

MA799 Assignment 1: The Flights Dataset

February 12 2015

Contents

Flights Data Description	1
General Overview	2
Descriptions of the rows/observations	4
Flights Dataset preparation	4
1. Create a copy of the dataset and save it to “flt1” to avoid manipulating the original data	4
2. Convert variables from character to factor (categorical)	4
3. Add new variables that contain the names of the airports and airlines	5
4. Convert new variables from character to factor (categorical)	5
5. View structure of the data set after modifications	6
6. Create categorical variable for departure delay (dep_delay)	6
7. Create categorical variable for arrival delay (arr_delay)	7
8. Verify that the new factors created match the original data using visualisation	7
Flights variable summaries	8
1. Obtain statistical data summaries for departure delay	8
2. Get cross-tabs for several categorical variables	10
Relationships between variables	20
1. Produce data visualizations	20
Conclusion	33

Flights Data Description

Flights data description can be “googled” and available at:

- <http://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf>

Alternatively, description of the dataset flights can be obtained with the command help:

```
help(flights)
```

General Overview

The dataset consists of on-time data for all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013.

Usage: Flights ; **Format:** Data frame with columns

The **Column Names** and **Column Descriptions** for the dataset are given below:

year, month, day - Date of departure

dep_time, arr_time - Departure and arrival times, local time zone

dep_delay, arr_delay - Departure and arrival delays, in minutes. Negative times represent early departures/arrivals

hour, minute - Time of departure broken in to hour and minutes

carrier - Two letter carrier abbreviation. See **airlines** dataset to get the names

tailnum - Plane tail number

flight - Flight number

origin, dest - Origin and destination. See **airports** dataset for additional metadata

air_time - Amount of time spent in the air

distance - Distance flown

Sourceof the original data: RITA, Bureau of transportation statistics

- http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

The flights dataset is just one of the datasets in the nycflights13 package. R comes with a handy collection of datasets which are used in some demos and tests. To see a list of provided datasets, we submitted the command:

```
data()
```

Datasets in package ‘nycflights13’:

airlines - Airline names

airports - Airport metadata

flights - Flights data

planes - Plane metadata

weather - Hourly weather data

Let us load flights dataset (Flights data) by passing the dataset name to data() function:

```
data(flights)
```

We can inspect the top of the dataset with Unix like function head() and the bottom with function tail()

```
library(nycflights13)
tail(flights)
```

```

##      year month day dep_time dep_delay arr_time arr_delay carrier
## 336771 2013    9   30       NA       NA       NA       NA     EV
## 336772 2013    9   30       NA       NA       NA       NA     9E
## 336773 2013    9   30       NA       NA       NA       NA     9E
## 336774 2013    9   30       NA       NA       NA       NA     MQ
## 336775 2013    9   30       NA       NA       NA       NA     MQ
## 336776 2013    9   30       NA       NA       NA       NA     MQ
##      tailnum flight origin dest air_time distance hour minute
## 336771 N740EV   5274   LGA   BNA       NA     764   NA   NA
## 336772           3393   JFK   DCA       NA     213   NA   NA
## 336773           3525   LGA   SYR       NA     198   NA   NA
## 336774 N535MQ   3461   LGA   BNA       NA     764   NA   NA
## 336775 N511MQ   3572   LGA   CLE       NA     419   NA   NA
## 336776 N839MQ   3531   LGA   RDU       NA     431   NA   NA

```

or

```

library(nycflights13)
head(flights, 2)

```

```

##      year month day dep_time dep_delay arr_time arr_delay carrier tailnum
## 1 2013    1   1     517        2     830       11      UA  N14228
## 2 2013    1   1     533        4     850       20      UA  N24211
##      flight origin dest air_time distance hour minute
## 1    1545   EWR   IAH     227    1400      5     17
## 2    1714   LGA   IAH     227    1416      5     33

```

The dataset has 336,776 observations (rows) and 16 variables (columns)

```

library(nycflights13)
dim(flights)

```

```

## [1] 336776     16

```

The str command lists the variables of the dataset along with their datatype and a few of the initial values.

```

library(nycflights13)
str(flights)

```

```

## Classes 'tbl_df', 'tbl' and 'data.frame': 336776 obs. of 16 variables:
## $ year      : int 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
## $ month     : int 1 1 1 1 1 1 1 1 1 ...
## $ day       : int 1 1 1 1 1 1 1 1 1 ...
## $ dep_time  : int 517 533 542 544 554 554 555 557 557 558 ...
## $ dep_delay : num 2 4 2 -1 -6 -4 -5 -3 -2 ...
## $ arr_time  : int 830 850 923 1004 812 740 913 709 838 753 ...
## $ arr_delay: num 11 20 33 -18 -25 12 19 -14 -8 8 ...
## $ carrier   : chr "UA" "UA" "AA" "B6" ...
## $ tailnum   : chr "N14228" "N24211" "N619AA" "N804JB" ...
## $ flight    : int 1545 1714 1141 725 461 1696 507 5708 79 301 ...
## $ origin    : chr "EWR" "LGA" "JFK" "JFK" ...

```

```

## $ dest      : chr  "IAH" "IAH" "MIA" "BQN" ...
## $ air_time  : num  227 227 160 183 116 150 158 53 140 138 ...
## $ distance  : num  1400 1416 1089 1576 762 ...
## $ hour      : num  5 5 5 5 5 5 5 5 5 ...
## $ minute    : num  17 33 42 44 54 54 55 57 58 ...

```

Descriptions of the rows/observations

Descriptions of the variables were obtained from the “googled” links mentioned above and the data types for all variables were obtained from running the commands `str(flights)` and `summary(flights)`. The results are presented in a Table 1 below.

Each observation (row) in the data.frame (table) `flights` represents a separate flight originated from the state of New York during 2013. Every flight has attributes shown in the Column Descriptions list in above. For example, row (record) #13 can be displayed with the command `flights [13,]`:

```

library(nycflights13)
flights [13, ]

```

```

##   year month day dep_time dep_delay arr_time arr_delay carrier tailnum
## 13 2013     1    1      558        -2       924         7     UA  N29129
##   flight origin dest air_time distance hour minute
## 13     194    JFK   LAX      345     2475      5      58

```

The interpretation of this particular record is as follows. The flight United Airlines (UA) # 194 with the tail number N29129 originated from JFK airport in New York, NY (JFK) flew to Los Angeles (LAX) airport in California on January 1, 2013 departing at 5:58 am (EST) and arriving at 9:24 am (PST). It flew a distance of 2,475 miles for 5 hours and 58 minutes, spending in the air 345 minutes and leaving JFK airport earlier than the scheduled time by 2 minutes.

Flights Dataset preparation

We took the following steps in dataset preparation.

1. Create a copy of the dataset and save it to “`flt1`” to avoid manipulating the original data

```

library(nycflights13)
flt1 <- flights

```

2. Convert variables from character to factor (categorical)

```

library(nycflights13)
flt1$carrier<-factor(flt1$carrier)
flt1$tailnum<-factor(flt1$tailnum)
flt1$origin<-factor(flt1$origin)
flt1$dest<-factor(flt1$dest)

```

3. Add new variables that contain the names of the airports and airlines

```
library(nycflights13)
library(dplyr)

## 
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

flt1<-left_join(flt1, airlines, by = NULL, copy = false)

## Joining by: "carrier"

## Warning: joining factors with different levels, coercing to character
## vector

names(flt1)[names(flt1) == 'name'] <- 'carrier_name'

flt1<-left_join(flt1, airports, by = c("origin" = "faa"), copy = false)

## Warning: joining character vector and factor, coercing into character
## vector

names(flt1)[names(flt1) == 'name'] <- 'departure_airport'

flt1<-left_join(flt1, airports, by = c("dest" = "faa"), copy = false)

## Warning: joining character vector and factor, coercing into character
## vector

names(flt1)[names(flt1) == 'name'] <- 'arrival_airport'
```

4. Convert new variables from character to factor (categorical)

```
library(nycflights13)
flt1$carrier_name<-factor(flt1$carrier_name)
flt1$departure_airport<-factor(flt1$departure_airport)
flt1$origin<-factor(flt1$origin)
flt1$arrival_airport<-factor(flt1$arrival_airport)
```

5. View structure of the data set after modifications

```
library(nycflights13)
str(flt1)

## Classes 'tbl_df', 'tbl' and 'data.frame': 336776 obs. of 29 variables:
## $ year : int 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
## $ month : int 1 1 1 1 1 1 1 1 1 ...
## $ day : int 1 1 1 1 1 1 1 1 1 ...
## $ dep_time : int 517 533 542 544 554 554 555 557 557 558 ...
## $ dep_delay : num 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
## $ arr_time : int 830 850 923 1004 812 740 913 709 838 753 ...
## $ arr_delay : num 11 20 33 -18 -25 12 19 -14 -8 8 ...
## $ carrier : chr "UA" "UA" "AA" "B6" ...
## $ tailnum : Factor w/ 4044 levels "", "D942DN", "NOEGMQ", ... : 181 525 2402 3205 2662 1143 1830 ...
## $ flight : int 1545 1714 1141 725 461 1696 507 5708 79 301 ...
## $ origin : Factor w/ 3 levels "EWR", "JFK", "LGA": 1 3 2 2 3 1 1 3 2 3 ...
## $ dest : chr "IAH" "IAH" "MIA" "BQN" ...
## $ air_time : num 227 227 160 183 116 150 158 53 140 138 ...
## $ distance : num 1400 1416 1089 1576 762 ...
## $ hour : num 5 5 5 5 5 5 5 5 5 5 ...
## $ minute : num 17 33 42 44 54 54 55 57 57 58 ...
## $ carrier_name : Factor w/ 16 levels "AirTran Airways Corporation", ... : 14 14 3 10 4 14 10 7 10 ...
## $ departure_airport: Factor w/ 3 levels "John F Kennedy Intl", ... : 3 2 1 1 2 3 3 2 1 2 ...
## $ lat.x : num 40.7 40.8 40.6 40.6 40.8 ...
## $ lon.x : num -74.2 -73.9 -73.8 -73.8 -73.9 ...
## $ alt.x : int 18 22 13 13 22 18 18 22 13 22 ...
## $ tz.x : num -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 ...
## $ dst.x : chr "A" "A" "A" "A" ...
## $ arrival_airport : Factor w/ 101 levels "Akron Canton Regional Airport", ... : 34 34 59 NA 38 19 30 ...
## $ lat.y : num 30 30 25.8 NA 33.6 ...
## $ lon.y : num -95.3 -95.3 -80.3 NA -84.4 ...
## $ alt.y : int 97 97 8 NA 1026 668 9 313 96 668 ...
## $ tz.y : num -6 -6 -5 NA -5 -6 -5 -5 -6 ...
## $ dst.y : chr "A" "A" "A" NA ...
```

