

SQL: Exercises

You can download this .qmd file from [here](#). Just hit the Download Raw File button.

The code in [15_SQL.qmd](#) walked us through many of the examples in MDSR Chapter 15; now, we present a set of practice exercises in converting from the tidyverse to SQL.

```
library(tidyverse)
library(mdsr)
library(dbplyr)
library(DBI)

# connect to the database which lives on a remote server maintained by
#   St. Olaf's IT department

library(RMariaDB)
con <- dbConnect(
  MariaDB(), host = "mdb.stolaf.edu",
  user = "ruser", password = "ruserpass",
  dbname = "flight_data"
)
```

On Your Own - Adapting 164 Code

These problems are based on class exercises from SDS 164, so you've already solved them in R! Now we're going to try to duplicate those solutions in SQL (but with 2023 data instead of 2013).

```
# Read in 2013 NYC flights data
library(nycflights13)
flights_nyc13 <- nycflights13::flights
planes_nyc13 <- nycflights13::planes
```

1. Summarize carriers flying to MSP by number of flights and proportion that are cancelled (assuming that a missing arrival time indicates a cancelled flight). [This was #4 in 17_longer_pipelines.Rmd.]

```
# Original solution from SDS 164
flights_nyc13 |>
  mutate(carrier = fct_collapse(carrier, "Delta +" = c("DL", "9E"),
                                "American +" = c("AA", "MQ"),
                                "United +" = c("EV", "00", "UA"))) |>
  filter(dest == "MSP") |>
  group_by(origin, carrier) |>
  summarize(n_flights = n(),
            num_cancelled = sum(is.na(arr_time)),
            prop_cancelled = mean(is.na(arr_time)))
```

```
# A tibble: 5 x 5
# Groups:   origin [3]
  origin carrier    n_flights num_cancelled prop_cancelled
  <chr>   <fct>         <int>         <int>         <dbl>
1 EWR    Delta +           598             10         0.0167
2 EWR    United +        1779             105         0.0590
3 JFK    Delta +        1095              41         0.0374
4 LGA    Delta +        2420              25         0.0103
5 LGA    American +      1293              62         0.0480
```

First duplicate the output above, then check trends in 2023 across all origins. Here are a few hints:

- use `flightdata` instead of `flights_nyc13`
- remember that `flights_nyc13` only contained 2013 and 3 NYC origin airports (EWR, JFK, LGA)
- `is.na` can be replaced with `CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END` or with `CASE WHEN cancelled = 1 THEN 1 ELSE 0 END`
- `CASE WHEN` can also be used replace `fct_collapse`

```
flights_23 <- dbGetQuery(con,
'SELECT
  origin,
  COUNT(*) AS numflights,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) AS num_cancelled,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS prop_cancelled,
  CASE WHEN Reporting_Airline IN ('\DL\','\9E\') THEN \'delta+\\' WHEN Reporting_Airline IN
```

```

FROM flightdata AS o
WHERE year = 2023 AND origin IN ('\JFK\' , \'EWR\' , \'LGA\' )
GROUP BY origin, carrier_type
LIMIT 0, 30')

```

```

flights_23_all <- dbGetQuery(con,
'SELECT
  origin,
  COUNT(*) AS numflights,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) AS num_cancelled,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS prop_cancelled,
  CASE WHEN Reporting_Airline IN ('\DL\' , \'9E\' ) THEN \'delta+\' WHEN Reporting_Airline IN
FROM flightdata AS o
WHERE year = 2023
GROUP BY origin, carrier_type
LIMIT 0, 30')

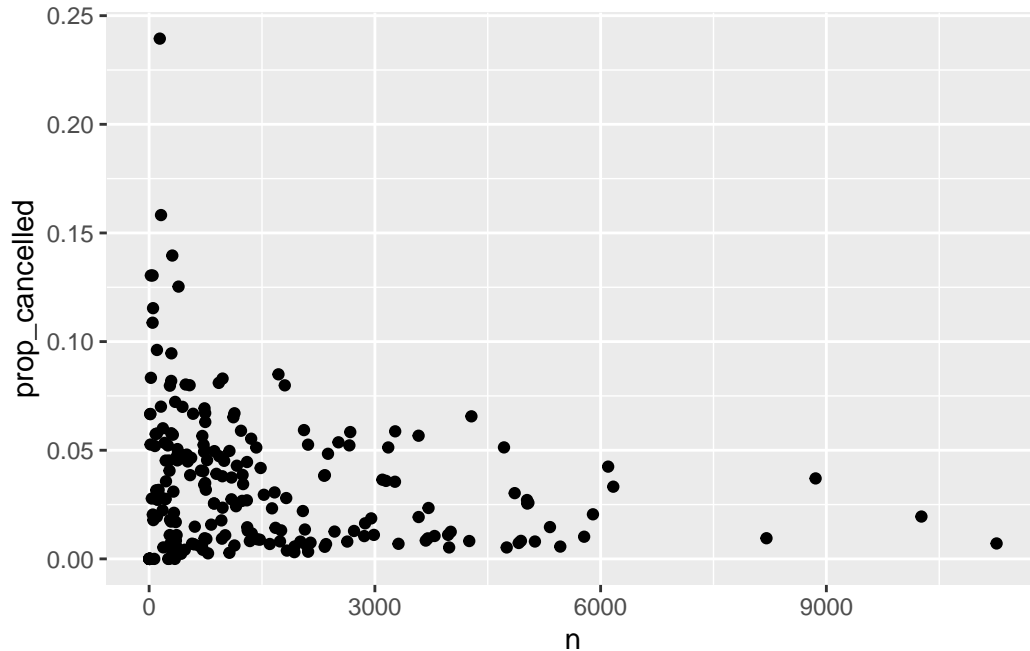
```

2. Plot number of flights vs. proportion cancelled for every origin-destination pair (assuming that a missing arrival time indicates a cancelled flight). [This was #7 in 17_longer_pipelines.Rmd.]

```

# Original solution from SDS 164
flights_nyc13 |>
  group_by(origin, dest) |>
  summarize(n = n(),
            prop_cancelled = mean(is.na(arr_time))) |>
  filter(prop_cancelled < 1) |>
  ggplot(aes(n, prop_cancelled)) +
  geom_point()

```



First duplicate the plot above for 2023 data, then check trends across all origins. Do all of the data wrangling in SQL. Here are a few hints:

- use `flightdata` instead of `flights_nyc13`
- remember that `flights_nyc13` only contained 2013 and 3 NYC origin airports (EWR, JFK, LGA)
- use an `sql` chunk and an `r` chunk
- include `connection =` and `output.var =` in your `sql` chunk header (this doesn't seem to work with `dbGetQuery()`...)

```
SELECT
  origin,
  dest,
  COUNT(*) AS numflights,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) AS num_cancelled,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS prop_cancelled
FROM flightdata AS o
WHERE year = 2023 AND origin IN ('JFK', 'EWR', 'LGA')
GROUP BY origin, dest
HAVING prop_cancelled > 0
LIMIT 0, 300
```

flights_cancelled

	origin	dest	numflights	num_cancelled	prop_cancelled
1	EWR	ALB	668	35	0.0524
2	EWR	ANC	95	4	0.0421
3	EWR	ATL	5915	118	0.0199
4	EWR	AUS	2175	63	0.0290
5	EWR	AVL	589	27	0.0458
6	EWR	AVP	145	5	0.0345
7	EWR	BGR	668	50	0.0749
8	EWR	BNA	2209	60	0.0272
9	EWR	BOS	4387	165	0.0376
10	EWR	BQN	365	10	0.0274
11	EWR	BTX	902	37	0.0410
12	EWR	BUF	1108	47	0.0424
13	EWR	BZN	200	6	0.0300
14	EWR	CHS	1646	45	0.0273
15	EWR	CLE	2252	62	0.0275
16	EWR	CLT	4028	106	0.0263
17	EWR	CMH	2048	95	0.0464
18	EWR	CVG	2232	81	0.0363
19	EWR	DCA	869	31	0.0357
20	EWR	DEN	3251	76	0.0234
21	EWR	DFW	3775	108	0.0286
22	EWR	DSM	88	1	0.0114
23	EWR	DTW	3409	96	0.0282
24	EWR	EGE	95	2	0.0211
25	EWR	EYW	665	8	0.0120
26	EWR	FLL	5502	163	0.0296
27	EWR	GRR	837	48	0.0573
28	EWR	GSO	963	48	0.0498
29	EWR	GSP	919	37	0.0403
30	EWR	HHH	217	3	0.0138
31	EWR	HNL	273	4	0.0147
32	EWR	IAD	1890	41	0.0217
33	EWR	IAH	3631	77	0.0212
34	EWR	ILM	589	34	0.0577
35	EWR	IND	2326	107	0.0460
36	EWR	JAC	194	14	0.0722
37	EWR	JAX	1369	29	0.0212
38	EWR	LAS	3048	59	0.0194
39	EWR	LAX	5906	116	0.0196

40	EWR	MCI	687	18	0.0262
41	EWR	MCO	7262	167	0.0230
42	EWR	MDT	510	39	0.0765
43	EWR	MEM	686	41	0.0598
44	EWR	MHT	617	29	0.0470
45	EWR	MIA	4673	95	0.0203
46	EWR	MKE	907	37	0.0408
47	EWR	MSN	399	19	0.0476
48	EWR	MSP	2381	53	0.0223
49	EWR	MSY	1426	23	0.0161
50	EWR	MYR	1021	32	0.0313
51	EWR	OAK	215	2	0.0093
52	EWR	ORD	5540	140	0.0253
53	EWR	ORF	1141	60	0.0526
54	EWR	PBI	3260	66	0.0202
55	EWR	PDX	888	18	0.0203
56	EWR	PHL	263	25	0.0951
57	EWR	PHX	2637	33	0.0125
58	EWR	PIT	2889	133	0.0460
59	EWR	PVD	656	31	0.0473
60	EWR	PWM	1274	56	0.0440
61	EWR	RDU	2813	71	0.0252
62	EWR	RIC	182	4	0.0220
63	EWR	ROC	832	27	0.0325
64	EWR	RSW	2185	44	0.0201
65	EWR	SAN	1681	38	0.0226
66	EWR	SAT	700	6	0.0086
67	EWR	SAV	1052	33	0.0314
68	EWR	SDF	1056	44	0.0417
69	EWR	SEA	2096	37	0.0177
70	EWR	SFO	4211	75	0.0178
71	EWR	SJU	2171	41	0.0189
72	EWR	SLC	1052	21	0.0200
73	EWR	SMF	365	14	0.0384
74	EWR	SNA	1081	38	0.0352
75	EWR	SRQ	1090	22	0.0202
76	EWR	STL	689	14	0.0203
77	EWR	SYR	749	24	0.0320
78	EWR	TPA	3084	71	0.0230
79	EWR	TVC	73	4	0.0548
80	EWR	VPS	109	2	0.0183
81	JFK	ABQ	217	10	0.0461
82	JFK	ACK	444	18	0.0405

