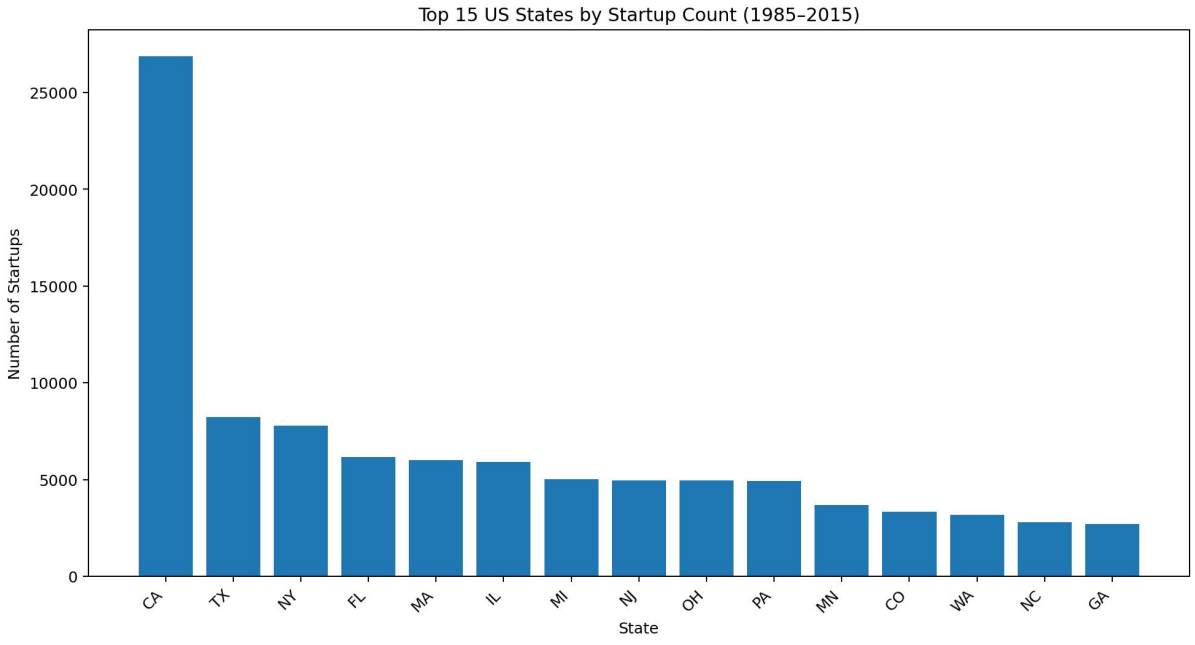
Objective:

We calculated the number of startups founded in each U.S. state between 1985 and 2015 based on their first patent grant date (which is assumed to be the startup's founding year).

Assumptions:

1. A startup is defined as an assignee (company) whose first patent grant (patent with earliest grant date) is considered its founding year. Each assignee's first patent is identified using the first_patents function.
2. If a startup's first patent lacks location information (i.e., missing location_id), it is excluded from the analysis, as the state cannot be determined without the location data.
3. We define a U.S. startup as an assignee whose first patent has at least one U.S. location.
4. U.S. locations are filtered using the state abbreviations to ensure proper mapping of states.

Results:



/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task3.py

[Info] Startups without location_id on their first patent: 2533

=== Top US States by Startup Count (Founding years 1985-2015) ===

	state	num_startups
0	CA	26887
1	TX	8241
2	NY	7766
3	FL	6161
4	MA	5993
5	IL	5921
6	MI	5014
7	NJ	4968
8	OH	4959
9	PA	4936

*** State with the most startups: CA (26887 startups) ***

Saved files:

- TXT: /Users/zhangshuhai/Desktop/Programming Test
Task3/task3_console_output_20250923-115213.txt

- FIG: /Users/zhangshuhai/Desktop/Programming Test
Task3/task3_top_states_20250923-115213.png

Process finished with exit code 0

Suggestions:

If a first patent has multiple assignees, pick one deterministic primary state (e.g., the assignee with the lowest assignee_id or earliest address record) to avoid multi-state double counting; log the count of multi-state cases.

Add a per-capita view (normalize by 1990/2000 population) or at least show share (%) next to counts; annotate bars with values to make rank differences obvious.

Task 4 CPC Classes with the Largest Number of US Startups (1985–2015)

Objective:

Identify the CPC technology classes associated with the largest number of U.S. startups founded between 1985 and 2015, where a "startup" is defined as an assignee whose first granted patent (grant date = founding date) falls in the window. The output is a ranking of CPC classes by the count of distinct startups.

Assumptions:

1. Founding = first grant date of the first patent (as specified in the assignment).
2. US status is determined by the assignee's location on the first patent; other co-assignees or later relocations are ignored.
3. If CPC primary is available it is authoritative; if not, we treat the first patent as a multi-label record but prevent double counting of the same firm within a class.
4. If a startup's first patent lacks any CPC entry, we cannot credibly assign a technology class, so it is omitted from the CPC tally (but its count is reported).

