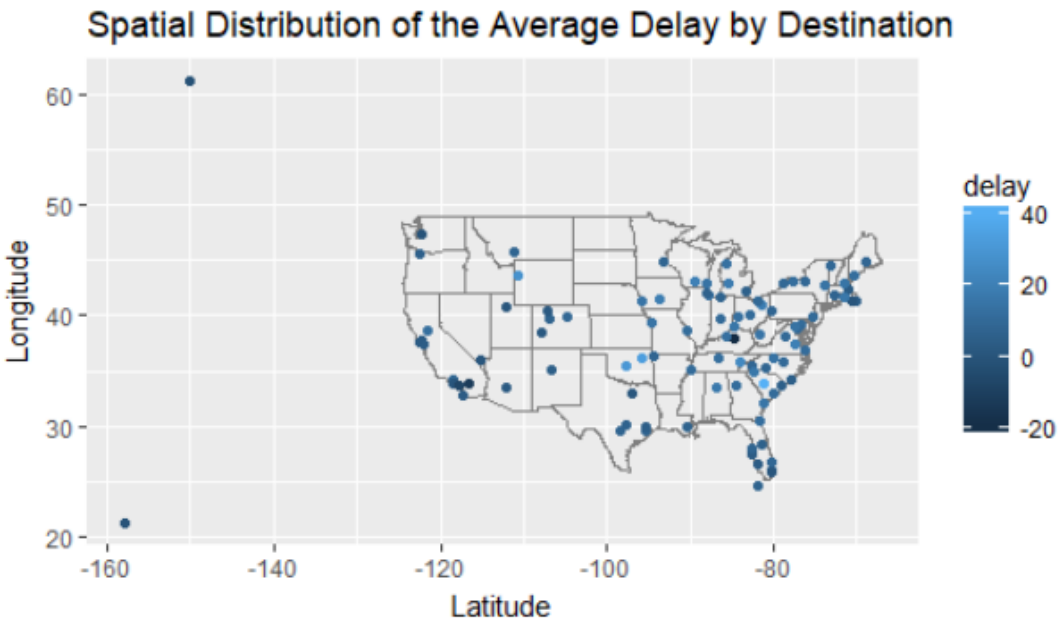
Casey Carr
04/23/2018
HW4 Math 7608

Q1. Compute the average delay by destination, then join on the airports data frame so you can show the spatial distribution of delays.

```
airports %>%
  semi_join(flights, c("faa" = "dest")) %>%
  ggplot(aes(lon, lat)) +
  borders("state") +
  geom_point() +
  coord_quickmap()

avg_dest_delays <-
  flights %>% group_by(dest) %>%
  # arrival delay NA's are cancelled flights
  summarise(delay = mean(arr_delay, na.rm = TRUE)) %>%
  inner_join(airports, by = c(dest = "faa"))

# must input: install.packages("maps")
avg_dest_delays %>%
  ggplot(aes(lon, lat, color = delay)) + borders("state") +
  geom_point() + coord_quickmap() +
  xlab("Latitude") + ylab("Longitude")
```



One could argue that, on average, intermediate distance flights have the longest delays, while short- and long-distance flights typically have no delays.

Q2. What does it mean for a flight to have a missing tail number? What do the tail numbers that don't have a matching record in planes have in common? (Hint: find one variable explains ~90% of the problems.)

By entering the following code, one can see that flights that have a missing tail number are those that were cancelled, that is, they do not have departure time information.

```
flights %>%
  filter(is.na(tailnum))
```

	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time
1	NA	1545	NA	NA	1910	NA	AA	133	NA	JFK	LAX	NA
2	NA	1601	NA	NA	1735	NA	UA	623	NA	EWR	ORD	NA
3	NA	857	NA	NA	1209	NA	UA	714	NA	EWR	MIA	NA
4	NA	645	NA	NA	952	NA	UA	719	NA	EWR	DFW	NA
5	NA	845	NA	NA	1015	NA	9E	3405	NA	JFK	DCA	NA
6	NA	1830	NA	NA	2044	NA	9E	3716	NA	EWR	DTW	NA
7	NA	840	NA	NA	1001	NA	9E	3422	NA	JFK	BOS	NA
8	NA	820	NA	NA	958	NA	9E	3317	NA	JFK	BUF	NA
9	NA	1645	NA	NA	1838	NA	US	123	NA	EWR	CLT	NA
10	NA	755	NA	NA	1012	NA	9E	4023	NA	EWR	CVG	NA
11	NA	1251	NA	NA	1602	NA	UA	421	NA	LGA	IAH	NA
12	NA	1500	NA	NA	1639	NA	UA	685	NA	LGA	ORD	NA
13	NA	700	NA	NA	1007	NA	UA	719	NA	EWR	DFW	NA
14	NA	1700	NA	NA	1813	NA	US	2136	NA	LGA	BOS	NA
15	NA	900	NA	NA	1020	NA	US	2120	NA	LGA	BOS	NA
16	NA	629	NA	NA	805	NA	UA	297	NA	EWR	ORD	NA
17	NA	2045	NA	NA	2216	NA	9E	3395	NA	JFK	DCA	NA
18	NA	2029	NA	NA	2140	NA	9E	3609	NA	JFK	PHL	NA
19	NA	1610	NA	NA	1732	NA	9E	3689	NA	JFK	PHL	NA
20	NA	955	NA	NA	1100	NA	9E	3667	NA	JFK	PHL	NA