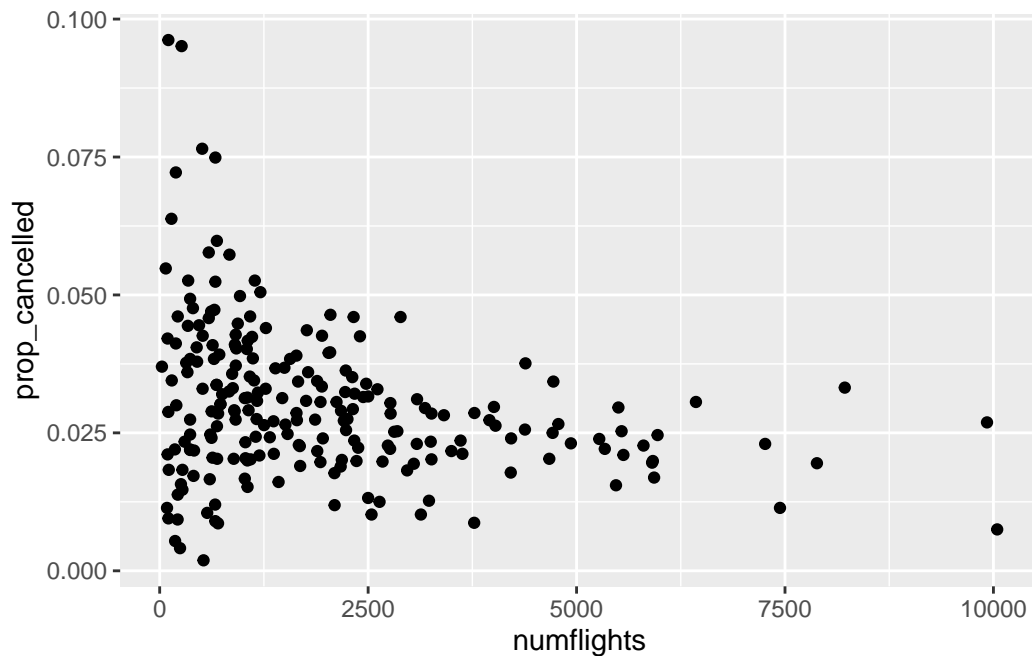
83	JFK	ATL	3772	33	0.0087
84	JFK	AUS	2673	53	0.0198
85	JFK	BGR	516	22	0.0426
86	JFK	BNA	2310	81	0.0351
87	JFK	BOS	6432	197	0.0306
88	JFK	BQN	623	18	0.0289
89	JFK	BTX	1660	57	0.0343
90	JFK	BUF	3087	96	0.0311
91	JFK	BUR	365	18	0.0493
92	JFK	BWI	937	42	0.0448
93	JFK	CHS	1756	54	0.0308
94	JFK	CLE	1888	65	0.0344
95	JFK	CLT	3499	76	0.0217
96	JFK	CMH	2225	72	0.0324
97	JFK	CVG	1864	51	0.0274
98	JFK	DCA	2770	79	0.0285
99	JFK	DEN	1508	40	0.0265
100	JFK	DFW	1928	38	0.0197
101	JFK	DTW	2501	79	0.0316
102	JFK	EGE	105	1	0.0095
103	JFK	FLL	3955	108	0.0273
104	JFK	HNL	667	6	0.0090
105	JFK	HYA	104	10	0.0962
106	JFK	IAD	1029	21	0.0204
107	JFK	IAH	406	7	0.0172
108	JFK	IND	2121	65	0.0306
109	JFK	ITH	715	28	0.0392
110	JFK	JAX	1929	59	0.0306
111	JFK	LAS	2968	54	0.0182
112	JFK	LAX	10045	75	0.0075
113	JFK	MCI	729	22	0.0302
114	JFK	MCO	4932	114	0.0231
115	JFK	MIA	5930	100	0.0169
116	JFK	MKE	729	22	0.0302
117	JFK	MSP	1196	25	0.0209
118	JFK	MSY	1152	28	0.0243
119	JFK	MVY	338	15	0.0444
120	JFK	ONT	365	8	0.0219
121	JFK	ORD	2737	62	0.0227
122	JFK	ORF	2338	75	0.0321
123	JFK	ORH	683	23	0.0337
124	JFK	PBI	2235	57	0.0255
125	JFK	PDX	1054	16	0.0152

126	JFK	PHX	2541	26	0.0102
127	JFK	PIT	1780	64	0.0360
128	JFK	PSE	333	12	0.0360
129	JFK	PSP	185	1	0.0054
130	JFK	PWM	1208	61	0.0505
131	JFK	RDU	4722	162	0.0343
132	JFK	RIC	1066	31	0.0291
133	JFK	RNO	141	9	0.0638
134	JFK	ROC	1945	65	0.0334
135	JFK	RSW	1252	33	0.0264
136	JFK	SAN	2500	33	0.0132
137	JFK	SAT	601	10	0.0166
138	JFK	SAV	1641	47	0.0286
139	JFK	SEA	3133	32	0.0102
140	JFK	SFO	7440	85	0.0114
141	JFK	SJC	194	8	0.0412
142	JFK	SJU	3230	41	0.0127
143	JFK	SLC	2098	25	0.0119
144	JFK	SMF	364	8	0.0220
145	JFK	SRQ	365	9	0.0247
146	JFK	STT	526	1	0.0019
147	JFK	SYR	1564	60	0.0384
148	JFK	TPA	2334	55	0.0236
149	LGA	ACK	472	21	0.0445
150	LGA	ALB	913	34	0.0372
151	LGA	ATL	7883	154	0.0195
152	LGA	AVL	1028	24	0.0233
153	LGA	BDL	701	20	0.0285
154	LGA	BGM	651	25	0.0384
155	LGA	BGR	1085	50	0.0461
156	LGA	BHM	912	39	0.0428
157	LGA	BNA	4713	118	0.0250
158	LGA	BOS	8217	273	0.0332
159	LGA	BTV	894	26	0.0291
160	LGA	BUF	2045	81	0.0396
161	LGA	CAE	682	23	0.0337
162	LGA	CHO	1496	55	0.0368
163	LGA	CHS	1957	47	0.0240
164	LGA	CLE	2477	84	0.0339
165	LGA	CLT	5339	118	0.0221
166	LGA	CMH	3260	93	0.0285
167	LGA	CVG	2441	77	0.0315
168	LGA	DAL	1365	37	0.0271

169	LGA	DAY	607	15	0.0247
170	LGA	DCA	5803	132	0.0227
171	LGA	DEN	5272	126	0.0239
172	LGA	DFW	5972	147	0.0246
173	LGA	DSM	635	26	0.0409
174	LGA	DTW	4381	112	0.0256
175	LGA	EYW	244	1	0.0041
176	LGA	FLL	4782	127	0.0266
177	LGA	GRR	911	25	0.0274
178	LGA	GSO	2026	80	0.0395
179	LGA	GSP	1388	51	0.0367
180	LGA	HHH	255	4	0.0157
181	LGA	HOU	1021	17	0.0167
182	LGA	HYA	104	3	0.0288
183	LGA	IAD	412	9	0.0218
184	LGA	IAH	3609	85	0.0236
185	LGA	ILM	1765	77	0.0436
186	LGA	IND	2767	84	0.0304
187	LGA	JAX	1669	38	0.0228
188	LGA	LIT	515	17	0.0330
189	LGA	MCI	2317	68	0.0293
190	LGA	MCO	5562	117	0.0210
191	LGA	MDW	2612	86	0.0329
192	LGA	MEM	1470	46	0.0313
193	LGA	MIA	5473	85	0.0155
194	LGA	MKE	1271	42	0.0330
195	LGA	MSN	449	17	0.0379
196	LGA	MSP	2361	47	0.0199
197	LGA	MSY	1685	32	0.0190
198	LGA	MVY	342	18	0.0526
199	LGA	MYR	1168	36	0.0308
200	LGA	OKC	877	29	0.0331
201	LGA	OMA	1118	43	0.0385
202	LGA	ORD	9923	267	0.0269
203	LGA	ORF	1947	83	0.0426
204	LGA	ORH	299	7	0.0234
205	LGA	PBI	2764	61	0.0221
206	LGA	PIT	3182	94	0.0295
207	LGA	PVD	899	26	0.0289
208	LGA	PWM	1640	64	0.0390
209	LGA	RDU	4009	119	0.0297
210	LGA	RIC	2402	102	0.0425
211	LGA	ROA	634	13	0.0205

