```r
library(tidyverse)
```

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

```
## ✔ ggplot2 3.0.0     ✔ purrr   0.2.4
## ✔ tibble  1.4.2     ✔ dplyr   0.7.6
## ✔ tidyr   0.8.0     ✔ stringr 1.3.1
## ✔ readr   1.1.1     ✔ forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## Warning: package 'stringr' was built under R version 3.4.4
```

```
## ── Conflicts ────────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```r
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 3.4.4
```

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.4
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
library(rworldmap)
```

```
## Loading required package: sp
```

```
## Warning: package 'sp' was built under R version 3.4.4
```

```
## ### Welcome to rworldmap ###
```

```
## For a short introduction type :   vignette('rworldmap')
```

```r
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(knitr)
```

```r
ksp <- read.csv("~/Downloads/kickstarter-projects/ks-projects-201801.csv")
```

1 Data Cleaning

```
sum(is.na(ksp))
```

```
## [1] 3797
```

```
str(ksp)
```

```
## 'data.frame':    378661 obs. of  15 variables:
##  $ ID              : int  1000002330 1000003930 1000004038 1000007540 1000011046 1000014025 1000023410 1
000030581 1000034518 100004195 ...
##  $ name            : Factor w/ 375765 levels "","    IT'S A HOT CAPPUCCINO NIGHT  ",..: 332493 135633 36
4946 344770 77274 206067 293430 69281 284103 290686 ...
##  $ category        : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 56 124 59 42 114 40 ...
##  $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7 11 7 8 8 8 5 7 ...
##  $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14 14 14 14 14 14 14 14 14 ...
##  $ deadline        : Factor w/ 3164 levels "2009-05-03","2009-05-16",..: 2288 3042 1333 1017 2247 2463 19
96 2448 1790 1863 ...
##  $ goal            : num  1000 30000 45000 5000 19500 50000 1000 25000 125000 65000 ...
##  $ launched        : Factor w/ 378089 levels "1970-01-01 01:00:00",..: 243292 361975 80409 46557 235943
278600 187500 274014 139367 153766 ...
##  $ pledged         : num  0 2421 220 1 1283 ...
##  $ state           : Factor w/ 6 levels "canceled","failed",..: 2 2 2 2 1 4 4 2 1 1 ...
##  $ backers         : int  0 15 3 1 14 224 16 40 58 43 ...
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 23 23 23 23 23 23 23 23 ...
##  $ usd.pledged     : num  0 100 220 1 1283 ...
##  $ usd_pledged_real: num  0 2421 220 1 1283 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...
```

```
sapply(ksp, function(x) sum(is.na(x)))
```

```
##               ID             name         category    main_category 
##                0                0                0                0 
##         currency         deadline             goal         launched 
##                0                0                0                0 
##          pledged            state          backers          country 
##                0                0                0                0 
##      usd.pledged usd_pledged_real    usd_goal_real 
##             3797                0                0
```

```
sapply(ksp, function(x) sum(is.null(x)))
```

```
##               ID             name         category    main_category 
##                0                0                0                0 
##         currency         deadline             goal         launched 
##                0                0                0                0 
##          pledged            state          backers          country 
##                0                0                0                0 
##      usd.pledged usd_pledged_real    usd_goal_real 
##                0                0                0
```

```
#usd.pledged has 3797 missing values. I will just replace the value to the mean of its column.
```

