Artist Selection

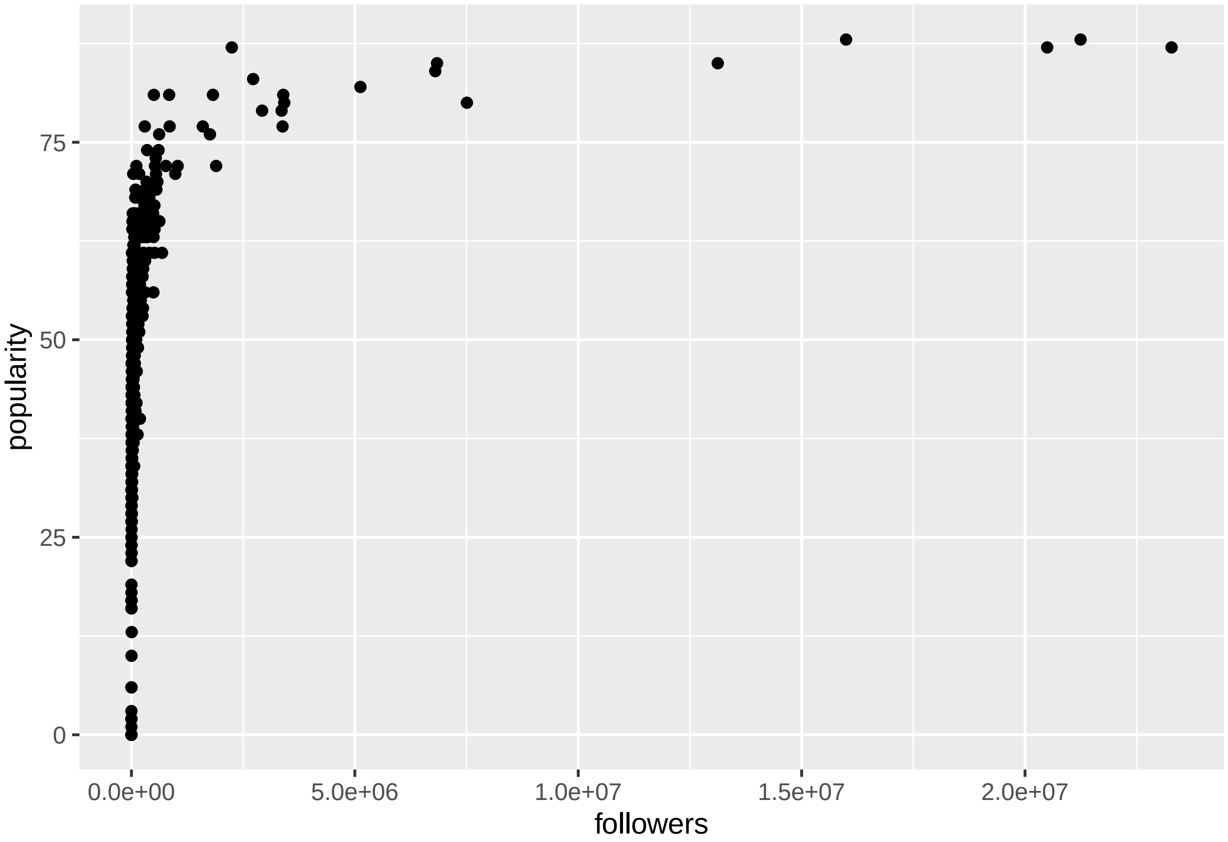
```r
artists <- read.csv("/Users/Anton/Desktop/DSU Project/scraped_data.csv")
summary(artists)
```

```
##      name             followers          popularity   
##  Length:412         Min.   :       3   Min.   : 0.00  
##  Class :character   1st Qu.:   14276   1st Qu.:40.00  
##  Mode  :character   Median :   46594   Median :52.00  
##                     Mean   :  461220   Mean   :49.92  
##                     3rd Qu.:  148534   3rd Qu.:61.00  
##                     Max.   :23280041   Max.   :88.00  
```

```r
str(artists)
```

