```
library(mosaic)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggstance
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
##
## New to ggformula?  Try the tutorials:
##   learnr::run_tutorial("introduction", package = "ggformula")
##   learnr::run_tutorial("refining", package = "ggformula")
```

```
## Loading required package: mosaicData
```

```
## Loading required package: Matrix
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
## 
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected
by this.
## 
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.
```

```
## 
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
## 
##     mean
```

```
## The following object is masked from 'package:ggplot2':
## 
##     stat
```

```
## The following objects are masked from 'package:dplyr':
## 
##     count, do, tally
```

```
## The following objects are masked from 'package:stats':
## 
##     binom.test, cor, cor.test, cov, fivenum, IQR, median,
##     prop.test, quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
## 
##     max, mean, min, prod, range, sample, sum
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────
─────────────── tidyverse 1.2.1 ──
```

```
## ✔ tibble  2.1.3     ✔ purrr   0.3.2
## ✔ tidyr   0.8.3     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ─────────────────────────────────────────────────────
────────── tidyverse_conflicts() ──
## ✖ mosaic::count()         masks dplyr::count()
## ✖ purrr::cross()          masks mosaic::cross()
## ✖ mosaic::do()            masks dplyr::do()
## ✖ tidyr::expand()         masks Matrix::expand()
## ✖ dplyr::filter()         masks stats::filter()
## ✖ ggstance::geom_errorbarh() masks ggplot2::geom_errorbarh()
## ✖ dplyr::lag()            masks stats::lag()
## ✖ mosaic::stat()          masks ggplot2::stat()
## ✖ mosaic::tally()         masks dplyr::tally()
```