```
ksp$usd.pledged <- ifelse(is.na(ksp$usd.pledged), mean(na.omit(ksp$usd.pledged)), ksp$usd.pledged)
sapply(ksp, function(x) sum(is.na(x)))
```

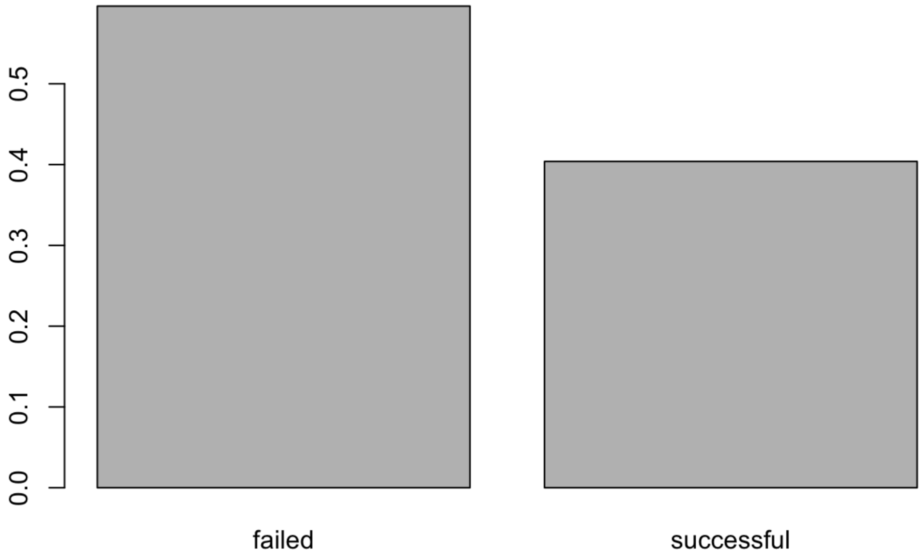
```
##                ID               name           category    main_category
##                 0                  0                  0                0
##          currency           deadline               goal         launched
##                 0                  0                  0                0
##           pledged              state            backers          country
##                 0                  0                  0                0
##       usd.pledged usd_pledged_real      usd_goal_real
##                 0                  0                  0
```

```
ksp$ID <- as.character(ksp$ID)
ksp$name <- as.character(ksp$name)

#Now I have no missing values in the dataset
```

```
ksp.new <- ksp[ksp$state == 'failed' | ksp$state == 'successful', ]
ksp.new$state <- as.character(ksp.new$state)
ksp.new$state <- as.factor(ksp.new$state)
prop.table(table(ksp.new$state))
```

```
##
##     failed successful
##  0.5961227  0.4038773
```

```
barplot(prop.table(table(ksp.new$state)))
```



```
#Since our target variable is state, I subsetted records that the state is either success or fail to make it
binary problem
#Success rate has been incresed to 40% (35% before) after dropping other states.
```

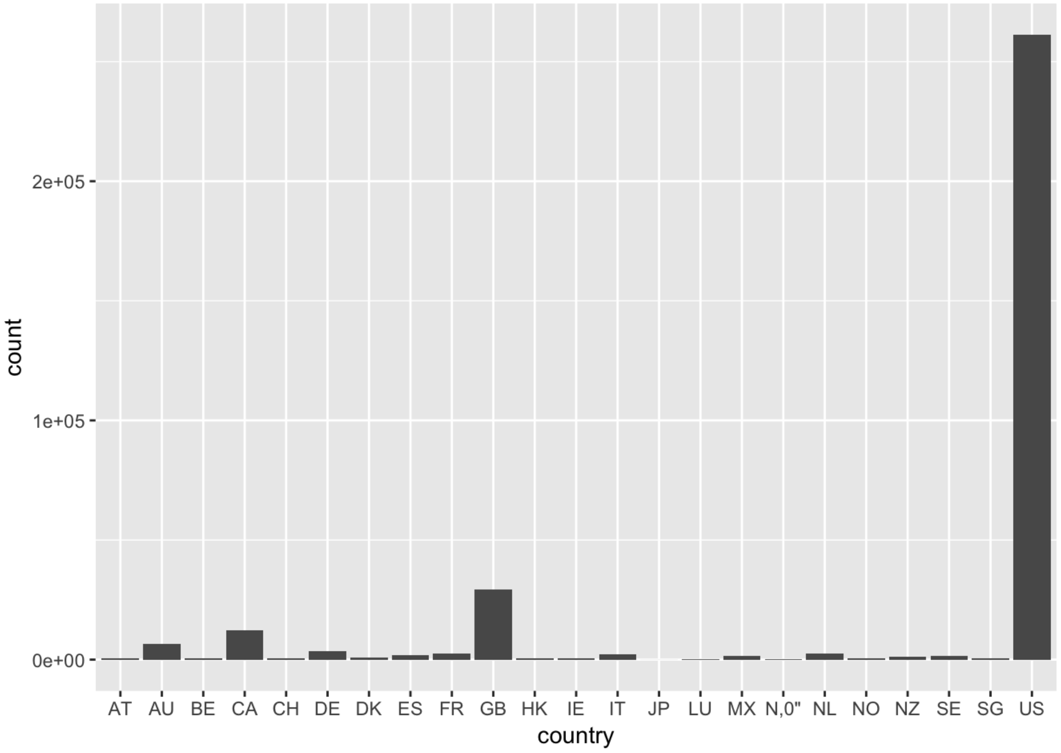
```
ksp.new$duration <- as.Date(ksp.new$deadline) - as.Date(ksp.new$launched)
ksp.new$duration <- as.numeric(ksp.new$duration)
#added a new variable called duration to understand how many days spent for each project
```

