

# Major Northeast Cities' Airbnb Listings Price Prediction

Yash Gudimalla, Pranavnath Bathula, Anish Jha



# The Problem

## The Issue:

- Airbnb pricing is highly subjective - hosts guess prices, and guests struggle to assess value
- Two Airbnbs can look identical in photos but have vastly different price tags with no apparent reason

## The Goal:

- Build a data-driven system to predict the "fair" nightly price of listings in Major Northeast Cities
- Move from "gut feeling" to "quantitative assessment"

# Pipeline (1/4) (Raw CSVs)

```
data > Washington.DC > listings.csv
1 id,listing_url,scraper_id,last_scraped,source,name,description,neighborhood,overview,picture_url,host_id,host_url,host_name,host_since,host_location,host_about,host_respon
2 3344,https://www.airbnb.com/rooms/3344,2825861745457,2025-06-17,city scrape,"White House-Center City, entire 2BR/2BA Condo","This listing is for full, exclusive-to-the
3 3688,https://www.airbnb.com/rooms/3688,2825861745457,2025-06-17,city scrape,Vista's Hideaway,"IMPORTANT NOTES-Br /> Carefully read and be sure that you understand the c
3943,https://www.airbnb.com/rooms/3943,2825861745457,2025-06-17,city scrape,Historic Rowhouse Near Monuments,"You will be staying in high ceiling bedroom with the pri
4
5 I have been living and working in DC for the last 15 years. I am knowledgeable about the city.
6 I enjoy being outdoors - hiking in the city, working out, gardening, exploring new places and sailing."within an hour,100%,100%,f,https://a8.muscache.com/in/pictures/user/4597,2825861745457,2025-06-17,city scrape,Capitol Hill Bedroom walk to Metro,"This is the middle bedroom upstairs in a restored Victor
4529,https://www.airbnb.com/rooms/4529,2825861745457,2025-06-17,previous scrape,Bertina's House Part One,"This is large private bedroom with plenty of closet space.
7 "A/N/A,N/A,N/A",https://a8.muscache.com/in/pictures/user/5083/original/786d418b-f4c-d9b1-288f83b5c2c1.jpg?aki_policy=profile_small,https://a8.muscache.com/in/pic
11 5569,https://www.airbnb.com/rooms/5569,2825861745457,2025-06-17,city scrape,Cozy apt in Adams Morgan,"This is a 1 br (bedroom + living room in Adams Morgan on a quiet st
12
13 I love Washington, DC, and am happy to share the city with guests."within an hour,100%,0%,f,https://a8.muscache.com/in/users/6537/profile_pic/1417196379/original.jpg?ak
14 7183,https://www.airbnb.com/rooms/7183,2825861745457,2025-06-17,city scrape,Lovely guest suite in a quiet but close-in neighborhood,"Private guest suite with cathedral
15
16 A bit about me...I am a former Californian who, after 16 years in DC, is now an honorary Washington native. Being well traveled both domestically and internationally has
17
18 We very much enjoy hosting visitors to our Nation's Capital and ensuring that they are able to make the most of their visit.
19
20 Charlotte Perry
21 Founder, LUXNB
22 Your short-term rental source in Washington, DC."within an hour,100%,100%,f,https://a8.muscache.com/in/pictures/user/ec8b33c3-9b0b-4f93-44d1-6d4f6d8e872.jpg?aki_policy=
23 11785,https://www.airbnb.com/rooms/11785,2825861745457,2025-06-17,city scrape,Sanctuary near Cathedral,"Our neighborhood is informally known as Cathedral Heights, giv
24
25 We have our own pet care business, which I am happy to share the city with guests."within an hour,100%,100%,f,https://a8.muscache.com/in/pictures/user/58f3c7af-a6a3-4a88-8f16-efdc6ac7c3bc
26 12442,https://www.airbnb.com/rooms/12442,2825861745457,2025-06-17,city scrape,Peaches & Cream near Cathedral,"There is so much to love in Cathedral Heights! The trees
27
28 We have our own pet care business, which I am happy to share the city with guests."within an hour,100%,100%,f,https://a8.muscache.com/in/pictures/user/58f3c7af-a6a3-4a88-8f16-efdc6ac7c3bc
29 13744,https://www.airbnb.com/rooms/13744,2825861745457,2025-06-17,city scrape,Heart of the City -HalfBlock to METRO, Restaurants,"Please do NOT use Instant booking.
30
31 I've traveled all over the world - backpacking for 2-3 years at a time, so I really value having a good "home base.""
32
33 Now I'm a full-time writer, with seven published novels. (The best known is "Shackleton's Stowaway.") Writing is a pretty solitary occupation, so I especially like
34
35 I have over 150 - 5 star - reviews from guests on other sites, but sorry - can't post them here!"within a day,100%,0%,f,https://a8.muscache.com/in/users/53202/profile_pic/14218
36 14218,https://www.airbnb.com/rooms/14218,2825861745457,2025-06-17,city scrape,Cozy Comfort near Nat'l Cathedral,"https://a8.muscache.com/in/pictures/user/53202/profile_pic/14218
```