```r
library(ggplot2)
library(ggthemes)
```

```
##
## Attaching package: 'ggthemes'
```

```
## The following object is masked from 'package:mosaic':
##
##     theme_map
```

```r
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
flight = read.csv('ABIA.csv')
attach(flight)
names(flight)
```

```
##  [1] "Year"               "Month"              "DayofMonth"
##  [4] "DayOfWeek"          "DepTime"            "CRSDepTime"
##  [7] "ArrTime"            "CRSArrTime"         "UniqueCarrier"
## [10] "FlightNum"          "TailNum"            "ActualElapsedTime"
## [13] "CRSElapsedTime"     "AirTime"            "ArrDelay"
## [16] "DepDelay"           "Origin"             "Dest"
## [19] "Distance"           "TaxiIn"             "TaxiOut"
## [22] "Cancelled"          "CancellationCode"   "Diverted"
## [25] "CarrierDelay"       "WeatherDelay"       "NASDelay"
## [28] "SecurityDelay"      "LateAircraftDelay"
```
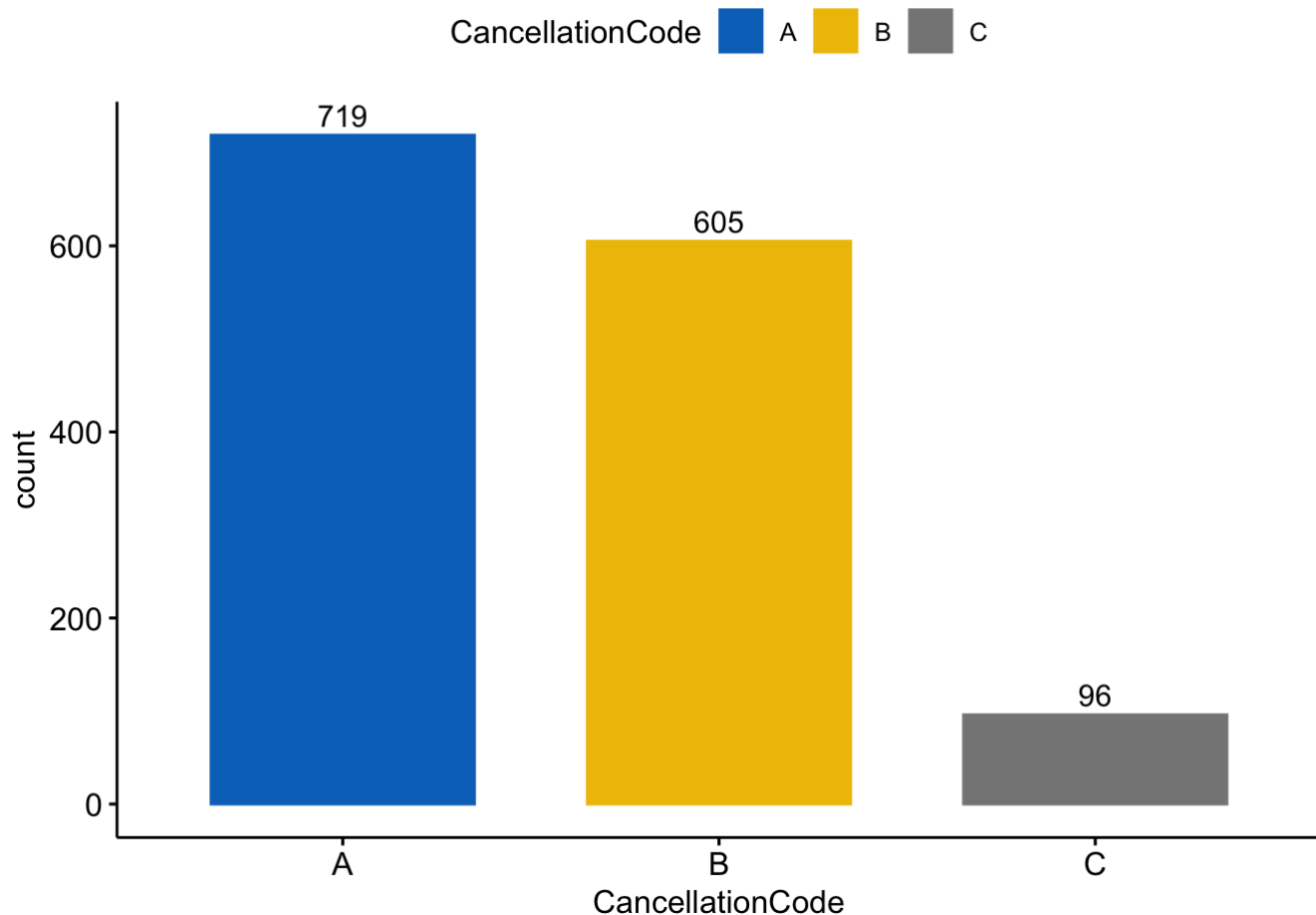
```
cancel = subset(flight, Cancelled == 1)
```

```
table1 = xtabs(~Cancelled + CancellationCode, data = cancel)

a1 = cancel %>%
  group_by(CancellationCode) %>%
  summarize(count = sum(Cancelled))
a1
```

```
## # A tibble: 3 x 2
##   CancellationCode count
##   <fct>            <int>
## 1 A                  719
## 2 B                  605
## 3 C                   96
```

```
ggbarplot(a1, x = 'CancellationCode', y = 'count',
    fill = 'CancellationCode', color = 'CancellationCode', palette = "jco", label = round
(a1$count))
```

There are a totle of 1420 cancellations. 719 of them are type A cancellations which is cancellations related to the unique carriers,. 605 of them are cancellations due to wheater (Type B), and 96 of them are NAS(Type C) which are delays or cancellations within the control of National Airspace System.

