EDA, Titanic training data

Brooke Anderson

December 14, 2015

Load some packages I'll be using:

```r
library(ggplot2)
library(ggthemes)
library(stringr)
library(dplyr)
library(stats)
library(tidyr)
```

Load the Titanic training data (I have it in a `data` directory in the parent directory for this file):

```r
train <- read.csv("../data/train.csv")
```

There are 891 observations.

```r
nrow(train)
```

```
## [1] 891
```

Each observation is a passenger on the Titanic. The features for each passenger are:

```r
colnames(train)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

The `PassengerId` is a unique identifier for each passenger.

```r
head(train$PassengerId)
```

```
## [1] 1 2 3 4 5 6
```

```r
length(unique(train$PassengerId)) == nrow(train)  # Check for duplicates
```

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

`Survived` is a binary variable of whether the passenger survived (`1`) or died (`0`). In the training data, about 38% of the passengers survived.

```r
table(train$Survived)
```

```
##
##   0   1
## 549 342
```

```
round(100 * prop.table(table(train$Survived)))
```

```
##
##  0  1
## 62 38
```

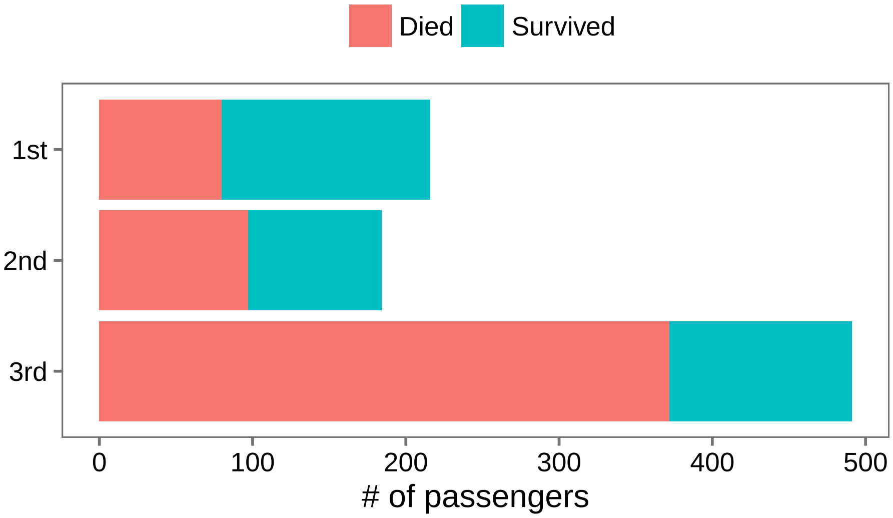
Pclass gives the passenger's ticket class. There are three options: 1st, 2nd, and 3rd class:

```
table(train$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

More of the passengers in train were in 3rd class than 1st or 2nd class. Most of the passengers in 3rd class died, most in the 1st class survived, and about an even number in the 2nd class died and survived.

```
ggplot(train, aes(x = factor(Pclass, levels = c(1, 2, 3),
                             labels = c("1st", "2nd", "3rd")),
                  fill = factor(Survived, levels = c(0, 1),
                                labels = c("Died", "Survived")))) +
        geom_bar() +
        coord_flip() +
        scale_x_discrete("",
                         limits=c("3rd","2nd","1st")) +
        ylab("# of passengers") +
        theme_few() + #  Uses `ggtheme` package
        theme(legend.title = element_blank(),
              legend.position = "top")
```



Name gives the passenger's name:

```
train$Name <- as.character(train$Name) # No reason for these to be factors
sample(train$Name, 5)
```

```
## [1] "Cacic, Mr. Luka"
## [2] "Andersson, Miss. Ebba Iris Alfrida"
## [3] "Henry, Miss. Delia"
## [4] "Heininen, Miss. Wendla Maria"
## [5] "Tomlin, Mr. Ernest Portage"
```