6. Create categorical variable for departure delay (dep_delay)

```
library(nycflights13)
flt1$dep_status <- flt1$dep_delay

flt1$dep_status<- ifelse(flt1$dep_status < 0, 'Early',
                           ifelse(flt1$dep_status < 15, 'On time', 'Late'))

flt1$dep_status<-factor(flt1$dep_status)

str(flt1$dep_status)

## Factor w/ 3 levels "Early", "Late", ... : 3 3 3 1 1 1 1 1 1 1 ...
```

7. Create categorical variable for arrival delay (arr_delay)

```
library(nycflights13)
flt1$arr_status <- flt1$arr_delay

flt1$arr_status<- ifelse(flt1$arr_status < 0,'Early',
                         ifelse(flt1$arr_status < 15,'On time','Late'))

flt1$arr_status<-factor(flt1$arr_status)

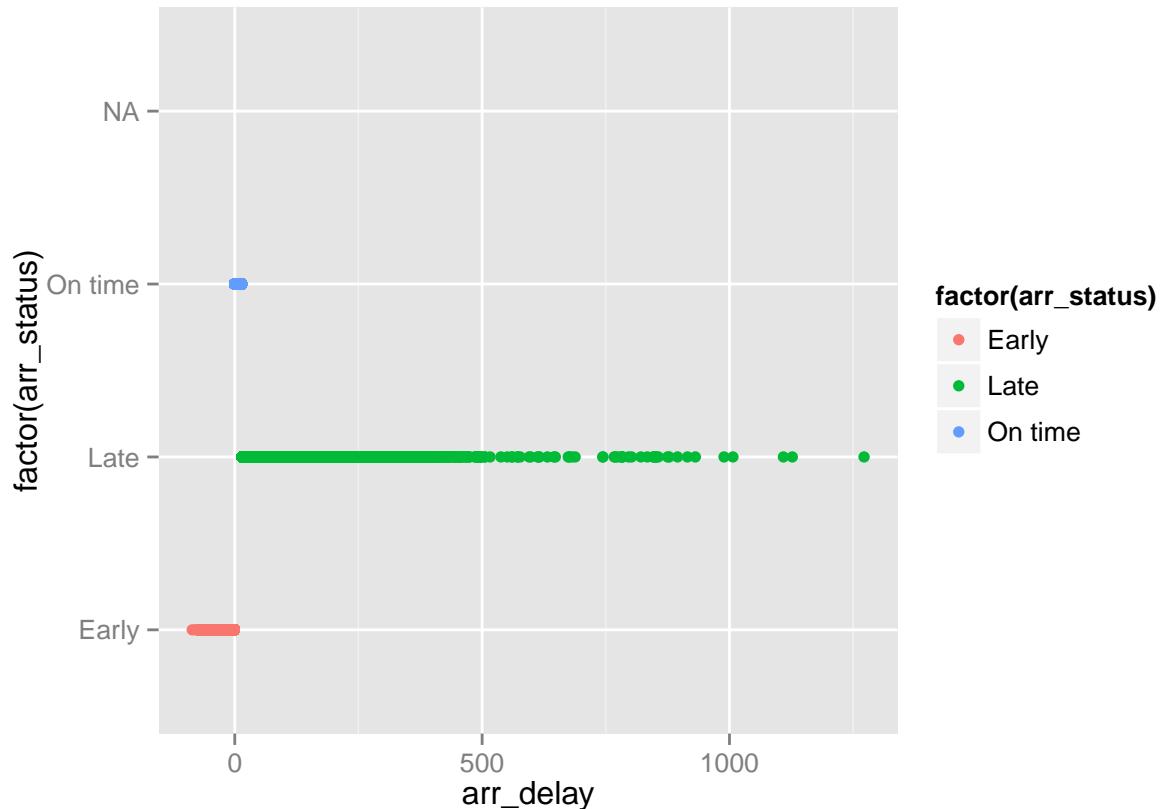
str(flt1$arr_status)

##  Factor w/ 3 levels "Early","Late",...: 3 2 2 1 1 3 2 1 1 3 ...
```

8. Verify that the new factors created match the original data using visualisation

```
library(nycflights13)
library(ggplot2)
ggplot(data=flt1, aes(x=arr_delay,
                      y=factor(arr_status))) +
  geom_point(aes(color=factor(arr_status)))

## Warning: Removed 9430 rows containing missing values (geom_point).
```

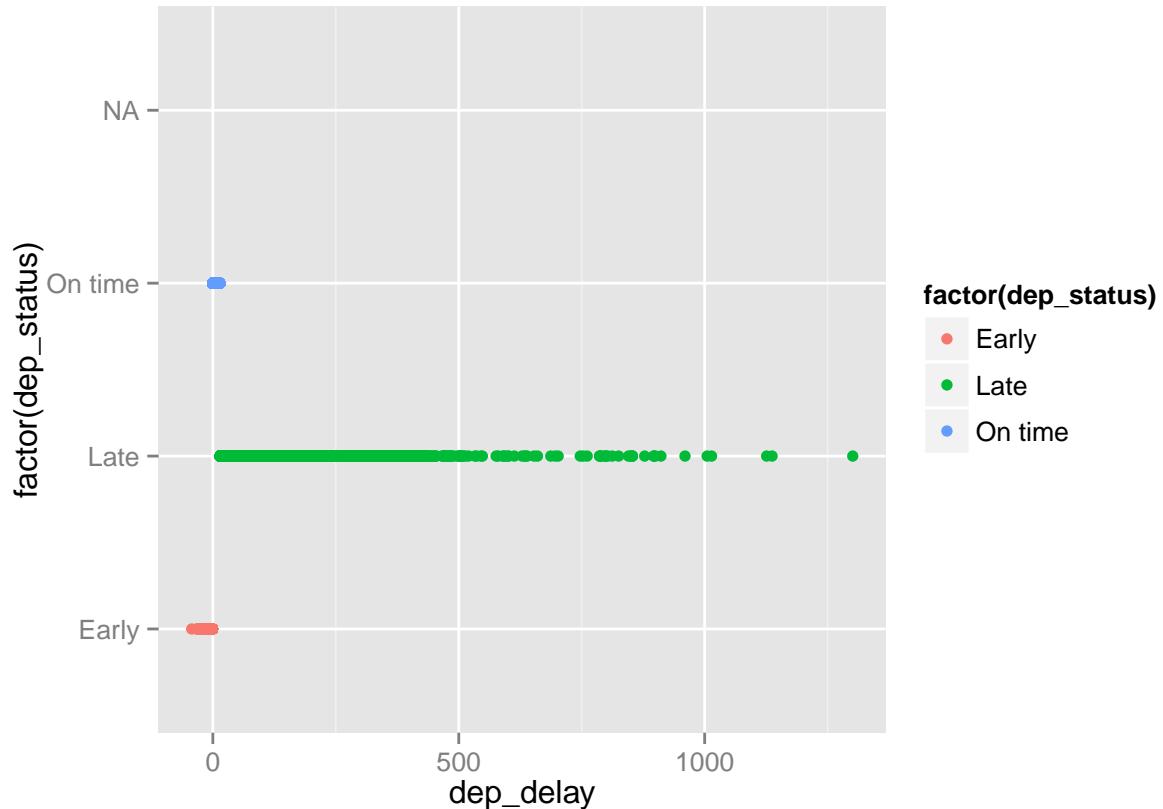


```

ggplot(data=flt1, aes(x=dep_delay,
                      y=factor(dep_status))) +
  geom_point(aes(color=factor(dep_status)))

## Warning: Removed 8255 rows containing missing values (geom_point).

```



Flights variable summaries

1. Obtain statistical data summaries for departure delay

```

library(nycflights13)
summarize(group_by(flt1, month),
          min=min(dep_delay,na.rm=TRUE),
          max=max(dep_delay,na.rm=TRUE),
          mean=mean(dep_delay,na.rm=TRUE),
          median=median(dep_delay,na.rm=TRUE),
          standard_dev=sd(dep_delay,na.rm=TRUE))

```

```

## Source: local data frame [12 x 6]
##
##    month min  max      mean median standard_dev

```

```

## 1      1 -30 1301 10.036665      -2    36.39031
## 2      2 -33 853 10.816843      -2    36.26655
## 3      3 -25 911 13.227076      -1    40.13097
## 4      4 -21 960 13.938038      -2    42.96626
## 5      5 -24 878 12.986859      -1    39.35283
## 6      6 -21 1137 20.846332      0    51.45694
## 7      7 -22 1005 21.727787      0    51.61608
## 8      8 -26 520 12.611040      -1    37.66692
## 9      9 -24 1014 6.722476      -3    35.61480
## 10     10 -25 702 6.243988      -3    29.67176
## 11     11 -32 798 5.435362      -3    27.58836
## 12     12 -43 896 16.576688      0    41.87681

summarize(group_by(flt1, departure_airport),
          min=min(dep_delay,na.rm=TRUE),
          max=max(dep_delay,na.rm=TRUE),
          mean=mean(dep_delay,na.rm=TRUE),
          median=median(dep_delay,na.rm=TRUE),
          standard_dev=sd(dep_delay,na.rm=TRUE))

## Source: local data frame [3 x 6]
##
##   departure_airport min   max     mean median standard_dev
## 1 John F Kennedy Intl -43 1301 12.11216      -1    39.03507
## 2           La Guardia -33  911 10.34688      -3    39.99302
## 3 Newark Liberty Intl -25 1126 15.10795      -1    41.32370

summarize(group_by(flt1, carrier_name),
          min=min(dep_delay,na.rm=TRUE),
          max=max(dep_delay,na.rm=TRUE),
          mean=mean(dep_delay,na.rm=TRUE),
          median=median(dep_delay,na.rm=TRUE),
          standard_dev=sd(dep_delay,na.rm=TRUE))

## Source: local data frame [16 x 6]
##
##   carrier_name min   max     mean median standard_dev
## 1 AirTran Airways Corporation -22  602 18.726075    1.0    52.66160
## 2 Alaska Airlines Inc. -21  225  5.804775   -3.0    31.36303
## 3 American Airlines Inc. -24 1014  8.586016   -3.0    37.35486
## 4 Delta Air Lines Inc. -33  960  9.264505   -2.0    39.73505
## 5 Endeavor Air Inc. -24  747 16.725769   -2.0    45.90604
## 6 Envoy Air -26 1137 10.552041   -3.0    39.18457
## 7 ExpressJet Airlines Inc. -32  548 19.955390   -1.0    46.55235
## 8 Frontier Airlines Inc. -27  853 20.215543    0.5    58.36265
## 9 Hawaiian Airlines Inc. -16 1301  4.900585   -4.0    74.10990
## 10 JetBlue Airways -43  502 13.022522   -1.0    38.50337
## 11 Mesa Airlines Inc. -16  387 18.996330   -2.0    49.17227
## 12 SkyWest Airlines Inc. -14  154 12.586207   -6.0    43.06599
## 13 Southwest Airlines Co. -13  471 17.711744    1.0    43.34435
## 14 United Air Lines Inc. -20  483 12.106073    0.0    35.71660
## 15 US Airways Inc. -19  500  3.782418   -4.0    28.05633
## 16 Virgin America -20  653 12.869421    0.0    44.81510