Results:

/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task4.py

[Info] Startups without location_id on their first patent: 2533

[Info] Non-US startups dropped: 122102

[Info] Startups without CPC class on their first patent: 100

=== CPC classes with the largest number of US startups (1985–2015) ===

cpc_class	num_startups
A61	14834
G06	12589
H04	8380
G01	7154
H01	4668

A63	4432
B65	4377
A47	4109
A01	3589
F16	3420

[Saved] Console summary → /Users/zhangshuhai/Desktop/Programming Test Task4/task4_console_output_20250923-120711.txt

Saved files:

- task4_startups_by_cpc_class.csv
- task4_top10_cpc_class.csv

[Sanity] US startups in window (input): 134,347

[Sanity] US startups with mappable CPC: 133,786

[Info] CPC counting used deduplicated multi-label mode (startup may appear in multiple classes, but counted at most once per class).

Process finished with exit code 0

Suggestions:

Sensitivity analyses: (a) use the first 3–5 years of patents per startup to define a majority CPC; (b) compare primary-only vs. multi-label outcomes.

State robustness: when location_id is missing on the first patent, fall back to the earliest available assignee location within ±1–2 years.

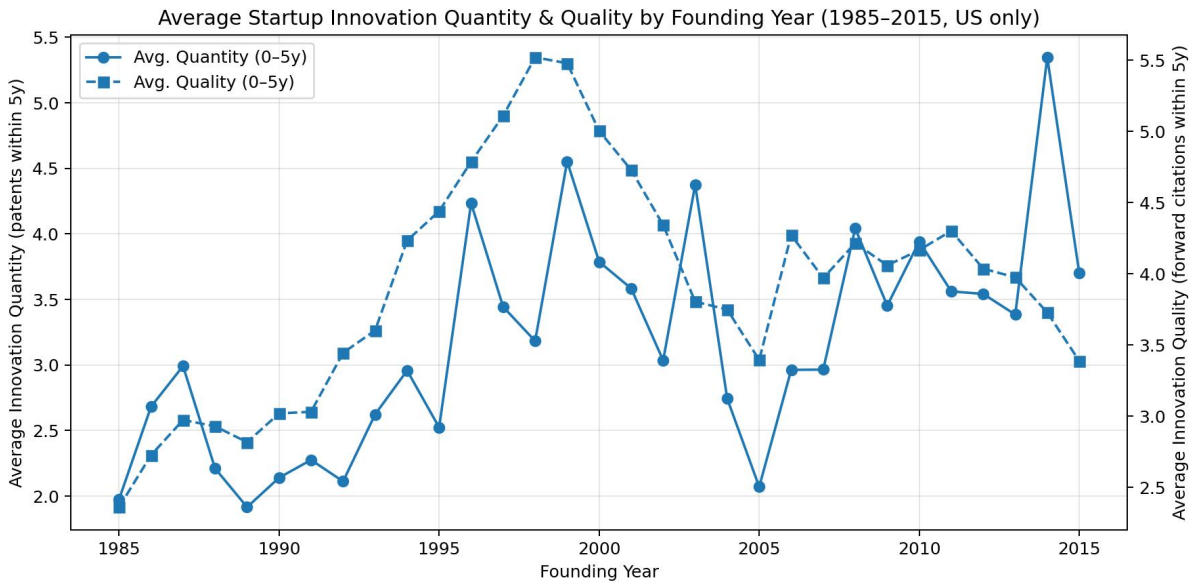
Task5 Average innovation quantity & quality by founding year (US only)

Objective: Visualize the average quantity and quality of startups' innovative output by founding year.

Assumptions:

1. Startup definition & timing. A startup is an assignee; its founding year is the grant year of its first patent. This aligns with Tasks 1–4.
2. US filter. We keep only US startups, determined from the assignee row on the startup's first patent. Non-US or missing-location first patents are excluded to keep the state logic consistent with earlier tasks.
3. Windows. Innovation quantity counts the startup's patents within 0–5 years after the first grant (same-assignee filings). Innovation quality counts forward citations to the first patent within 0–5 years from grant.
4. Citations input. Citations are streamed from the two large files to avoid memory issues; only citations that land within 5 years of the first grant are counted.
5. Aggregation. For each founding year we compute means across startups (not per-patent averages). Missing metrics are treated as 0 so cohorts include startups with no follow-on patents and/or no early citations.

Results:



/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task5.py

[START] Load g_patent, g_assignee_disambiguated, g_location ...
[DONE] Load g_patent, g_assignee_disambiguated, g_location in 73.9s
pat shape: (7403393, 2) | asg shape: (6926219, 3) | loc shape: (77755, 3)

[START] Identify first patent per assignee ...
[DONE] Identify first patent per assignee in 8.6s

[START] US-filter on the assignee row of the FIRST patent ...
[DONE] US-filter on the assignee row of the FIRST patent in 2.9s

[START] Apply safe cohort window (full 5y coverage) ...
Founding years used: 1985–2015 (full 5y windows)
[DONE] Apply safe cohort window (full 5y coverage) in 0.2s

[START] Compute innovation QUANTITY (0–5y) ...
[DONE] Compute innovation QUANTITY (0–5y) in 2.4s

[START] Compute innovation QUALITY via streaming citations ...
[DONE] Compute innovation QUALITY via streaming citations in 246.7s

[START] Assemble per-startup table and aggregate means by year ...

	founding_year	avg_innov_quantity	avg_innov_quality
0	1985	1.973184	2.362525
1	1986	2.680872	2.727331
2	1987	2.991128	2.974242
3	1988	2.209946	2.930316
4	1989	1.914483	2.818092