```
ksp.new <- ksp.new %>%
  separate(col = "deadline", into = c("deadline_year", "deadline_month", "deadline_day"), sep = "-") %>%
  separate(col = "launched", into = c("launched_year", "launched_month", "launched_day"), sep = "-")
#broke down the date variables to year, month and day
```

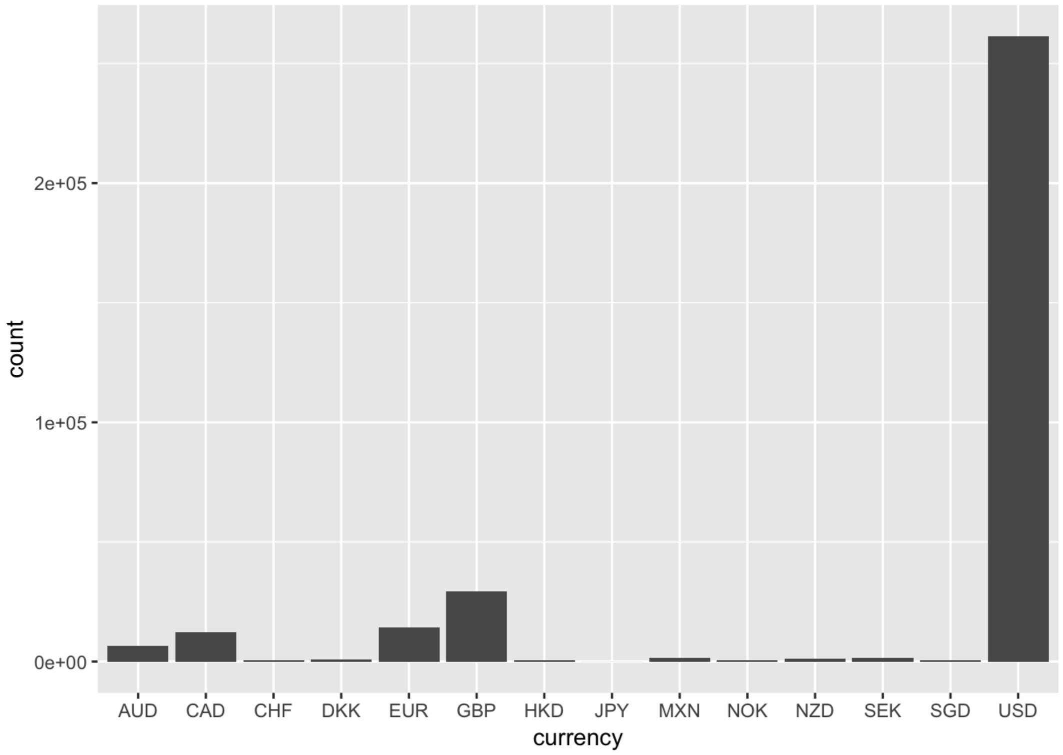
```
str(ksp.new)
```

```
## 'data.frame':    331675 obs. of  20 variables:
##  $ ID              : chr  "1000002330" "1000003930" "1000004038" "1000007540" ...
##  $ name            : chr  "The Songs of Adelaide & Abullah" "Greeting From Earth: ZGAC Arts Capsule For E
T" "Where is Hank?" "ToshiCapital Rekordz Needs Help to Complete Album" ...
##  $ category        : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 124 59 42 96 73 33 ...
##  $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7 11 8 8 8 13 11 3 ...
##  $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14 14 14 14 14 14 2 14 14 ...
##  $ deadline_year   : chr  "2015" "2017" "2013" "2012" ...
##  $ deadline_month  : chr  "10" "11" "02" "04" ...
##  $ deadline_day    : chr  "09" "01" "26" "16" ...
##  $ goal            : num  1000 30000 45000 5000 50000 1000 25000 2500 12500 5000 ...
##  $ launched_year   : chr  "2015" "2017" "2013" "2012" ...
##  $ launched_month  : chr  "08" "09" "01" "03" ...
##  $ launched_day    : chr  "11 12:12:28" "02 04:43:57" "12 00:20:50" "17 03:24:11" ...
##  $ pledged         : num  0 2421 220 1 52375 ...
##  $ state           : Factor w/ 2 levels "failed","successful": 1 1 1 1 2 2 1 1 2 1 ...
##  $ backers         : int  0 15 3 1 224 16 40 0 100 0 ...
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 23 23 23 23 23 4 23 23 ...
##  $ usd.pledged     : num  0 100 220 1 52375 ...
##  $ usd_pledged_real: num  0 2421 220 1 52375 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 50000 ...
##  $ duration        : num  59 60 45 30 35 20 45 30 30 30 ...
```