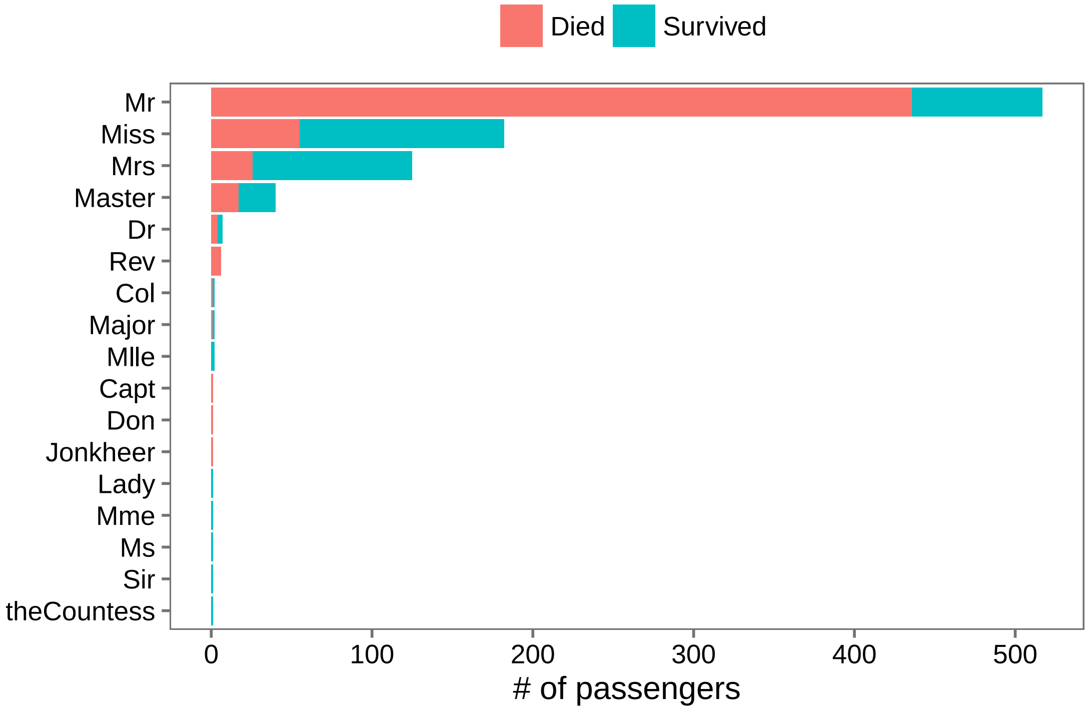
You can pull more out of this variable. For example, you can pull out each passenger's honorific and create a new column in `train` with that.

```
honorific <- str_extract(train$Name, ",\\ .+?\\.") # Uses `stringr` package
honorific <- gsub("[\\,\\.\\ ]", "", honorific)
train <- cbind(train, honorific)
```

```
(hon_count <- group_by(train, honorific) %>%  # Uses `dplyr` package
        summarize(n = n()) %>%
        arrange(desc(n)))
```

```
## Source: local data frame [17 x 2]
##
##      honorific     n
##         (fctr) (int)
## 1           Mr   517
## 2         Miss   182
## 3          Mrs   125
## 4       Master    40
## 5           Dr     7
## 6          Rev     6
## 7          Col     2
## 8        Major     2
## 9         Mlle     2
## 10        Capt     1
## 11         Don     1
## 12    Jonkheer     1
## 13        Lady     1
## 14         Mme     1
## 15          Ms     1
## 16         Sir     1
## 17 theCountess     1
```

```
ggplot(train, aes(x = factor(honorific, levels = rev(hon_count$honorific)),
                  fill = factor(Survived, levels = c(0, 1),
                                labels = c("Died", "Survived")))) +
        geom_bar() +
        xlab("") +
        ylab("# of passengers") +
        coord_flip() +
        theme_few() +
        theme(legend.title = element_blank(),
              legend.position = "top")
```



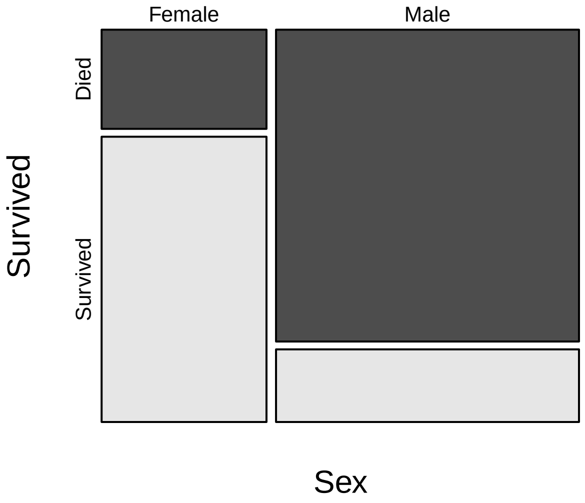
`Sex` gives the passenger's sex. In this `train` dataset, about two-thirds of passengers were male.

```
round(100 * prop.table(table(train$Sex)))
```

```
##
## female   male
##     35     65
```

Here's a mosaic plot of the distribution of survival by sex for the passengers in `train`:

```
train2 <- mutate(train,
                Sex = factor(Sex, levels = c("female", "male"),
                             labels = c("Female", "Male")),
                Survived = factor(Survived, levels = c(0, 1),
                                  labels = c("Died", "Survived")))
mosaicplot(~ Sex + Survived, data = train2, color = TRUE,
           main = "")
```