```
b1 = cancel %>%
  group_by(DayOfWeek) %>%
  summarize(total = sum(Cancelled))
b1
```

```
## # A tibble: 7 x 2
##   DayOfWeek total
##       <int> <int>
## 1         1   183
## 2         2   289
## 3         3   224
## 4         4   218
## 5         5   199
## 6         6   151
## 7         7   156
```

```
ggbarplot(b1, x = 'DayOfWeek', y = 'total',
   fill ='DayOfWeek', color = 'DayOfWeek', label = round(b1$total))
```
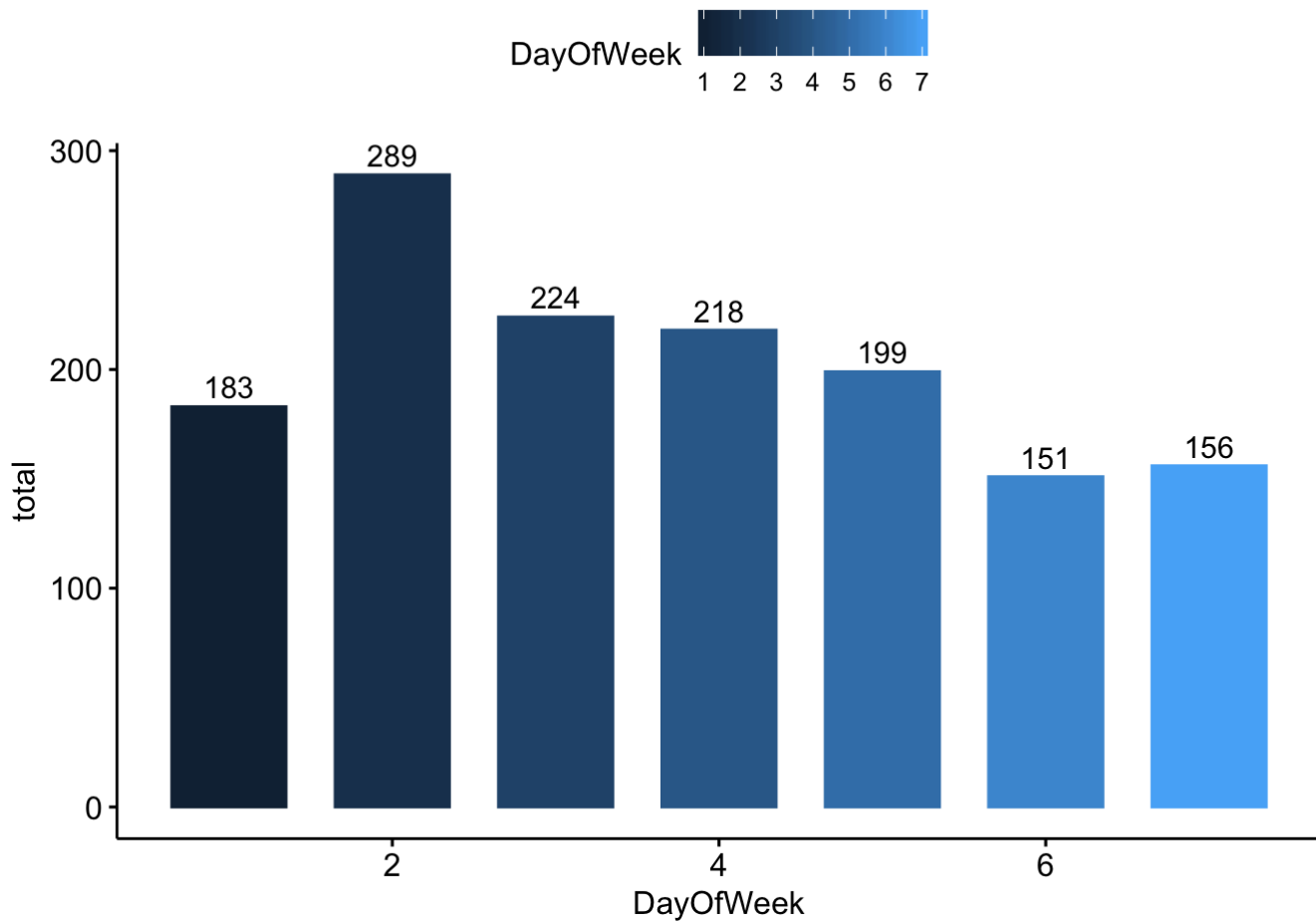
Next we plotted cancellations by days of week and found that cancellations mostly happen on Tuesdays.