[DONE] Assemble per-startup table and aggregate means by year in 0.3s
=== Average Startup Innovation Quantity & Quality by Founding Year (1985–2015, US only) ===
(Computed on cohorts 1985–2015 with a 5-year window after the first patent)

founding_year	avg_innov_quantity	avg_innov_quality
1985	1.973184	2.362525
1986	2.680872	2.727331
1987	2.991128	2.974242
1988	2.209946	2.930316
1989	1.914483	2.818092
1990	2.138688	3.020348
1991	2.274329	3.032088
1992	2.110756	3.445421
1993	2.621880	3.603558
1994	2.956616	4.234273

1995	2.522881	4.440458
1996	4.233782	4.787271
1997	3.442491	5.110246
1998	3.181853	5.520587
1999	4.552005	5.479009
2000	3.786214	5.004272
2001	3.582857	4.729524
2002	3.033818	4.342898
2003	4.374415	3.805369
2004	2.742544	3.749561
2005	2.069113	3.395828
2006	2.961771	4.272570
2007	2.963822	3.971058
2008	4.041625	4.215898
2009	3.451820	4.059713
2010	3.940980	4.167421
2011	3.560649	4.301623
2012	3.541020	4.036360
2013	3.383439	3.976086
2014	5.346337	3.728671
2015	3.699541	3.388907

[Saved] Files:
  - TXT: /Users/zhangshuhai/Desktop/Programming Test Task5/task5_console_output_20250923-163748.txt
  - FIG: /Users/zhangshuhai/Desktop/Programming Test Task5/task5_avg_qty_quality_20250923-163748.png

Process finished with exit code 0

Suggestions:

Self-citation check. Exclude assignee self-citations to the first patent as a robustness test.

Include vs exclude first patent. Show both versions of quantity (±1) in a small compare table.

This setup keeps Task 5 aligned with Tasks 1–4 (same startup definition, US filter, and first-patent anchor), while guarding against right-truncation and mechanical inflation of quantity.

Task6 Cohort Comparison (US-only): 1995–1999 vs 2003–2007

Objective:

Compare the quantity and quality of startup innovation by startups founded between 1995 and 1999 and by startups founded between 2003 and 2007.

Assumptions(aligned with Tasks 1–5):

1. Startup definition: a unique assignee; founding year is the grant year of its first patent.
2. US filter: The assignee is treated as US if, on the specific assignee row of its first patent, the linked location maps to disambiguated_country ∈ {US, USA} and has a non-missing state. Startups with missing location_id at the first patent are excluded (can't be attributed to US).
3. Windows: Fixed 0–5 years after the first patent grant for both metrics.
4. Quantity: We exclude the first patent from the 0–5y count, avoiding mechanical inflation of year-0 activity.
5. Quality (citations): All forward citations to the first patent within 5 years are counted; missing or none → 0. (Self-citations not removed unless explicitly flagged in source; see suggestions.)
6. Robustness: Because patent/citation distributions are heavy-tailed, we supplement mean-based tests with Mann–Whitney U and robust summaries.

Results:

/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task6.py

[START] Load g_patent, g_assignee_disambiguated, g_location ...

[DONE] Load g_patent, g_assignee_disambiguated, g_location in 74.5s

[START] Identify first patents per assignee (startups) ...

[DONE] Identify first patents per assignee (startups) in 8.7s

[START] US-filter on the assignee row of the FIRST patent ...

[DONE] US-filter on the assignee row of the FIRST patent in 3.2s

[START] Apply safe 5y window (avoid right truncation) ...

Cohorts used (safe): A=1995-1999, B=2003-2007

[DONE] Apply safe 5y window (avoid right truncation) in 0.2s

[START] Compute innov_quantity (0–5y patents) ...

[DONE] Compute innov_quantity (0–5y patents) in 2.3s

[START] Compute innov_quality (0–5y citations to FIRST patent) ...

[DONE] Compute innov_quality (0–5y citations to FIRST patent) in 238.4s

[START] Assemble per-startup dataset (US-only) ...

[DONE] Assemble per-startup dataset (US-only) in 0.1s

[START] Cohort summary & significance tests ...

=== Task 6: Cohort Comparison (US-only; 1995–1999 vs 2003–2007) ===

metric	A_mean	A_sd	A_n	B_mean	B_sd	B_n	Welch_t	Welch_p	MWU_U	MWU_p
innov_quantity (0–5y patents)	3.629394	48.236560	22447	3.070517	61.603381	22094	1.064921	2.869177e-01	253795855.5	2.555631e-06
innov_quality (0–5y citations)	5.119481	8.786917	22447	3.848058	7.048689	22094	16.857559	1.481491e-63	292499516.5	2.185627e-241

[DONE] Cohort summary & significance tests in 0.0s

=== Robust Checks (median/IQR, 5% trimmed mean, mean log(1+x)) ===

metric	cohort	median	IQR	trimmed_mean_5pct	mean_log1p	n
innov_quantity (0–5y)	A	0.0	2.0	1.285552	0.654723	22447
innov_quantity (0–5y)	B	0.0	2.0	1.152268	0.611914	22094
innov_quality (0–5y)	A	3.0	5.0	3.822997	1.340648	22447
innov_quality (0–5y)	B	2.0	4.0	2.722317	1.068847	22094

[START] Save outputs ...