212	LGA	ROC	1178	38	0.0323
213	LGA	RSW	1321	32	0.0242
214	LGA	SAV	1163	32	0.0275
215	LGA	SCE	273	5	0.0183
216	LGA	SDF	1534	38	0.0248
217	LGA	SLC	27	1	0.0370
218	LGA	SRQ	570	6	0.0105
219	LGA	STL	4215	101	0.0240
220	LGA	SYR	1045	42	0.0402
221	LGA	TPA	2851	72	0.0253
222	LGA	TUL	318	12	0.0377
223	LGA	TYS	1132	39	0.0345
224	LGA	XNA	622	15	0.0241

```
ggplot(flights_cancelled, aes(x = numflights, prop_cancelled)) +
  geom_point()
```



```
SELECT
  origin,
  dest,
  COUNT(*) AS numflights,
  SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) AS num_cancelled,
```

```

SUM(CASE WHEN ArrTime IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS prop_cancelled
FROM flightdata AS o
WHERE year = 2023
GROUP BY origin, dest
HAVING prop_cancelled > 0
LIMIT 0, 500

```

```
flights_cancelled_all
```

	origin	dest	numflights	num_cancelled	prop_cancelled
1	ABE	ATL	994	3	0.0030
2	ABE	BNA	139	1	0.0072
3	ABE	CLT	1098	9	0.0082
4	ABE	DEN	58	1	0.0172
5	ABE	FLL	95	2	0.0211
6	ABE	ORD	674	6	0.0089
7	ABE	PGD	204	2	0.0098
8	ABE	SFB	614	1	0.0016
9	ABI	DFW	1359	17	0.0125
10	ABQ	ATL	1001	5	0.0050
11	ABQ	AUS	1250	13	0.0104
12	ABQ	BUR	355	2	0.0056
13	ABQ	BWI	167	1	0.0060
14	ABQ	DAL	1725	20	0.0116
15	ABQ	DEN	2764	43	0.0156
16	ABQ	DFW	2255	30	0.0133
17	ABQ	HOU	940	2	0.0021
18	ABQ	IAH	607	7	0.0115
19	ABQ	JFK	217	12	0.0553
20	ABQ	LAS	1985	19	0.0096
21	ABQ	MCI	245	1	0.0041
22	ABQ	MDW	392	6	0.0153
23	ABQ	MSP	206	3	0.0146
24	ABQ	ORD	815	7	0.0086
25	ABQ	PDX	155	3	0.0194
26	ABQ	PHX	2986	13	0.0044
27	ABQ	SAN	617	3	0.0049
28	ABQ	SEA	466	4	0.0086
29	ABQ	SFO	478	4	0.0084
30	ABQ	SLC	1046	5	0.0048
31	ABR	MSP	725	14	0.0193
32	ABY	ATL	814	2	0.0025

33	ACK	BOS	184	9	0.0489
34	ACK	DCA	281	15	0.0534
35	ACK	HPN	139	9	0.0647
36	ACK	JFK	444	29	0.0653
37	ACK	LGA	472	29	0.0614
38	ACK	PHL	75	5	0.0667
39	ACT	DFW	1068	22	0.0206
40	ACV	DEN	314	5	0.0159
41	ACV	LAX	354	5	0.0141
42	ACV	SFO	997	25	0.0251
43	ACY	FLL	523	9	0.0172
44	ACY	MCO	1001	17	0.0170
45	ACY	MYR	308	2	0.0065
46	ACY	PBI	366	7	0.0191
47	ACY	RSW	396	11	0.0278
48	ACY	TPA	439	12	0.0273
49	ADK	ANC	104	8	0.0769
50	ADQ	ANC	790	27	0.0342
51	AEX	ATL	729	4	0.0055
52	AEX	DFW	1079	19	0.0176
53	AGS	ATL	2031	7	0.0034
54	AGS	CLT	793	2	0.0025
55	AGS	DCA	364	10	0.0275
56	AGS	DFW	376	2	0.0053
57	AGS	ORD	12	1	0.0833
58	AKN	ANC	215	2	0.0093
59	ALB	ATL	997	10	0.0100
60	ALB	BNA	108	1	0.0093
61	ALB	BWI	1664	17	0.0102
62	ALB	CLT	1051	9	0.0086
63	ALB	DCA	1258	34	0.0270
64	ALB	DEN	136	1	0.0074
65	ALB	DTW	1331	15	0.0113
66	ALB	EWR	669	41	0.0613
67	ALB	FLL	347	9	0.0259
68	ALB	LGA	913	28	0.0307
69	ALB	MCO	920	11	0.0120
70	ALB	MDW	664	8	0.0120
71	ALB	MIA	21	1	0.0476
72	ALB	MYR	103	2	0.0194
73	ALB	ORD	1033	14	0.0136
74	ALB	TPA	227	1	0.0044
75	ALO	ORD	75	4	0.0533

