Assignment 2

Derek G. Nokes

Saturday, January 31, 2015

Contents

Connect to the PostgreSQL Database

To anwser assignment questions we connect to the PostgreSQL database as follows:

```
#
library('RPostgreSQL')
```

```
## Loading required package: DBI
```

```
# connect to PoststgreSQL using RPostgreSQL
driverHandle <- dbDriver("PostgreSQL")

# establish connection to 'flights' database
connectionHandle <- dbConnect(driverHandle,
                              dbname="flights",
                              host="localhost",
                              port=5432,user="postgres",
                              password ="tgdnrx78")
```

Q1

Which destination in the flights database is the furthest distance away?

Anwser

We join the 'flights' and 'airports' tables:

```
# construct the query
# join flights and airports tables to match origin and destination
# FAA codes with airport names, latitude, and longitude
query<-'SELECT DISTINCT flights.origin as origin_faa,airports.name as
origin_name,airports.lat as origin_lat,airports.lon as origin_lon,
flights.dest as dest_faa,a.name as dest_name,a.lat as dest_lat,a.lon
as dest_lon,flights.distance FROM flights INNER JOIN airports ON
flights.origin = airports.faa JOIN airports a ON flights.dest = a.faa
GROUP BY origin_faa,dest_faa,flights.distance,origin_name,origin_lat,
origin_lon,dest_name,dest_lat,dest_lon ORDER BY flights.distance DESC
LIMIT 1'

# fetch data
data<-dbGetQuery(connectionHandle,query)
knitr::kable(t(data), caption = 'Destination Furthest Distance Away')
```

	1
origin_faa	JFK
origin_name	John F Kennedy Intl
origin_lat	40.63975

	1
origin_lon	-73.77893
dest_faa	HNL
dest_name	Honolulu Intl
dest_lat	21.31868
dest_lon	-157.9224
distance	4983

Table 1: Destination Furthest Distance Away

Honolulu Intl (HNL) is the destination that is the furthest distance away from John F Kennedy Intl (JFK) at 4983 miles.

Q2

What are the different numbers of engines in the planes table? For each number of engines, which aircraft have the most number of seats?

Anwser

We can fetch the maxiumum number of seats for each distinct number of engines:

```
# fetch the distinct number of engines and join on
query<-'SELECT DISTINCT planes.engines,MAX(planes.seats) as max_seats
FROM planes GROUP BY engines ORDER BY engines DESC'
numberOfEngines<-dbGetQuery(connectionHandle,query)

knitr::kable(numberOfEngines, caption = 'Maximum Number of Seats by
             Number of Engines')
```

engines	max_seats
4	450
3	379
2	400
1	16

Table 2: Maximum Number of Seats by Number of Engines

We can then loop over each combination of number of engines and maximum number of seats and fetch the details of the corresponding planes:

3

```
#

for(i in seq_along(numberOfEngines[,1])) {
    numberOfEnginesAndSeats<-numberOfEngines[i,]
    engines<-numberOfEnginesAndSeats$engines
    seats<-numberOfEnginesAndSeats$max_seats

    # build the query
    query<-paste0('SELECT DISTINCT model,manufacturer,type,engines,seats
                  FROM planes WHERE engines=',engines,' AND seats=',seats,
                  ' ORDER BY seats')
    # execute the query
    data<-dbGetQuery(connectionHandle,query)

    # create the results table
    print(data)
    knitr::kable(data)
}
```

```
##      model manufacturer                    type engines seats
## 1 747-451       BOEING Fixed wing multi engine       4   450
##      model manufacturer                    type engines seats
## 1 A330-223      AIRBUS Fixed wing multi engine       3   379
##      model manufacturer                    type engines seats
## 1 777-224       BOEING Fixed wing multi engine       2   400
## 2 777-222       BOEING Fixed wing multi engine       2   400
## 3 777-232       BOEING Fixed wing multi engine       2   400
## 4 777-200       BOEING Fixed wing multi engine       2   400
##        model manufacturer                     type engines seats
## 1 OTTER DHC-3  DEHAVILLAND Fixed wing single engine       1    16
```

We can also use a slightly more complex single query to return the plane details associated with the maximum number of seats for each number of engines group:

```
# build the query
query<-'SELECT DISTINCT t.manufacturer,t.model,t.type,t.engine,t.engines,
t.seats FROM planes t INNER JOIN (SELECT DISTINCT engines as numberOfEngines,
max(seats) as max_seats FROM planes GROUP BY engines) group_t ON t.seats =
group_t.max_seats AND t.engines = group_t.numberOfEngines ORDER BY t.engines
DESC,t.seats DESC'

# execute the query
data<-dbGetQuery(connectionHandle,query)

# create the results table
knitr::kable(data, caption = 'Model with Maximum Number of Seats by Number
             of Engines')
```

manufacturer	model	type	engine	engines	seats
BOEING	747-451	Fixed wing multi engine	Turbo-jet	4	450
AIRBUS	A330-223	Fixed wing multi engine	Turbo-fan	3	379

manufacturer	model	type	engine	engines	seats
BOEING	777-200	Fixed wing multi engine	Turbo-jet	2	400
BOEING	777-222	Fixed wing multi engine	Turbo-fan	2	400
BOEING	777-224	Fixed wing multi engine	Turbo-fan	2	400
BOEING	777-232	Fixed wing multi engine	Turbo-jet	2	400
DEHAVILLAND	OTTER DHC-3	Fixed wing single engine	Reciprocating	1	16

Table 3: Model with Maximum Number of Seats by Number of Engines

Q3

What weather conditions are associated with New York City departure delays?

Anwser

We join the 'flights' and 'weather' tables so that we can examine the relationship between different weather variables and the departure delays:

```
# build the query (join the weather and flights tables on date and hour)
query<-"SELECT w.w_date_hour,f.dep_delay,f.carrier,f.tailnum,f.flight,
f.dest,f.air_time,f.distance,f.origin,w.origin,w.temp,w.dewp,w.humid,
w.wind_dir,w.wind_speed,w.wind_gust,w.precip,w.pressure,w.visib FROM
flights f INNER JOIN (SELECT concat_ws('-',weather.year,weather.month,
weather.day,weather.hour) as w_date_hour,weather.* FROM weather) w
ON concat_ws('-',f.year,f.month,f.day,f.hour) = w.w_date_hour AND
f.origin = w.origin ORDER BY f.dep_delay DESC"

# execute the query
data<-dbGetQuery(connectionHandle,query)
```