```

```

summarize(group_by(flt1, carrier_name),
      min_dep=min(dep_delay,na.rm=TRUE),
      max_dep=max(dep_delay,na.rm=TRUE),
      mean_dep=mean(dep_delay,na.rm=TRUE),
      median_dep=median(dep_delay,na.rm=TRUE),
      standard_dev_dep=sd(dep_delay,na.rm=TRUE),
      min_arr=min(arr_delay,na.rm=TRUE),
      max_arr=max(arr_delay,na.rm=TRUE),
      mean_arr=mean(arr_delay,na.rm=TRUE),
      median_arr=median(arr_delay,na.rm=TRUE),
      standard_dev_arr=sd(arr_delay,na.rm=TRUE))

```

```

## Source: local data frame [16 x 11]
##
## #> #>   carrier_name min_dep max_dep  mean_dep median_dep
## #> 1 AirTran Airways Corporation    -22     602 18.726075     1.0
## #> 2 Alaska Airlines Inc.       -21     225  5.804775    -3.0
## #> 3 American Airlines Inc.     -24    1014  8.586016    -3.0
## #> 4 Delta Air Lines Inc.      -33     960  9.264505    -2.0
## #> 5 Endeavor Air Inc.        -24     747 16.725769    -2.0
## #> 6 Envoy Air                 -26    1137 10.552041    -3.0
## #> 7 ExpressJet Airlines Inc.   -32     548 19.955390    -1.0
## #> 8 Frontier Airlines Inc.    -27     853 20.215543     0.5
## #> 9 Hawaiian Airlines Inc.   -16    1301  4.900585    -4.0
## #> 10 JetBlue Airways         -43     502 13.022522    -1.0
## #> 11 Mesa Airlines Inc.       -16     387 18.996330    -2.0
## #> 12 SkyWest Airlines Inc.    -14     154 12.586207    -6.0
## #> 13 Southwest Airlines Co.   -13     471 17.711744     1.0
## #> 14 United Air Lines Inc.   -20     483 12.106073     0.0
## #> 15 US Airways Inc.          -19     500  3.782418    -4.0
## #> 16 Virgin America          -20     653 12.869421     0.0
## #> #> Variables not shown: standard_dev_dep (dbl), min_arr (dbl), max_arr (dbl),
## #> #>   mean_arr (dbl), median_arr (dbl), standard_dev_arr (dbl)

```

2. Get cross-tabs for several categorical variables

```

library(nycflights13)

xtabs(formula = ~ carrier_name + departure_airport, data = flt1)

```

	carrier_name	departure_airport
##		John F Kennedy Intl La Guardia
##	AirTran Airways Corporation	0 3260
##	Alaska Airlines Inc.	0 0
##	American Airlines Inc.	13783 15459
##	Delta Air Lines Inc.	20701 23067
##	Endeavor Air Inc.	14651 2541
##	Envoy Air	7193 16928
##	ExpressJet Airlines Inc.	1408 8826
##	Frontier Airlines Inc.	0 685
##	Hawaiian Airlines Inc.	342 0

```

## JetBlue Airways          42076    6002
## Mesa Airlines Inc.       0        601
## SkyWest Airlines Inc.    0        26
## Southwest Airlines Co.   0       6087
## United Air Lines Inc.   4534     8044
## US Airways Inc.         2995    13136
## Virgin America           3596      0
##                                     departure_airport
## carrier_name                Newark Liberty Intl
## AirTran Airways Corporation 0
## Alaska Airlines Inc.       714
## American Airlines Inc.    3487
## Delta Air Lines Inc.      4342
## Endeavor Air Inc.         1268
## Envoy Air                  2276
## ExpressJet Airlines Inc.   43939
## Frontier Airlines Inc.    0
## Hawaiian Airlines Inc.    0
## JetBlue Airways            6557
## Mesa Airlines Inc.         0
## SkyWest Airlines Inc.      6
## Southwest Airlines Co.    6188
## United Air Lines Inc.    46087
## US Airways Inc.           4405
## Virgin America             1566

```

```
xtabs(formula = ~ arrival_airport + departure_airport, data = flt1)
```

	arrival_airport	departure_airport
		John F Kennedy Intl La Guardia
## Akron Canton Regional Airport		0 864
## Albany Intl		0 0
## Albuquerque International Sunport		254 0
## Asheville Regional Airport		0 10
## Austin Bergstrom Intl		1471 0
## Baltimore Washington Intl		1221 15
## Bangor Intl		0 375
## Birmingham Intl		1 296
## Blue Grass		0 1
## Bob Hope		371 0
## Bradley Intl		0 0
## Buffalo Niagara Intl		3582 126
## Burlington Intl		1364 294
## Charleston Afb Intl		960 500
## Charlotte Douglas Intl		2870 6168
## Charlottesville-Albemarle		0 52
## Cherry Capital Airport		0 77
## Chicago Midway Intl		0 2070
## Chicago Ohare Intl		2326 8857
## Cincinnati Northern Kentucky Intl		997 271
## Cleveland Hopkins Intl		707 2112
## Columbia Metropolitan		0 12
## Dallas Fort Worth Intl		732 4858
## Dane Co Rgnl Truax Fld		0 218

## Denver Intl	703	3704
## Des Moines Intl	0	158
## Detroit Metro Wayne Co	1166	5040
## Eagle Co Rgnl	103	0
## Eppley Afld	0	95
## Fort Lauderdale Hollywood Intl	4254	4008
## Gallatin Field	0	0
## General Edward Lawrence Logan Intl	5898	4283
## General Mitchell Intl	183	1525
## George Bush Intercontinental	274	2951
## Gerald R Ford Intl	0	46
## Greater Rochester Intl	1679	224
## Greenville-Spartanburg International	0	104
## Hartsfield Jackson Atlanta Intl	1930	10263
## Honolulu Intl	342	0
## Indianapolis Intl	689	87
## Jackson Hole Airport	2	0
## Jacksonville Intl	1299	179
## James M Cox Dayton Intl	0	391
## John Wayne Arpt Orange Co	0	0
## Kansas City Intl	276	376
## Key West Intl	0	17
## La Guardia	0	0
## Lambert St Louis Intl	1	1822
## Long Beach	668	0
## Los Angeles Intl	11262	0
## Louis Armstrong New Orleans Intl	1633	1011
## Louisville International Airport	46	217
## Manchester Regional Airport	0	142
## Martha\\\\'s Vineyard	221	0
## Mc Carran Intl	3987	0
## Mc Ghee Tyson	0	308
## Memphis Intl	1	862
## Metropolitan Oakland Intl	312	0
## Miami Intl	3314	5781
## Minneapolis St Paul Intl	1095	3713
## Montrose Regional Airport	0	0
## Myrtle Beach Intl	0	3
## Nantucket Mem	265	0
## Nashville Intl	730	3267
## Norfolk Intl	746	206
## Norman Y Mineta San Jose Intl	329	0
## NW Arkansas Regional	0	745
## Orlando Intl	5464	3677
## Palm Beach Intl	1739	2464
## Palm Springs Intl	19	0
## Philadelphia Intl	976	607
## Phoenix Sky Harbor Intl	1933	0
## Piedmont Triad	0	486
## Pittsburgh Intl	1249	1067
## Port Columbus Intl	742	2058
## Portland Intl	783	0
## Portland Intl Jetport	1304	278
## Raleigh Durham Intl	3100	3581

##	Richmond Intl	249	538
##	Ronald Reagan Washington Natl	3270	4716
##	Sacramento Intl	284	0
##	Salt Lake City Intl	2113	0
##	San Antonio Intl	356	0
##	San Diego Intl	1603	0
##	San Francisco Intl	8204	0
##	Sarasota Bradenton Intl	474	737
##	Savannah Hilton Head Intl	0	68
##	Seattle Tacoma Intl	2092	0
##	South Bend Rgnl	0	6
##	Southwest Florida Intl	1339	761
##	Syracuse Hancock Intl	1311	293
##	Tampa Intl	2987	2145
##	Ted Stevens Anchorage Intl	0	0
##	Theodore Francis Green State	0	0
##	Tulsa Intl	0	0
##	Washington Dulles Intl	2661	1803
##	Will Rogers World	0	0
##	William P Hobby	714	425
##	Wilmington Intl	0	110
##	Yampa Valley	0	0
##	Yeager	0	138
##	departure_airport		
##	arrival_airport	Newark Liberty Intl	
##	Akron Canton Regional Airport	0	
##	Albany Intl	439	
##	Albuquerque International Sunport	0	
##	Asheville Regional Airport	265	
##	Austin Bergstrom Intl	968	
##	Baltimore Washington Intl	545	
##	Bangor Intl	0	
##	Birmingham Intl	0	
##	Blue Grass	0	
##	Bob Hope	0	
##	Bradley Intl	443	
##	Buffalo Niagara Intl	973	
##	Burlington Intl	931	
##	Charleston Afb Intl	1424	
##	Charlotte Douglas Intl	5026	
##	Charlottesville-Albemarle	0	
##	Cherry Capital Airport	24	
##	Chicago Midway Intl	2043	
##	Chicago Ohare Intl	6100	
##	Cincinnati Northern Kentucky Intl	2673	
##	Cleveland Hopkins Intl	1754	
##	Columbia Metropolitan	104	
##	Dallas Fort Worth Intl	3148	
##	Dane Co Rgnl Truax Fld	354	
##	Denver Intl	2859	
##	Des Moines Intl	411	
##	Detroit Metro Wayne Co	3178	
##	Eagle Co Rgnl	110	
##	Eppley Afld	754	