Raw Washington D.C. csv file

```
data > NYC > listings.csv
1 id,listing_url,scraper_id,last_scraped,source,name,description,neighborhood,overview,picture_url,host_id,host_url,host_name,host_since,host_location,host_about,host_respon
2 48024319,https://www.airbnb.com/rooms/48024319,20251001171547,2025-10-02,city scrape,Room close to Manhattan for FEMALE guests,"This cozy spacious room includes a twin
48033186,https://www.airbnb.com/rooms/48033186,20251001171547,2025-10-02,previous scrape,Soho LES East village private room downtown,"https://a8.muscache.com/pictures/a
48037137,https://www.airbnb.com/rooms/48037137,20251001171547,2025-10-02,previous scrape,Sunset Park - Quiet and close to subway,"Cozy, lovely bedroom with a comfortable
48038018,https://www.airbnb.com/rooms/48038018,20251001171547,2025-10-02,previous scrape,Cozy One Bedroom in Clinton Hill,"This sunny one-bedroom apartment is located in
48039416,https://www.airbnb.com/rooms/48039416,20251001171547,2025-10-02,city scrape,xL dojo in shared green yogi palace apt 6,"New York City Living at its best. A mas
48043908,https://www.airbnb.com/rooms/48043908,20251001171547,2025-10-01,city scrape,Cozy 2 Bedroom Spacious Apartment near Manhattan,"This 2 bed, furnished apt on the 2
48044212,https://www.airbnb.com/rooms/48044212,20251001171547,2025-10-02,previous scrape,Beautiful apartment on quiet side of Bushwick,"Modern and cozy apartment in Bush
48043430,https://www.airbnb.com/rooms/48043430,20251001171547,2025-10-02,city scrape,Cozy room in Williamsburg,"This place is located in Williamsburg, close to popular d
Also a host and I love Traveling."N/A,N/A,68%,f,https://a8.muscache.com/in/pictures/user/72729938-ce02-4cfc-93c5-071b0d46658d.jpg?aki_policy=profile_small,https://a8.m
48025740,https://www.airbnb.com/rooms/48025740,20251001171547,2025-10-02,city scrape,House of Oye - A Historic Brownstone Mansion,"Located on the prestigious St. Marks W
12
13 I am a former Wall St professional turned Entrepreneur. I have a Nomadic Soul and have been to 65 countries. Hoping to reach 100 by 75.
14
15 My motto: Go!See!Do!Share!
16
17 House0fOye,"a few days or more,0%,0%,f,https://a8.muscache.com/in/pictures/user/ed5fe198-bca2-47e7-b877-d28ac454e27.jpg?aki_policy=profile_small,https://a8.muscache.co
18 2595,https://www.airbnb.com/rooms/2595,20251001171547,2025-10-02,city scrape,Skylit Studio Oasis | Midtown Manhattan Sanctuary,"Prime Midtown | Spacious 500 Sq Ft | Pyra
19
20 I am a Sound Therapy Practitioner and Kundalini Yoga & Meditation teacher. I work with energy and sound for relaxation and healing, using Synchronic song, singing bowls,
21
22 Any questions, please text or call Jennifer at 646.498.8710,"a few days or more,45%,24%,f,https://a8.muscache.com/in/pictures/user/58f3c7af-a6a3-4a88-8f16-efdc6ac7c3bc
23 6848,https://www.airbnb.com/rooms/6848,20251001171547,2025-10-02,city scrape,Only 2 steps to Manhattan studio,"Comfortable studio apartment with super comfortable king b
24
25 Want to take a break from the city we'll tell you about wilderness canoeing on the Delaware River - just an hour and a half's drive. We can tell you where to find the g
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2
```