We extract each of the weather variables and create data frames for plotting:

```
delayInHours<-data$dep_delay/60
delayIndex<-floor(delayInHours)>0 & floor(delayInHours)<5
tempDF<-data.frame(delayInHours[delayIndex],data$temp[delayIndex])
windSpeedIndex<-data$wind_speed<200 & floor(delayInHours)>0 & floor(delayInHours)<5
windSpeedDF<-data.frame(delayInHours[windSpeedIndex],data$wind_speed[windSpeedIndex])
windGustIndex<-data$wind_speed<200 & floor(delayInHours)>0 & floor(delayInHours)<5
windGustDF<-data.frame(delayInHours[windGustIndex],data$wind_gust[windGustIndex])
precipDF<-data.frame(delayInHours[delayIndex],data$precip[delayIndex])
pressureDF<-data.frame(delayInHours[delayIndex],data$pressure[delayIndex])
visibDF<-data.frame(delayInHours[delayIndex],data$visib[delayIndex])
```

We create 2D histograms to examine the relationship between departure delays and temperature, wind speed, wind gust, precipitation, pressure, and visibility:
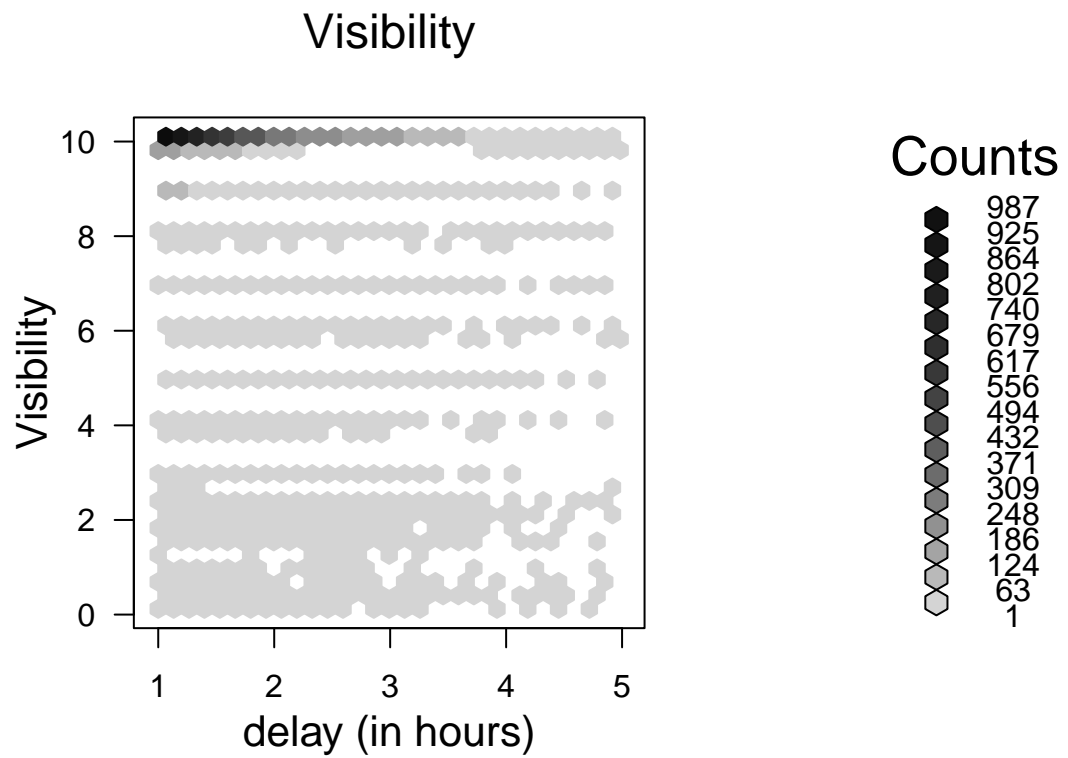
There is no clear relationship between temperature and departure delays, but high and low temperature appear as though they have more delays

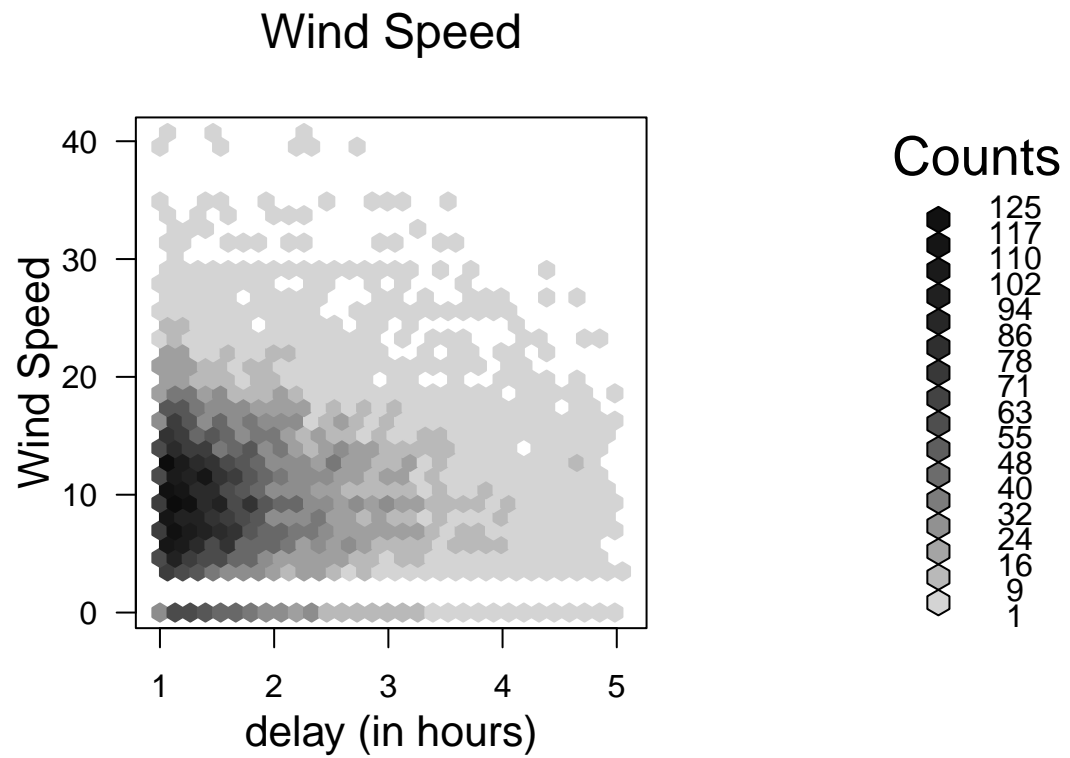```
plot(h1,main='Temperature')
```

## Temperature
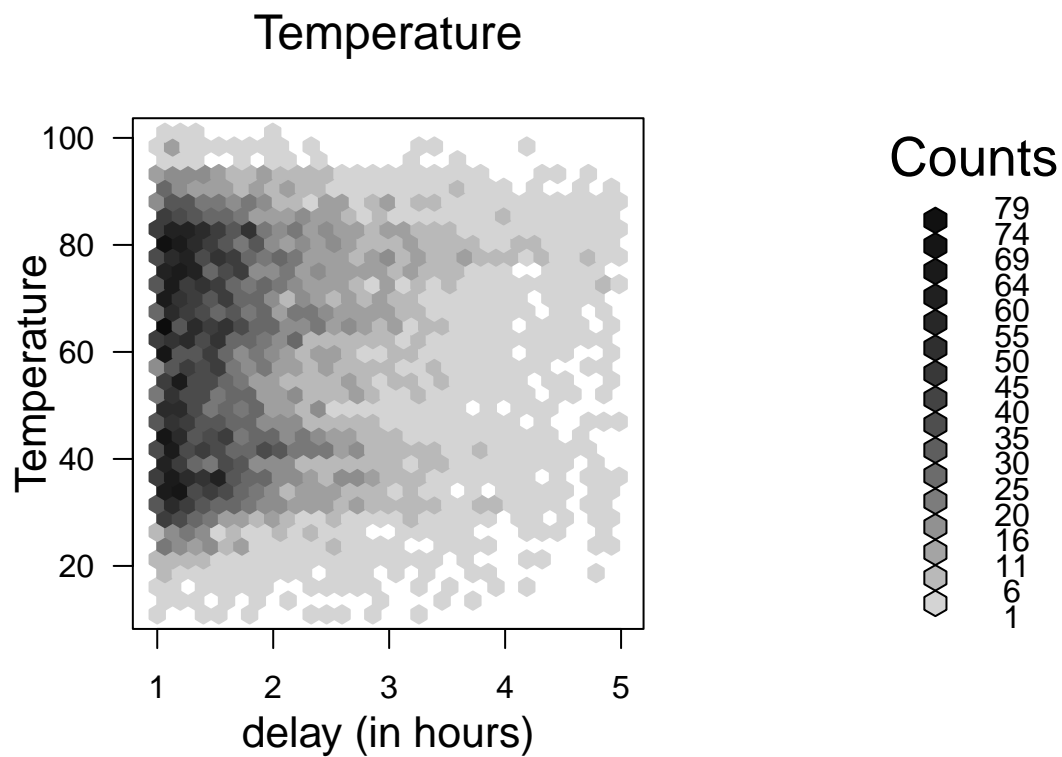


There is no clear relationship between visibility and departure delays