## Fort Lauderdale Hollywood Intl	3793
## Gallatin Field	36
## General Edward Lawrence Logan Intl	5327
## General Mitchell Intl	1094
## George Bush Intercontinental	3973
## Gerald R Ford Intl	719
## Greater Rochester Intl	513
## Greenville-Spartanburg International	745
## Hartsfield Jackson Atlanta Intl	5022
## Honolulu Intl	365
## Indianapolis Intl	1301
## Jackson Hole Airport	23
## Jacksonville Intl	1242
## James M Cox Dayton Intl	1134
## John Wayne Arpt Orange Co	825
## Kansas City Intl	1356
## Key West Intl	0
## La Guardia	1
## Lambert St Louis Intl	2516
## Long Beach	0
## Los Angeles Intl	4912
## Louis Armstrong New Orleans Intl	1155
## Louisville International Airport	894
## Manchester Regional Airport	867
## Martha\\\\'s Vineyard	0
## Mc Carran Intl	2010
## Mc Ghee Tyson	323
## Memphis Intl	926
## Metropolitan Oakland Intl	0
## Miami Intl	2633
## Minneapolis St Paul Intl	2377
## Montrose Regional Airport	15
## Myrtle Beach Intl	56
## Nantucket Mem	0
## Nashville Intl	2336
## Norfolk Intl	584
## Norman Y Mineta San Jose Intl	0
## NW Arkansas Regional	291
## Orlando Intl	4941
## Palm Beach Intl	2351
## Palm Springs Intl	0
## Philadelphia Intl	49
## Phoenix Sky Harbor Intl	2723
## Piedmont Triad	1120
## Pittsburgh Intl	559
## Port Columbus Intl	724
## Portland Intl	571
## Portland Intl Jetport	770
## Raleigh Durham Intl	1482
## Richmond Intl	1667
## Ronald Reagan Washington Natl	1719
## Sacramento Intl	0
## Salt Lake City Intl	354
## San Antonio Intl	330

```

## San Diego Intl          1134
## San Francisco Intl      5127
## Sarasota Bradenton Intl     0
## Savannah Hilton Head Intl   736
## Seattle Tacoma Intl       1831
## South Bend Rgnl            4
## Southwest Florida Intl     1437
## Syracuse Hancock Intl      157
## Tampa Intl                  2334
## Ted Stevens Anchorage Intl    8
## Theodore Francis Green State 376
## Tulsa Intl                  315
## Washington Dulles Intl     1236
## Will Rogers World           346
## William P Hobby                976
## Wilmington Intl                 0
## Yampa Valley                      15
## Yeager                           0

```

```
xtabs(formula = ~ arrival_airport, data = flt1)
```

```

## arrival_airport
## Akron Canton Regional Airport          Albany Intl
##                               864                   439
## Albuquerque International Sunport        Asheville Regional Airport
##                               254                   275
## Austin Bergstrom Intl                  Baltimore Washington Intl
##                               2439                  1781
## Bangor Intl                            Birmingham Intl
##                               375                   297
## Blue Grass                             Bob Hope
##                               1                   371
## Bradley Intl                           Buffalo Niagara Intl
##                               443                   4681
## Burlington Intl                        Charleston Afb Intl
##                               2589                  2884
## Charlotte Douglas Intl                 Charlottesville-Albemarle
##                               14064                  52
## Cherry Capital Airport                  Chicago Midway Intl
##                               101                   4113
## Chicago Ohare Intl                    Cincinnati Northern Kentucky Intl
##                               17283                  3941
## Cleveland Hopkins Intl                 Columbia Metropolitan
##                               4573                   116
## Dallas Fort Worth Intl                Dane Co Rgnl Truax Fld
##                               8738                   572
## Denver Intl                            Des Moines Intl
##                               7266                   569
## Detroit Metro Wayne Co                Eagle Co Rgnl
##                               9384                   213
## Eppley Afld                           Fort Lauderdale Hollywood Intl
##                               849                   12055
## Gallatin Field                         General Edward Lawrence Logan Intl
##                               36                   15508

```

##	General Mitchell Intl	George Bush Intercontinental
##	2802	7198
##	Gerald R Ford Intl	Greater Rochester Intl
##	765	2416
##	Greenville-Spartanburg International	Hartsfield Jackson Atlanta Intl
##	849	17215
##	Honolulu Intl	Indianapolis Intl
##	707	2077
##	Jackson Hole Airport	Jacksonville Intl
##	25	2720
##	James M Cox Dayton Intl	John Wayne Arpt Orange Co
##	1525	825
##	Kansas City Intl	Key West Intl
##	2008	17
##	La Guardia	Lambert St Louis Intl
##	1	4339
##	Long Beach	Los Angeles Intl
##	668	16174
##	Louis Armstrong New Orleans Intl	Louisville International Airport
##	3799	1157
##	Manchester Regional Airport	Martha\\\\'s Vineyard
##	1009	221
##	Mc Carran Intl	Mc Ghee Tyson
##	5997	631
##	Memphis Intl	Metropolitan Oakland Intl
##	1789	312
##	Miami Intl	Minneapolis St Paul Intl
##	11728	7185
##	Montrose Regional Airport	Myrtle Beach Intl
##	15	59
##	Nantucket Mem	Nashville Intl
##	265	6333
##	Norfolk Intl	Norman Y Mineta San Jose Intl
##	1536	329
##	NW Arkansas Regional	Orlando Intl
##	1036	14082
##	Palm Beach Intl	Palm Springs Intl
##	6554	19
##	Philadelphia Intl	Phoenix Sky Harbor Intl
##	1632	4656
##	Piedmont Triad	Pittsburgh Intl
##	1606	2875
##	Port Columbus Intl	Portland Intl
##	3524	1354
##	Portland Intl Jetport	Raleigh Durham Intl
##	2352	8163
##	Richmond Intl	Ronald Reagan Washington Natl
##	2454	9705
##	Sacramento Intl	Salt Lake City Intl
##	284	2467
##	San Antonio Intl	San Diego Intl
##	686	2737
##	San Francisco Intl	Sarasota Bradenton Intl
##	13331	1211

```

##          Savannah Hilton Head Intl      Seattle Tacoma Intl
##                               804           3923
##          South Bend Rgnl      Southwest Florida Intl
##                               10            3537
##          Syracuse Hancock Intl      Tampa Intl
##                               1761           7466
##          Ted Stevens Anchorage Intl    Theodore Francis Green State
##                               8              376
##          Tulsa Intl      Washington Dulles Intl
##                               315            5700
##          Will Rogers World      William P Hobby
##                               346            2115
##          Wilmington Intl      Yampa Valley
##                               110             15
##          Yeager
##                               138

```

```
xtabs(formula = ~ departure_airport + dep_status, data = flt1)
```

```

##          dep_status
## departure_airport Early Late On time
## John F Kennedy Intl 61146 23347 24923
## La Guardia          63129 19792 18588
## Newark Liberty Intl 59300 29775 28521

```

```
xtabs(formula = ~ arrival_airport + arr_status, data = flt1)
```

##	arr_status
## arrival_airport	Early Late On time
## Akron Canton Regional Airport	374 272 196
## Albany Intl	229 130 59
## Albuquerque International Sunport	145 78 31
## Asheville Regional Airport	134 62 65
## Austin Bergstrom Intl	1401 615 395
## Baltimore Washington Intl	1001 448 238
## Bangor Intl	216 101 41
## Birmingham Intl	144 89 36
## Blue Grass	1 0 0
## Bob Hope	197 103 70
## Bradley Intl	264 102 46
## Buffalo Niagara Intl	2708 1072 790
## Burlington Intl	1446 641 423
## Charleston Afb Intl	1520 726 513
## Charlotte Douglas Intl	7575 3307 2792
## Charlottesville-Albemarle	27 12 7
## Cherry Capital Airport	59 29 7
## Chicago Midway Intl	2073 1069 883
## Chicago Ohare Intl	10146 3973 2447
## Cincinnati Northern Kentucky Intl	1987 1126 612
## Cleveland Hopkins Intl	2540 1090 764
## Columbia Metropolitan	17 67 22
## Dallas Fort Worth Intl	5389 1675 1324
## Dane Co Rgnl Truax Fld	264 197 95

##	Denver Intl	3819	1966	1384
##	Des Moines Intl	257	184	82
##	Detroit Metro Wayne Co	5627	1933	1471
##	Eagle Co Rgnl	113	53	41
##	Eppley Afld	420	273	124
##	Fort Lauderdale Hollywood Intl	6496	3200	2201
##	Gallatin Field	18	8	9
##	General Edward Lawrence Logan Intl	10068	2855	2099
##	General Mitchell Intl	1346	817	546
##	George Bush Intercontinental	4078	1633	1374
##	Gerald R Ford Intl	332	252	144
##	Greater Rochester Intl	1378	624	356
##	Greenville-Spartanburg International	395	270	125
##	Hartsfield Jackson Atlanta Intl	8541	4417	3879
##	Honolulu Intl	446	115	140
##	Indianapolis Intl	1104	516	361
##	Jackson Hole Airport	3	11	7
##	Jacksonville Intl	1360	793	470
##	James M Cox Dayton Intl	744	416	239
##	John Wayne Arpt Orange Co	564	131	117
##	Kansas City Intl	934	615	336
##	Key West Intl	6	5	6
##	La Guardia	0	0	0
##	Lambert St Louis Intl	2248	1124	770
##	Long Beach	416	150	95
##	Los Angeles Intl	9790	3365	2871
##	Louis Armstrong New Orleans Intl	2169	934	612
##	Louisville International Airport	591	313	200
##	Manchester Regional Airport	505	287	140
##	Martha\\\\'s Vineyard	147	40	23
##	Mc Carran Intl	3745	1178	1029
##	Mc Ghee Tyson	261	228	89
##	Memphis Intl	912	490	284
##	Metropolitan Oakland Intl	191	77	41
##	Miami Intl	7570	2184	1839
##	Minneapolis St Paul Intl	4047	1656	1226
##	Montrose Regional Airport	9	3	2
##	Myrtle Beach Intl	37	16	5
##	Nantucket Mem	155	52	57
##	Nashville Intl	3244	1729	1111
##	Norfolk Intl	792	408	234
##	Norman Y Mineta San Jose Intl	198	84	46
##	NW Arkansas Regional	522	265	205
##	Orlando Intl	8199	3192	2576
##	Palm Beach Intl	3508	1740	1239
##	Palm Springs Intl	12	1	5
##	Philadelphia Intl	850	441	250
##	Phoenix Sky Harbor Intl	2711	1001	894
##	Piedmont Triad	786	469	237
##	Pittsburgh Intl	1624	681	441
##	Port Columbus Intl	1842	894	590
##	Portland Intl	765	369	208
##	Portland Intl Jetport	1272	643	373
##	Raleigh Durham Intl	4232	2059	1479

```

## Richmond Intl                1115   858   373
## Ronald Reagan Washington Natl 4933  2320  1858
## Sacramento Intl                  131    100    51
## Salt Lake City Intl              1578   465   408
## San Antonio Intl                  395    169    95
## San Diego Intl                   1563   659   487
## San Francisco Intl                 8020  2982  2171
## Sarasota Bradenton Intl            713    246   242
## Savannah Hilton Head Intl            376    225   148
## Seattle Tacoma Intl                 2560   773   552
## South Bend Rgnl                      5     3     2
## Southwest Florida Intl                2043   742   717
## Syracuse Hancock Intl                 1005   432   270
## Tampa Intl                         4204  1816  1370
## Ted Stevens Anchorage Intl                  3     1     4
## Theodore Francis Green State            173    121    64
## Tulsa Intl                           99    145    50
## Washington Dulles Intl                2916  1477  990
## Will Rogers World                     111    163    41
## William P Hobby                        1152   546   385
## Wilmington Intl                         64    32    11
## Yampa Valley                            6     2     6
## Yeager                                68    48    18