Regarding the tail numbers that don't have a matching record in planes, it seems most of them come from the same two carriers. For example, American Airlines (AA) and Envoy Airlines (MQ). It could be that these two carriers simply do not report tail numbers.

```
flights %>%
  anti_join(planes, by = "tailnum") %>%
  count(carrier, sort = TRUE)
```

	carrier	n
1	MQ	25397
2	AA	22558
3	UA	1693
4	9E	1044
5	B6	830
6	US	699
7	FL	187
8	DL	110
9	F9	50
10	WN	38

Q3. Find the 48 hours (over the course of the whole year) that have the worst delays. Cross-reference it with the weather data. Can you see any patterns?

```
frequency_delay <-
  flights %>%
  group_by(month, day) %>%
  summarize(avg_delay = sum(arr_delay + dep_delay, na.rm = TRUE)) %>%
  mutate(twoday_delay = avg_delay + lag(avg_delay)) %>%
  arrange(-twoday_delay)

weather_patterns <-
  weather %>%
  group_by(month, day) %>%
  summarize_at(vars(humid, precip, temp), mean, na.rm = TRUE)

frequency_delay %>%
  left_join(weather_patterns) %>%
  arrange(desc(twoday_delay))
```

TOP 10 WORST DELAYS OVER 48 HOUR PERIOD

	month	day	avg_delay	twoday_delay	humid	precip	temp
1	7	23	80220	174111	78.37958	0.0087500000	79.76500
2	3	8	135269	167538	85.20417	0.0133333333	34.14750
3	6	25	80303	166288	58.99597	0.0000000000	81.39750
4	8	9	72632	164316	83.17708	0.0012500000	77.80500
5	6	28	81320	156325	75.85444	0.0045833333	77.68250
6	7	10	93755	153565	73.02694	0.0011111111	80.26000
7	4	19	81667	149167	85.25806	0.0000000000	59.62000
8	3	9	3192	138461	47.93694	0.0000000000	43.08250
9	5	24	51525	136655	85.96000	0.0065277778	60.18250
10	6	14	39307	135693	74.38472	0.0229166667	61.49750

BOTTOM 10 WORST DELAYS OVER 48 HOUR PERIOD

344	9	10	-5251	-16029	71.65819	0.0000000000	74.47000
345	8	27	-8109	-16638	67.63028	0.0012500000	78.41500
346	8	26	-8529	-16868	56.40958	0.0000000000	74.61000
347	9	9	-10778	-19304	50.44528	0.0000000000	65.04500
348	9	18	-9503	-19423	53.20444	0.0000000000	60.57000
349	9	8	-8526	-22679	53.34264	0.0000000000	73.92000
350	9	5	-15381	-29234	51.04944	0.0000000000	73.30000
351	9	7	-14153	-31764	50.01903	0.0000000000	67.47750
352	10	2	-13615	-31911	60.33597	0.0000000000	72.50500
353	9	6	-17611	-32992	43.72083	0.0000000000	64.54000

Precipitation was clearly higher in the Top 10 Worst Delays, and therefore reasonably explains the higher number of delays as compared to the Bottom 10 Worst Delays.

Q4. Imagine you wanted to draw (approximately) the route each plane flies from its origin to its destination. What variables would you need? What tables would you need to combine? Draw (approximately) the route each plane flies from its origin to its TOP 10 destinations.