Just for fun, here are the honorifics by sex. It looks like there was a female doctor on board:

```r
table(train$honorific, train$Sex)
```

```
##
##                female male
##   Capt             0    1
##   Col              0    2
##   Don              0    1
##   Dr               1    6
##   Jonkheer         0    1
##   Lady             1    0
##   Major            0    2
##   Master           0   40
##   Miss           182    0
##   Mlle             2    0
##   Mme              1    0
##   Mr               0  517
##   Mrs            125    0
##   Ms               1    0
##   Rev              0    6
##   Sir              0    1
##   theCountess      1    0
```

```r
train[train$honorific == "Dr" & train$Sex == "female", ]
```

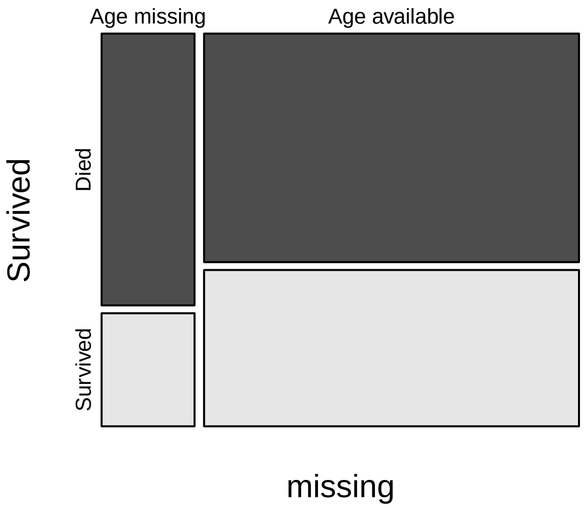
```
##     PassengerId Survived Pclass                        Name    Sex Age
## 797         797        1      1 Leader, Dr. Alice (Farnham) female  49
##     SibSp Parch Ticket    Fare Cabin Embarked honorific
## 797     0     0  17465 25.9292   D17        S        Dr
```

`Age` is the passenger’s age. For about 20% of passengers, this is missing. It looks like a higher percentage of passengers that had age data available survived compared to passengers missing age.

```r
prop.table(table(is.na(train$Age)))
```

```
##
##     FALSE      TRUE
## 0.8013468 0.1986532
```

```
train2 <- mutate(train,
                 Survived = factor(Survived, levels = c(0, 1),
                                   labels = c("Died", "Survived")),
                 missing = factor(is.na(Age), levels = c(TRUE, FALSE),
                                  labels = c("Age missing", "Age available")))
mosaicplot(~ missing + Survived, data = train2, color = TRUE,
           main = "")
```



For passengers with age data available, there was a large range of ages.

```
range(train$Age, na.rm = TRUE)
```

```
## [1]  0.42 80.00
```

For children below 1, it looks like age was given in months (which was then converted to a fraction).

```
filter(train, Age < 1) %>%
        select(Age, Name, Survived, Pclass) %>%
        arrange(Age) %>%
        mutate(months = round(Age * 12))
```

```
##    Age                            Name Survived Pclass months
## 1 0.42 Thomas, Master. Assad Alexander        1      3      5
## 2 0.67       Hamalainen, Master. Viljo        1      2      8
## 3 0.75   Baclini, Miss. Helene Barbara        1      3      9
## 4 0.75          Baclini, Miss. Eugenie        1      3      9
## 5 0.83   Caldwell, Master. Alden Gates        1      2     10
## 6 0.83 Richards, Master. George Sibley        1      2     10
## 7 0.92  Allison, Master. Hudson Trevor        1      1     11
```

For passengers above 1, for the most part, it looks like Age was always given as a whole number, with half years (.5) occasionally included.

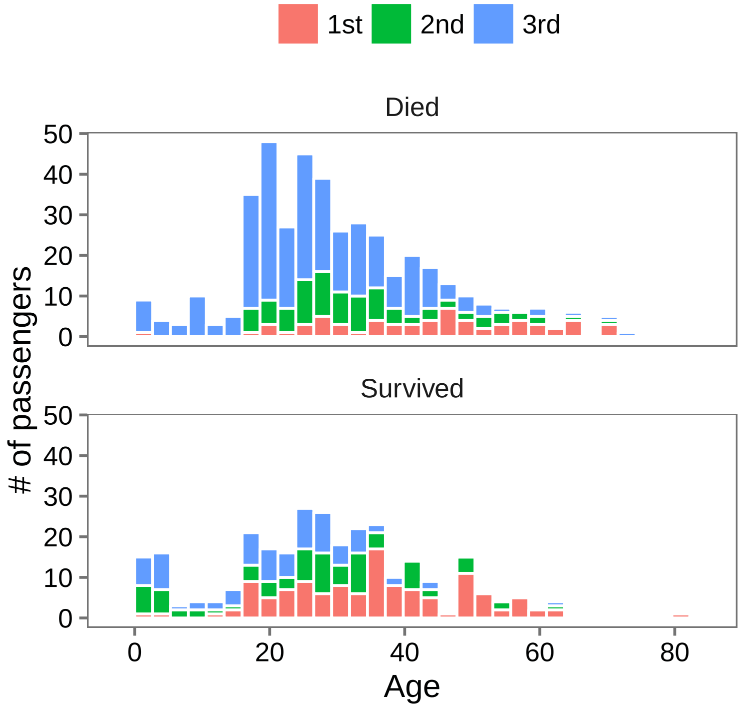
```
sample(unique(train$Age), 20)
```

```
##  [1]  0.42 12.00  1.00 23.00 66.00 51.00 30.50 26.00 64.00 29.00 17.00
## [12] 70.00 46.00 25.00  9.00 10.00 45.00 32.00 45.50 21.00
```