Saved to Desktop (or fallback) folder:

- /Users/zhangshuhai/Desktop/Programming Test Task6/task6_console_output_20250923-165040.txt

Saved:

- task6_per_startup_metrics_us_only.csv
- task6_cohort_comparison_summary_us_only.csv
- task6_robust_checks_us_only.csv

[DONE] Save outputs in 0.1s

Process finished with exit code 0

Suggestions:

Industry controls: Compare within CPC buckets or weight cohorts to a common CPC mix.

Survival & size: Condition on survival (still patenting after y years) or control for early portfolio size.

Distributional visuals: Add ECDFs/quantile plots or violin plots to illustrate heavy tails.

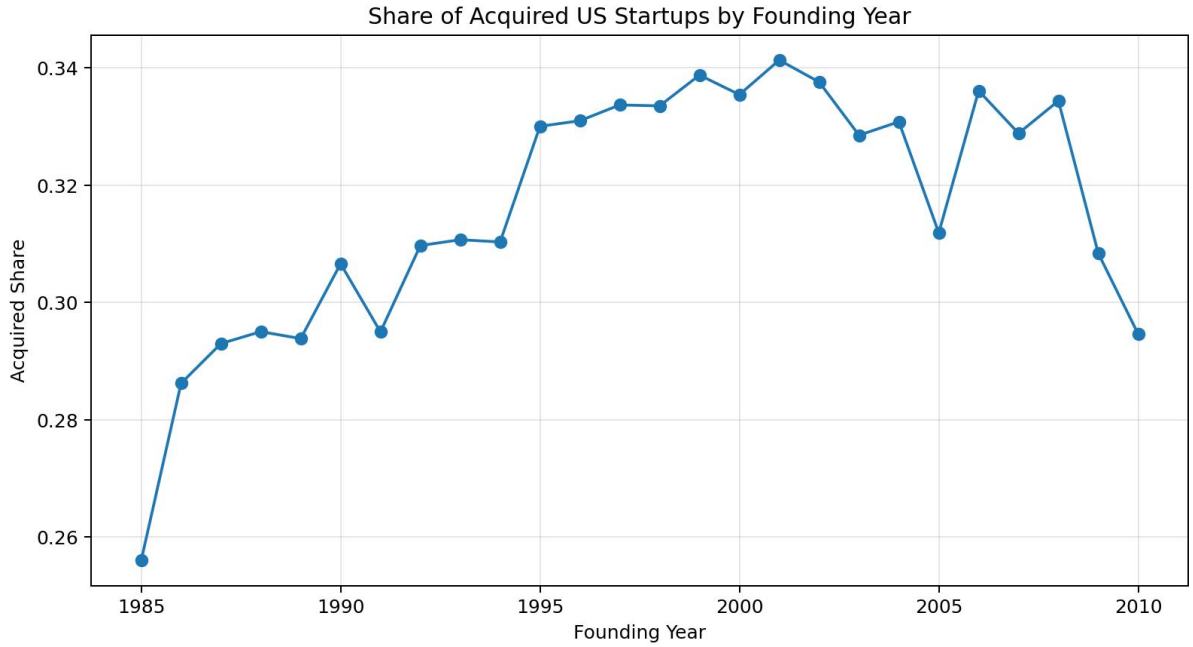
Task 7: Share of Acquired Startups by Founding Year (1985–2010)

Objective:
Visualize the share of acquired startups by founding year.

Assumptions (Aligned with 1-6):

1. Founding proxy: first-grant date ≈ startup founding year.
2. US attribution: US status is determined at first patent (assignee row); later relocations are ignored.
3. Acquisition proxy: USPTO assignment/merger filings within 10y capture acquisitions. We assume our keyword/regex cleaning removes most non-M&A admin filings.
4. Event coverage: if an acquisition lacks a corresponding USPTO assignment filing (or if it's miscoded), it will be missed.
5. Window integrity: cohorts after safe end (2010) are excluded to avoid under-measuring the 10-year horizon.
6. One startup, one outcome: if multiple qualifying events occur, the startup is counted once as acquired (no double counting).

Results:



/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task7.py

[START] Load patents, assignees, locations ...

[DONE] Load patents, assignees, locations in 79.2s

[START] Identify first patents per assignee ...

[DONE] Identify first patents per assignee in 8.8s

[START] US-only on the FIRST-patent assignee row ...

[DONE] US-only on the FIRST-patent assignee row in 3.1s

[START] Scan pa.csv for max exec_dt (safe window) ...

[DONE] Scan pa.csv for max exec_dt (safe window) in 2.5s

[START] Build per-startup patent set within 0–10y window ...

[DONE] Build per-startup patent set within 0–10y window in 5.1s

[START] Flag acquisitions from pa.csv (streaming) ...

[DONE] Flag acquisitions from pa.csv (streaming) in 4.3s

[START] Aggregate acquisition share by founding year ...