We would need to combine the following tables: airports and flights. Then, match both origin and dest in flights with faa in airports.

```
flights %>%
  left_join(airports, by = c("origin" = "faa")) %>%
  left_join(airports, by = c("dest" = "faa")) %>% filter(!is.na(name.y)) %>% filter(!is.na(name.x)) %>%
  group_by(tailnum)
```

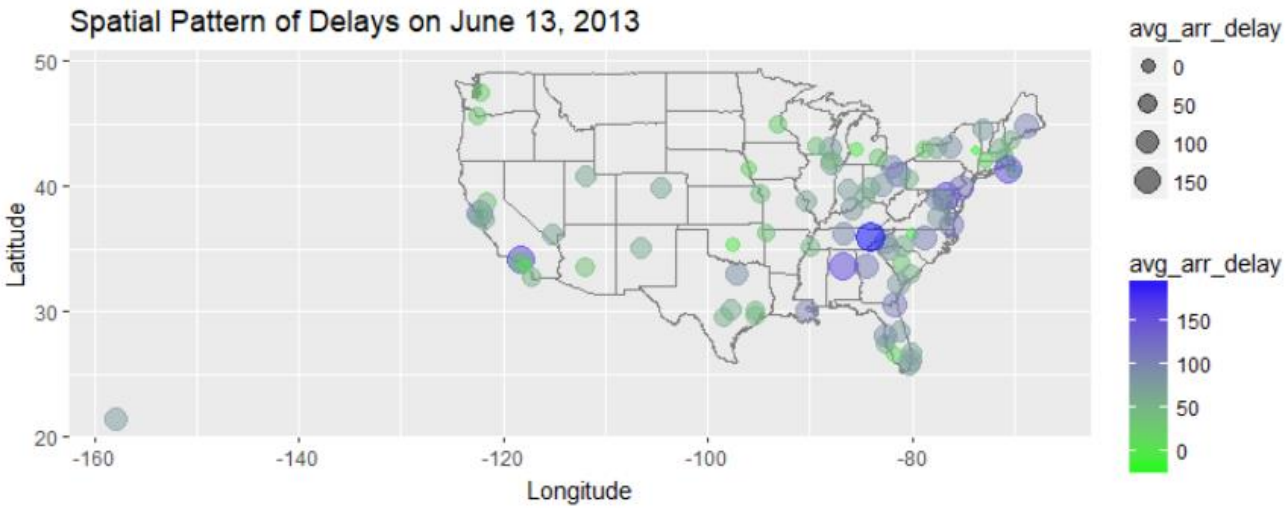
I could not get my graph to properly display the airplane route, however I will note that the code (above) produces a chart containing each flight's origin (latitude and longitude) and destination (latitude and longitude). Theoretically, one should be able to "summarise" each flight's tail number and then maintain a continuous map, by-date, from the Jan 1 NYC airport until the Dec 31 final destination. Below is the chart:

	year	month	day	carrier	flight	tailnum	origin	dest	hour	minute	time_hour	name.x	lat.x	lon.x	name.y	lat.y	lon.y
1	2013	1	1	UA	1545	N14228	EWR	IAH	5	15	2013-01-01 05:00:00	Newark Liberty Intl	40.69250	-74.16867	George Bush Intercontinental	29.98443	-95.34144
2	2013	1	1	UA	1714	N24211	LGA	IAH	5	29	2013-01-01 05:00:00	La Guardia	40.77725	-73.87261	George Bush Intercontinental	29.98443	-95.34144
3	2013	1	1	AA	1141	N619AA	JFK	MIA	5	40	2013-01-01 05:00:00	John F Kennedy Intl	40.63975	-73.77893	Miami Intl	25.79325	-80.29056
4	2013	1	1	DL	461	N668DN	LGA	ATL	6	0	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Hartsfield Jackson Atlanta Intl	33.63672	-84.42807
5	2013	1	1	UA	1696	N39463	EWR	ORD	5	58	2013-01-01 05:00:00	Newark Liberty Intl	40.69250	-74.16867	Chicago Ohare Intl	41.97860	-87.90484
6	2013	1	1	B6	507	N516JB	EWR	FLL	6	0	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	Fort Lauderdale Hollywood Intl	26.07258	-80.15275
7	2013	1	1	EV	5708	N829AS	LGA	IAD	6	0	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Washington Dulles Intl	38.94453	-77.45581
8	2013	1	1	B6	79	N593JB	JFK	MCO	6	0	2013-01-01 06:00:00	John F Kennedy Intl	40.63975	-73.77893	Orlando Intl	28.42939	-81.30899
9	2013	1	1	AA	301	N3ALAA	LGA	ORD	6	0	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Chicago Ohare Intl	41.97860	-87.90484
10	2013	1	1	B6	49	N793JB	JFK	PBI	6	0	2013-01-01 06:00:00	John F Kennedy Intl	40.63975	-73.77893	Palm Beach Intl	26.68316	-80.09559
11	2013	1	1	B6	71	N657JB	JFK	TPA	6	0	2013-01-01 06:00:00	John F Kennedy Intl	40.63975	-73.77893	Tampa Intl	27.97547	-82.53325
12	2013	1	1	UA	194	N29129	JFK	LAX	6	0	2013-01-01 06:00:00	John F Kennedy Intl	40.63975	-73.77893	Los Angeles Intl	33.94254	-118.40807
13	2013	1	1	UA	1124	N53441	EWR	SFO	6	0	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	San Francisco Intl	37.61897	-122.37489
14	2013	1	1	AA	707	N3DUAA	LGA	DFW	6	0	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Dallas Fort Worth Intl	32.89683	-97.03800
15	2013	1	1	B6	1806	N708JB	JFK	BOS	5	59	2013-01-01 05:00:00	John F Kennedy Intl	40.63975	-73.77893	General Edward Lawrence Logan Intl	42.36435	-71.00518
16	2013	1	1	UA	1187	N76515	EWR	LAS	6	0	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	Mc Carran Intl	36.08006	-115.15225
17	2013	1	1	B6	371	N595JB	LGA	FLL	6	0	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Fort Lauderdale Hollywood Intl	26.07258	-80.15275
18	2013	1	1	MQ	4650	N542MQ	LGA	ATL	6	0	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Hartsfield Jackson Atlanta Intl	33.63672	-84.42807
19	2013	1	1	B6	343	N644JB	EWR	PBI	6	0	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	Palm Beach Intl	26.68316	-80.09559
20	2013	1	1	DL	1919	N971DL	LGA	MSP	6	10	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Minneapolis St Paul Intl	44.88196	-93.22177
21	2013	1	1	MQ	4401	N730MQ	LGA	DTW	6	5	2013-01-01 06:00:00	La Guardia	40.77725	-73.87261	Detroit Metro Wayne Co	42.21244	-83.35339
22	2013	1	1	AA	1895	N633AA	EWR	MIA	6	10	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	Miami Intl	25.79325	-80.29056
23	2013	1	1	DL	1743	N3739P	JFK	ATL	6	10	2013-01-01 06:00:00	John F Kennedy Intl	40.63975	-73.77893	Hartsfield Jackson Atlanta Intl	33.63672	-84.42807
24	2013	1	1	UA	1077	N53442	EWR	MIA	6	7	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	Miami Intl	25.79325	-80.29056
25	2013	1	1	MQ	3768	N9EAMQ	EWR	ORD	6	0	2013-01-01 06:00:00	Newark Liberty Intl	40.69250	-74.16867	Chicago Ohare Intl	41.97860	-87.90484

Q5. What happened on June 13, 2013? Display the spatial pattern of delays, and then use Google to cross-reference with the weather.

On June 13, 2013, as can be seen on the graph below, the southeastern region of the US experienced a period of intense thunderstorms and violent weather. The largest delays are in Tennessee (Nashville), the Southeast, and the Midwest.

```
flights %>% filter(year == 2013, month == 6, day == 13) %>%
  group_by(dest) %>%
  summarize(avg_arr_delay = mean(arr_delay, na.rm = TRUE)) %>%
  left_join(airports, by = c('dest' = 'faa')) %>%
  ggplot(aes(x = lon, y = lat, size = avg_arr_delay, color = avg_arr_delay)) +
  borders('state') +
  geom_point(alpha = .5) +
  scale_color_continuous(low = 'green', high = 'blue') +
  coord_quickmap() +
  xlab("Longitude") + ylab("Latitude") +
  labs(title = paste("Spatial Pattern of Delays on June 13, 2013"))
```



Q6. Filter flights to only show flights with planes and carriers that have flown having most (top 10) and least (bottom 10) flights.

```
flights_flown <- flights %>%
  semi_join(count(flights, tailnum) %>% filter(!is.na(tailnum)))

ranked_flights <- flights_flown %>% group_by(carrier, tailnum) %>% summarise(n = n()) %>%
arrange(desc(n))
```

TOP 10

	carrier	tailnum	n
1	MQ	N725MQ	575
2	MQ	N722MQ	513
3	MQ	N723MQ	507
4	MQ	N711MQ	486
5	MQ	N713MQ	483
6	B6	N258JB	427
7	B6	N298JB	407
8	B6	N353JB	404
9	B6	N351JB	402
10	MQ	N735MQ	396

BOTTOM 10

4051	WN	N505SW	1
4052	WN	N510SW	1
4053	WN	N521SW	1
4054	WN	N636WN	1
4055	WN	N660SW	1
4056	WN	N669SW	1
4057	WN	N7713A	1
4058	WN	N7715E	1
4059	WN	N8618N	1
4060	WN	N8619F	1