```
## 'data.frame':    412 obs. of  3 variables:
##  $ name      : chr  "David Guetta" "The Chainsmokers" "Diplo" "Marshmello" ...
##  $ followers : int  21242602 15995277 2246640 23280041 20495926 6839766 13123138 6799590 2724034 5125
##  $ popularity: int  88 88 87 87 87 85 85 84 83 82 ...
```

```r
library(ggplot2)

ggplot(artists, aes(x = followers, y = popularity)) +
  geom_point()
```

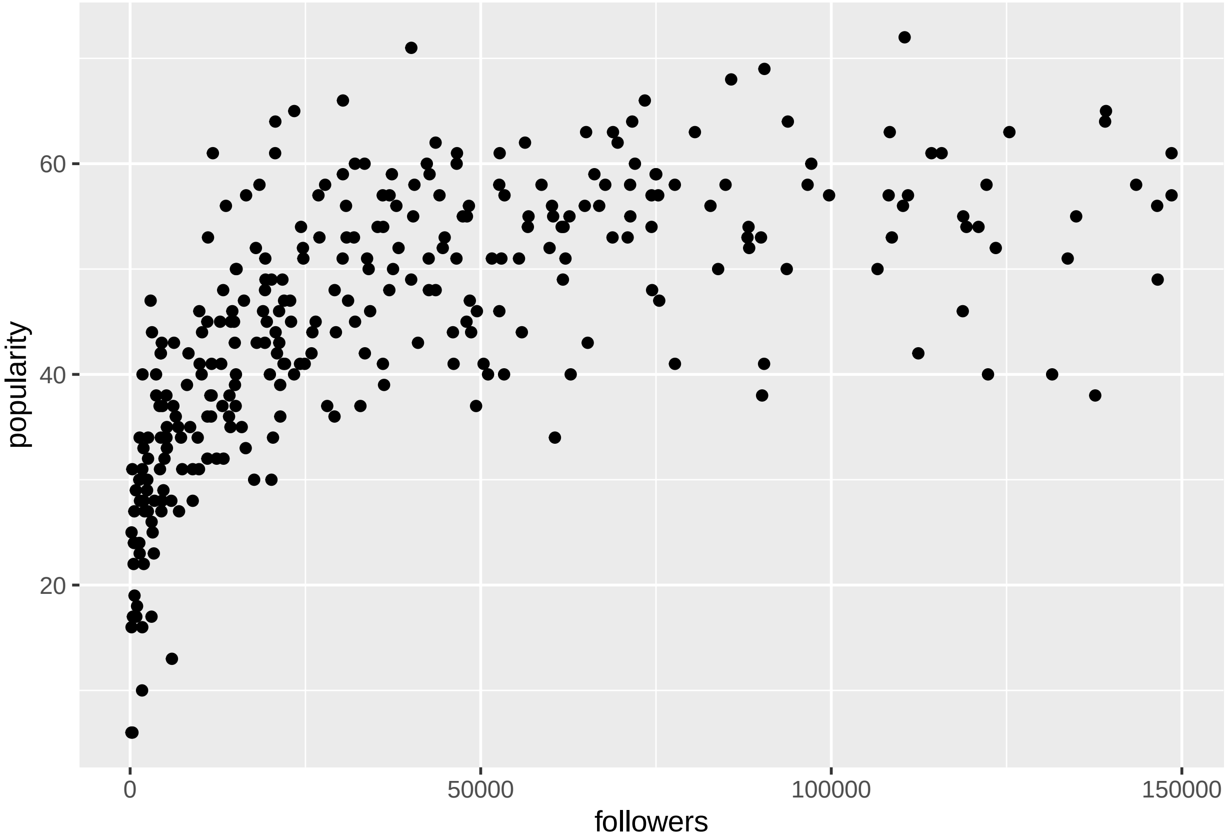


```r
head(artists[order(artists$followers, decreasing = TRUE),], 30)
```

```
##                  name followers popularity
## 4          Marshmello  23280041         87
## 1        David Guetta  21242602         88
## 5         Alan Walker  20495926         87
## 2    The Chainsmokers  15995277         88
## 7       Martin Garrix  13123138         85
## 16               Alok   7508404         80
## 6          A$AP Rocky   6839766         85
## 8            DJ Snake   6799590         84
## 10               Zedd   5125154         82
## 15   Armin van Buuren   3420318         80
## 12         Steve Aoki   3398204         81
## 