There was a pretty big break in passenger ages between adults (around 18, say) and children. While there were some young children, teenagers seemed pretty rare. There were particularly few children in the 1st class. Children were more generally more likely to survive, especially if they were in the 1st or second class.

```
train2 <- mutate(train,
                 Survived = factor(Survived, levels = c(0, 1),
                                   labels = c("Died", "Survived")),
                 Pclass = factor(Pclass, levels = c(1, 2, 3),
                                 labels = c("1st", "2nd", "3rd")))
ggplot(train2, aes(x = Age, fill = Pclass)) +
        geom_histogram(color = "white", position = "stack") +
        ylab("# of passengers") +
        theme_few() +
        facet_wrap(~ Survived, ncol = 1) +
        theme(legend.title = element_blank(),
              legend.position = "top")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Among adults, it looks like almost everyone over 65 died (although there weren't too many people that old). If you check into this, however, it seems to be an error. It looks like this guy was 80 when he died, but that wasn't until 1945. He was actually 45 when he was on the Titanic.

```
filter(train, Age >= 65) %>%
        select(Age, Survived, Pclass, Name) %>%
        arrange(Age)
```

```
##     Age Survived Pclass                           Name
## 1  65.0        0      1       Ostby, Mr. Engelhart Cornelius
## 2  65.0        0      3                   Duane, Mr. Frank
## 3  65.0        0      1            Millet, Mr. Francis Davis
## 4  66.0        0      2                Wheadon, Mr. Edward H
## 5  70.0        0      2           Mitchell, Mr. Henry Michael
## 6  70.0        0      1          Crosby, Capt. Edward Gifford
## 7  70.5        0      3                 Connors, Mr. Patrick
## 8  71.0        0      1            Goldschmidt, Mr. George B
## 9  71.0        0      1              Artagaveytia, Mr. Ramon
## 10 74.0        0      3                  Svensson, Mr. Johan
## 11 80.0        1      1 Barkworth, Mr. Algernon Henry Wilson
```

For children, it looks like several often shared the same last name (and so might have been siblings):

```
head(filter(train, Age < 16) %>%
        select(Age, Name, Survived) %>%
        arrange(Name), 20)
```

```
##      Age                                   Name Survived
## 1   0.92          Allison, Master. Hudson Trevor        1
## 2   2.00            Allison, Miss. Helen Loraine        0
## 3   4.00 Andersson, Master. Sigvard Harald Elias        0
## 4   6.00       Andersson, Miss. Ebba Iris Alfrida        0
## 5   2.00        Andersson, Miss. Ellis Anna Maria        0
## 6   9.00     Andersson, Miss. Ingeborg Constanzia        0
## 7  11.00        Andersson, Miss. Sigrid Elisabeth        0
## 8   9.00     Asplund, Master. Clarence Gustaf Hugo        0
## 9   3.00          Asplund, Master. Edvin Rojj Felix        1
## 10  5.00            Asplund, Miss. Lillian Gertrud        1
## 11 13.00                      Ayoub, Miss. Banoura        1
## 12  0.75                    Baclini, Miss. Eugenie        1
## 13  0.75              Baclini, Miss. Helene Barbara        1
## 14  5.00             Baclini, Miss. Marie Catherine        1
## 15  1.00                 Becker, Master. Richard F        1
## 16  4.00               Becker, Miss. Marion Louise        1
## 17  9.00                   Boulos, Miss. Nourelain        0
## 18  0.83              Caldwell, Master. Alden Gates        1
## 19 11.00        Carter, Master. William Thornton II        1
## 20 14.00                  Carter, Miss. Lucile Polk        1
```

For children under 16, it looks like siblings were definitely not independent in terms of their survival. First, siblings were pretty likely to all share the same survival status. Second, families with lots of children were likely to not have any survivors. None of the children in last name groups of four or more children, for example, survived (at least based on this measure of siblings).