```
plot(h2,main='Visibility')
```

6

Visibility

There is no clear relationship between wind speed and departure delays
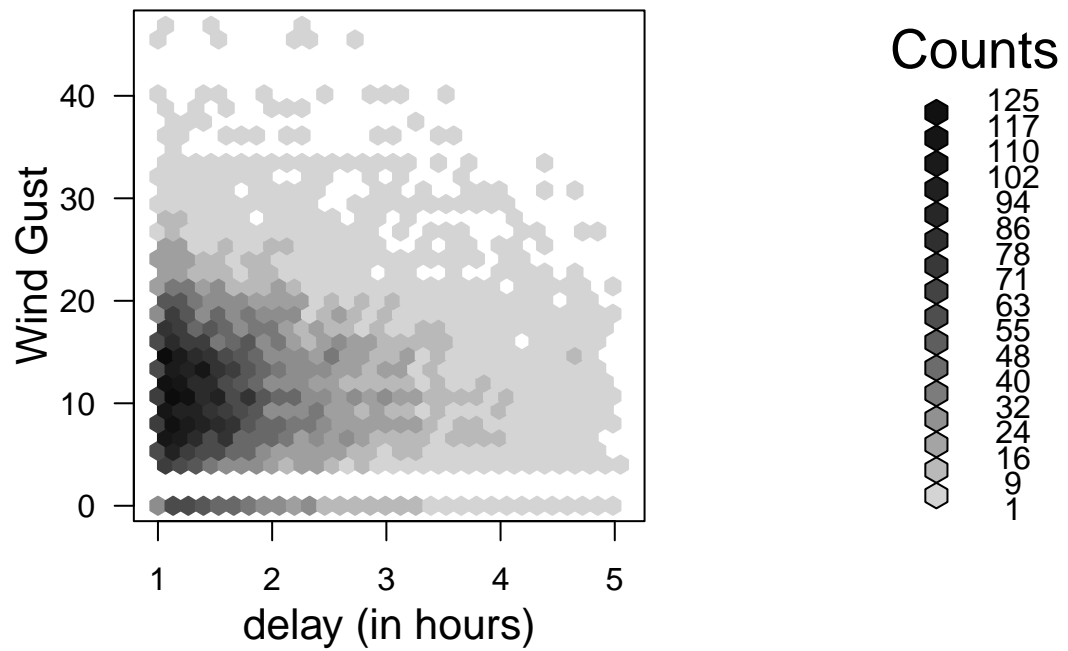
```
plot(h3,main='Wind Speed')
```

Wind Speed

There is no clear relationship between wind gust and departure delays
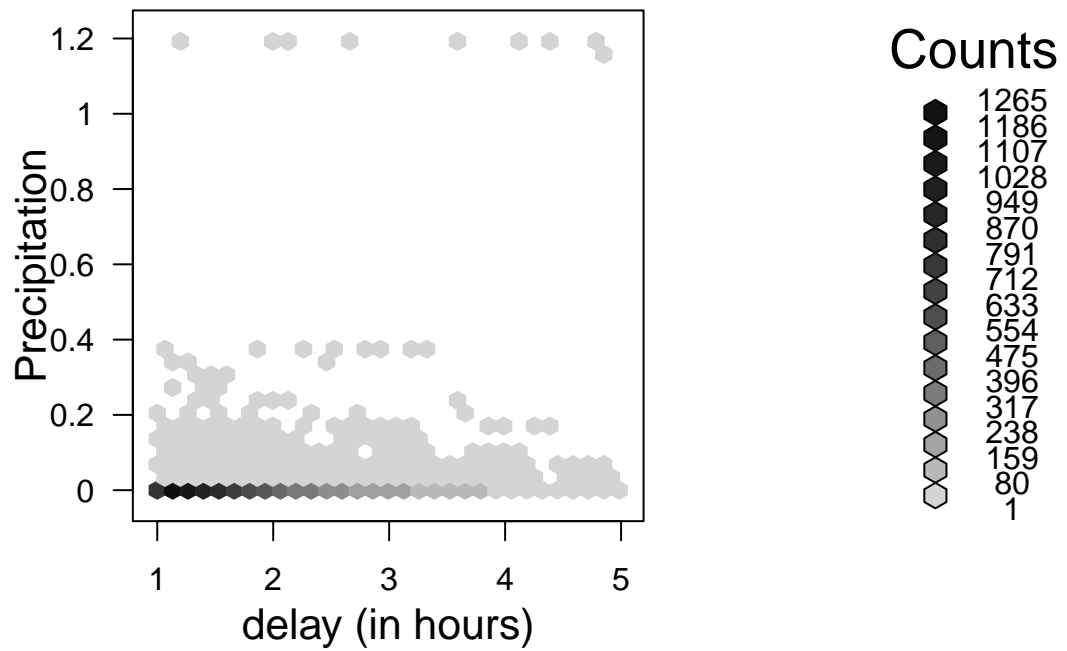
```
plot(h4,main='Wind Gust')
```

Wind Gust

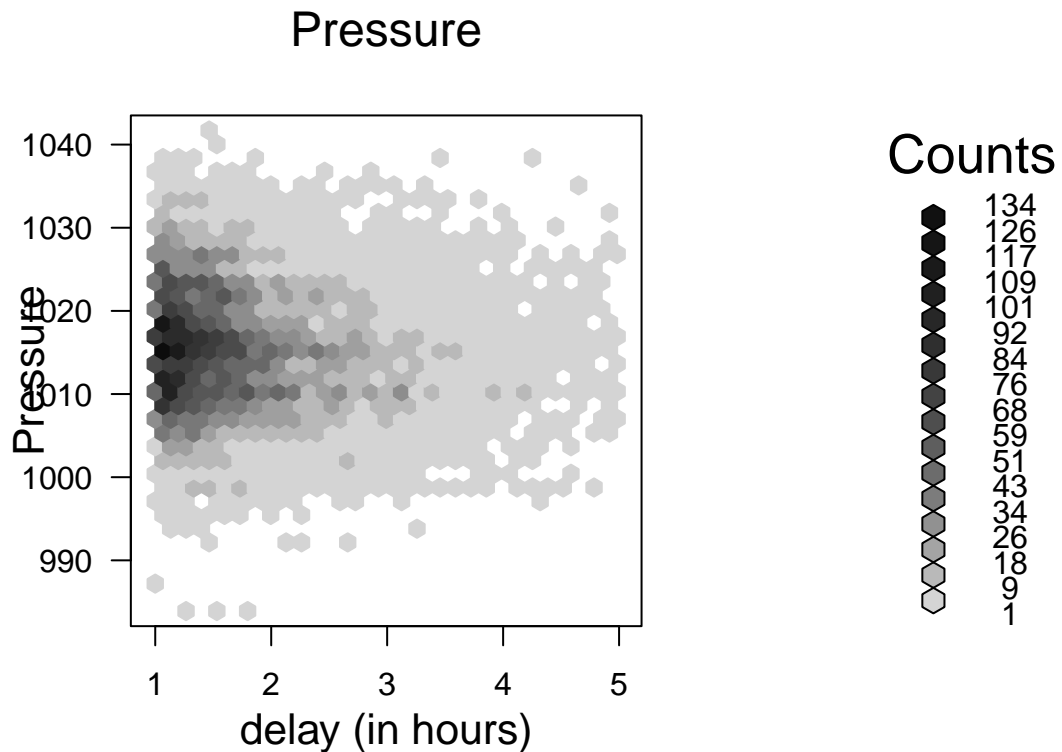There is no clear relationship between precipitation and departure delays

