01. Overview

- a. I used Postgres Pro's public demo database on Russian airline flight data with 3 months of data from 2017. I have always been very interested in the logistics involved in keeping airlines running on time and balancing having full flights without overbooking. Given there are so many dependencies (getting an aircraft or flight crew to an airport in time for the next flight), it seems like a great optimization problem.
- b. For the purposes of this project, I am a data engineer working with a data scientist who is creating a model to forecast air travel demand. The current data is organized at the flight level. However, to conduct her analysis, the data must be restructured at the route level. She also needs some additional variables to feed into her model, such as the distance, occupancy and duration of the route. My hypothesis is structuring the data at the route level will make it easier to identify patterns across flights and reduce the noisiness of the data. A route in this case is all the flights that start and end at the same airport. Only one aircraft type flew each route.

Example: data listed at the flight level, and there are multiple flights for route ABA-ARH

scheduled_departure	1	scheduled_arrival		departure_airport						
2017-07-29 01:40:00-07								Arrived		319
2017-06-24 01:40:00-07	ı	2017-06-24 05:40:00-07	I	ABA	ı	ARH	I	Arrived		319
2017-07-01 01:40:00-07	ı	2017-07-01 05:40:00-07	I	ABA	I	ARH	I	Arrived		319
2017-07-15 01:40:00-07	ı	2017-07-15 05:40:00-07	I	ABA	ı	ARH	I	Arrived		319
2017-08-26 01:40:00-07	ı	2017-08-26 05:40:00-07	I	ABA	ı	ARH	I	Scheduled		319
2017-08-19 01:40:00-07	ı	2017-08-19 05:40:00-07	I	ABA	ı	ARH	I	Scheduled		319
2017-06-17 01-40-00-07		2017-06-17 05-40-00-07		ADA		ADLI		Annivad	1	310

02. Model Structure

- a. Getting my dbt project hooked up to my SQL database took a long time, but writing the data pipeline itself probably took only 3-4 hours. The majority of the time was testing it in SQL and redoing early pieces of the pipeline as I realized what additional data I needed/what data was extraneous. The pipeline takes 14.60 seconds to run. 13.31 seconds of that time was for the incremental view. These measurements were in the dev environment but there were no significant run time differences between dev and prod.
- b. Below is a breakdown of each model. The <u>last page</u> has the full table produced by my 5 level deep pipeline. Table snippets added where it felt relevant to understanding the steps taken.

Step1: Allow for some distance based estimates at the route level Steps:

- Rough point distance estimate between airports (not perfect distance of route flown, but approximation)
- Calculate duration from arrival and departure times
- Identify if timezone change occurs via a boolean if arrival timezone = departure timezone

aircraft_d	code l	depart	ure_airport	arri	val_airport	S	scheduled_duration	I		Ü	timezone_change
SU9		AAQ		EG0			 00:50:00		391.3763978085819	-	
733	- 1	AAQ	1	NOZ		0	05:05:00	ı	2258.0695247517415	ı	t
733	- 1	AAQ	1	SV0		0	01:40:00	I	757.9922296343206	ı	f
319	- 1	ABA	1	ARH		0	04:00:00	I	1881.293217618737	I	t
319	- 1	ABA	1	DME		0	04:25:00	ı	2091.7054925222224	ı	t

Step2: Aggregate ticket and flight data at the route level Steps:

- Count unique flights
- Sum ticket info (seats purchased etc)
- Group by route (aircraft, departure airport, arrival airport)

	-) (_	F F -	-,	1 7		
9					departure_airport		· ·
121	T	Т	76414200.00	Т		Т-	EGO
120	12046	Ι	175036800.00		AAQ	١	SV0
35	2246	Ι	102332500.00		ABA	١	DME
114	986	ī	5718800 00	Ī	ΔΒΔ	I	OVR

Step3: Get available seats to determine occupancy

Steps: Use aircraft info to determine number of seats available per aircraft, multiply by number of flights to get total number of seats.