```
e1 = cancel %>%
  group_by(Origin) %>%
  summarize(count = sum(Cancelled))

subset(e1, Origin == 'AUS')
```

```
## # A tibble: 1 x 2
##   Origin count
##   <fct>  <int>
## 1 AUS      732
```

```
c1 = cancel %>%
  group_by(Origin, CancellationCode) %>%
  summarize(count = sum(Cancelled))
c1
```

```
## # A tibble: 71 x 3
## # Groups:   Origin [37]
##    Origin CancellationCode count
##    <fct>  <fct>           <int>
##  1 ABQ    A                   4
##  2 ABQ    B                   2
##  3 ATL    A                  17
##  4 ATL    B                  15
##  5 ATL    C                   2
##  6 AUS    A                 390
##  7 AUS    B                 294
##  8 AUS    C                  48
##  9 BOS    A                   1
## 10 BOS    B                   1
## # … with 61 more rows
```
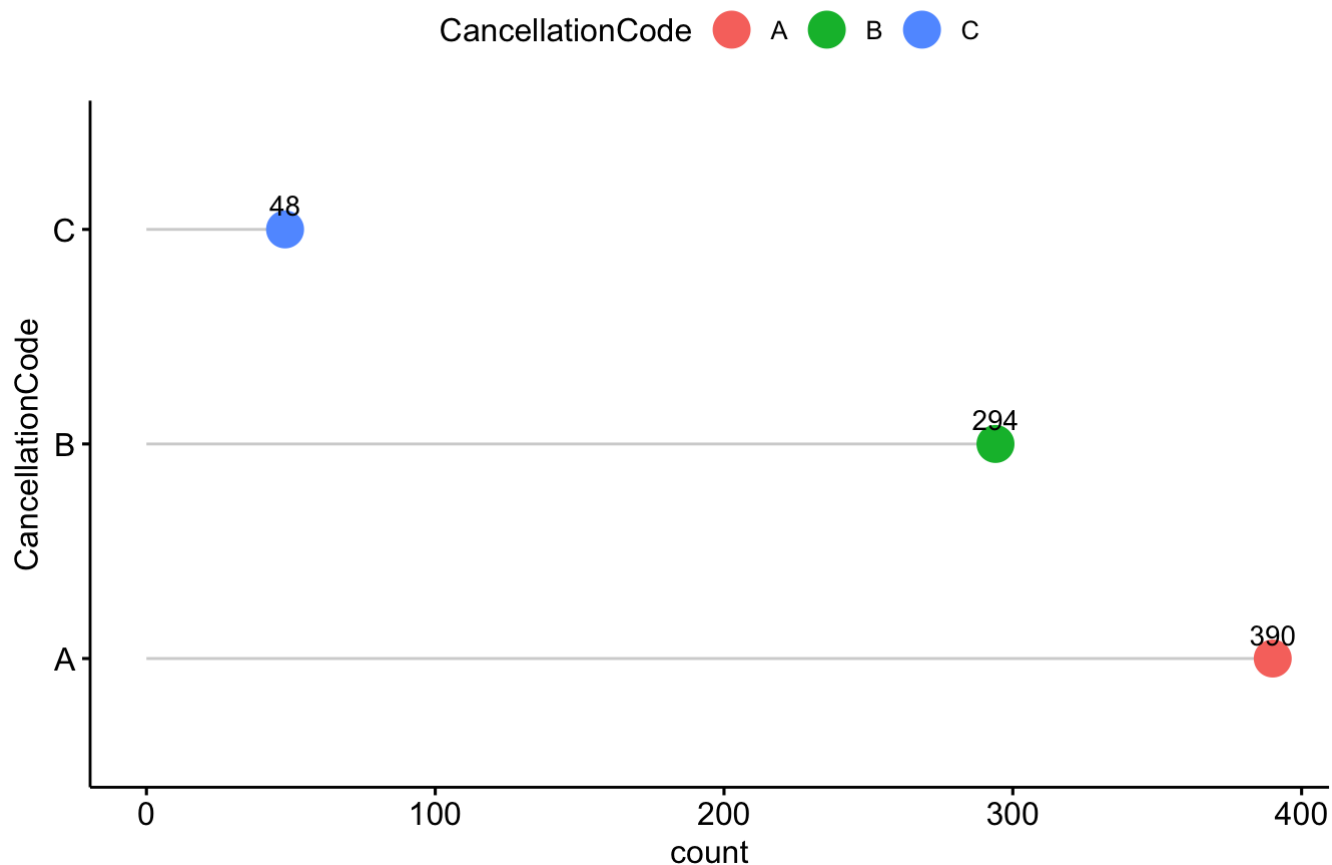
```
Austin = subset(c1, Origin == 'AUS')
Austin
```

```
## # A tibble: 3 x 3
## # Groups:   Origin [1]
##   Origin CancellationCode count
##   <fct>  <fct>           <int>
## 1 AUS    A                 390
## 2 AUS    B                 294
## 3 AUS    C                  48
```

```
ggdotchart(Austin, x = "CancellationCode", y = "count",
          color = "CancellationCode", add = 'segment', dot.size = 6, label = round(Aust
in$count),
          rotate = TRUE, title = 'Austin Airport Flight Cancellations')
```

Austin Airport Flight Cancellations

CancellationCode ● A ● B ● C

Since we are looking at datasets in the Austin Airport. We are also interested in looking at the amount of flights that got cancelled departing from the Austin Airport. The result shows that Austin Airport has a total of 732 cancellations and the most common cancellation type is Type A which is controlled by carrier.