```
plot(h5,main='Precipitation')
```

Precipitation

There is no clear relationship between pressure and departure delays

```
plot(h6,main='Pressure')
```

Pressure

There are no clear relationships between departure delays and any of the weather variables. This may be an artifact of the way the table join was done. Rather than join on the departure time, it might make more sense to join on the intended departure time (i.e., the departure time adjusted for the delay)

Q4

Are older planes more likely to be delayed?

Anwser

First we join the 'fights' and 'plane' tables and look at the delays binned by hour. A delay in hours of zero indicates a delay of between 1 minute and 59 minutes, a delay of one indicates a delay of between 1 hour and 1 hour and 59 minutes, and so forth:

```
query<-"SELECT p.year,(f.dep_delay/60) as delay_in_hours FROM flights f JOIN (SELECT * FROM planes) p O

# execute the query
data<-dbGetQuery(connectionHandle,query)
```

We create a 2D histogram to examine the relationship between departure delays and the age of the plane:

```
h1<- hexbin(data,xlab='year',ylab='delay (in hours)')
plot(h1,main='Delay (In Hours) By Year of Aircraft')
```

# Delay (In Hours) By Year of Aircraft
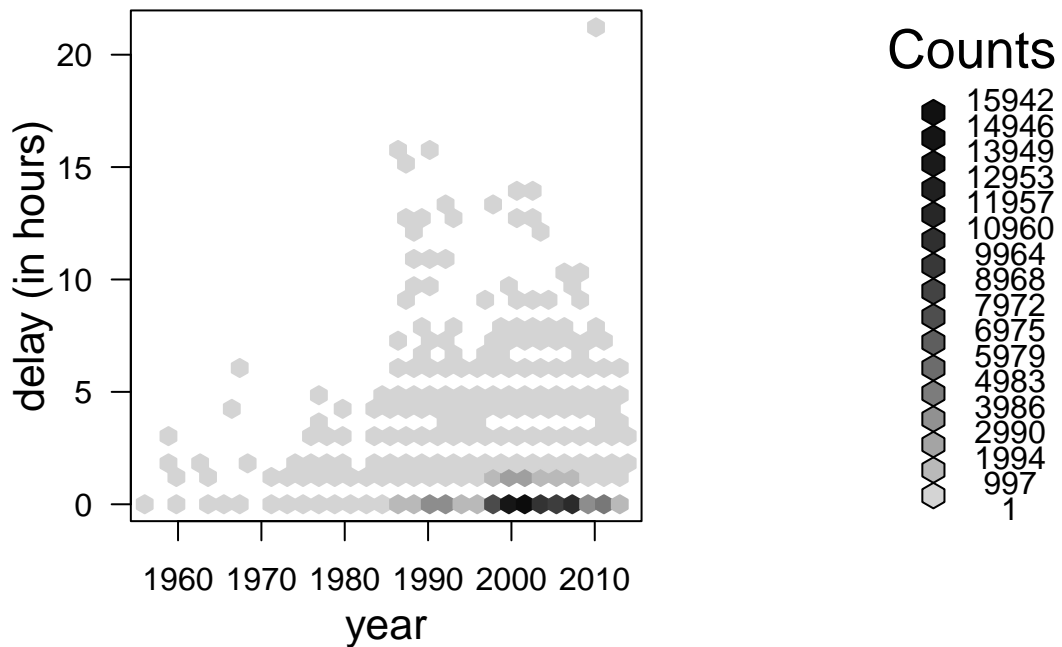
There is no clear relationship between the age of the plane and the aircraft model year.