Step4: Calculate occupancy per route

Steps: Add a column calculating occupancy (purchased seats / total seats)

Step5: Determine busiest average departure airports

Steps: Use a window function to partition avg num flights on departure airport

departure_airport						
AAQ	+	EG0	+	121	1	120.5000000000000000
AAQ	i	SVO	i	120	i	120.50000000000000000
ABA		TOF		115	i	88.00000000000000000
ABA		OVB		114	I	88.00000000000000000
ABA		DME		35	I	88.00000000000000000
AER		VK0		121	I	79.666666666666667
AER		SV0		121	I	79.666666666666667
AER	Ī	KUF	ı	120	١	79.666666666666667

Step6: Make it easy to interpret data

Steps: Add english name of airport and city, extracted from JSON. For readability this only includes the new columns but these were added to all of the additional columns in step4.

departure_airport_name	departure_city	arrival_airport_name	arrival_city
"Anapa Vityazevo Airport"	l "Anapa"	"Belgorod International Airport"	"Belgorod"
"Anapa Vityazevo Airport"	l "Anapa"	"Sheremetyevo International Airport"	"Moscow"
"Abakan Airport"	l "Abakan"	"Domodedovo International Airport"	"Moscow"
"Abakan Airport"	l "Abakan"	"Tolmachevo Airport"	"Novosibirsk"

Lucrative_Routes: Find the routes with the highest fare per distance unit Steps: Select the top 10 routes with the highest fare per distance unit measurements

fare_per_dist_unit								scheduled_duration
						372.38015990111717		
732868.7975011024			1	"Moscow"	I	372.38015990111717	١	00:50:00
726205.6765946784	I	"Moscow"	١	"Novosibirsk"	ı	1733.4250344928032	١	03:25:00
718487.6833230691	I	"Moscow"	ı	"Yekaterinburg"	I	892.4608102316953	١	01:45:00
704560 4047004400		HAI	_	U14		4777 4750744070077		AA AE AA

Delays: Find the routes with max delays, among flights with actual data Steps: Compare scheduled with actual duration, take max per route

			_airport scheduled_dura		
CR2	+ OVB	PYJ	02:50:00	1 02:52:00	00:02:00
CR2	I CSY	I NBC	1 00:25:00	00:26:00	00:01:00
733	LED	I PYJ	1 05:25:00	1 05:33:00	00:08:00
CR2	I GDZ	I DME	01:45:00	01:48:00	1 00:03:00
CR2	l NFG	I VKO	1 03:00:00	1 03:03:00	1 00:03:00
CR2	I ARH	I DME	01:25:00	01:27:00	1 00:02:00

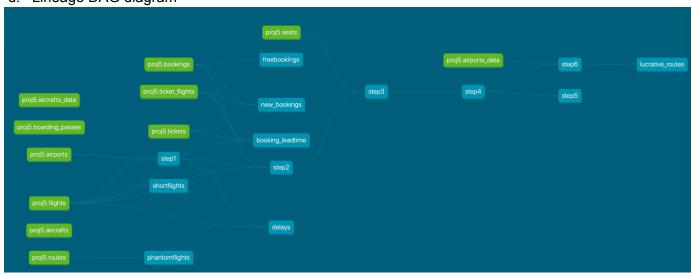
Booking_Leadtime: Create clear picture of how far in advance people purchased tickets Steps: Compare booking date to min flight departure date, calculate difference

		book		1				lead_time
								11 days 14:03:00
000012	I	2017-07-13	23:02:00-07	I	2017-07-28	07:05:00-07	ı	14 days 08:03:00
00002D	١	2017-05-20	08:45:00-07	I	2017-05-30	04:45:00-07	ı	9 days 20:00:00
000000		2017 00 15	04.37.00 07		2017 00 20	00.20.00 07		12 4 20-02-00