```
train$last_name <- gsub(",.*", "", train$Name)
sample(train$last_name, 20)
```

```
##  [1] "Samaan"    "Goodwin"   "Jarvis"    "Salonen"   "Keefe"    
##  [6] "Ahlin"     "Hegarty"   "Andersson" "Panula"    "Haas"     
## [11] "Goodwin"   "Elias"     "Persson"   "Lines"     "Becker"   
## [16] "Attalah"   "Horgan"    "Lobb"      "Sandstrom" "Hampe"
```

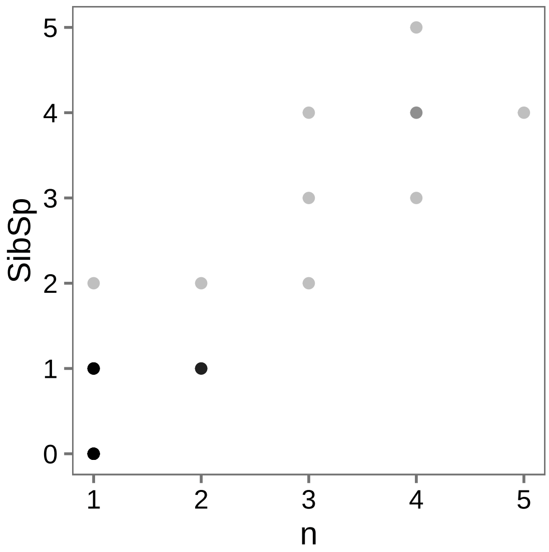
```
children <- filter(train, Age < 16) %>%
        select(last_name, Name, Survived, Pclass, SibSp) %>%
        group_by(last_name) %>%
        summarize(n = n(),
                  SibSp = SibSp[1],
                  Survived = sum(Survived),
                  pSurvived = round(Survived / n, 2),
                  Pclass = Pclass[1]) %>%
        arrange(desc(n), desc(Survived))

filter(children, n > 1)
```

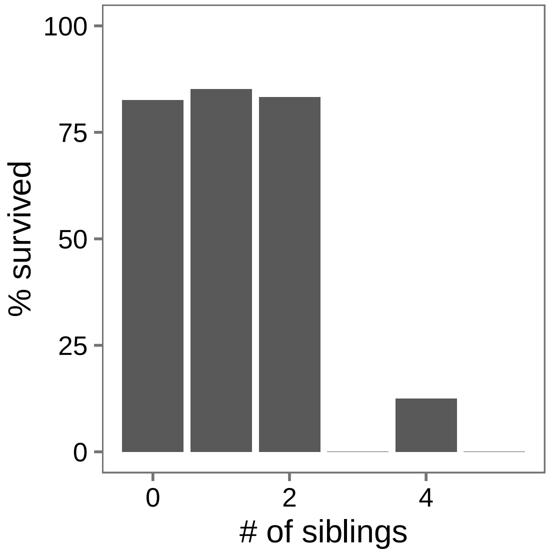
```
## Source: local data frame [16 x 6]
## 
##        last_name     n SibSp Survived pSurvived Pclass
##            (chr) (int) (int)    (int)     (dbl)  (int)
## 1      Andersson     5     4        0      0.00      3
## 2        Goodwin     4     5        0      0.00      3
## 3         Panula     4     4        0      0.00      3
## 4           Rice     4     4        0      0.00      3
## 5          Skoog     4     3        0      0.00      3
## 6        Baclini     3     2        3      1.00      3
## 7        Asplund     3     4        2      0.67      3
## 8        Palsson     3     3        0      0.00      3
## 9         Becker     2     2        2      1.00      2
## 10        Carter     2     1        2      1.00      1
## 11        Coutts     2     1        2      1.00      3
## 12       Johnson     2     1        2      1.00      3
## 13      Navratil     2     1        2      1.00      2
## 14 Nicola-Yarred     2     1        2      1.00      3
## 15      Richards     2     1        2      1.00      2
## 16       Allison     2     1        1      0.50      1
```

Finally, this way of measuring numbers of siblings is pretty well correlated (for children < 16, at least), with the next feature, `SibSp`, which gives the number of siblings and / or spouse aboard. Reassuringly, the metric based on last names always gives an equal or lower number of siblings (some of the siblings will be in the testing data).