```

```
xtabs(formula = ~ carrier_name + dep_status, data = flt1)
```

	dep_status			
carrier_name	Early	Late	On time	
AirTran Airways Corporation	1401	882	904	
Alaska Airlines Inc.	458	100	154	
American Airlines Inc.	20324	5270	6499	
Delta Air Lines Inc.	29654	8081	10026	
Endeavor Air Inc.	9718	4563	3135	
Envoy Air	15825	5481	3857	
ExpressJet Airlines Inc.	26558	15991	8807	
Frontier Airlines Inc.	297	202	183	
Hawaiian Airlines Inc.	259	24	59	
JetBlue Airways	29952	12711	11506	
Mesa Airlines Inc.	294	160	91	
SkyWest Airlines Inc.	20	6	3	
Southwest Airlines Co.	4538	3352	4193	
United Air Lines Inc.	27321	12655	18003	
US Airways Inc.	14461	2511	2901	
Virgin America	2495	925	1711	

```
xtabs(formula = ~ carrier_name + arr_status, data = flt1)
```

	arr_status			
carrier_name	Early	Late	On time	
AirTran Airways Corporation	1211	1100	864	
Alaska Airlines Inc.	512	104	93	
American Airlines Inc.	20769	6203	4975	
Delta Air Lines Inc.	30463	9038	8157	

```

## Endeavor Air Inc.          10430  4422   2442
## Envoy Air                  12856  6986   5195
## ExpressJet Airlines Inc.   25823 16399   8886
## Frontier Airlines Inc.    271    264    146
## Hawaiian Airlines Inc.    241    44     57
## JetBlue Airways            29514 14567   9968
## Mesa Airlines Inc.         280    179     85
## SkyWest Airlines Inc.      19     7      3
## Southwest Airlines Co.    6513   3094   2437
## United Air Lines Inc.     34642 13004  10136
## US Airways Inc.           12113  3733   3985
## Virgin America             3276   956    884

```

```
xtabs(formula = ~ dep_status + arr_status, data = flt1)
```

```

##                 arr_status
## dep_status Early Late On time
##   Early      144346 10585 28204
##   Late       4495  57877 10048
##   On time    40092 11638 20061

```

Relationships between variables

1. Produce data visualizations

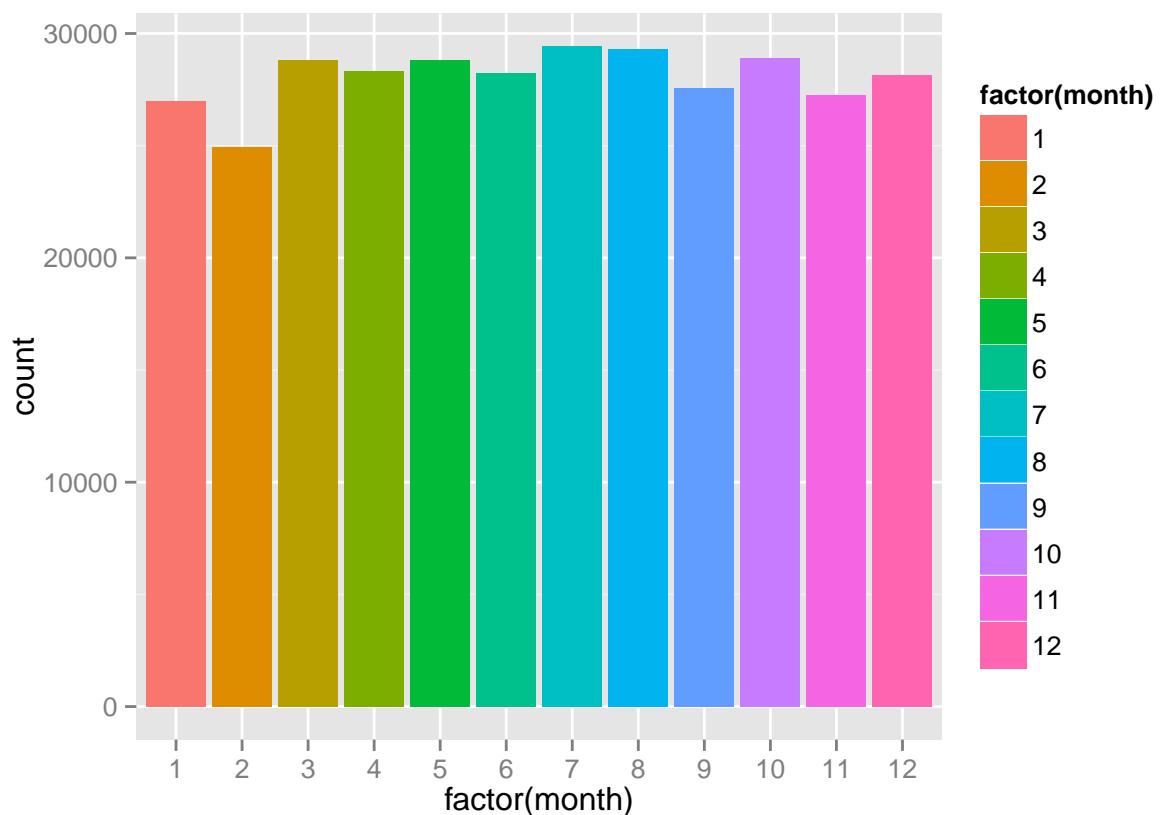
```

library(nycflights13)
library(dplyr)
library(ggplot2)
library(vcd)

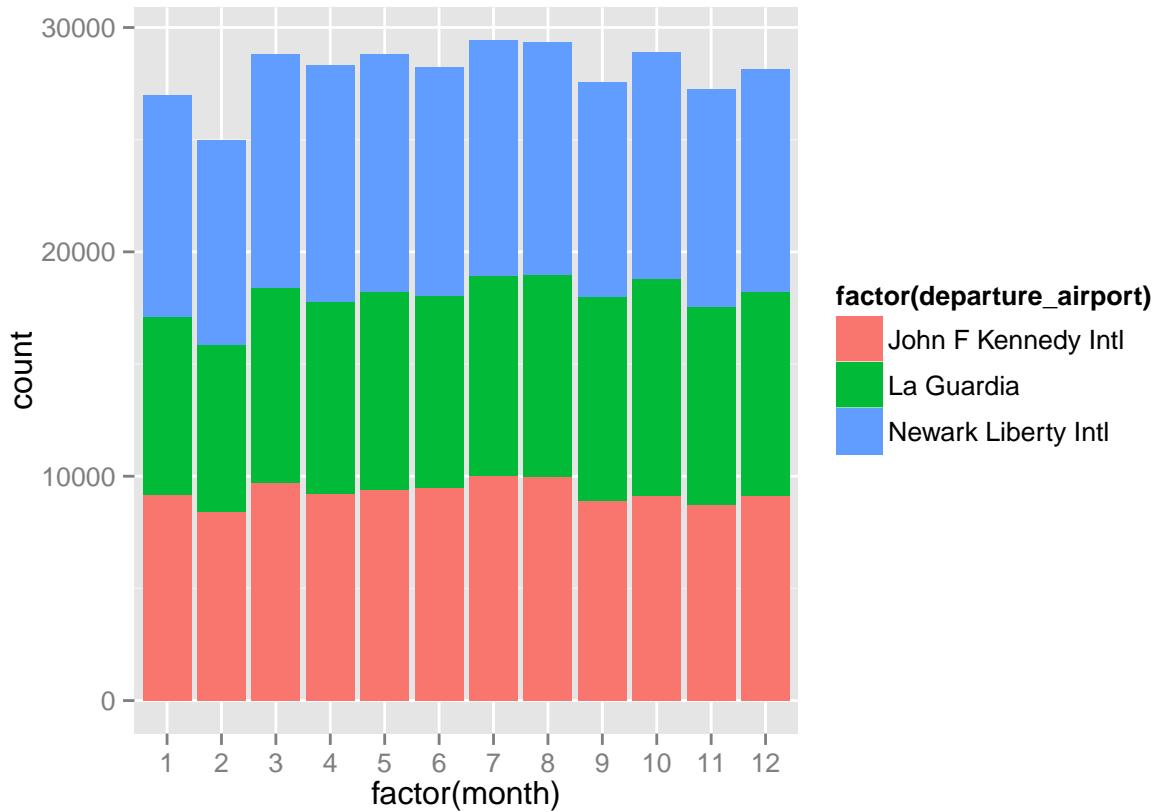
## Loading required package: grid

ggplot(data=flt1,
       aes(x=factor(month))) +
  geom_bar(aes(fill=factor(month)))

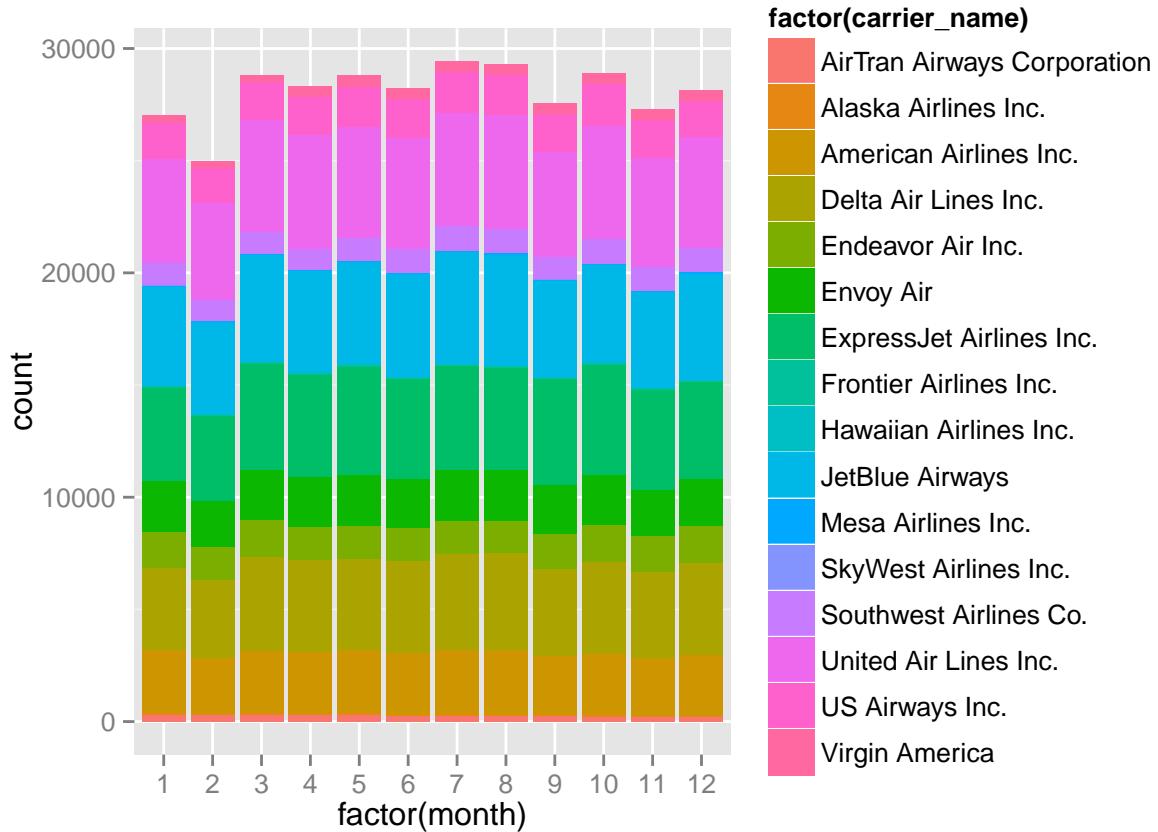
```