# Pipeline (2/4) (Cleaning CSVs to SQL DB)

```
# Process each city's listings
all_cleaned_data = []

for city in cities:
    city_dir = data_dir / city
    listings_path = city_dir / "listings.csv"

    if not listings_path.exists():
        print(f"Skipping {city}: {listings_path} not found")
        continue

    print(f"\n{'='*60}")
    print(f"Processing {city}...")
    print(f"{'='*60}")

    # Read the listings CSV
    df = pd.read_csv(listings_path)
    print(f"Original shape: {df.shape}")

    # Basic price cleaning
    df_clean = df.copy()

    # Remove $ and commas from price, convert to numeric
    price_str = (
        df_clean["price"]
        .astype(str)
        .str.replace("$", "", regex=False)
        .str.replace(",", "", regex=False)
        .str.strip()
    )

    df_clean["price"] = pd.to_numeric(price_str, errors="coerce")

    # Remove rows with missing or invalid prices
    original_count = len(df_clean)
    df_clean = df_clean[df_clean["price"].notna()]
    df_clean = df_clean[df_clean["price"] > 0]
```



```
-- SQLite Schema for Airbnb Price Predictor
-- Enable foreign key constraints
PRAGMA foreign_keys = ON;

CREATE TABLE neighbourhood (
    neighbourhood_id INTEGER PRIMARY KEY AUTOINCREMENT,
    borough TEXT,
    neighbourhood_name TEXT NOT NULL
);

CREATE TABLE listing (
    listing_id INTEGER PRIMARY KEY,

    neighbourhood_id INTEGER REFERENCES neighbourhood(neighbourhood_id),
    city TEXT, -- NYC, Boston, or Washington DC

    --Host Information
    host_id INTEGER,
    host_name TEXT,
    host_since DATE,
    host_is_superhost INTEGER, -- 0 = false, 1 = true

    --Property Information
    room_type TEXT,
    property_type TEXT,
    accommodates INTEGER,
    bedrooms INTEGER,
    beds INTEGER,
    bathrooms REAL,
    bathrooms_text TEXT,

    latitude REAL,
    longitude REAL,

    --Price and Availability
    price REAL,
    number_of_reviews INTEGER,
    availability_365 INTEGER,
    estimated_revenue REAL,

    --Review Information
    first_review DATE,
    last_review DATE,
    review_scores_rating REAL,

    instant_bookable INTEGER, -- 0 = false, 1 = true
    calculated_host_listings_count INTEGER,
    reviews_per_month REAL
);
```

This layer acts as the system's quality control. It changes raw, untyped CSVs into a structured relational database. By removing text artifacts, such as currency symbols, and applying strict SQL schema rules, we make sure that only clean, type-safe data goes to the modeling stage.