```
ggplot(children, aes(x = n, y = SibSp)) +
        geom_point(alpha = .25) +
        theme_few()
```

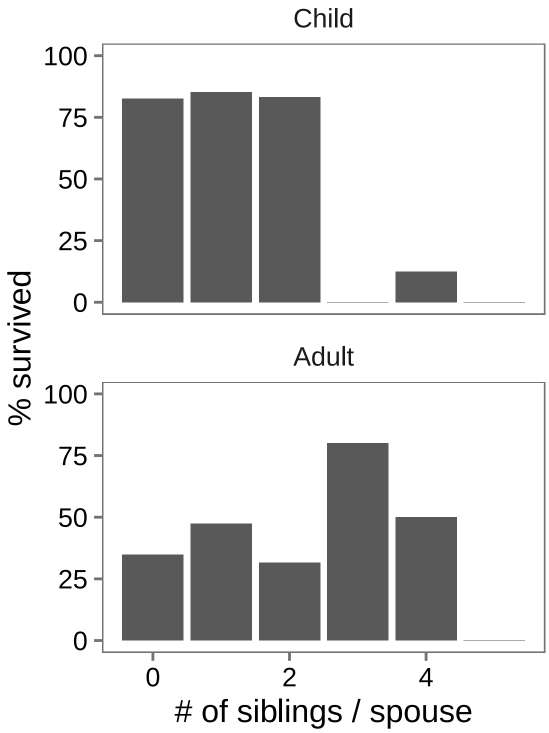


```
children2 <- filter(train, Age < 16) %>%
        group_by(SibSp) %>%
        summarize(n = n(),
                  Survived = sum(Survived),
                  pSurvived = round(100 * (Survived / n), 2))
ggplot(children2, aes(x = SibSp, y = pSurvived)) +
        geom_bar(stat = "identity") +
        ylim(c(0, 100)) +
        xlab("# of siblings") +
        ylab("% survived") +
        theme_few()
```



This pattern differs between children and adults.

```
children2 <- mutate(train,
                    child = factor(Age < 16, levels = c(TRUE, FALSE),
                                   labels = c("Child", "Adult"))) %>%
        filter(!is.na(Age)) %>%
        group_by(SibSp, child) %>%
        summarize(siblings = SibSp[1],
                  n = n(),
                  Survived = sum(Survived),
                  pSurvived = 100 * Survived / n)
ggplot(children2, aes(x = SibSp, y = pSurvived)) +
        facet_wrap(~ child, ncol = 1) +
        geom_bar(stat = "identity") +
        ylim(c(0, 100)) +
        xlab("# of siblings / spouse") +
        ylab("% survived") +
        theme_few()
```



`Parch` gives the number of parents or children that the person has on board. Most people have no parents or children. One person has six (presumably children).

```
table(train$Parch)
```

```
##
##   0   1   2   3   4   5   6
## 678 118  80   5   4   5   1
```

```
train[train$Parch >= 5, c("Name", "Pclass", "Parch", "Survived")]
```

```
##                                                             Name Pclass Parch
## 14                                   Andersson, Mr. Anders Johan      3     5
## 26  Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)      3     5
## 611 Andersson, Mrs. Anders Johan (Alfrida Konstantia Brogren)      3     5
## 639                      Panula, Mrs. Juha (Maria Emilia Ojala)      3     5
## 679                   Goodwin, Mrs. Frederick (Augusta Tyler)      3     6
## 886                      Rice, Mrs. William (Margaret Norton)      3     5
##     Survived
## 14         0
## 26         1
## 611        0
## 639        0
## 679        0
## 886        0
```

Evidently, if a child had `Parch == 0`, it meant they were traveling with a nanny or governess. None of these children traveled in first class. (I might be pushing a bit here including children as old as 15 in this subset.)

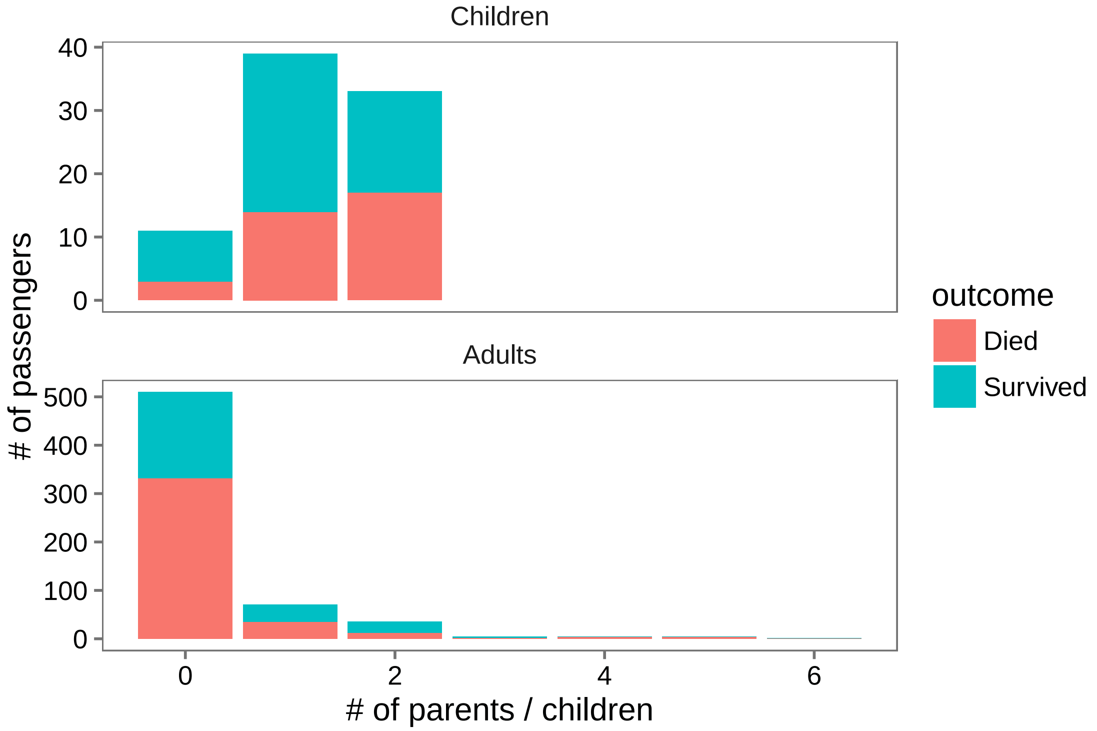
```r
(with_nanny <- filter(train, Age < 16 & Parch == 0) %>%
  select(Survived, Pclass, Name, Age))
```