```
a = sort(table(flight$Origin), decreasing = TRUE)
b =  sort(table(cancel$Origin), decreasing = TRUE)

all_air = data.frame(a)
cancel_air = data.frame(b)
merged = merge(all_air, cancel_air, by = "Var1")
merged$percentdec = merged$Freq.y/merged$Freq.x

merged$percent = (merged$percentdec)*100


merged1 = subset(merged, Freq.y != 0)
merged1
```

```
##      Var1 Freq.x Freq.y    percentdec    percent
## 1    ABQ     433      6 0.0138568129 1.38568129
## 2    ATL    2255     34 0.0150776053 1.50776053
## 3    AUS   49623    732 0.0147512242 1.47512242
## 6    BOS     368      2 0.0054347826 0.54347826
## 8    CLE     380      5 0.0131578947 1.31578947
## 9    CLT     660      2 0.0030303030 0.30303030
## 10   CVG     653      3 0.0045941807 0.45941807
## 11   DAL    5583    115 0.0205982447 2.05982447
## 12   DEN    2719      7 0.0025744759 0.25744759
## 13   DFW    5508    159 0.0288671024 2.88671024
## 14   ELP    1344      1 0.0007440476 0.07440476
## 15   EWR     939      9 0.0095846645 0.95846645
## 16   FLL     481      1 0.0020790021 0.20790021
## 17   HOU    2310     39 0.0168831169 1.68831169
## 18   HRL     366      7 0.0191256831 1.91256831
## 19   IAD     631     10 0.0158478605 1.58478605
## 20   IAH    3704     51 0.0137688985 1.37688985
## 21   IND     218      3 0.0137614679 1.37614679
## 22   JAX     229      2 0.0087336245 0.87336245
## 23   JFK    1356     33 0.0243362832 2.43362832
## 24   LAS    1232      3 0.0024350649 0.24350649
## 25   LAX    1732     12 0.0069284065 0.69284065
## 27   LGB     