[DONE] Aggregate acquisition share by founding year in 0.1s

=== Task 7: Share of Acquired US Startups by Founding Year (1985–2010) ===

Mode: ALL patents within 10y

(Safe-end based on pa.csv max exec_dt year = 2020)

founding_year	startups	acquired	acquired_share(%)
1985	2824	723	25.60
1986	2802	802	28.62
1987	3314	971	29.30
1988	3220	950	29.50
1989	3934	1156	29.38
1990	3630	1113	30.66
1991	3634	1072	29.50
1992	3613	1119	30.97
1993	3579	1112	31.07
1994	3854	1196	31.03
1995	3748	1237	33.00
1996	3982	1318	33.10
1997	3797	1267	33.37
1998	5148	1717	33.35
1999	5145	1743	33.88
2000	5059	1697	33.54
2001	5086	1736	34.13
2002	4769	1610	33.76
2003	4779	1570	32.85

2004	4504	1490	33.08
2005	3883	1211	31.19
2006	4567	1535	33.61
2007	3938	1295	32.88
2008	3801	1271	33.44
2009	3820	1178	30.84
2010	5020	1479	29.46

[Saved] Files:

- TXT: /Users/zhangshuhai/Desktop/Programming Test Task7/task7_console_output_20250923-170028.txt

- FIG: /Users/zhangshuhai/Desktop/Programming Test Task7/task7_acquisition_share_20250923-170028.png

Saved table: task7_acquisition_share_by_year.csv

Saved figure: task7_acquisition_share_by_year.png

Process finished with exit code 0

Suggestions:

Estimate Kaplan–Meier curves for time-to-acquisition and report cohort-specific hazards; avoids arbitrary fixed windows and handles censoring transparently.

Break out by state or CPC class (from Task 3/4), and by founding-year bins (pre-/post-1998, pre-/post-2001, etc.).

Cross-check a sample against Crunchbase/PitchBook/Orbis deal tags to quantify precision/recall of the USPTO-based proxy.

Task 8: Estimate linear regression models

Objective: explain which startup characteristics (measured from patents) correlate with being acquired. Outcome is a binary indicator for whether a startup is acquired within 10 years of its first patent grant.

Assumptions (aligned with Tasks 1–7):

1. Startup definition: a unique assignee; its founding year is the grant year of its first patent.
2. US-only cohort: keep startups whose first patent's assignee row maps to a U.S. location (country = "US/USA/United States" and a valid state).
3. Windows: innovation quantity and quality use a 0–5 year window after the first patent; acquisition uses 0–10 years after the first patent.
4. Acquisition signal: from pa.csv, we treat events whose normalized type contains "assignment" or "merger" as acquisitions; others (e.g., name changes, security interests) are ignored.
5. Timing guard: we check the max exec_dt in pa.csv and restrict founding years to ensure the full 10-year window is observable.
6. Missing values: innovation metrics are numeric; missing values are set to 0 before modeling.
7. Model: Linear Probability Model (OLS) with HC1 robust SE; include centered founding year (year_c), tech and state dummies (drop-first).
8. Tech field = top CPC class of first patent; rare classes collapsed to "Other".
9. State = state attached to the first-patent assignee row; rare states collapsed to "Other".

Results:

/Users/zhangshuhai/miniconda3/envs/research/bin/python
/Users/zhangshuhai/Desktop/Research Bias/PythonProject/task8.py

OLS Regression Results

Dep. Variable:	acquired	R-squared:	0.342
Model:	OLS	Adj. R-squared:	0.342
Method:	Least Squares	F-statistic:	1046.
Date:	Tue, 23 Sep 2025	Prob (F-statistic):	0.00
Time:	17:22:04	Log-Likelihood:	-81658.
No. Observations:	196035	AIC:	1.637e+05
Df Residuals:	195867	BIC:	1.654e+05
Df Model:	167		
Covariance Type:	HC1		

==========

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1160	0.027	4.367	0.000	0.064	0.168
innov_quantity	0.0004	0.000	2.344	0.019	6.25e-05	0.001
innov_quality	0.0033	0.000	13.878	0.000	0.003	0.004
year_c	0.0008	0.000	5.111	0.000	0.001	0.001
tech_slim_A21	0.6869	0.016	43.855	0.000	0.656	0.718
tech_slim_A22	0.1001	0.027	3.684	0.000	0.047	0.153
tech_slim_A23	0.1184	0.013	9.046	0.000	0.093	0.144
tech_slim_A24	0.0461	0.043	1.086	0.278	-0.037	0.129
tech_slim_A41	0.0734	0.014	5.356	0.000	0.047	0.100
tech_slim_A42	0.0026	0.021	0.125	0.900	-0.038	0.044
tech_slim_A43	0.0843	0.021	4.033	0.000	0.043	0.125
tech_slim_A44	-0.0013	0.021	-0.063	0.949	-0.043	0.040
tech_slim_A45	0.0352	0.012	2.944	0.003	0.012	0.059
tech_slim_A46	0.0055	0.020	0.270	0.787	-0.034	0.045
tech_slim_A47	-0.0054	0.007	-0.800	0.423	-0.019	0.008
tech_slim_A61	0.1417	0.006	24.037	0.000	0.130	0.153
tech_slim_A62	0.0090	0.015	0.611	0.541	-0.020	0.038
tech_slim_A63	-0.0412	0.006	-6.925	0.000	-0.053	-0.030
tech_slim_B01	0.1195	0.010	11.538	0.000	0.099	0.140