```
##    Survived Pclass                                    Name  Age
## 1         1      2     Nasser, Mrs. Nicholas (Adele Achem) 14.0
## 2         0      3    Vestrom, Miss. Hulda Amanda Adolfina 14.0
## 3         1      3             McGowan, Miss. Anna "Annie" 15.0
## 4         1      3             Nicola-Yarred, Miss. Jamila 14.0
## 5         0      3                    Zabour, Miss. Hileni 14.5
## 6         1      3            Nicola-Yarred, Master. Elias 12.0
## 7         0      3                Hassan, Mr. Houssein G N 11.0
## 8         1      3           Emanuel, Miss. Virginia Ethel  5.0
## 9         1      3                    Ayoub, Miss. Banoura 13.0
## 10        1      3 Yasbeck, Mrs. Antoni (Selini Alexander) 15.0
## 11        1      3        Najib, Miss. Adele Kiamie "Jane" 15.0
```

All children had a `Pchar` value of 2 or lower (i.e., no more than two parents, which makes sense). There were few children with `Pchar` of 0; the survival probability was highest for these children in the training data. Survival probability was lowest for chilrden with two parents onboard. Most adults traveled without any parents or children (`Pchar = 0`). Survival rates were lowest in this group.

```r
child <- mutate(train,
                child = factor(Age < 16,
                               levels = c(TRUE, FALSE),
                               labels = c("Children", "Adults"))) %>%
  filter(!is.na(Age)) %>%
  group_by(child, Parch) %>%
  summarize(Died = n() - sum(Survived),
            Survived = sum(Survived)) %>%
  gather(outcome, number, -child, -Parch)

ggplot(child, aes(x = Parch, y = number, fill = outcome)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ child, ncol = 1, scale = "free_y") +
  xlab("# of parents / children") +
  ylab("# of passengers") +
  theme_few()
```



`Ticket` gives the

```
train$Ticket <- as.character(train$Ticket) # Doesn't need to be a factor
head(train$Ticket, 20)
```

```
##  [1] "A/5 21171"        "PC 17599"         "STON/O2. 3101282"
##  [4] "113803"           "373450"           "330877"
##  [7] "17463"            "349909"           "347742"
## [10] "237736"           "PP 9549"          "113783"
## [13] "A/5. 2151"        "347082"           "350406"
## [16] "248706"           "382652"           "244373"
## [19] "345763"           "2649"
```

These vary a lot, but sometimes you'll have several people with the same `Ticket`. Often, it looks like these were all members of the same family.

```
table(train$Ticket)[table(train$Ticket) > 5]
```

```
##
##     1601  3101295   347082   347088  CA 2144 CA. 2343
##        7        6        7        6        6        7
```

```
common_tickets <- names(table(train$Ticket)[table(train$Ticket) > 5])
filter(train, Ticket %in% common_tickets) %>%
  select(Name, Ticket, Survived) %>%
  arrange(Ticket)
```