notebooks/clean\_csvs.ipynb

sql/schema/schema\_sqlite.sql

# Pipeline (3/4) (Loading DB to Creating Features)

```
# Load all city data
all_city_data = {}

for city_config in cities_config:
    city_folder = city_config["folder"]
    city_name = city_config["name"]

    listings_file = processed_dir / f"{city_folder}_listings_cleaned.csv"
    neighbourhoods_file = data_dir / city_folder / "neighbourhoods.csv"

    if not listings_file.exists():
        print(f"⚠ Skipping {city_name}: {listings_file} not found")
        continue

    if not neighbourhoods_file.exists():
        print(f"⚠ Skipping {city_name}: {neighbourhoods_file} not found")
        continue

    print(f"\n{'='*60}")
    print(f>Loading data for {city_name}...")
    print(f"{'='*60}")

    # Load neighbourhoods
    print(f>Loading neighbourhoods from {neighbourhoods_file}...")
    df_neighbourhoods = pd.read_csv(neighbourhoods_file)
    print(f"   Loaded {len(df_neighbourhoods)} neighbourhoods")

    # Load listings
    print(f>Loading listings from {listings_file}...")
    df_listings = pd.read_csv(listings_file)
    print(f"   Loaded {len(df_listings)} listings")

    all_city_data[city_name] = {
        "neighbourhoods": df_neighbourhoods,
        "listings": df_listings,
        "folder": city_folder
    }
```



```
# 3.1 Target transform: log price (for later modeling)
df_features["log_price"] = np.log1p(df_features["price"])

# 3.2 Price per capacity features (ratios)
accom = df_features["accommodates"].replace(0, np.nan)
beds = df_features["beds"].replace(0, np.nan)
bedrooms = df_features["bedrooms"].replace(0, np.nan)

df_features["price_per_accommodate"] = df_features["price"] / accom
df_features["price_per_bed"] = df_features["price"] / beds
df_features["price_per_bedroom"] = df_features["price"] / bedrooms

# 3.3 Availability-based features
df_features["available_days_365"] = df_features["availability_365"]
df_features["availability_rate_365"] = df_features["available_days_365"] / 365.0
df_features["blocked_or_booked_days_365"] = 365 - df_features["available_days_365"]
df_features["blocked_or_booked_rate_365"] = (
    df_features["blocked_or_booked_days_365"] / 365.0
)

# 3.4 Review-based transforms
df_features["log_number_of_reviews"] = np.log1p(df_features["number_of_reviews"])
rpm = df_features["reviews_per_month"].clip(lower=0)
df_features["log_reviews_per_month"] = np.log1p(rpm)

# 3.5 Legacy features (for backward compatibility)
df_features["availability_ratio"] = df_features["availability_rate_365"]
df_features["is_high_rating"] = (df_features["review_scores_rating"] >= 4.0).astype(int)
df_features["is_active_host"] = (df_features["reviews_per_month"] > 0).astype(int)

df_features.head()
```

This step connects static storage with active modeling. After filling the centralized database with standardized records from multiple cities, we programmatically create valuable features, such as price-per-bedroom ratios and log-transformed metrics. These features reveal complex market patterns that the raw data alone cannot show.

sql/etl/populate\_database.ipynb

notebooks/features.ipynb

# Pipeline (4/4) (Using Features to Create Models and Predict)

```
num_features = numeric_features + binary_features
cat_features = categorical_features

numeric_transformer = Pipeline([
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler()),
])

categorical_transformer = Pipeline([
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("encoder", OneHotEncoder(handle_unknown="ignore")),
])

preprocess = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, num_features),
        ("cat", categorical_transformer, cat_features),
    ]
)

# Build the full pipeline: preprocessing + linear regression
linreg_model = Pipeline(steps=[
    ("preprocess", preprocess),
    ("regressor", LinearRegression())
])

# Fit the model on the training data
linreg_model.fit(X_train, y_train)

# Predict on the test data
y_pred_lr = linreg_model.predict(X_test)

# Evaluate metrics for linear regression
rmse_lr = sqrt(mean_squared_error(y_test, y_pred_lr))
mae_lr = mean_absolute_error(y_test, y_pred_lr)
r2_lr = r2_score(y_test, y_pred_lr)

print("Global mean baseline RMSE: ", rmse_global)
print("Neighbourhood mean baseline RMSE: ", rmse_neigh)
print("Linear regression RMSE: ", rmse_lr)
print("Linear regression MAE: ", mae_lr)
print("Linear regression R^2: ", r2_lr)
```

Linear Regression model

```
# Random Forest Regressor (tree-based, non-linear model)
# Using the same preprocessing pipeline as linear regression
rf_model = Pipeline(steps=[
    ("preprocess", preprocess),
    ("regressor", RandomForestRegressor(
        n_estimators=200,
        max_depth=None,
        n_jobs=-1,
        random_state=42
    ))
])

# Fit on training data
rf_model.fit(X_train, y_train)

# Predict on test data
y_pred_rf = rf_model.predict(X_test)

# Evaluate metrics for Random Forest
rmse_rf = sqrt(mean_squared_error(y_test, y_pred_rf))
mae_rf = mean_absolute_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

print("Random Forest RMSE: ", rmse_rf)
print("Random Forest MAE: ", mae_rf)
print("Random Forest R^2: ", r2_rf)
```

Random Forest model

This final stage carries out the machine learning workflow by combining preprocessing, such as scaling and imputation, with modeling in one pipeline. We train both Linear Regression and Random Forest algorithms on our engineered features to create price predictions. We then compare their performance using metrics like RMSE and  $R^2$ .