tech_slim_B02	0.0090	0.020	0.457	0.648	-0.030 0.048
tech_slim_B03	0.0848	0.035	2.402	0.016	0.016 0.154
tech_slim_B04	0.1563	0.057	2.722	0.006	0.044 0.269
tech_slim_B05	-0.1619	0.007	-23.059	0.000	-0.176 -0.148
tech_slim_B07	-0.0031	0.026	-0.119	0.905	-0.054 0.048
tech_slim_B08	0.0893	0.026	3.453	0.001	0.039 0.140
tech_slim_B09	0.3194	0.034	9.471	0.000	0.253 0.385
tech_slim_B21	-0.0009	0.013	-0.073	0.942	-0.026 0.024
tech_slim_B22	0.1171	0.027	4.362	0.000	0.064 0.170
tech_slim_B23	0.0809	0.012	7.005	0.000	0.058 0.103
tech_slim_B24	-0.0811	0.010	-8.072	0.000	-0.101 -0.061
tech_slim_B25	0.6440	0.009	75.170	0.000	0.627 0.661
tech_slim_B26	0.0199	0.014	1.388	0.165	-0.008 0.048
tech_slim_B27	-0.0163	0.019	-0.846	0.398	-0.054 0.021
tech_slim_B28	0.0235	0.024	0.986	0.324	-0.023 0.070
tech_slim_B29	0.1616	0.014	11.811	0.000	0.135 0.188
tech_slim_B30	-0.0282	0.026	-1.070	0.285	-0.080 0.023
tech_slim_B31	0.1409	0.045	3.118	0.002	0.052 0.230
tech_slim_B32	0.1897	0.020	9.412	0.000	0.150 0.229
tech_slim_B41	0.1657	0.019	8.880	0.000	0.129 0.202
tech_slim_B42	0.0031	0.016	0.190	0.849	-0.028 0.035
tech_slim_B43	-0.0367	0.017	-2.189	0.029	-0.070 -0.004

tech_slim_B44	0.0082	0.021	0.394	0.694	-0.033 0.049
tech_slim_B60	0.0527	0.008	6.554	0.000	0.037 0.068
tech_slim_B61	-0.0341	0.019	-1.768	0.077	-0.072 0.004
tech_slim_B62	-0.0575	0.007	-7.935	0.000	-0.072 -0.043
tech_slim_B63	0.0512	0.012	4.160	0.000	0.027 0.075
tech_slim_B64	0.1286	0.019	6.781	0.000	0.091 0.166
tech_slim_B65	0.0636	0.007	8.732	0.000	0.049 0.078
tech_slim_B66	0.1121	0.019	5.828	0.000	0.074 0.150
tech_slim_B67	0.0593	0.019	3.057	0.002	0.021 0.097
tech_slim_B68	-0.0583	0.018	-3.270	0.001	-0.093 -0.023
tech_slim_B82	0.0875	0.039	2.249	0.025	0.011 0.164
tech_slim_C01	0.0417	0.021	1.963	0.050	6.17e-05 0.083
tech_slim_C02	0.0843	0.013	6.420	0.000	0.059 0.110
tech_slim_C03	0.0750	0.028	2.664	0.008	0.020 0.130
tech_slim_C04	0.0630	0.021	2.995	0.003	0.022 0.104
tech_slim_C05	-0.1442	0.011	-13.525	0.000	-0.165 -0.123
tech_slim_C07	0.1440	0.011	12.539	0.000	0.121 0.167
tech_slim_C08	0.1402	0.017	8.190	0.000	0.107 0.174
tech_slim_C09	0.1421	0.019	7.601	0.000	0.105 0.179
tech_slim_C10	0.1367	0.023	5.944	0.000	0.092 0.182
tech_slim_C11	0.0220	0.020	1.117	0.264	-0.017 0.061
tech_slim_C12	0.1281	0.014	9.446	0.000	0.102 0.155

tech_slim_C21	-0.0913	0.009	-9.856	0.000	-0.109	-0.073
tech_slim_C22	0.1716	0.029	5.975	0.000	0.115	0.228
tech_slim_C23	0.1279	0.025	5.089	0.000	0.079	0.177
tech_slim_C25	-0.1609	0.009	-17.716	0.000	-0.179	-0.143
tech_slim_C30	0.1750	0.051	3.407	0.001	0.074	0.276
tech_slim_D01	0.0873	0.041	2.137	0.033	0.007	0.167
tech_slim_D02	0.1649	0.061	2.700	0.007	0.045	0.285
tech_slim_D04	-0.0406	0.008	-5.228	0.000	-0.056	-0.025
tech_slim_D05	0.0345	0.026	1.338	0.181	-0.016	0.085
tech_slim_D06	0.0611	0.020	3.109	0.002	0.023	0.100
tech_slim_D21	0.2043	0.035	5.824	0.000	0.136	0.273
tech_slim_E01	0.0435	0.015	2.991	0.003	0.015	0.072
tech_slim_E02	0.0359	0.014	2.495	0.013	0.008	0.064
tech_slim_E03	-0.0162	0.014	-1.118	0.264	-0.044	0.012
tech_slim_E04	0.2259	0.009	24.721	0.000	0.208	0.244
tech_slim_E05	0.0085	0.012	0.696	0.487	-0.015	0.032
tech_slim_E06	-0.0217	0.010	-2.255	0.024	-0.041	-0.003
tech_slim_E21	0.1687	0.014	12.127	0.000	0.141	0.196
tech_slim_F01	0.0624	0.015	4.137	0.000	0.033	0.092
tech_slim_F02	0.4814	0.013	36.978	0.000	0.456	0.507
tech_slim_F03	0.2438	0.031	7.891	0.000	0.183	0.304
tech_slim_F04	-0.0494	0.010	-4.871	0.000	-0.069	-0.030