```
query<-"SELECT p.year,(f.dep_delay/60) as delay_in_hours FROM flights f
JOIN (SELECT * FROM planes) p
ON f.tailnum = p.tailnum
WHERE p.year >0
ORDER BY p.year"

# execute the query
flights<-dbGetQuery(connectionHandle,query)
```

We can see that there are considerably more flights for aircraft models of particular years:

```
library('ggplot2')
qplot(factor(year), data=flights, geom="bar",
      xlab='total number of flights',ylab='year',
      main='Total Number of Flights By Aircraft Model
      Year') + coord_flip()
```

Total Number of Flights By Aircraft Model
Year
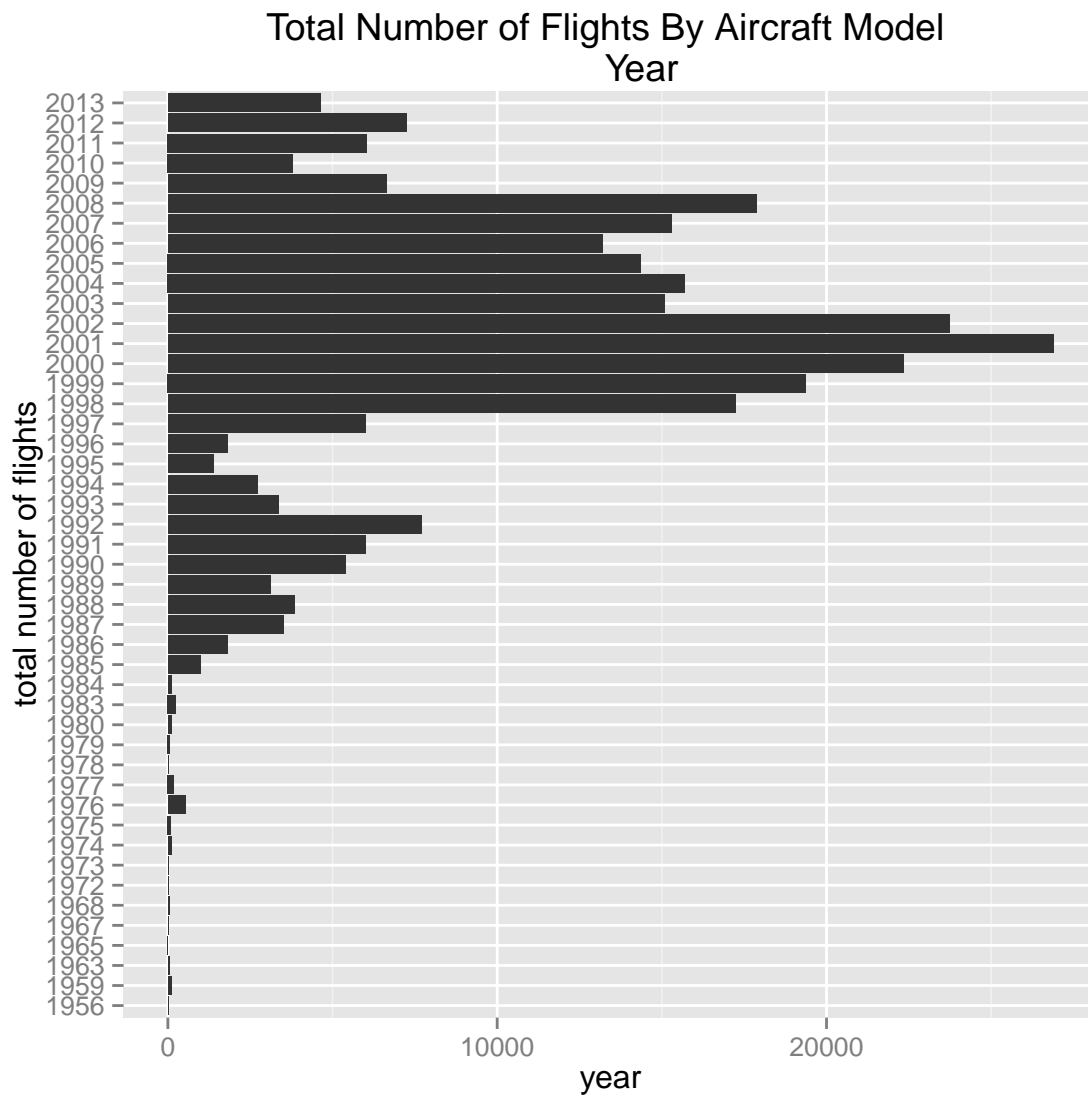
There are far fewer total flights by very old planes.

When we group by aircraft model year and look at the delayed flights as a percentage of total flights, there is again no clear relationship

```
query<-"SELECT p.year,COUNT(f.dep_delay) FROM flights f
JOIN (SELECT * FROM planes) p ON f.tailnum = p.tailnum
WHERE p.year >0 GROUP BY p.year
ORDER BY p.year"