New_Bookings: Identify new bookings that haven't yet been analyzed Steps: Create incremental view of most recent bookings

passenger_id		book_date 				fare_conditions		
		2017-05-02 00:24:00-07						18500.00
3889 683019	I	2017-04-30 19:50:00-07	I	0005432000861	I	Economy	I	6200.00
3554 024596	I	2017-04-30 19:50:00-07	I	0005432000862	١	Business	I	18500.00
2026 425060		2017 05 01 10 40 00 07		0005433000000		Farmer		C200 00

- c. See schemas.yml for descriptions
- d. Lineage DAG diagram



e. Luckily for me, the creators of this dataset included a number of constraints that meant the data was pretty clean when it got to me. As such, I tried to focus on situations that might be valid data types but not logical (things a human would immediately find suspicious but a computer might not). I came up with three tests to validate the data behaved how we'd expect of an airline. With more time, I'd like to add some similar tests for newly created

fields. For example, occupancy rate should never exceed 1, and booking lead time should never be negative.

- Phantom flights: ensuring no flights took off and landed from the same airport
- Free bookings: no tickets were booked for \$0
- Short flights: there were no flights under 10 minutes. This could throw off duration calculations down the line.

03. Interesting Findings (Optional)

a. I was very interested to see how the actual duration differed from the scheduled duration. However, for most flights this data was null. For the data available, the greatest difference was 16 minutes. That struck me as surprisingly low given some of the flights were 8 hours long. I excluded this data from my 5 depth model pipeline given the amount of missing data and this bizarre result, which made me think the data available might be systematically different from the data that was missing.

04. Reflection

- a. Originally I envisioned this data being used by a data scientist, but I underestimated how much less data I'd have after rolling it up (70k rows to 450). More likely, this data would be useful to an operations or business analyst trying to get a birds eye view of the airline's opportunities. I also took the data schema at its word, and found it was occasionally aspirational. Data in the schema was missing or incomplete, so that caused me to change some plans (for example, not calculating the median delay because the actual flight data was spotty and the status field contradictory).
- b. Now that I have this pipeline in place I'd love to dig into the analysis. How far in advance do people typically buy tickets (bookings_leadtime) and does this vary based on the length of the trip? Additionally, I'd like to do more with the seat class do people buy business class tickets farther in advance than economy seats? Also looking at occupancy by class would be very cool.
- c. I enjoyed DBT a lot. In the past I've used SQL stored procedures to manage basic data manipulations and this was definitely more modular and easier to iterate with. Also the doc serve functionality was delightful. I never worked out a great solution for building in dbt and testing in SQL without rewriting a lot of code, so if I were to use it again I'd probably want a better SQL solution than just running in the command line, but otherwise I would definitely use it again.
- d. I would prefer to build and write these models and specs in the command line/Visual Studio rather than a GUI for the flexibility it provides. However, I do see value in a GUI particularly for analyst level users whose priority is exploring the existing data rather than creating new pipelining. And I think there would be value to a low-code solution that retained the flexibility of the command line but brought in more visuals.

05. Link to Code

a. Github https://github.com/kailinkoch/airline-dbt

06. Sources

- a. Dataset Information: https://postgrespro.com/docs/postgrespro/10/demodb-bookings
- b. I used the medium dataset here
- c. Other resources

INFO 290-T | Project 5 | Kailin Koch

- i. https://www.postgresql.org/docs/13/app-psql.html
- ii. http://postgresguide.com/utilities/psql.html
- iii. https://stackoverflow.com/questions/40865564/why-command-dt-gives-no-relations-found
- iv. https://stackoverflow.com/questions/3393961/how-to-import-existing-sql-fil es-in-postgresql-8-4
- v. https://stackoverflow.com/questions/1213430/how-to-fully-delete-a-git-rep-ository-created-with-init

Appendix Images on Next Pages

Final Pipeline Full Table (Rotate and zoom to see clearer)