```
ggplot(ksp.new, aes(country)) + geom_bar()
```



```
ggplot(ksp.new, aes(currency)) + geom_bar()
```

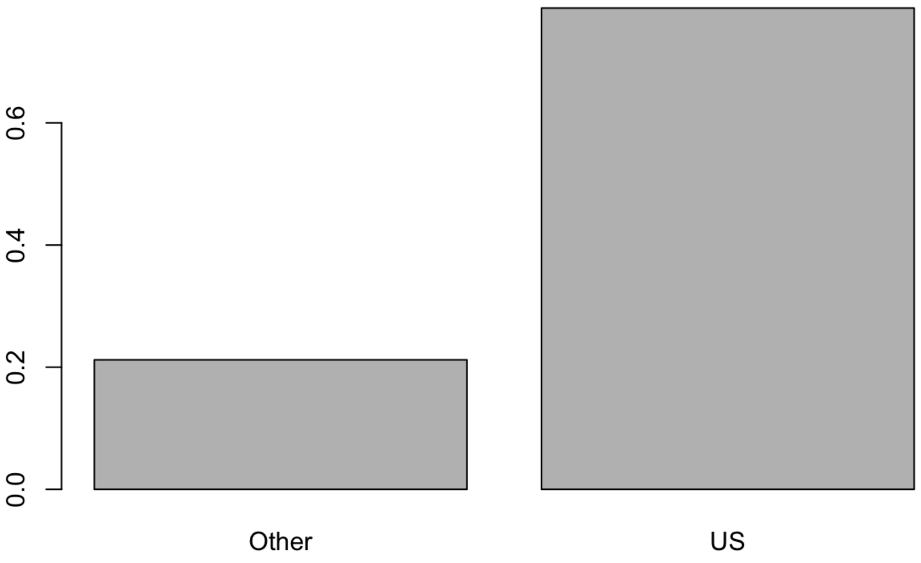


```
#when you see the graph below, most of the projects are took place in US. To reduce the level of columns, I'
m going to make it binary either us or not. Same for currency.
```

```
ksp.new$country <- as.character(ksp.new$country)
ksp.new$country[ksp.new$country %in% c("JP", "LU", "AT", "HK", "SG", "BE", "CH", "IE", "NO", "DK",
                                       "MX", "NZ", "SE", "ES", "IT", "NL", "FR", "DE","AU","CA","GB",'N,0"'
)] <- "Other"
ksp.new$country <- as.factor(ksp.new$country)
prop.table(table(ksp.new$country))
```

```
##
##     Other        US
## 0.2119997 0.7880003
```

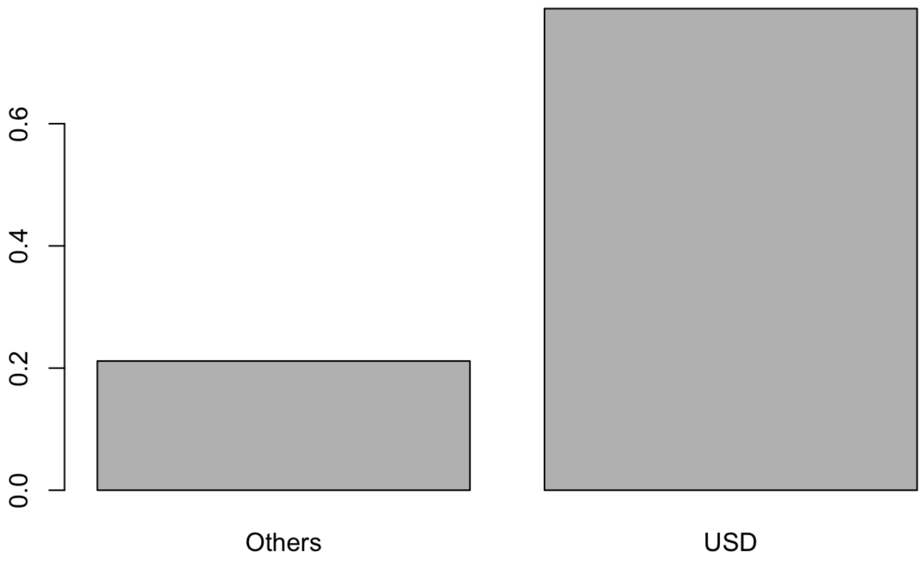
```
barplot(prop.table(table(ksp.new$country)))
```