76	ALW	SEA	343	13	0.0379
77	AMA	AUS	318	2	0.0063
78	AMA	DAL	1354	20	0.0148
79	AMA	DEN	707	8	0.0113
80	AMA	DFW	2064	33	0.0160
81	AMA	IAH	465	10	0.0215
82	AMA	LAS	365	2	0.0055
83	ANC	ADK	104	8	0.0769
84	ANC	ADQ	790	33	0.0418
85	ANC	AKN	215	2	0.0093
86	ANC	BET	722	20	0.0277
87	ANC	BRW	363	9	0.0248
88	ANC	CDV	362	5	0.0138
89	ANC	DEN	542	6	0.0111
90	ANC	DLG	145	4	0.0276
91	ANC	EWR	95	5	0.0526
92	ANC	FAI	677	2	0.0030
93	ANC	HNL	356	2	0.0056
94	ANC	JNU	1185	17	0.0143
95	ANC	LAX	363	1	0.0028
96	ANC	MSP	611	5	0.0082
97	ANC	OME	537	26	0.0484
98	ANC	ORD	758	4	0.0053
99	ANC	OTZ	166	13	0.0783
100	ANC	PDX	498	3	0.0060
101	ANC	PHX	294	1	0.0034
102	ANC	SCC	415	10	0.0241
103	ANC	SEA	7518	45	0.0060
104	ANC	SFO	214	3	0.0140
105	APN	DTW	386	8	0.0207
106	APN	PLN	65	7	0.1077
107	ASE	ATL	216	21	0.0972
108	ASE	AUS	206	13	0.0631
109	ASE	DEN	2416	182	0.0753
110	ASE	DFW	1021	73	0.0715
111	ASE	IAH	544	36	0.0662
112	ASE	LAX	934	76	0.0814
113	ASE	ORD	841	61	0.0725
114	ASE	PHX	105	17	0.1619
115	ASE	SFO	308	28	0.0909
116	ATL	ABE	994	3	0.0030
117	ATL	ABQ	1001	5	0.0050
118	ATL	ABY	814	2	0.0025

119	ATL	AEX	730	4	0.0055
120	ATL	AGS	2032	7	0.0034
121	ATL	ALB	997	10	0.0100
122	ATL	ASE	216	19	0.0880
123	ATL	ATW	367	2	0.0054
124	ATL	AUS	4197	38	0.0091
125	ATL	AVL	2101	12	0.0057
126	ATL	BDL	2033	15	0.0074
127	ATL	BHM	2760	37	0.0134
128	ATL	BMI	727	2	0.0028
129	ATL	BNA	3861	23	0.0060
130	ATL	BOI	447	4	0.0089
131	ATL	BOS	5066	44	0.0087
132	ATL	BQK	729	3	0.0041
133	ATL	BTR	1491	7	0.0047
134	ATL	BTW	365	1	0.0027
135	ATL	BUF	1275	14	0.0110
136	ATL	BWI	6447	55	0.0085
137	ATL	BZN	329	1	0.0030
138	ATL	CAE	2012	12	0.0060
139	ATL	CHA	2285	13	0.0057
140	ATL	CHO	1055	4	0.0038
141	ATL	CHS	2708	22	0.0081
142	ATL	CID	365	3	0.0082
143	ATL	CLE	3389	32	0.0094
144	ATL	CLT	4938	31	0.0063
145	ATL	CMH	3056	24	0.0079
146	ATL	CRW	989	6	0.0061
147	ATL	CSG	948	6	0.0063
148	ATL	CVG	2723	28	0.0103
149	ATL	DAB	1515	17	0.0112
150	ATL	DAL	3504	40	0.0114
151	ATL	DAY	1299	2	0.0015
152	ATL	DCA	6954	73	0.0105
153	ATL	DEN	6176	65	0.0105
154	ATL	DFW	6359	90	0.0142
155	ATL	DHN	730	1	0.0014
156	ATL	DSM	1077	5	0.0046
157	ATL	DTW	5354	29	0.0054
158	ATL	ECP	1839	15	0.0082
159	ATL	EGE	133	1	0.0075
160	ATL	ELP	684	5	0.0073
161	ATL	EVV	868	1	0.0012

162	ATL	EWR	5913	136	0.0230
163	ATL	EYW	1486	8	0.0054
164	ATL	FAY	730	11	0.0151
165	ATL	FLL	8040	104	0.0129
166	ATL	FSD	365	9	0.0247
167	ATL	FWA	893	2	0.0022
168	ATL	GEG	447	5	0.0112
169	ATL	GNV	1647	10	0.0061
170	ATL	GPT	1196	6	0.0050
171	ATL	GRB	364	5	0.0137
172	ATL	GRR	1317	12	0.0091
173	ATL	GSO	2252	13	0.0058
174	ATL	GSP	3074	10	0.0033
175	ATL	GTR	802	3	0.0037
176	ATL	HDN	108	4	0.0370
177	ATL	HHH	203	2	0.0099
178	ATL	HNL	365	1	0.0027
179	ATL	HOU	3096	17	0.0055
180	ATL	HPN	1769	7	0.0040
181	ATL	HSV	2080	17	0.0082
182	ATL	IAD	3490	23	0.0066
183	ATL	IAH	4747	56	0.0118
184	ATL	ICT	1042	3	0.0029
185	ATL	ILM	1235	6	0.0049
186	ATL	IND	3384	22	0.0065
187	ATL	JAC	222	2	0.0090
188	ATL	JAN	3130	10	0.0032
189	ATL	JAX	4338	32	0.0074
190	ATL	JFK	3781	39	0.0103
191	ATL	LAS	5144	50	0.0097
192	ATL	LAX	5545	43	0.0078
193	ATL	LEX	1888	11	0.0058
194	ATL	LFT	992	2	0.0020
195	ATL	LGA	7883	156	0.0198
196	ATL	LIT	2307	11	0.0048
197	ATL	MCI	3272	24	0.0073
198	ATL	MCO	9039	55	0.0061
199	ATL	MDW	3597	60	0.0167
200	ATL	MEM	3538	32	0.0090
201	ATL	MGM	1363	10	0.0073
202	ATL	MIA	6799	53	0.0078
203	ATL	MKE	2404	26	0.0108
204	ATL	MLB	1415	8	0.0057