# Feature Engineering

- Started with the cleaned listing and neighbourhood tables. Constructed per-listing features (size, availability, reviews, host behaviour, log-transformed price and review counts) using a join.
- Developed neighbourhood and city-level context features (average ratings, number of listings, rates for superhosts, and entire homes) to constitute each listing's local market context.
- Bucketed continuous variables (capacity, host listings count, rating bands), encoded the room/property type and city as categorical variables; eliminated price-related leakage columns; and exported the feature engineered listings to an CSV file for use in developing models in the future.

```
city_env = (
    df_features
    .groupby("city")
    .agg(
        city_listing_count=("city", "size"),
        city_superhost_rate=("host_is_superhost", "mean"),
        city_avg_rating=("review_scores_rating", "mean"),
        city_avg_reviews_per_month=("reviews_per_month", "mean"),
    )
    .reset_index()
)

if "is_entire_home" in df_features.columns:
    entire_share = (
        df_features
        .groupby("city")["is_entire_home"]
        .mean()
        .rename("city_entire_home_share")
        .reset_index()
    )
    city_env = city_env.merge(entire_share, on="city", how="left")
else:
    entire_share = (
        df_features
        .assign(is_entire_home=(df_features["room_type"] == "Entire home/apt").astype(int))
        .groupby("city")["is_entire_home"]
        .mean()
        .rename("city_entire_home_share")
        .reset_index()
    )
    city_env = city_env.merge(entire_share, on="city", how="left")
```

# Our Results & Performance

=== Model comparison on test set ===

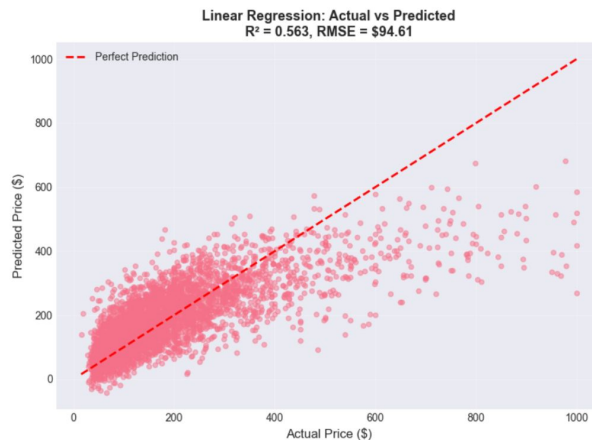
Global mean baseline RMSE:	143.09
Neighbourhood mean baseline RMSE:	124.57
Linear Regression RMSE:	94.61
Random Forest RMSE:	75.69

MAE scores:

Linear Regression MAE:	63.27
Random Forest MAE:	46.61

R<sup>2</sup> scores:

Linear Regression R <sup>2</sup> :	0.563
Random Forest R <sup>2</sup> :	0.720



Our analysis shows that the Random Forest Regressor is the best model. It achieved an R<sup>2</sup> of 0.720 and an RMSE of \$75.69, which is a 47% reduction in error compared to the neighborhood baseline. The scatter plots illustrate that the Random Forest predictions (green) cluster closely along the perfect prediction line. This demonstrates a much better ability to capture complex non-linear pricing factors than the scattered Linear Regression model.

# Conclusion

This project shows the potential of combining data engineering, machine learning, and SQL to tackle important market challenges, such as evaluating fair property pricing in major Northeast cities: NYC, Boston, and Washington D.C. By carefully removing data leakage and focusing on valid features, our Random Forest model acts as a dependable support tool. It explains about 72% of pricing variance ( $R^2$ : 0.72) and achieves an RMSE of \$75.69, which is a 47% improvement over neighborhood baselines. In the end, this complete pipeline demonstrates how managing structured data, with over 20,000 listings in a normalized SQLite schema, and targeted algorithms can work together. This effort creates repeatable solutions that directly enhance decision-making and transparency for both hosts and guests.