"Tomsk"	"Bogashevo Airport"	"Arkhangelsk"	14096 "Talagi Airport"	8869.20506250	34.062500000000000000000	40178 t	1577.491996340178	83:40:00	I TOF	I ARH	545 CR2	9	13991100.0	3	160
"Tyumen"	"Roshchino International Airport"	"Arkhangelsk"	4548 "Talagi Airport"	8300.790339274548	21.82978723404255319100	82162 t	963.317969882162	02:15:00	MCT I	I ARH	513 CR2		7996300.00	-	235
"Perm"	"Bolshoye Savino Airport"	"Arkhangelsk"	9615 "Talagi Airport"	16553.39173608061	75.21929824561403508800	03106 t	683.7873579303106	1 03:35:00	I PEE	I ARH	029 CN1) 1	11319000.00	3 1	136
"Naryan-Mar"	"Naryan Mar Airport"	"Arkhangelsk"	\$1152 "Talagi Airport"	74136.374859811	76.37288135593220339000	78704 f	404.53690994078704	1 00:55:00	I NAM	I ARH	506 CR2	9 4	29990900.00) 11	590
"Khanty-Mansiysk"	"Khanty Mansiysk Airport"	"Arkhangelsk"	"5539 "Talagi Airport"	17660.11854775539	63.64705882352941176500	19309 t	920.1976734219309	1 02:10:00	I HMA	I ARH	082 CR2) 1	16250800.00	9 1 3	170
"Moscow"	"Domodedovo International Airport"	"Arkhangelsk"	LS436 "Talagi Airport"	64398.186298415436	68.26890756302521008400	38257 f	642.2882751438257	01:25:00	I DME	I ARH	862 CR2	9 4	41362200.00) 11	595
"Moscow"	"Vnukovo International Airport"	l "Sochi"	3 "Sochi International Airport"	352308.6208017173	64.17243689970962698200	80565 f	847.399928280565	1 01:45:00	I VKO	I AER	238 763		298546300.00	2 12	2686;
"Moscow"	"Sheremetyevo International Airport"	l "Sochi"	I "Sochi International Airport"	665256.5343876096	68.89519345421652070200	42186 f	872.5016140342186	1 01:45:00	I SVO	I AER	33512 773		588437400.00	2 1 12	4864;
"Rostov"	"Rostov-on-Don Airport"	l "Sochi"	"Sochi International Airport"	108106.21761430737	49.65942562592847128188	97891 f	263.2133528297091	1 00:35:00	I ROV	I AER	5395 SU9		28455000.00	11	10864
"Novokuznetsk"	"Spichenkovo Airport"	"Sochi"	3 "Sochi International Airport"	12689.91546055592	15.13574660633484162900	37263 t	2212.7176565737263	1 84:55:00	ZON I	I AER	669 733		28079200.00	9 1 3	4426
"Samara"	"Kurumoch International Airport"	" I "Sochi" I	: "Sochi International Airport	255650.1385256342	85.83333333333333333	03678 t	835.5907852503678	1 01:50:00	I KUF	I AER	13390 733		213618900.00) 12	15600