tech_slim_F15	0.1080	0.042	2.581	0.010	0.026	0.190
tech_slim_F16	-0.0163	0.008	-2.143	0.032	-0.031	-0.001
tech_slim_F17	0.1334	0.041	3.244	0.001	0.053	0.214
tech_slim_F21	0.0134	0.014	0.983	0.325	-0.013	0.040
tech_slim_F22	0.0815	0.060	1.362	0.173	-0.036	0.199
tech_slim_F23	0.1577	0.019	8.145	0.000	0.120	0.196
tech_slim_F24	0.2193	0.015	15.040	0.000	0.191	0.248
tech_slim_F25	0.1976	0.022	8.941	0.000	0.154	0.241
tech_slim_F26	0.6668	0.013	51.081	0.000	0.641	0.692
tech_slim_F27	0.0869	0.039	2.238	0.025	0.011	0.163
tech_slim_F28	0.0684	0.023	2.930	0.003	0.023	0.114
tech_slim_F41	-0.0826	0.007	-11.139	0.000	-0.097	-0.068
tech_slim_F42	-0.0017	0.020	-0.083	0.934	-0.042	0.038
tech_slim_G01	0.0740	0.006	11.525	0.000	0.061	0.087
tech_slim_G02	0.1212	0.012	10.063	0.000	0.098	0.145
tech_slim_G03	0.0741	0.018	4.165	0.000	0.039	0.109
tech_slim_G04	-0.0053	0.030	-0.175	0.861	-0.064	0.054
tech_slim_G05	0.0677	0.015	4.514	0.000	0.038	0.097
tech_slim_G06	0.2179	0.007	30.224	0.000	0.204	0.232
tech_slim_G07	0.1221	0.014	8.657	0.000	0.094	0.150
tech_slim_G08	0.4071	0.013	32.190	0.000	0.382	0.432
tech_slim_G09	0.0305	0.010	2.969	0.003	0.010	0.051

tech_slim_G10	0.0912	0.018	5.013	0.000	0.056 0.127
tech_slim_G11	0.1288	0.015	8.530	0.000	0.099 0.158
tech_slim_G16	0.1187	0.031	3.861	0.000	0.058 0.179
tech_slim_G21	0.0607	0.028	2.137	0.033	0.005 0.116
tech_slim_H01	0.1308	0.008	15.651	0.000	0.114 0.147
tech_slim_H02	0.0754	0.012	6.499	0.000	0.053 0.098
tech_slim_H03	0.1655	0.017	9.996	0.000	0.133 0.198
tech_slim_H04	0.6071	0.006	100.188	0.000	0.595 0.619
tech_slim_H05	0.1354	0.014	9.483	0.000	0.107 0.163
tech_slim_Other	0.1335	0.033	4.024	0.000	0.068 0.198
state_slim_AL	-0.0830	0.031	-2.682	0.007	-0.144 -0.022
state_slim_AR	0.0879	0.033	2.668	0.008	0.023 0.153
state_slim_AZ	-0.0554	0.028	-2.003	0.045	-0.110 -0.001
state_slim_CA	0.1024	0.026	3.881	0.000	0.051 0.154
state_slim_CO	0.0331	0.028	1.197	0.231	-0.021 0.087
state_slim_CT	0.0075	0.028	0.267	0.789	-0.047 0.062
state_slim_DC	-0.0990	0.038	-2.628	0.009	-0.173 -0.025
state_slim_DE	0.0776	0.031	2.509	0.012	0.017 0.138
state_slim_FL	0.0508	0.027	1.897	0.058	-0.002 0.103
state_slim_GA	0.0044	0.028	0.158	0.875	-0.050 0.059
state_slim_HI	-0.1297	0.032	-4.068	0.000	-0.192 -0.067
state_slim_IA	0.1112	0.031	3.607	0.000	0.051 0.172

state_slim_ID	0.1291	0.032	3.974	0.000	0.065	0.193
state_slim_IL	0.0015	0.027	0.054	0.957	-0.051	0.054
state_slim_IN	-0.0265	0.028	-0.958	0.338	-0.081	0.028
state_slim_KS	0.0657	0.028	2.373	0.018	0.011	0.120
state_slim_KY	-0.0638	0.029	-2.183	0.029	-0.121	-0.007
state_slim_LA	0.1243	0.030	4.178	0.000	0.066	0.183
state_slim_MA	0.0061	0.027	0.226	0.821	-0.047	0.059
state_slim_MD	-0.0400	0.028	-1.435	0.151	-0.095	0.015
state_slim_ME	0.0449	0.038	1.183	0.237	-0.030	0.119
state_slim_MI	0.0033	0.027	0.122	0.903	-0.049	0.056
state_slim_MN	0.0136	0.027	0.500	0.617	-0.040	0.067
state_slim_MO	0.0595	0.028	2.129	0.033	0.005	0.114
state_slim_MS	-0.0438	0.032	-1.388	0.165	-0.106	0.018
state_slim_MT	0.3476	0.038	9.165	0.000	0.273	0.422
state_slim_NC	-0.0907	0.027	-3.356	0.001	-0.144	-0.038
state_slim_ND	-0.1554	0.034	-4.597	0.000	-0.222	-0.089
state_slim_NE	-0.0897	0.029	-3.071	0.002	-0.147	-0.032
state_slim_NH	-0.0023	0.031	-0.076	0.939	-0.062	0.058
state_slim_NJ	-0.0370	0.027	-1.374	0.169	-0.090	0.016
state_slim_NM	-0.0332	0.031	-1.065	0.287	-0.094	0.028
state_slim_NV	0.0091	0.030	0.307	0.759	-0.049	0.067
state_slim_NY	-0.0685	0.027	-2.586	0.010	-0.120	-0.017