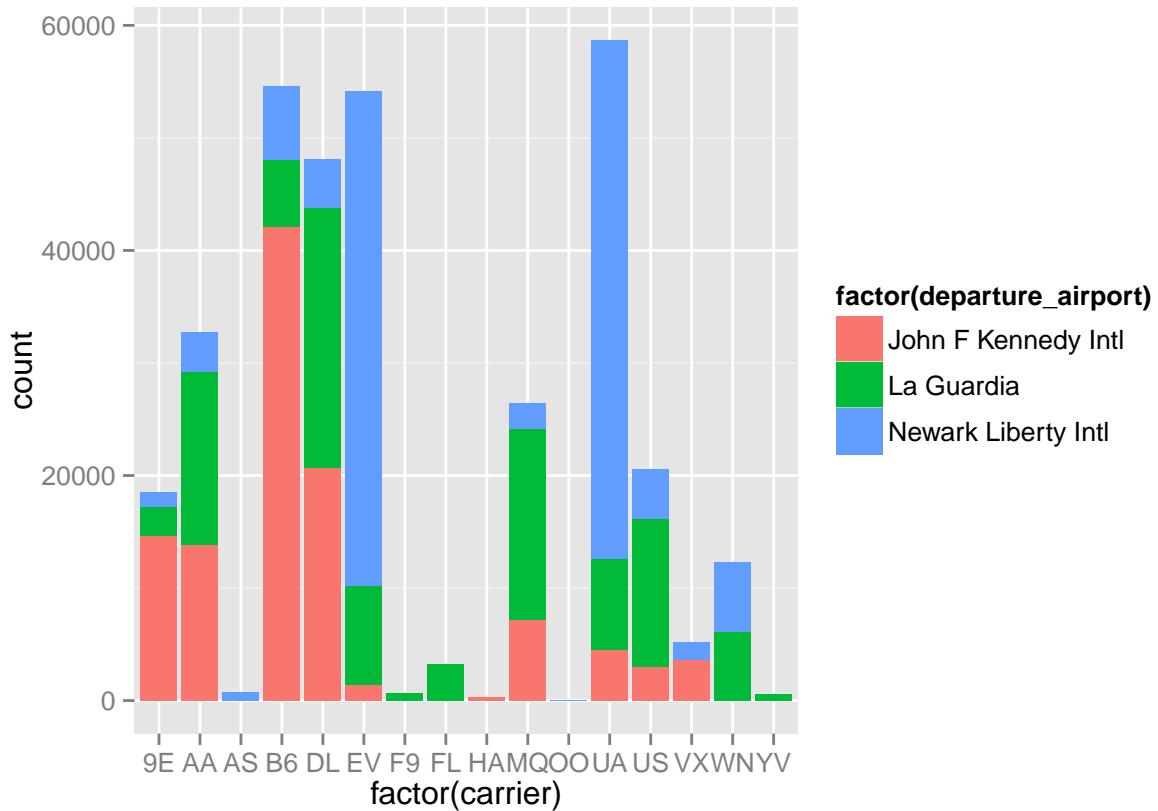
```
ggplot(data=flt1,
       aes(x=factor(month))) +
  geom_bar(aes(fill=factor(departure_airport)))
```



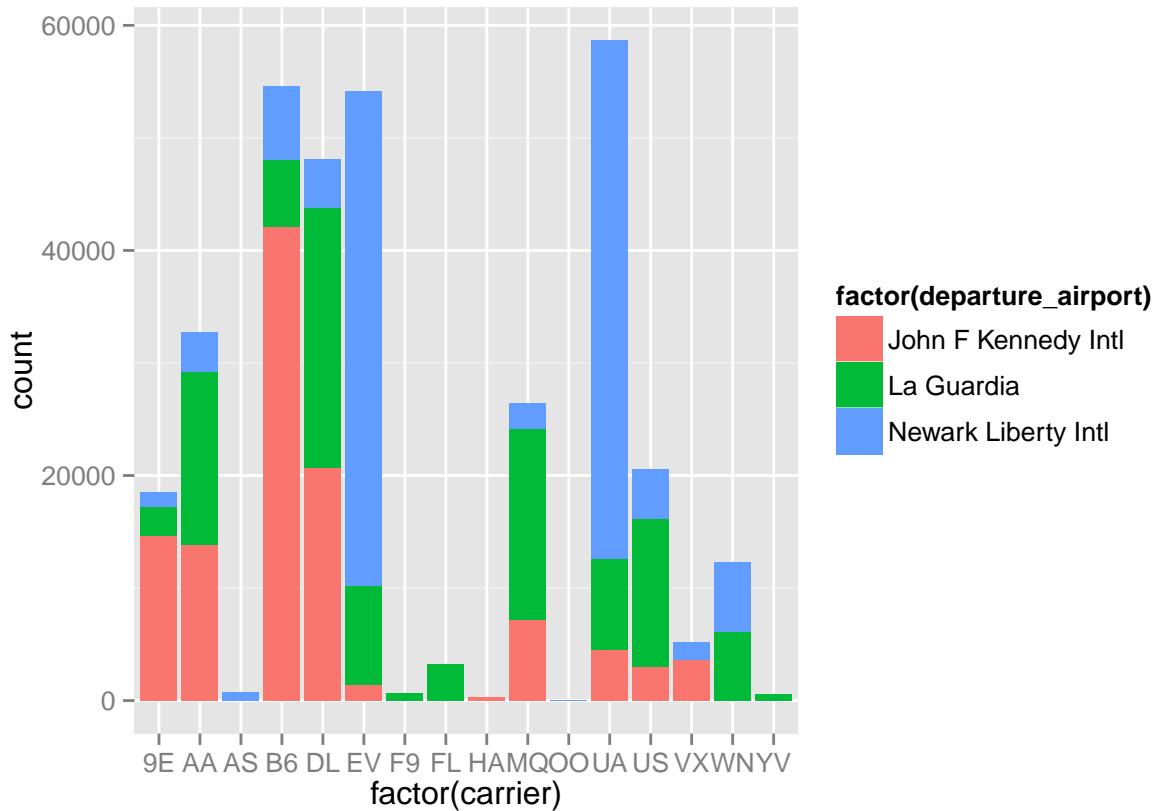
```
ggplot(data=flt1,
       aes(x=factor(month))) +
  geom_bar(aes(fill=factor(carrier_name)))
```



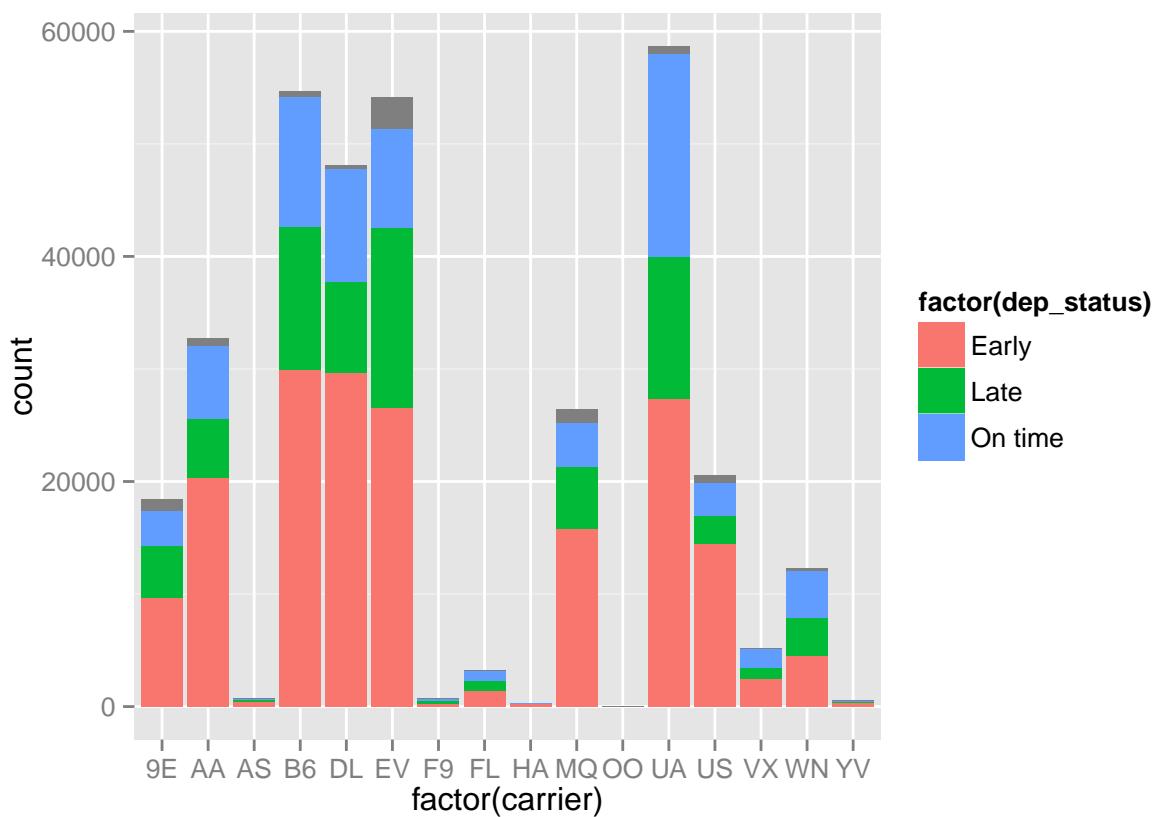
```
ggplot(data=flt1,
       aes(x=factor(carrier))) +
  geom_bar(aes(fill=factor(departure_airport)))
```



```
ggplot(data=flt1,
       aes(x=factor(carrier))) +
  geom_bar(aes(fill=factor(departure_airport)))
```