205	ATL	MLI	729	2	0.0027
206	ATL	MLU	752	1	0.0013
207	ATL	MOB	1398	10	0.0072
208	ATL	MSN	1033	3	0.0029
209	ATL	MSP	3789	28	0.0074
210	ATL	MSY	4567	30	0.0066
211	ATL	MTJ	102	2	0.0196
212	ATL	MYR	1430	10	0.0070
213	ATL	OAJ	730	2	0.0027
214	ATL	OKC	2079	8	0.0038
215	ATL	OMA	2136	13	0.0061
216	ATL	ONT	1070	20	0.0187
217	ATL	ORD	6178	70	0.0113
218	ATL	ORF	2361	12	0.0051
219	ATL	PBI	4047	35	0.0086
220	ATL	PDX	1253	12	0.0096
221	ATL	PHL	5815	46	0.0079
222	ATL	PHX	3649	25	0.0069
223	ATL	PIT	2765	15	0.0054
224	ATL	PNS	2372	23	0.0097
225	ATL	PSP	143	2	0.0140
226	ATL	PVD	997	10	0.0100
227	ATL	PWM	521	6	0.0115
228	ATL	RDU	4480	22	0.0049
229	ATL	RIC	3414	21	0.0062
230	ATL	ROA	1076	12	0.0112
231	ATL	ROC	998	10	0.0100
232	ATL	RSW	3132	27	0.0086
233	ATL	SAN	2172	11	0.0051
234	ATL	SAT	3149	26	0.0083
235	ATL	SAV	2539	28	0.0110
236	ATL	SBN	915	2	0.0022
237	ATL	SDF	2982	23	0.0077
238	ATL	SEA	3525	18	0.0051
239	ATL	SFO	3279	35	0.0107
240	ATL	SGF	1031	3	0.0029
241	ATL	SHV	1023	5	0.0049
242	ATL	SJC	784	2	0.0026
243	ATL	SJU	1916	14	0.0073
244	ATL	SLC	3012	16	0.0053
245	ATL	SMF	930	6	0.0065
246	ATL	SNA	1060	7	0.0066
247	ATL	SRQ	2781	18	0.0065

248	ATL	STL	3456	23	0.0067
249	ATL	STX	106	1	0.0094
250	ATL	SYR	996	8	0.0080
251	ATL	TLH	1830	14	0.0077
252	ATL	TPA	6698	75	0.0112
253	ATL	TRI	1344	2	0.0015
254	ATL	TTN	316	13	0.0411
255	ATL	TUL	1454	6	0.0041
256	ATL	TUS	799	2	0.0025
257	ATL	TYS	2480	13	0.0052
258	ATL	VLD	819	3	0.0037
259	ATL	VPS	1913	21	0.0110
260	ATL	XNA	1460	5	0.0034
261	ATW	ATL	367	3	0.0082
262	ATW	AZA	192	1	0.0052
263	ATW	CLT	150	1	0.0067
264	ATW	DEN	444	6	0.0135
265	ATW	DTW	731	8	0.0109
266	ATW	FLL	56	1	0.0179
267	ATW	LAS	146	2	0.0137
268	ATW	MSP	827	7	0.0085
269	ATW	ORD	1296	17	0.0131
270	ATW	PGD	140	2	0.0143
271	ATW	PIE	137	2	0.0146
272	ATW	SFB	151	2	0.0132
273	AUS	ABQ	1250	11	0.0088
274	AUS	AMA	318	2	0.0063
275	AUS	ASE	206	6	0.0291
276	AUS	ATL	4192	36	0.0086
277	AUS	BNA	2690	30	0.0112
278	AUS	BOS	1877	29	0.0155
279	AUS	BUR	356	3	0.0084
280	AUS	BWI	1530	17	0.0111
281	AUS	CLT	1914	12	0.0063
282	AUS	CMH	355	4	0.0113
283	AUS	CVG	720	5	0.0069
284	AUS	DAL	2956	65	0.0220
285	AUS	DCA	365	5	0.0137
286	AUS	DEN	4450	71	0.0160
287	AUS	DFW	4850	89	0.0184
288	AUS	DTW	1477	15	0.0102
289	AUS	EGE	62	2	0.0323
290	AUS	ELP	1958	21	0.0107

291	AUS	EWR	2176	75	0.0345
292	AUS	FLL	1254	24	0.0191
293	AUS	FSD	54	1	0.0185
294	AUS	HOU	1757	22	0.0125
295	AUS	HRL	620	7	0.0113
296	AUS	IAD	1250	11	0.0088
297	AUS	IAH	2308	54	0.0234
298	AUS	IND	1021	7	0.0069
299	AUS	JAX	460	2	0.0043
300	AUS	JFK	2672	51	0.0191
301	AUS	LAS	3373	46	0.0136
302	AUS	LAX	4365	54	0.0124
303	AUS	LBB	329	2	0.0061
304	AUS	LGB	638	5	0.0078
305	AUS	MAF	316	3	0.0095
306	AUS	MCI	990	9	0.0091
307	AUS	MCO	3043	42	0.0138
308	AUS	MDW	1615	27	0.0167
309	AUS	MEM	318	4	0.0126
310	AUS	MIA	1923	19	0.0099
311	AUS	MSP	1484	16	0.0108
312	AUS	MSY	2459	22	0.0089
313	AUS	OAK	372	1	0.0027
314	AUS	OKC	731	8	0.0109
315	AUS	ONT	356	3	0.0084
316	AUS	ORD	3327	48	0.0144
317	AUS	PDX	358	5	0.0140
318	AUS	PHL	848	5	0.0059
319	AUS	PHX	3290	37	0.0112
320	AUS	PSP	174	3	0.0172
321	AUS	PVU	102	1	0.0098
322	AUS	RDU	1561	8	0.0051
323	AUS	RNO	311	4	0.0129
324	AUS	SAN	2098	28	0.0133
325	AUS	SEA	2025	16	0.0079
326	AUS	SFO	2028	22	0.0108
327	AUS	SJC	1576	14	0.0089
328	AUS	SLC	1575	13	0.0083
329	AUS	SMF	678	2	0.0029
330	AUS	SNA	845	8	0.0095
331	AUS	SRQ	65	1	0.0154
332	AUS	STL	1096	15	0.0137
333	AUS	TPA	1406	13	0.0092