# execute the query
flights<-dbGetQuery(connectionHandle,query)

query<-"SELECT p.year,COUNT(f.dep_delay) FROM flights f
JOIN (SELECT * FROM planes) p ON f.tailnum = p.tailnum
WHERE p.year >0 AND dep_delay > 0 GROUP BY p.year
ORDER BY p.year"
```
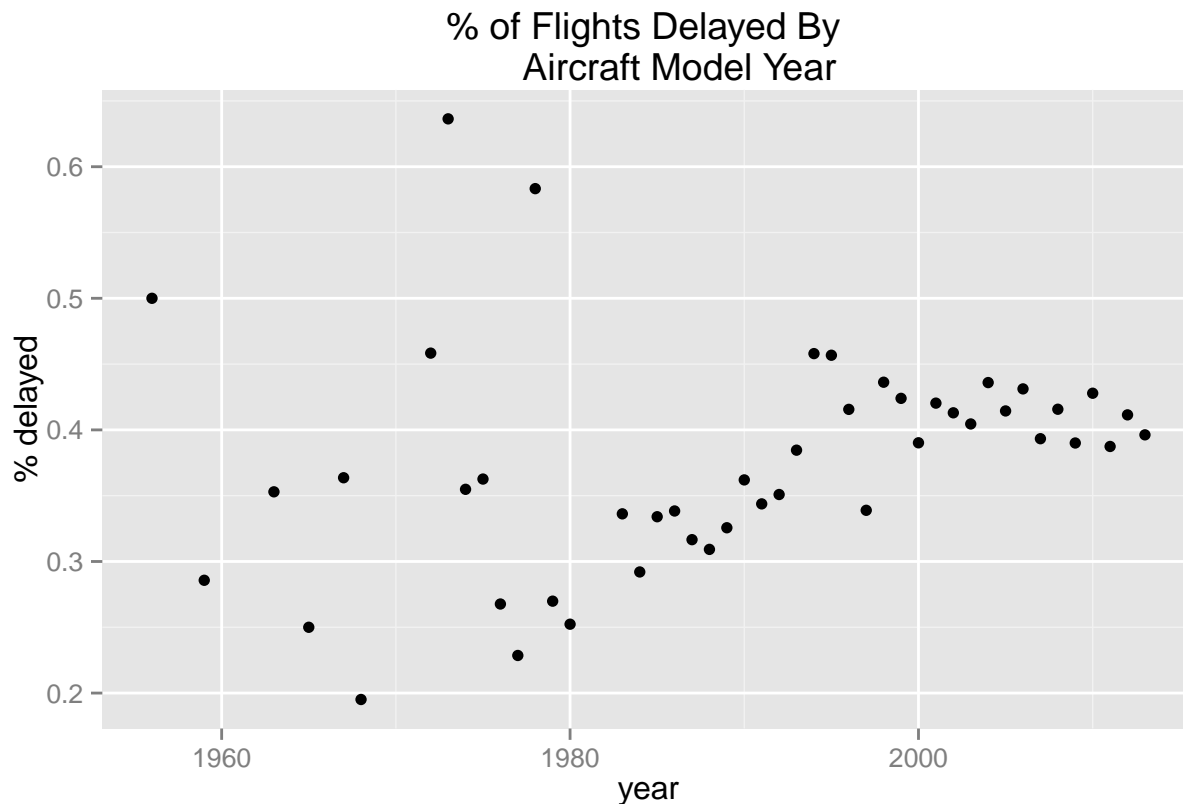
```
# execute the query
delays<-dbGetQuery(connectionHandle,query)

qplot(delays$year,delays$count/flights$count,xlab='year',
      ylab='% delayed',main='% of Flights Delayed By
      Aircraft Model Year')
```


% of Flights Delayed By Aircraft Model Year

Theefore, the age of the aircraft does not consistently appear to cause more delays

Q5

Ask (and if possible answer) a question that also requires joining information from two or more tables in the flights database, and/or assumes that additional information can be collected in advance of answering your question.

Anwser

Are departure delays related to destination?

```
# construct the query
# join flights and airports tables to match origin and destination
# FAA codes with airport names, latitude, and longitude
query<-'SELECT DISTINCT flights.dest as dest_faa,a.name as dest_name,
```

```
a.lat as dest_lat,a.lon as dest_lon,AVG(dep_delay) as delay,COUNT(dep_delay)
as number_of_flights,AVG(flights.distance) as distance FROM flights INNER JOIN airports
ON flights.origin = airports.faa JOIN airports a
ON flights.dest = a.faa GROUP BY dest_faa,dest_name,dest_lat,dest_lon
ORDER BY number_of_flights,delay'

# fetch data
data<-dbGetQuery(connectionHandle,query)
```

Here is a plot of the destinations:

```
library('rworldmap')
```

```
## Loading required package: sp
## ### Welcome to rworldmap ###
## For a short introduction type :   vignette('rworldmap')
```

```
newmap <- getMap(resolution = "high")
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'rworldxtra'
```

```
## Warning in getMap(resolution = "high"): for resolution='high' option you
## need to install package rworldxtra, using low resolution version for now
```

```
plot(newmap,xlim = c(-139.3, -58.8),ylim = c(13.5, 55.7))
points(data$dest_lon, data$dest_lat, col = "red", cex = .6)
```
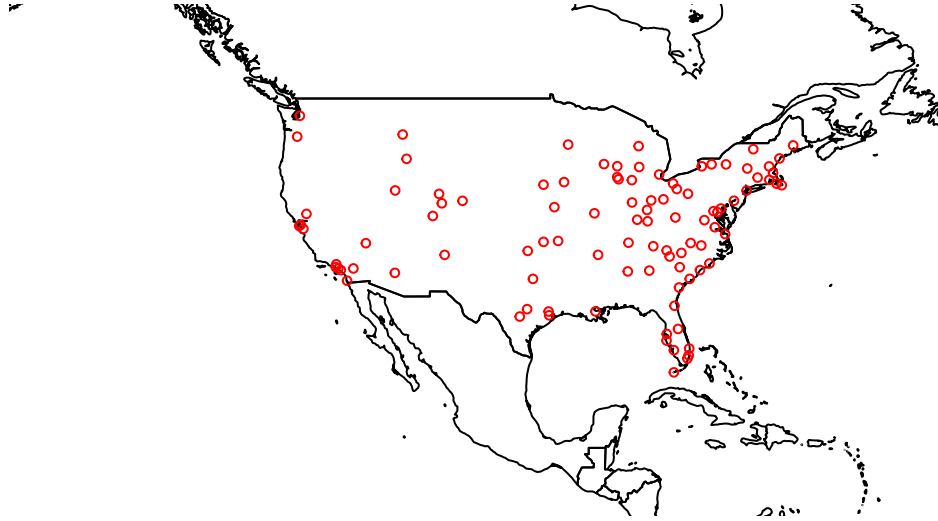
Here are the results