state_slim_OH	0.0142	0.027	0.525	0.599	-0.039 0.067
state_slim_OK	0.0242	0.029	0.819	0.413	-0.034 0.082
state_slim_OR	-0.0395	0.027	-1.476	0.140	-0.092 0.013
state_slim_PA	-0.0208	0.027	-0.774	0.439	-0.074 0.032
state_slim_PR	-0.0920	0.056	-1.640	0.101	-0.202 0.018
state_slim_RI	0.0973	0.034	2.823	0.005	0.030 0.165
state_slim_SC	0.0160	0.030	0.541	0.589	-0.042 0.074
state_slim_SD	-0.0316	0.042	-0.754	0.451	-0.114 0.050
state_slim_TN	-0.0151	0.028	-0.539	0.590	-0.070 0.040
state_slim_TX	0.0828	0.026	3.130	0.002	0.031 0.135
state_slim_UT	-0.0641	0.027	-2.377	0.017	-0.117 -0.011
state_slim_VA	-0.0642	0.028	-2.305	0.021	-0.119 -0.010
state_slim_VT	0.0044	0.039	0.114	0.909	-0.071 0.080
state_slim_WA	0.0736	0.027	2.735	0.006	0.021 0.126
state_slim_WI	-0.0228	0.027	-0.840	0.401	-0.076 0.030
state_slim_WV	-0.0749	0.030	-2.461	0.014	-0.135 -0.015
state_slim_WY	-0.0368	0.043	-0.867	0.386	-0.120 0.046

Omnibus:	19771.646	Durbin-Watson:	1.657
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26380.167
Skew:	0.855	Prob(JB):	0.00
Kurtosis:	3.551	Cond. No.	6.97e+03

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 6.97e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Key coefficients (Δ acquisition probability; 'coef(pp)' = percentage points):

term	coef	se	coef(pp)
const	0.1160	0.0266	11.6050
innov_quantity	0.0004	0.0002	0.0381
innov_quality	0.0033	0.0002	0.3277
year_c	0.0008	0.0002	0.0841

[Saved] Programming Test Task8 outputs:
- /Users/zhangshuhai/Desktop/Programming Test Task8/task8_key_coefficients_20250923-172204.csv
- /Users/zhangshuhai/Desktop/Programming Test Task8/task8_model_summary_20250923-172204.txt
- /Users/zhangshuhai/Desktop/Programming Test Task8/task8_design_matrix_cols_20250923-172204.txt

Process finished with exit code 0

Interpretations:

We analyzed which startups get acquired and how that relates to their early innovation, the year, technology field, and state. The model is a linear probability model, so each coefficient is the change in acquisition probability measured in percentage points, holding other factors constant. It explains a meaningful share of variation for a yes/no outcome ($R^2 \approx 0.34$). The baseline probability for a reference startup is about 11.6%.

Innovation quality matters far more than quantity. Each additional forward citation to a startup's early patent(s) is associated with about +0.33 percentage points higher chance of acquisition (coef ≈ 0.0033). Ten more citations correspond to roughly +3.3 points. In contrast, each extra patent adds only about +0.04 points (coef ≈ 0.0004), so five more patents add about +0.2 points. Buyers seem to value impactful ideas—signaled by citations—more than simply having more patents.

Acquisition likelihood rises slightly over time (about +0.08 points per calendar year; coef ≈ 0.0008), consistent with a long-run increase in M&A activity and market depth. Technology fields differ substantially even after controlling for innovation: some sectors show very large positive associations, suggesting hotter M&A demand there. States also vary, implying local ecosystems and networks matter beyond industry mix.

Uncertainty is quantified with heteroscedasticity-robust (HC1) standard errors. The high condition number indicates multicollinearity among some variables, which can make certain estimates noisier, though the broad patterns—quality over quantity, modest upward trend, clear tech and location differences—are robust.

In short, a startup with more highly cited early patents is meaningfully more likely to be acquired; piling up additional patents helps much less. Timing, sector, and place still shape the odds.

Suggestions:

Link-function robustness: replicate with Logit/Probit and report marginal effects; compare signs and magnitudes with LPM.

Variance robustness: compute clustered SE (e.g., by state or founding year) to account for within-group correlation.

Distributional robustness: re-estimate with winsorized or log(1+x) versions of quantity/quality; also report median/IQR contrasts.

Event-type sensitivity: (i) mergers only; (ii) assignment+merger excluding strings like “name change”/“security”; check stability of coefficients.

Heterogeneity: interact quality × tech field and quantity × founding-year to see if effects differ across technologies and over time.