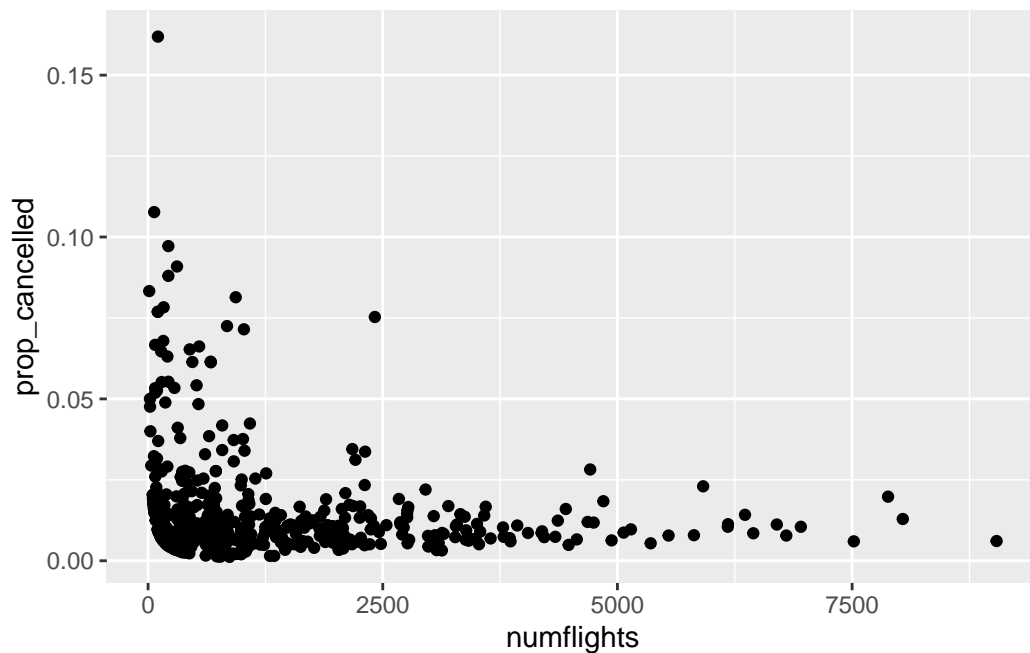
334	AUS	TUL	633	8	0.0126
335	AUS	VPS	104	2	0.0192
336	AVL	ATL	2101	13	0.0062
337	AVL	BOS	210	1	0.0048
338	AVL	BWI	88	2	0.0227
339	AVL	CLT	2110	15	0.0071
340	AVL	DCA	486	8	0.0165
341	AVL	DFW	601	5	0.0083
342	AVL	EWR	590	15	0.0254
343	AVL	EYW	105	3	0.0286
344	AVL	FLL	635	11	0.0173
345	AVL	LGA	1028	35	0.0340
346	AVL	MIA	145	2	0.0138
347	AVL	MSP	203	1	0.0049
348	AVL	ORD	848	12	0.0142
349	AVL	PBI	115	1	0.0087
350	AVL	PHX	63	1	0.0159
351	AVL	PIE	430	2	0.0047
352	AVP	CLT	1063	9	0.0085
353	AVP	EWR	145	8	0.0552
354	AVP	ORD	835	10	0.0120
355	AZA	BIS	193	1	0.0052
356	AZA	BZN	126	1	0.0079
357	AZA	CVG	105	1	0.0095
358	AZA	DSM	170	2	0.0118
359	AZA	FAR	279	1	0.0036
360	AZA	FSD	297	4	0.0135
361	AZA	FWA	68	1	0.0147
362	AZA	GEG	104	1	0.0096
363	AZA	GFK	61	1	0.0164
364	AZA	IDA	190	1	0.0053
365	AZA	MOT	80	1	0.0125
366	AZA	PSC	115	1	0.0087
367	AZA	PVU	637	9	0.0141
368	AZA	RAP	152	1	0.0066
369	AZA	RFD	139	1	0.0072
370	AZA	STC	77	2	0.0260
371	AZO	DTW	810	6	0.0074
372	BDL	ATL	2035	16	0.0079
373	BDL	BNA	365	3	0.0082
374	BDL	BWI	1678	23	0.0137
375	BDL	CLT	1625	7	0.0043
376	BDL	DCA	1896	36	0.0190

377	BDL	DEN	593	7	0.0118
378	BDL	DFW	523	6	0.0115
379	BDL	DTW	1225	9	0.0073
380	BDL	FLL	571	12	0.0210
381	BDL	IAD	694	11	0.0159
382	BDL	LAS	49	1	0.0204
383	BDL	LGA	702	12	0.0171
384	BDL	MCO	2185	37	0.0169
385	BDL	MDW	674	13	0.0193
386	BDL	MIA	529	6	0.0113
387	BDL	MSP	861	7	0.0081
388	BDL	MYR	278	1	0.0036
389	BDL	ORD	1827	18	0.0099
390	BDL	PBI	430	7	0.0163
391	BDL	PHL	503	5	0.0099
392	BDL	RDU	77	4	0.0519
393	BDL	RSW	212	1	0.0047
394	BDL	SJU	1035	17	0.0164
395	BDL	TPA	831	11	0.0132
396	BET	ANC	722	20	0.0277
397	BFF	DEN	607	20	0.0329
398	BFL	DEN	365	2	0.0055
399	BFL	DFW	646	13	0.0201
400	BFL	PHX	1039	5	0.0048
401	BFL	SFO	567	4	0.0071
402	BGM	LGA	650	25	0.0385
403	BGR	CLT	217	2	0.0092
404	BGR	DCA	1011	38	0.0376
405	BGR	EWR	667	41	0.0615
406	BGR	FLL	20	1	0.0500
407	BGR	JFK	517	28	0.0542
408	BGR	LGA	1085	46	0.0424
409	BGR	ORD	302	6	0.0199
410	BGR	PHL	525	13	0.0248
411	BGR	PIE	162	2	0.0123
412	BGR	SFB	195	4	0.0205
413	BHM	ATL	2759	40	0.0145
414	BHM	CLT	1968	11	0.0056
415	BHM	DAL	731	10	0.0137
416	BHM	DCA	1007	6	0.0060
417	BHM	DEN	1078	9	0.0083
418	BHM	DFW	1836	25	0.0136
419	BHM	DTW	455	4	0.0088