```
ksp.new$currency <- as.character(ksp.new$currency)
ksp.new$currency[ksp.new$currency %in% c("AUD", "CAD","CHF","DKK","EUR","GBP","HKD","JPY","MXN","NOK","NZD",
"SEK","SGD")] <- "Others"
ksp.new$currency <- as.factor(ksp.new$currency)
prop.table(table(ksp.new$currency))
```

```
##
##    Others       USD
## 0.2115444 0.7884556
```

```
barplot(prop.table(table(ksp.new$currency)))
```

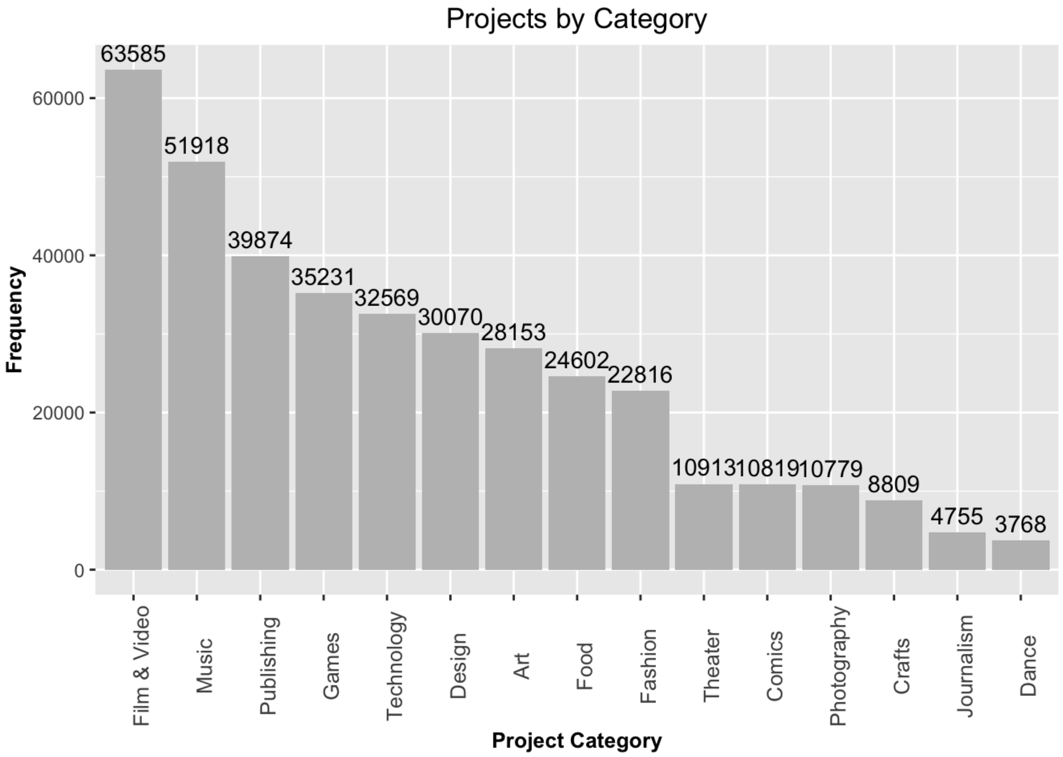