```
##                                                           Name   Ticket
## 1                                              Bing, Mr. Lee     1601
## 2                                              Ling, Mr. Lee     1601
## 3                                             Lang, Mr. Fang     1601
## 4                                            Foo, Mr. Choong     1601
## 5                                               Lam, Mr. Ali     1601
## 6                                               Lam, Mr. Len     1601
## 7                                            Chip, Mr. Chang     1601
## 8                                 Panula, Master. Juha Niilo  3101295
## 9                               Panula, Master. Eino Viljami  3101295
## 10                                 Panula, Mr. Ernesti Arvid  3101295
## 11                    Panula, Mrs. Juha (Maria Emilia Ojala)  3101295
## 12                                  Panula, Mr. Jaako Arnold  3101295
## 13                              Panula, Master. Urho Abraham  3101295
## 14                               Andersson, Mr. Anders Johan   347082
## 15                         Andersson, Miss. Ellis Anna Maria   347082
## 16                      Andersson, Miss. Ingeborg Constanzia   347082
## 17                         Andersson, Miss. Sigrid Elisabeth   347082
## 18 Andersson, Mrs. Anders Johan (Alfrida Konstantia Brogren)   347082
## 19                        Andersson, Miss. Ebba Iris Alfrida   347082
## 20                   Andersson, Master. Sigvard Harald Elias   347082
## 21                                     Skoog, Master. Harald   347088
## 22           Skoog, Mrs. William (Anna Bernhardina Karlsson)   347088
## 23                                        Skoog, Mr. Wilhelm   347088
## 24                                        Skoog, Miss. Mabel   347088
## 25                             Skoog, Miss. Margit Elizabeth   347088
## 26                              Skoog, Master. Karl Thorsten   347088
## 27                        Goodwin, Master. William Frederick  CA 2144
## 28                                Goodwin, Miss. Lillian Amy  CA 2144
## 29                           Goodwin, Master. Sidney Leonard  CA 2144
## 30                            Goodwin, Master. Harold Victor  CA 2144
## 31                   Goodwin, Mrs. Frederick (Augusta Tyler)  CA 2144
## 32                               Goodwin, Mr. Charles Edward  CA 2144
## 33                                Sage, Master. Thomas Henry CA. 2343
## 34                              Sage, Miss. Constance Gladys CA. 2343
## 35                                       Sage, Mr. Frederick CA. 2343
## 36                                  Sage, Mr. George John Jr CA. 2343
## 37                                   Sage, Miss. Stella Anna CA. 2343
## 38                                  Sage, Mr. Douglas Bullen CA. 2343
## 39                         Sage, Miss. Dorothy Edith "Dolly" CA. 2343
##    Survived
## 1         1
## 2         0
## 3         1
## 4         1
## 5         1
## 6         0
## 7         1
## 8         0
## 9         0
## 10        0
## 11        0
## 12        0
## 13        0
## 14        0
## 15        0
## 16        0
## 17        0
## 18        0
## 19        0
## 20        0
## 21        0
## 22        0
## 23        0
## 24        0
## 25        0
## 26        0
## 27        0
## 28        0
## 29        0
## 30        0
## 31        0
## 32        0
## 33        0
## 34        0
## 35        0
## 36        0
## 37        0
## 38        0
## 39        0
```

Based on this, it looks like survival rates tended to be pretty low for large families (same last name and all on the same ticket). It's possible to set family (last name and ticket number) as an additional feature.

```
family <- mutate(train,
                 last_name = gsub(",.*", "", Name),
                 family = paste(last_name, Ticket, sep = "-")) %>%
  select(Survived, family, Pclass) %>%
  arrange(family)

head(rev(sort(table(family$family))), 10)
```

```
## 
##    Sage-CA. 2343 Andersson-347082     Skoog-347088   Panula-3101295 
##                7                7                6                6 
##   Goodwin-CA 2144      Rice-382652   Palsson-349909     Lefebre-4133 
##                6                5                4                4 
##    Fortune-19950  Ford-W./C. 6608 
##                4                4
```

```
family_num <- group_by(family, family) %>%
  summarize(n = n(),
            Survived = sum(Survived),
            Pclass = Pclass[1]) %>%
  arrange(desc(n))
head(family_num, 15)
```

```
## Source: local data frame [15 x 4]
```

```
##
##                family     n Survived Pclass
##                 (chr) (int)    (int)  (int)
## 1    Andersson-347082     7        0      3
## 2        Sage-CA. 2343     7        0      3
## 3     Goodwin-CA 2144     6        0      3
## 4      Panula-3101295     6        0      3
## 5        Skoog-347088     6        0      3
## 6         Rice-382652     5        0      3
## 7      Asplund-347077     4        3      3
## 8        Baclini-2666     4        4      3
## 9       Carter-113760     4        4      1
## 10    Ford-W./C. 6608     4        0      3
## 11      Fortune-19950     4        2      1
## 12       Lefebre-4133     4        0      3
## 13     Palsson-349909     4        0      3
## 14     Allison-113781     3        1      1
## 15 Collyer-C.A. 31921     3        2      2
```