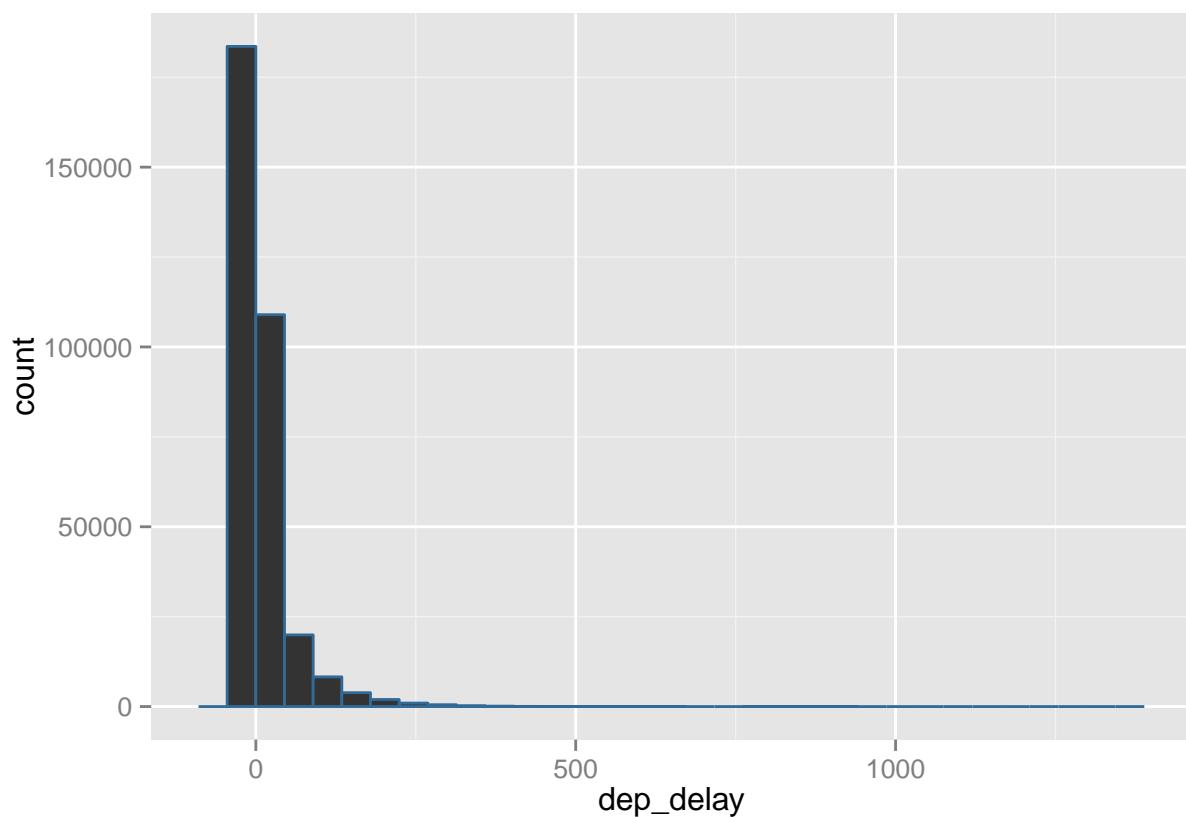


```
ggplot(data=flt1,
       aes(x=factor(carrier))) +
  geom_bar(aes(fill=factor(dep_status)))
```



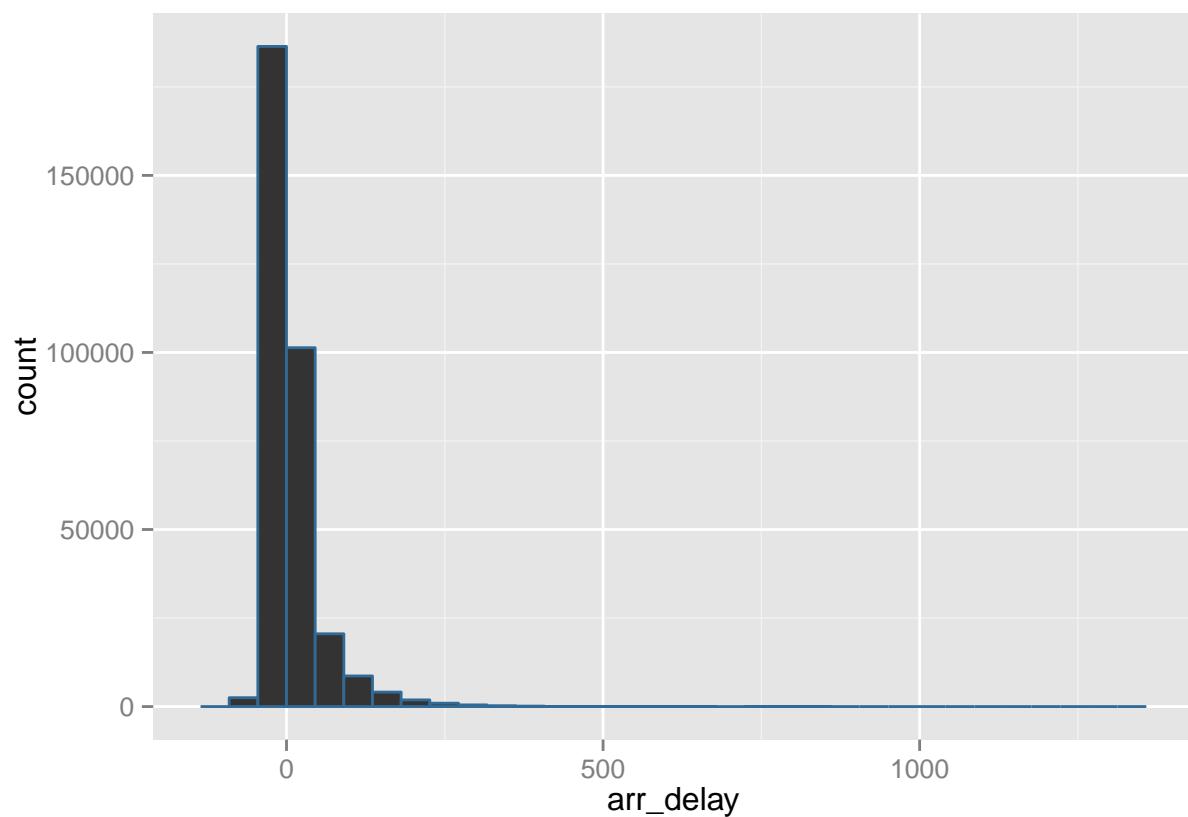
```
ggplot(data=flt1, aes(x=dep_delay)) +
  geom_bar(aes(color=1)) +
  guides(color=FALSE)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

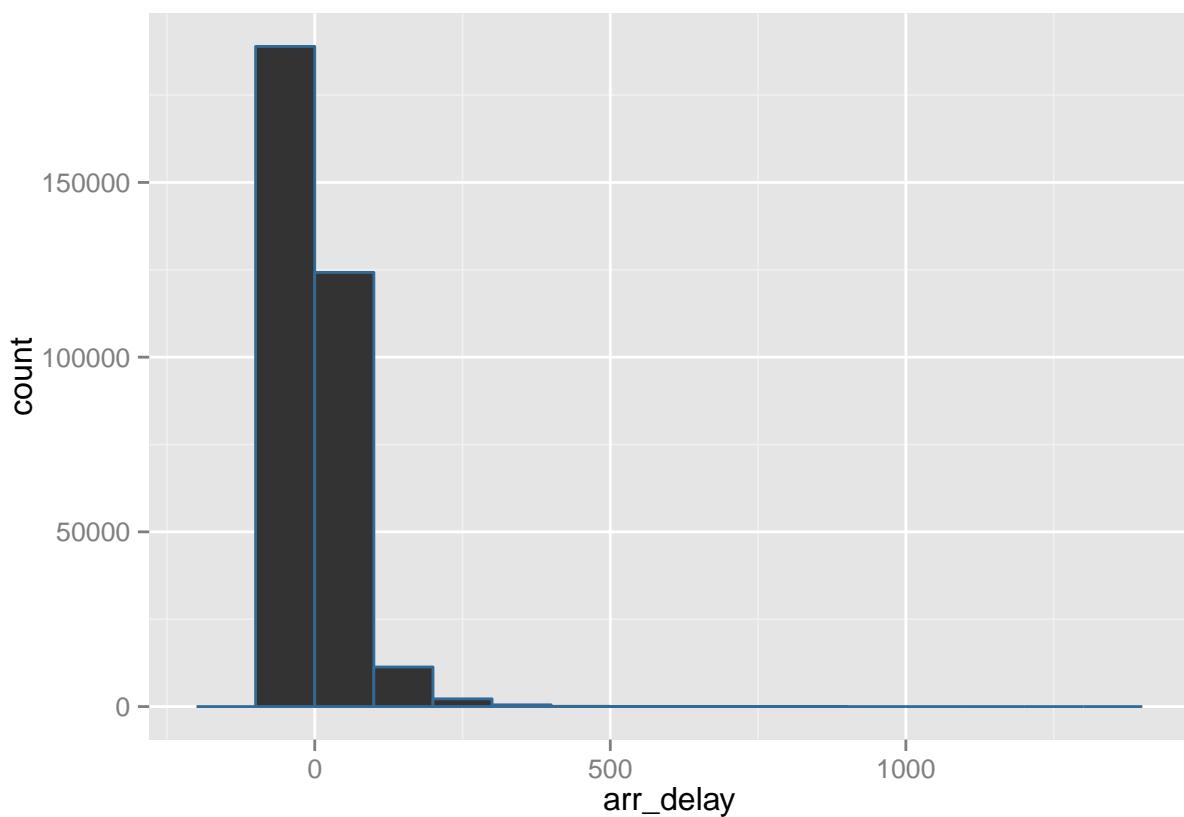


```
ggplot(data=flt1, aes(x=arr_delay)) +  
  geom_bar(aes(color=1)) +  
  guides(color=FALSE)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