```
#approximately 80% of projects are held in US and 20% are held in other countries
```

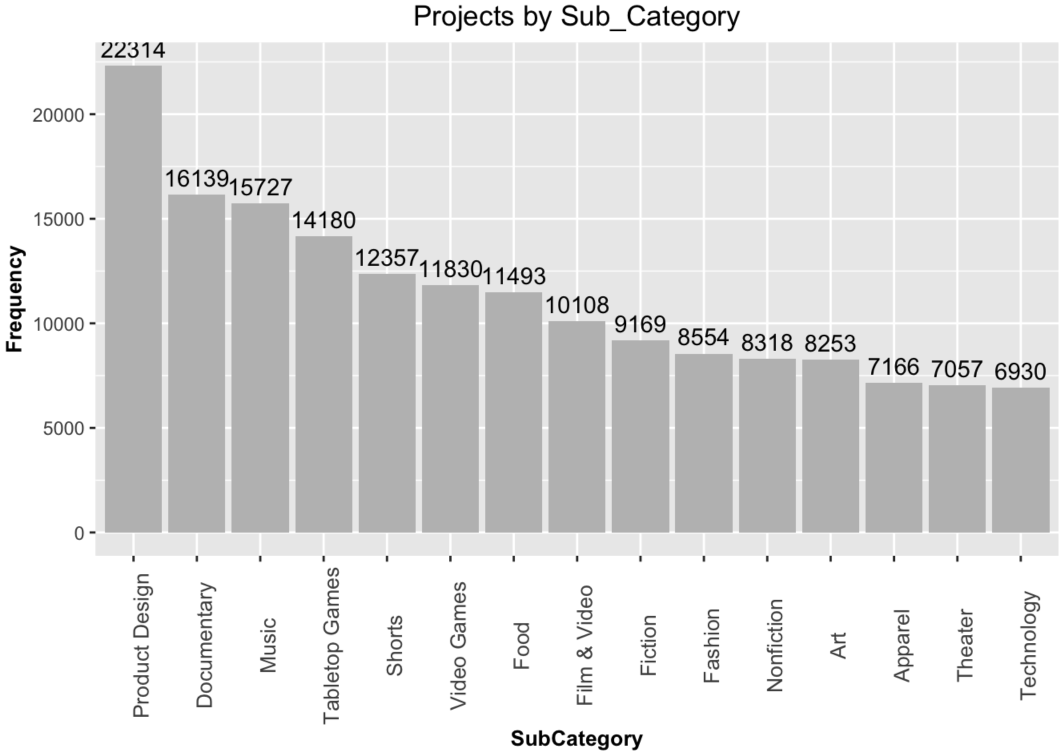
```
cat.freq <- ksp %>%
  group_by(main_category) %>%
  summarize(count=n()) %>%
  arrange(desc(count))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
cat.freq$main_category <- factor(cat.freq$main_category, levels=cat.freq$main_category)


ggplot(cat.freq, aes(main_category, count, fill=count)) + geom_bar(stat="identity") +
    ggtitle("Projects by Category") + xlab("Project Category") + ylab("Frequency") +
    geom_text(aes(label=count), vjust=-0.5) +
    theme(plot.title=element_text(hjust=0.5), axis.title=element_text(size=10, face="bold"),
          axis.text.x=element_text(size=10, angle=90), legend.position="null") +
    scale_fill_gradient(low="grey", high="grey")
```



```
subcat.freq <- ksp %>%
  group_by(category) %>%
  summarize(count=n()) %>%
  arrange(desc(count))

subcat.freq$category <- factor(subcat.freq$category, levels=subcat.freq$category)

ggplot(head(subcat.freq, 15), aes(category, count, fill=count)) + geom_bar(stat="identity") +
    ggtitle("Projects by Sub_Category") + xlab("SubCategory") + ylab("Frequency") +
    geom_text(aes(label=count), vjust=-0.5) +
    theme(plot.title=element_text(hjust=0.5), axis.title=element_text(size=10, face="bold"),
          axis.text.x=element_text(size=10, angle=90), legend.position="null") +
    scale_fill_gradient(low="grey", high="grey")
```



```
prop.table(table((ksp.new$state[ksp.new$main_category == 'Film & Video'])))
```

```
##
##     failed successful
##  0.5820935  0.4179065
```

```
prop.table(table((ksp.new$state[ksp.new$main_category == 'Music'])))
```

```
##
##     failed successful
##  0.4733944  0.5266056
```

```
prop.table(table((ksp.new$state[ksp.new$main_category == 'Publishing'])))
```

```
##
##     failed successful
##  0.6529835  0.3470165
```

```
#Music industry has a highest success rate at 52% among three categories.
```