```
knitr::kable(data, caption = 'Average Departure Delay By Destination')
```

Table: Average Departure Delay By Destination

dest_faa	dest_name	dest_lat	dest_lon	delay	number_of_flights	distance
LGA	La Guardia	40.77725	-73.87261	NA	0	17.00000
LEX	Blue Grass	38.03650	-84.60589	-9.000000	1	604.00000
ANC	Ted Stevens Anchorage Intl	61.17436	-149.99636	12.875000	8	3370.00000
SBN	South Bend Rgnl	41.70866	-86.31725	21.100000	10	645.40000
HDN	Yampa Valley	40.48118	-107.21766	12.285714	14	1728.00000
MTJ	Montrose Regional Airport	38.50979	-107.89424	17.642857	14	1795.00000
EYW	Key West Intl	24.55611	-81.75956	3.647059	17	1207.00000
PSP	Palm Springs Intl	33.82967	-116.50669	-2.944444	18	2378.00000
JAC	Jackson Hole Airport	43.60733	-110.73775	26.545455	22	1875.60000
BZN	Gallatin Field	45.77764	-111.16015	11.457143	35	1882.00000
CHO	Charlottesville-Albemarle	38.13864	-78.45286	21.391304	46	305.00000
MYR	Myrtle Beach Intl	33.67975	-78.92833	15.758621	58	550.66102
TVC	Cherry Capital Airport	44.74144	-85.58223	22.083333	96	652.38614
CAE	Columbia Metropolitan	33.93883	-81.11953	35.570093	107	603.55172

dest_faa	dest_name	dest_lat	dest_lon	delay	number_of_flights	distance
ILM	Wilmington Intl	34.27061	-77.90257	19.416667	108	500.00000
CRW	Yeager	38.37315	-81.59319	17.000000	137	444.00000
EGE	Eagle Co Rgnl	39.64256	-106.91769	15.495192	208	1735.70892
MVY	Martha\'s Vineyard	41.39167	-70.61528	7.051643	213	173.00000
ABQ	Albuquerque International Sunport	35.04022	-106.60919	13.740158	254	1826.00000
AVL	Asheville Regional Airport	35.43619	-82.54181	8.190114	263	583.58182
ACK	Nantucket Mem	41.25305	-70.06018	6.456604	265	199.00000
BHM	Birmingham Intl	33.56294	-86.75355	29.694853	272	865.99663
SMF	Sacramento Intl	38.69542	-121.59078	18.691489	282	2521.00000
TUL	Tulsa Intl	36.19839	-95.88811	34.906354	299	1215.00000
OAK	Metropolitan Oakland Intl	37.72128	-122.22072	13.344051	311	2576.00000
OKC	Will Rogers World	35.39309	-97.60073	30.568807	327	1325.00000
SJC	Norman Y Mineta San Jose Intl	37.36260	-121.92902	10.103658	328	2569.00000
PVD	Theodore Francis Green State	41.73258	-71.42038	21.765363	358	160.00000
BGR	Bangor Intl	44.80744	-68.82814	19.475000	360	378.00000
BUR	Bob Hope	34.20067	-118.35867	13.475676	370	2465.00000
BDL	Bradley Intl	41.93889	-72.68322	17.720874	412	116.00000
ALB	Albany Intl	42.74827	-73.80169	23.620525	419	143.00000
DSM	Des Moines Intl	41.53397	-93.66308	26.232955	528	1020.88752
MSN	Dane Co Rgnl Truax Fld	43.13986	-89.33751	23.580071	562	803.95455
TYS	Mc Ghee Tyson	35.81097	-83.99403	28.493955	579	638.80983
LGB	Long Beach	33.81772	-118.15161	11.183735	664	2465.00000
SAT	San Antonio Intl	29.53369	-98.46978	20.737463	678	1578.34111
HNL	Honolulu Intl	21.31868	-157.92243	9.289362	705	4972.67468
GRR	Gerald R Ford Intl	42.88083	-85.52281	19.533333	735	605.78170
SAV	Savannah Hilton Head Intl	32.12758	-81.20214	18.260292	753	709.18408
GSP	Greenville-Spartanburg International	34.89556	-82.21889	19.261965	794	595.95995
SNA	John Wayne Arpt Orange Co	33.67567	-117.86822	7.761905	819	2434.00000
OMA	Eppley Afld	41.30317	-95.89407	20.205596	822	1135.56655
CAK	Akron Canton Regional Airport	40.91608	-81.44219	20.820878	843	397.00000
MHT	Manchester Regional Airport	42.93256	-71.43567	21.024678	932	207.02973
XNA	NW Arkansas Regional	36.28187	-94.30681	6.464886	1011	1142.50579
SDF	Louisville International Airport	38.17409	-85.73650	16.426141	1117	645.98358
SRQ	Sarasota Bradenton Intl	27.39544	-82.55439	7.256027	1203	1044.65153
PDX	Portland Intl	45.58872	-122.59750	16.255935	1348	2445.56573
DAY	James M Cox Dayton Intl	39.90237	-84.21937	17.498573	1402	537.10230

| dest_faa	dest_name	dest_lat	dest_lon	delay	number_of_flights	distance
ORF	Norfolk Intl	36.89461	-76.20122	17.576389	1440	288.52344
GSO	Piedmont Triad	36.09775	-79.93731	19.398000	1500	449.84184
PHL	Philadelphia Intl	39.87194	-75.24114	11.998709	1549	94.32353
MEM	Memphis Intl	35.04242	-89.97667	15.658796	1694	954.20123
BWI	Baltimore Washington Intl	39.17536	-76.66833	16.396816	1696	179.41830
SYR	Syracuse Hancock Intl	43.11119	-76.10631	14.444769	1711	205.92164
MCI	Kansas City Intl	39.29761	-94.71390	20.326477	1896	1097.69522
IND	Indianapolis Intl	39.71733	-86.29438	14.046710	1991	652.26288
HOU	William P Hobby	29.64542	-95.27889	14.340191	2090	1420.15508
PWM	Portland Intl Jetport	43.64616	-70.30928	16.455773	2295	276.12840
RIC	Richmond Intl	37.50517	-77.31967	23.639847	2349	281.40465
ROC	Greater Rochester Intl	43.11887	-77.67239	16.243861	2362	259.25083
AUS	Austin Bergstrom Intl	30.19453	-97.66989	13.025641	2418	1514.25297
SLC	Salt Lake City Intl	40.78839	-111.97777	9.026851	2458	1986.98662
BTV	Burlington Intl	44.47186	-73.15328	13.605651	2513	265.09154
JAX	Jacksonville Intl	30.49406	-81.68786	16.484434	2634	824.67610
MKE	General Mitchell Intl	42.94722	-87.89658	18.760118	2718	733.38151
SAN	San Diego Intl	32.73356	-117.18967	11.110866	2724	2437.29923
PIT	Pittsburgh Intl	40.49147	-80.23287	13.704966	2759	334.06122
CHS	Charleston Afb Intl	32.89865	-80.04053	14.694775	2775	632.91678
CMH	Port Columbus Intl	39.99797	-82.89189	12.223787	3338	476.55505
RSW	Southwest Florida Intl	26.53617	-81.75517	8.275577	3509	1072.85327
MSY	Louis Armstrong New Orleans Intl	29.99339	-90.25803	14.249463	3724	1177.70571
CVG	Cincinnati Northern Kentucky Intl	39.04884	-84.66782	19.520321	3740	575.15986
SEA	Seattle Tacoma Intl	47.44900	-122.30931	10.725922	3904	2412.66531
MDW	Chicago Midway Intl	41.78597	-87.75242	18.589763	4044	718.04595
STL	Lambert St Louis Intl	38.74870	-90.37003	16.010557	4168	878.72321
CLE	Cleveland Hopkins Intl	41.41169	-81.84979	13.387931	4408	414.17428
BUF	Buffalo Niagara Intl	42.94053	-78.73217	13.423951	4576	296.80837
PHX	Phoenix Sky Harbor Intl	33.43428	-112.01158	10.412926	4611	2141.30326
IAD	Washington Dulles Intl	38.94453	-77.45581	16.982934	5391	224.84684
LAS	Mc Carran Intl	36.08006	-115.15225	9.418819	5962	2240.96148
BNA	Nashville Intl	36.12447	-86.67819	15.982470	6104	758.21348
PBI	Palm Beach Intl	26.68316	-80.09559	12.992918	6495	1028.83811
MSP	Minneapolis St Paul Intl	44.88196	-93.22177	13.324806	6958	1017.40167
IAH	George Bush Intercontinental	29.98443	-95.34144	10.842179	7103	1407.20672

dest_faa	dest_name	dest_lat	dest_lon	delay	number_of_flights	distance
DEN	Denver Intl	39.86166	-104.67318	15.156228	7201	1614.67836
TPA	Tampa Intl	27.97547	-82.53325	12.135007	7407	1003.93557
RDU	Raleigh Durham Intl	35.87764	-78.78747	12.445870	7796	426.75769
DFW	Dallas Fort Worth Intl	32.89683	-97.03800	8.681909	8463	1383.04303
DTW	Detroit Metro Wayne Co	42.21244	-83.35339	11.812252	9060	498.12852
DCA	Ronald Reagan Washington Natl	38.85208	-77.03772	10.293000	9157	211.00618
MIA	Miami Intl	25.79325	-80.29056	8.876558	11633	1091.55244
FLL	Fort Lauderdale Hollywood Intl	26.07258	-80.15275	12.731104	11934	1070.06877
SFO	San Francisco Intl	37.61897	-122.37489	12.866289	13230	2577.92356
CLT	Charlotte Douglas Intl	35.21400	-80.94314	9.222879	13698	538.02730
MCO	Orlando Intl	28.42939	-81.30899	11.275998	13982	943.11057
BOS	General Edward Lawrence Logan Intl	42.36435	-71.00518	8.730613	15049	190.63696
LAX	Los Angeles Intl	33.94254	-118.40807	9.401344	16076	2468.62236
ORD	Chicago Ohare Intl	41.97860	-87.90484	13.570484	16642	729.00081
ATL	Hartsfield Jackson Atlanta Intl	33.63672	-84.42807	12.509824	16898	757.10822

Table 4: Average Departure Delay By Destination

There does not appear to be a relationship between the departure delay and the destination distance.