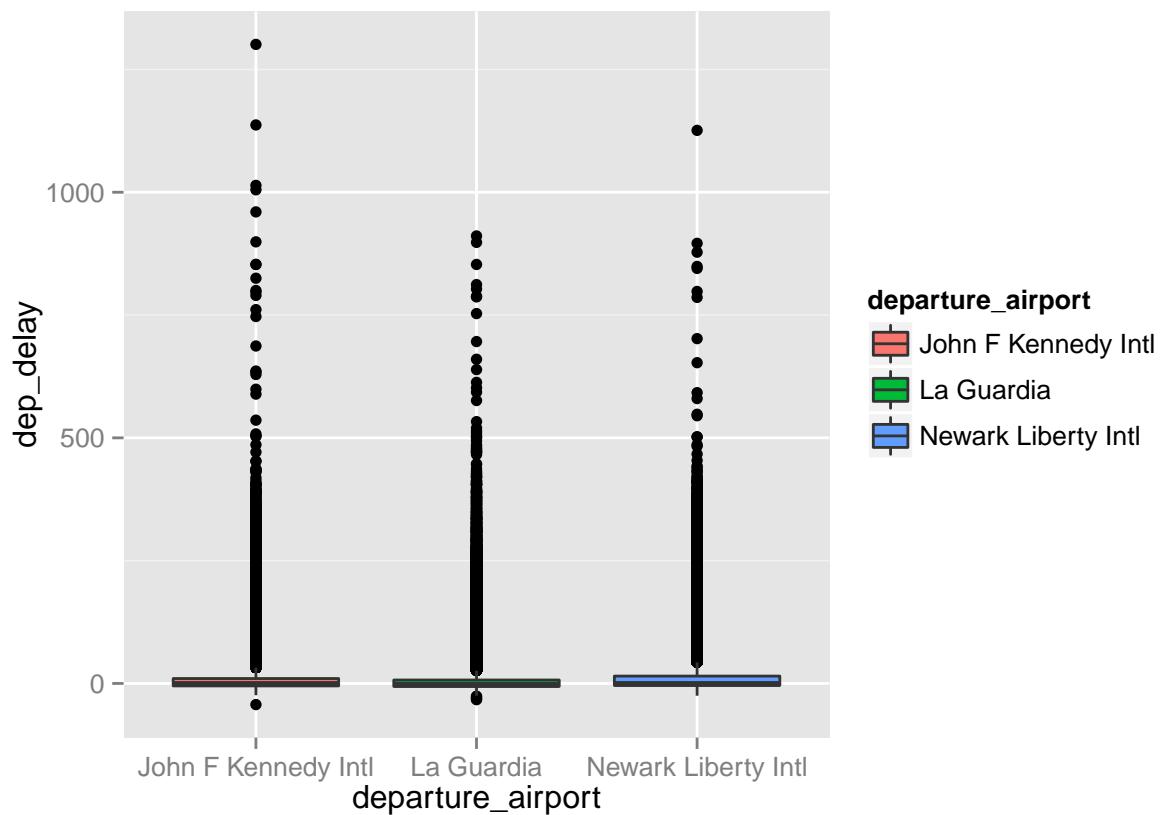


```
ggplot(data=flt1, aes(x=arr_delay)) +  
  geom_bar(aes(color=1),  
           binwidth =100) +  
  guides(color=FALSE)
```



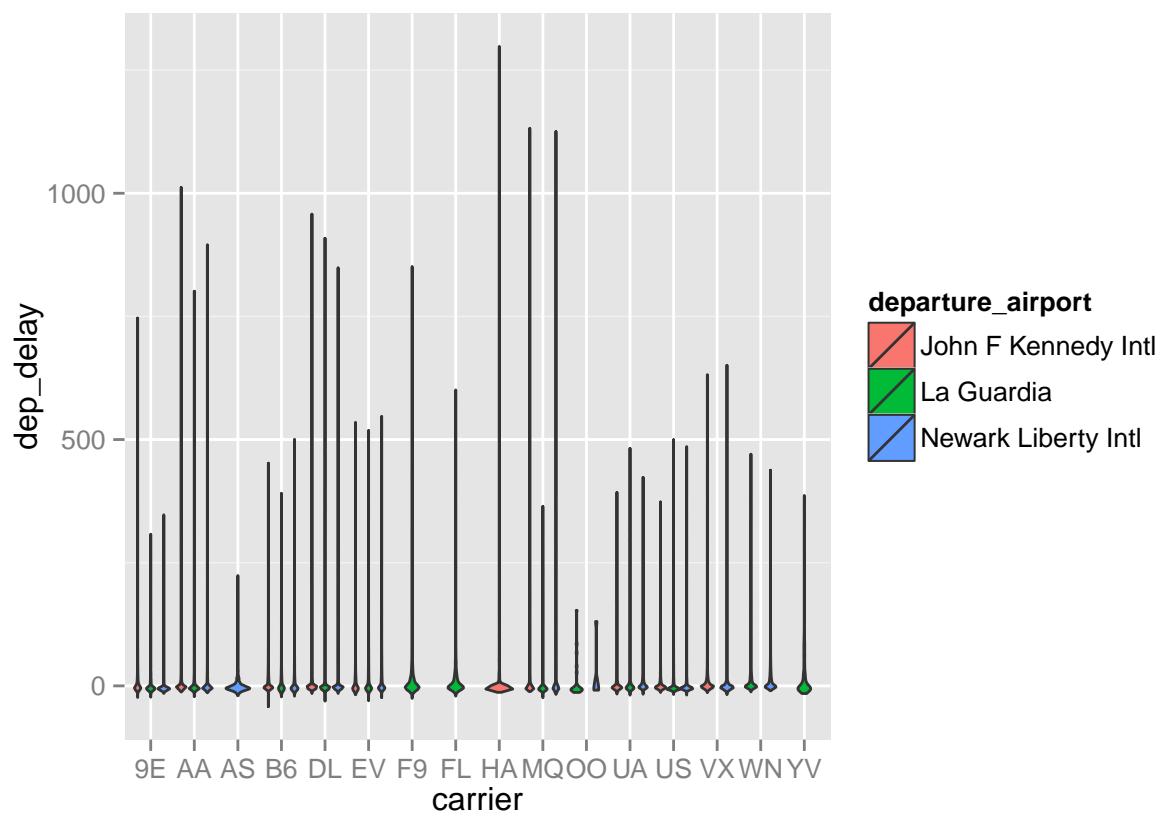
```
ggplot(data=flt1, aes(x=departure_airport,  
                      y=dep_delay)) +  
  geom_boxplot(aes(fill=departure_airport))
```

```
## Warning: Removed 8255 rows containing non-finite values (stat_boxplot).
```

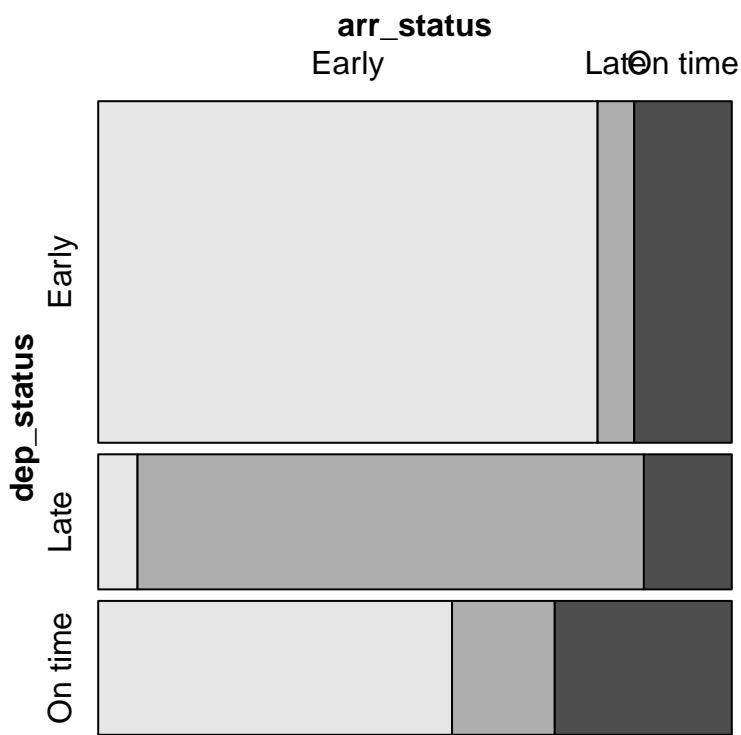


```
ggplot(data=flt1, aes(x=carrier,
                      y=dep_delay)) +
  geom_violin(aes(fill=departure_airport),
              scale='area')
```

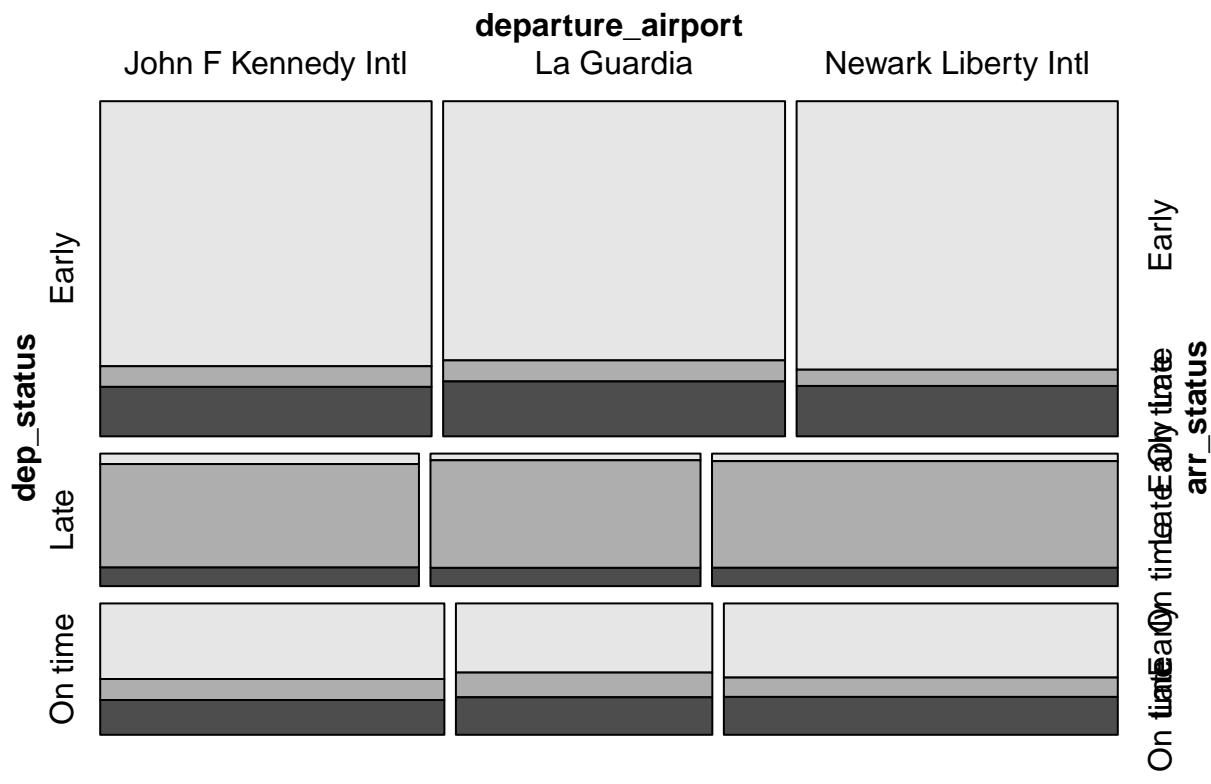
```
## Warning: Removed 8255 rows containing non-finite values (stat_ydensity).
```



```
mosaic(formula= arr_status ~ dep_status ,data=flt1)
```



```
mosaic(formula= arr_status ~ dep_status + departure_airport,data=flt1)
```



Conclusion