420	BHM	HOU	623	4	0.0064
421	BHM	IAH	440	1	0.0023
422	BHM	LGA	912	34	0.0373
423	BHM	MCO	650	4	0.0062
424	BHM	MDW	669	4	0.0060
425	BHM	MIA	635	4	0.0063
426	BHM	ORD	821	10	0.0122
427	BHM	PHL	433	5	0.0115
428	BHM	TPA	410	3	0.0073
429	BIH	DEN	95	3	0.0316
430	BIH	SFO	162	11	0.0679
431	BIL	DEN	988	23	0.0233
432	BIL	DFW	575	5	0.0087
433	BIL	LAS	114	1	0.0088
434	BIL	MSP	420	7	0.0167
435	BIL	PHX	142	1	0.0070
436	BIL	SLC	1025	7	0.0068
437	BIS	AZA	193	1	0.0052
438	BIS	DEN	1143	29	0.0254
439	BIS	DFW	557	4	0.0072
440	BIS	MSP	1415	20	0.0141
441	BJI	MSP	711	16	0.0225
442	BLI	LAS	863	2	0.0023
443	BLI	OAK	773	1	0.0013
444	BLI	SEA	601	7	0.0116
445	BLV	SAV	25	1	0.0400
446	BLV	SFB	125	2	0.0160
447	BLV	VPS	322	5	0.0155
448	BMI	ATL	727	2	0.0028
449	BMI	DFW	625	8	0.0128
450	BMI	DTW	66	1	0.0152
451	BMI	MCO	34	1	0.0294
452	BMI	ORD	52	1	0.0192
453	BMI	SFB	113	1	0.0088
454	BNA	ABE	139	1	0.0072
455	BNA	ATL	3853	27	0.0070
456	BNA	AUS	2687	32	0.0119
457	BNA	BDL	365	3	0.0082
458	BNA	BOS	2345	33	0.0141
459	BNA	BUR	167	1	0.0060
460	BNA	BWI	1965	21	0.0107
461	BNA	CAK	105	1	0.0095
462	BNA	CHS	750	6	0.0080

463	BNA	CID	157	1	0.0064
464	BNA	CLE	741	6	0.0081
465	BNA	CLT	4220	31	0.0073
466	BNA	CMH	724	5	0.0069
467	BNA	DAL	2259	38	0.0168
468	BNA	DCA	3581	50	0.0140
469	BNA	DEN	3933	43	0.0109
470	BNA	DFW	2772	46	0.0166
471	BNA	DSM	234	2	0.0085
472	BNA	DTW	2076	21	0.0101
473	BNA	ECP	1002	5	0.0050
474	BNA	EWR	2209	69	0.0312
475	BNA	FAR	192	1	0.0052
476	BNA	FLL	2102	44	0.0209
477	BNA	FNT	124	2	0.0161
478	BNA	FSD	192	2	0.0104
479	BNA	HOU	1700	10	0.0059
480	BNA	IAD	1034	4	0.0039
481	BNA	IAH	2374	31	0.0131
482	BNA	ISP	108	2	0.0185
483	BNA	JAX	975	6	0.0062
484	BNA	JFK	2312	78	0.0337
485	BNA	LAS	1929	18	0.0093
486	BNA	LAX	2155	14	0.0065
487	BNA	LGA	4710	133	0.0282
488	BNA	LGB	365	1	0.0027
489	BNA	MCI	1095	7	0.0064
490	BNA	MCO	3199	54	0.0169
491	BNA	MDT	129	2	0.0155
492	BNA	MDW	2144	37	0.0173
493	BNA	MIA	2266	16	0.0071
494	BNA	MKE	776	5	0.0064
495	BNA	MSP	1631	20	0.0123
496	BNA	MSY	1123	6	0.0053
497	BNA	MYR	702	5	0.0071
498	BNA	ORD	4683	56	0.0120
499	BNA	ORF	399	1	0.0025
500	BNA	PGD	206	1	0.0049

```
ggplot(flights_cancelled_all, aes(x = numflights, prop_cancelled)) +
  geom_point()
```



3. Produce a table of weighted plane age by carrier, where weights are based on number of flights per plane. [This was #6 in 26_more_joins.Rmd.]

```
# Original solution from SDS 164
flights_nyc13 |>
  left_join(planes_nyc13, join_by(tailnum)) |>
  mutate(plane_age = 2013 - year.y) |>
  group_by(carrier) |>
  summarize(unique_planes = n_distinct(tailnum),
            mean_weighted_age = mean(plane_age, na.rm =TRUE),
            sd_weighted_age = sd(plane_age, na.rm =TRUE)) |>
  arrange(mean_weighted_age)
```

A tibble: 16 x 4

	carrier	unique_planes	mean_weighted_age	sd_weighted_age
	<chr>	<int>	<dbl>	<dbl>
1	HA	14	1.55	1.14
2	AS	84	3.34	3.07
3	VX	53	4.47	2.14
4	F9	26	4.88	3.67
5	B6	193	6.69	3.29
6	00	28	6.84	2.41
7	9E	204	7.10	2.67

8	US	290	9.10	4.88
9	WN	583	9.15	4.63
10	YV	58	9.31	1.93
11	EV	316	11.3	2.29
12	FL	129	11.4	2.16
13	UA	621	13.2	5.83
14	DL	629	16.4	5.49
15	AA	601	25.9	5.42
16	MQ	238	35.3	3.13

First duplicate the output above for 2023, then check trends across all origins. Do all of the data wrangling in SQL. Here are a few hints:

- use flightdata instead of flights_nyc13
- remember that flights_nyc13 only contained 2013 and 3 NYC origin airports (EWR, JFK, LGA)
- you'll have to merge the flights dataset with the planes dataset
- you can use DISTINCT inside a COUNT()
- investigate SQL clauses for calculating a standard deviation
- you cannot use a derived variable inside a summary clause in SELECT

For bonus points, also merge the airlines dataset and include the name of each carrier and not just the abbreviation!

```
SELECT
  o.Reporting_Airline AS carrier,
  COUNT(DISTINCT o.TAIL_NUMBER) AS unique_planes,
  AVG(2023 - a.year) AS mean_plane_age,
  STDDEV(2023 - a.year) AS sd_plane_age
FROM flightdata AS o
LEFT JOIN planes AS a ON o.TAIL_NUMBER = a.tailnum
WHERE o.year = 2023 AND origin IN ('JFK', 'EWR', 'LGA')
GROUP BY carrier
ORDER BY mean_plane_age
LIMIT 0, 10
```

```
SELECT
  o.Reporting_Airline AS carrier,
  COUNT(DISTINCT o.TAIL_NUMBER) AS unique_planes,
  AVG(2023 - a.year) AS mean_plane_age,
  STDDEV(2023 - a.year) AS sd_plane_age
FROM flightdata AS o
LEFT JOIN planes AS a ON o.TAIL_NUMBER = a.tailnum
```



```
WHERE o.year = 2023
GROUP BY carrier
ORDER BY mean_plane_age
LIMIT 0, 10
```

```
dbDisconnect(con)
```