```
qplot(data$distance,data$delay,xlab='distance',
      ylab='delay',main='Delay By Destination Distance')
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

Delay By Destination Distance
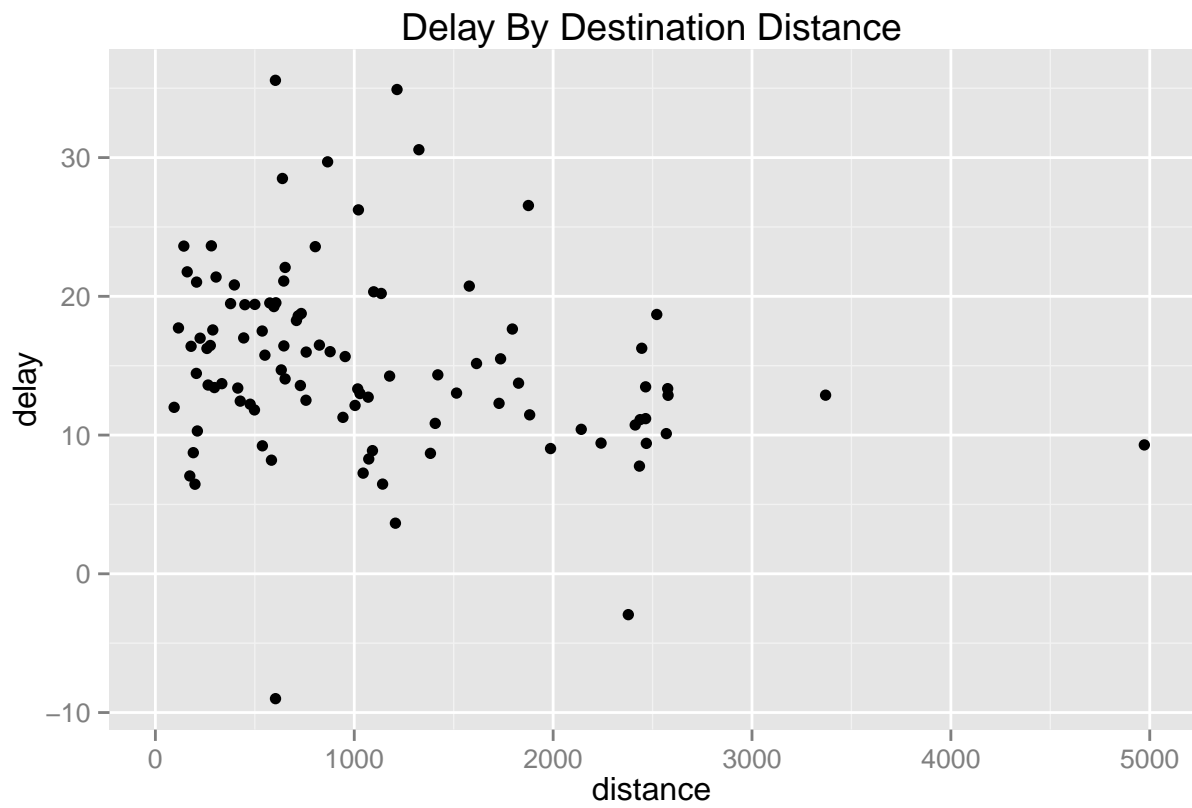
Disconnect from the PostgreSQL Database

```
# Close PostgreSQL connection
dbDisconnect(connectionHandle)
```

```
## [1] TRUE
```