245      2 0.0081632653 0.81632653
## 28   MAF     471      2 0.0042462845 0.42462845
## 29   MCI     459     12 0.0261437908 2.61437908
## 31   MDW     713      4 0.0056100982 0.56100982
## 32   MEM     835     19 0.0227544910 2.27544910
## 34   MSY     443      2 0.0045146727 0.45146727
## 38   ORD    2515     90 0.0357852883 3.57852883
## 40   PHX    2786      7 0.0025125628 0.25125628
## 41   RDU     231      2 0.0086580087 0.86580087
## 44   SEA     147      2 0.0136054422 1.36054422
## 45   SFO     609      4 0.0065681445 0.65681445
## 46   SJC     968     30 0.0309917355 3.09917355
## 47   SLC     550      1 0.0018181818 0.18181818
## 49   STL      95      6 0.0631578947 6.31578947
## 51   TUL      90      1 0.0111111111 1.11111111
```

```
merged1 = merged1[order(merged1$percent),]
merged1$Var1=factor(merged1$Var1,levels=merged1$Var1)
ggplot(data = merged1) +
  geom_bar(mapping = aes(x=factor(Var1), y=percent), stat='identity', ascending=FALSE) +
  coord_flip()
```

```
## Warning: Ignoring unknown parameters: ascending
```
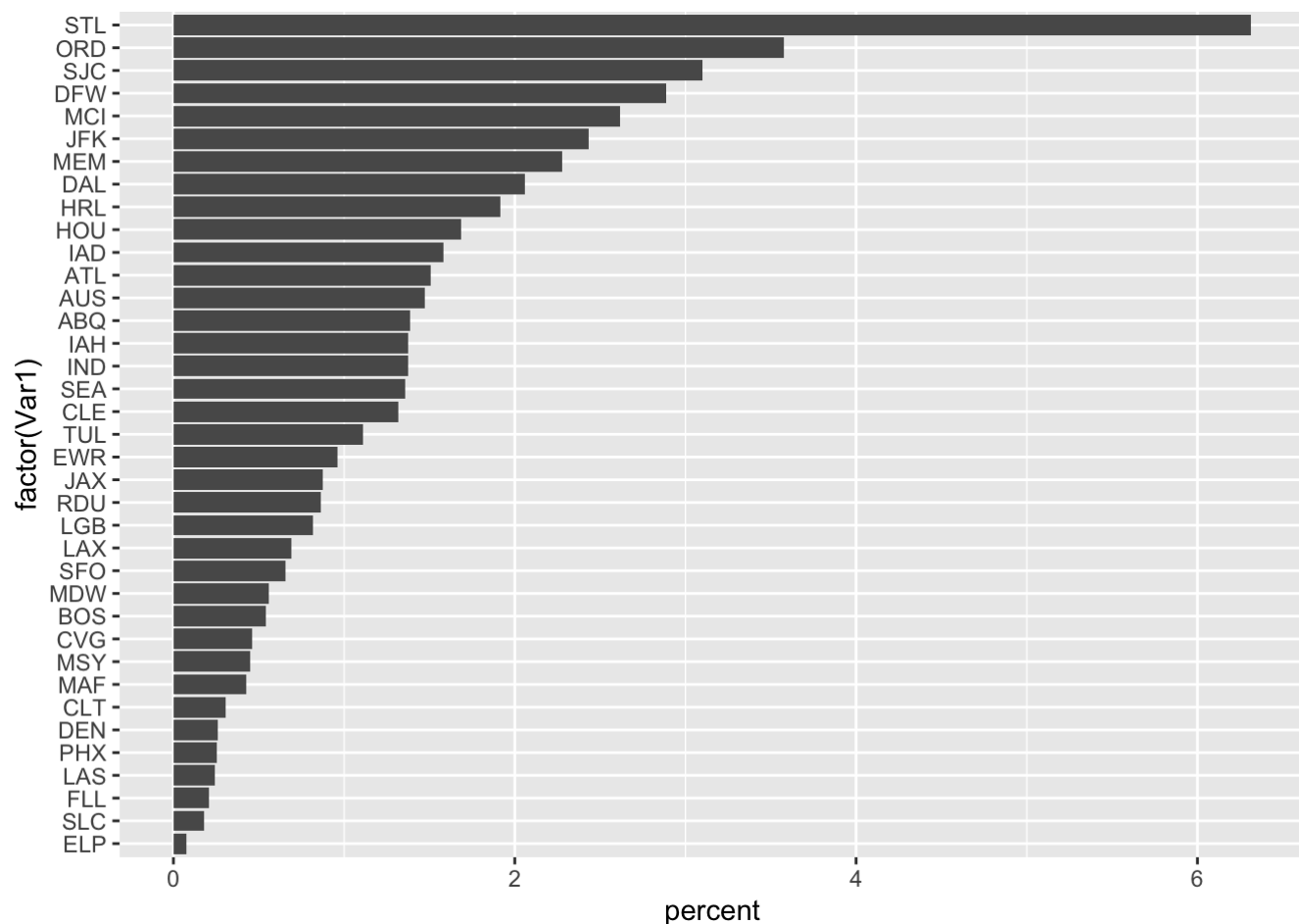
Finally, we are interested in looking at how Austin Airport's cancellations is compare to other airports. We normalized the number of cancellations into percentage of cancellation out of all flights in each airport, and we graphed the result. In turns out that St. Louis Airport is most frequent in canceling its flights with 6.32% cancelled flights out of all flights. Austin is actually ranked number 12th with 1.48% cancelled flights out of all flights.