"Krasnoyarsk"	"Yemelyanovo Airport"	l "Sochi"	i "Sochi International Airport	27911.71176799846	32.80933062880324543600	71626 t	2432.2191546071626	1 05:10:00	I KJA	I AER	1294 319		67887400.00	1 3	3944
"Kogalym"	"Kogalym International Airport"	"Sochi"	i "Sochi International Airport"	4882.912948287556	12.11538461538461538500	36395 t	1898.8050981436395	1 84:15:00	I KGP	I AER	252 733		9271700.00		2086
ort" "Nizhniy Novgorod"	"Nizhny Novgorod Strigino International Airport"	:" "Sochi"	I "Sochi International Airport	7617.553942799728	18.392156862745@98@39@@	77302 f	898.989367377302	1 02:05:00	I GOJ	I AER	469 CR2		6848100.00	9 1 5	255
"Belgorod"	"Belgorod International Airport"	Sochi"	"Sochi International Airport	7263.559104236956	34.799382716@49382716@0	28941 f	521.5624937628941	1 02:45:00	I EGO	I AER	451 CN1		3788400.00	5 I 1e	1296
"Tomsk"	"Bogashevo Airport"	l "Abakan"	92869 "Abakan Airport"	17699.64506292869	79.78268869565217391366	97685 f	394.8927223697685	1 01:35:00	I TOF	I ABA	1101 CN1		5394900.00)	1386
"Novosibirsk"	"Tolmachevo Airport"	l "Abakan"	i8878 "Abakan Airport"	15794.858630058878	72.07602339181286549700	50054 t	362.0671848950054	1 01:55:00	I OVB	I ABA	986 CN1		5718800.00		1368
"Moscow"	"Domodedovo International Airport"	l "Abakan"	18699 "Abakan Airport"	1 48922.9962658869	55.32019704433497536900	22224 t	2091.7054925222224	1 84:25:00	I DME	I ABA	246 319	3 2	102332500.00	9 1 3	406
"Moscow"	"Sheremetyevo International Airport"	l "Anapa"	1374 "Anapa Vityazevo Airport"	230921.6284242	77.21794871794871794900	43206 f	757.9922296343206	1 01:40:00	I SVO	I AAQ	12046 733		175036800.00		1560
"Belgorod"	"Belgorod International Airport"	l "Anapa"	195244.783353986 "Anapa Vityazevo Airport"	195244.783353	82.42310641560875862700	85819 f	391.3763978085819	80:50:00	I EGO	J AAQ	674 SU9	9 9	76414200.00	7 12	1173
arrival_city	arrival_airport_name	departure_city	fare_per_dist_unit departure_airport_name	fare_per_dist_	occupancy_rate	timezone_change	distance	total_seats num_flights total_fare purchased_seats aircraft_code departure_airport arrival_airport scheduled_duretion	re_airport arrival_airp	ode departur	ats aircraft_co	purchased_se	total_fare	; num_flight	total_seat:

Working in Dev and Prod

```
- proj5 git:(master) # dbt run
Running mith dbt-0.20.0-b1
Found 10 models, 3 tests, 0 snapshots, 0 analyses, 144 macros, 0 operations, 0 seed files, 11 sources, 0 exposures
15:56:56 | Concurrency: 1 threads (target='prod')
15:56:56 | 1 of 10 START view model bookings2.step1 [RIN]
15:56:56 | 1 of 10 OK created view model bookings2.step1 [CGEATE VIEW in 0.08s]
15:56:56 | 2 of 10 START view model bookings2.step1 [CGEATE VIEW in 0.08s]
15:56:56 | 2 of 10 START view model bookings2.booking_leadtime [RIN]
15:56:56 | 2 of 10 OK created view model bookings2.booking_leadtime [CGEATE VIEW in 0.04s]
15:57:10 | 3 of 10 START incremental model bookings2.new_bookings [RIN]
15:57:10 | 3 of 10 START view model bookings2.step2 [RIN]
15:57:11 | 4 of 10 OK created view model bookings2.step2 [RIN]
15:57:11 | 5 of 10 START view model bookings2.step2 [CGEATE VIEW in 0.83s]
15:57:11 | 5 of 10 START view model bookings2.step2 [CGEATE VIEW in 0.04s]
15:57:11 | 6 of 10 START view model bookings2.step3 [RIN]
15:57:11 | 7 of 10 START view model bookings2.step3 [RIN]
15:57:11 | 7 of 10 START view model bookings2.step3 [RIN]
15:57:11 | 7 of 10 START view model bookings2.step4 [RIN]
15:57:11 | 7 of 10 START view model bookings2.step4 [RIN]
15:57:11 | 7 of 10 START view model bookings2.step5 [RIN]
15:57:11 | 7 of 10 START view model bookings2.step6 [CGEATE VIEW in 0.03s]
15:57:11 | 9 of 10 START view model bookings2.step6 [CGEATE VIEW in 0.03s]
15:57:11 | 9 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
15:57:11 | 1 of 10 START view model bookings2.step5 [CGEATE VIEW in 0.03s]
```

```
ound 10 models, 3 tests, 0 snapshots, 0 analyses, 144 macros, 0 operations, 0 seed files, 11 sources, 0 expos
 15:55:10 | Concurrency: 1 threads (target='dey')
 15:55:10 | 1 of 10 START view model bookings2.step1.
15:55:10 | 1 of 10 OK created view model bookings2.step1.
15:55:10 | 2 of 10 START view model bookings2.booking_leadtime.
15:55:10 | 2 of 10 OK created view model bookings2.booking_leadtime.
15:55:10 | 3 of 10 START view model bookings2.new_bookings.
15:55:25 | 3 of 10 OK created incremental model bookings2.new_bookings.
15:55:25 | 4 of 10 START view model bookings2.step2.
15:55:25 | 4 of 10 START view model bookings2.step2.
15:55:25 | 5 of 10 START view model bookings2.delays.
15:55:25 | 6 of 10 OK created view model bookings2.delays.
15:55:25 | 6 of 10 OK created view model bookings2.step3.
15:55:25 | 6 of 10 START view model bookings2.step3.
15:55:25 | 7 of 10 START view model bookings2.step3.
15:55:25 | 7 of 10 START view model bookings2.step4.
15:55:25 | 7 of 10 START view model bookings2.step4.
15:55:25 | 8 of 10 START view model bookings2.step4.
                                                                                                                                                                                                                          TE VIEW in 0 09s7
                                                                                                                                                                                                                         TE VIEW in 0.04s]
                                                                                                                                                                                                             [RUN]
                                                                                                                                                                                                                           RT 0 2360335 in 14.50s]
                                                                                                                                                                                                             [INSER
                                                                                                                                                                                                                         ATE VIEW in 0.07s]
                                                                                                                                                                                                            [RUN]
                                                                                                                                                                                                            [CREAT
                                                                                                                                                                                                                         TE VIEW in 0.04s]
                                                                                                                                                                                                             [CREATE VIEW in 0.03s]
                                                                                                                                                                                                            [RUN]
 | 15:55:25 | 7 of 10 NC reacted view model bookings2.step4. |
| 15:55:25 | 8 of 10 START view model bookings2.step6. |
| 15:55:25 | 9 of 10 START view model bookings2.step5. |
| 15:55:25 | 9 of 10 OK created view model bookings2.step5. |
| 15:55:25 | 10 of 10 START view model bookings2.step5. |
| 15:55:25 | 10 of 10 START view model bookings2.lucrative_routes. |
                                                                                                                                                                                                            [RUN]
                                                                                                                                                                                                            [CREAT
                                                                                                                                                                                                                           TE VIEW in 0.03s]
                                                                                                                                                                                                            [CREATE VIEW in 0.03s]
[RUN]
[CREATE VIEW in 0.07s]
   L5:55:25 | Finished running 9 view models, 1 incremental model in 15.76s.
     ompleted successfully
```

```
→ proj5 git:(master) x dbt test

Running with dbt=0.20.0-b1

Found 10 models, 3 tests, 0 snapshots, 0 analyses, 144 macros, 0 operations, 0 seed files, 11 sources, 0 exposures

16:00:52 | Concurrency: 1 threads (target='dev')

16:00:52 | 1 of 3 START test freebookings. [RUN]

16:00:52 | 1 of 3 START test freebookings. [PASS in 0.09s]

16:00:52 | 2 of 3 START test phantomflights. [RUN]

16:00:52 | 2 of 3 PASS phantomflights. [PASS in 0.04s]

16:00:52 | 3 of 3 START test shortflights. [RUN]

16:00:52 | 3 of 3 PASS shortflights. [RUN]

16:00:52 | 5 of 3 PASS shortflights. [PASS in 0.04s]

16:00:52 | Finished running 3 tests in 1.02s.
```