22           Afrojack   3383058         77
## 17             Alesso   3360291         79
## 18           Galantis   2924045         79
## 9          Jonas Blue   2724034         83
## 3              Diplo   2246640         87
## 32       Nicky Romero   1894627         72
## 14       Metro Boomin   1824417         81
## 24   Lost Frequencies   1757519         76
## 19         Don Diablo   1597822         77
## 28             Deorro   1034957         72
## 34     Dillon Francis    981660         71
## 21     Oliver Heldens    854988         77
## 11           ILLENIUM    844490         81
## 29          Kaskade    770270         72
## 104          Zomboy    684442         61
## 71     Flux Pavilion    625582         65
## 23     Timmy Trumpet    619942         76
## 26           Matoma    607155         74
## 37    Above & Beyond    581916         70
```

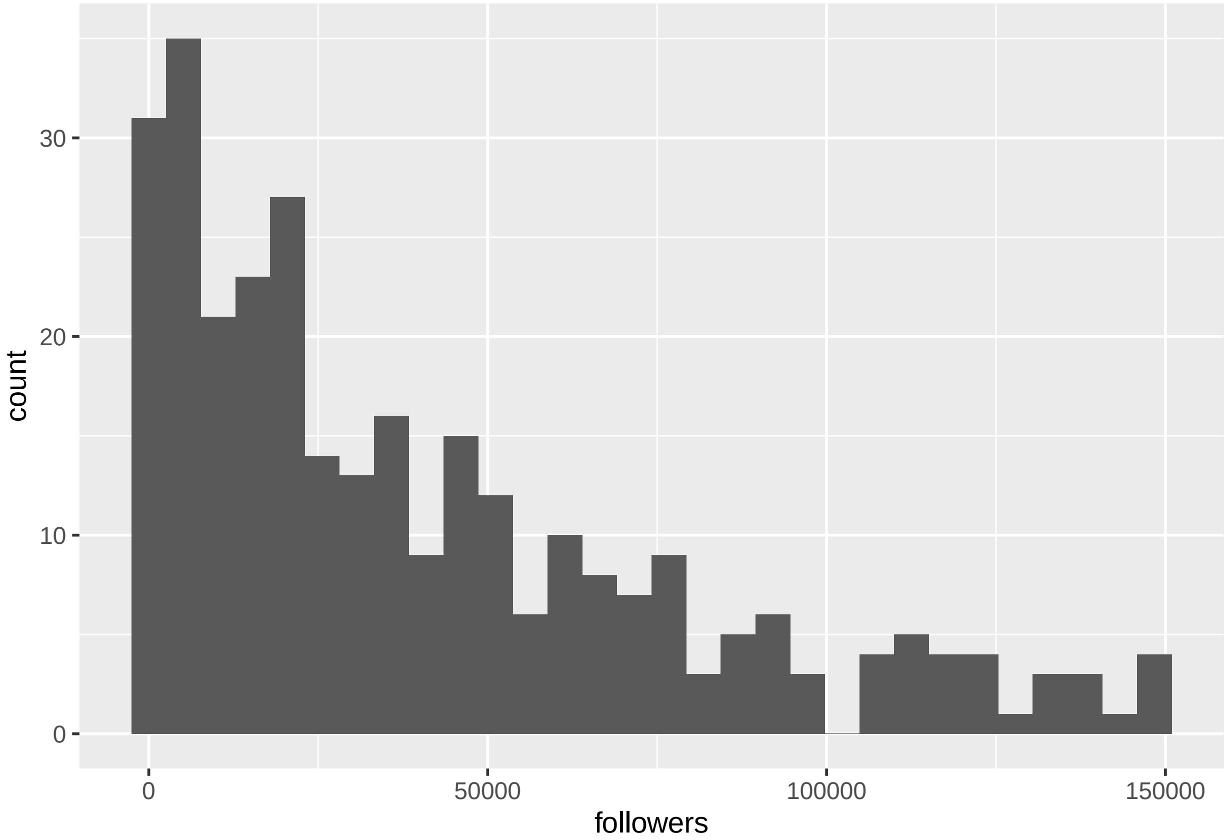
```
artists_small <- artists[artists$followers < 150000 & artists$popularity > 5,]
```

```
ggplot(artists_small, aes(x = followers, y = popularity)) +
  geom_point()
```

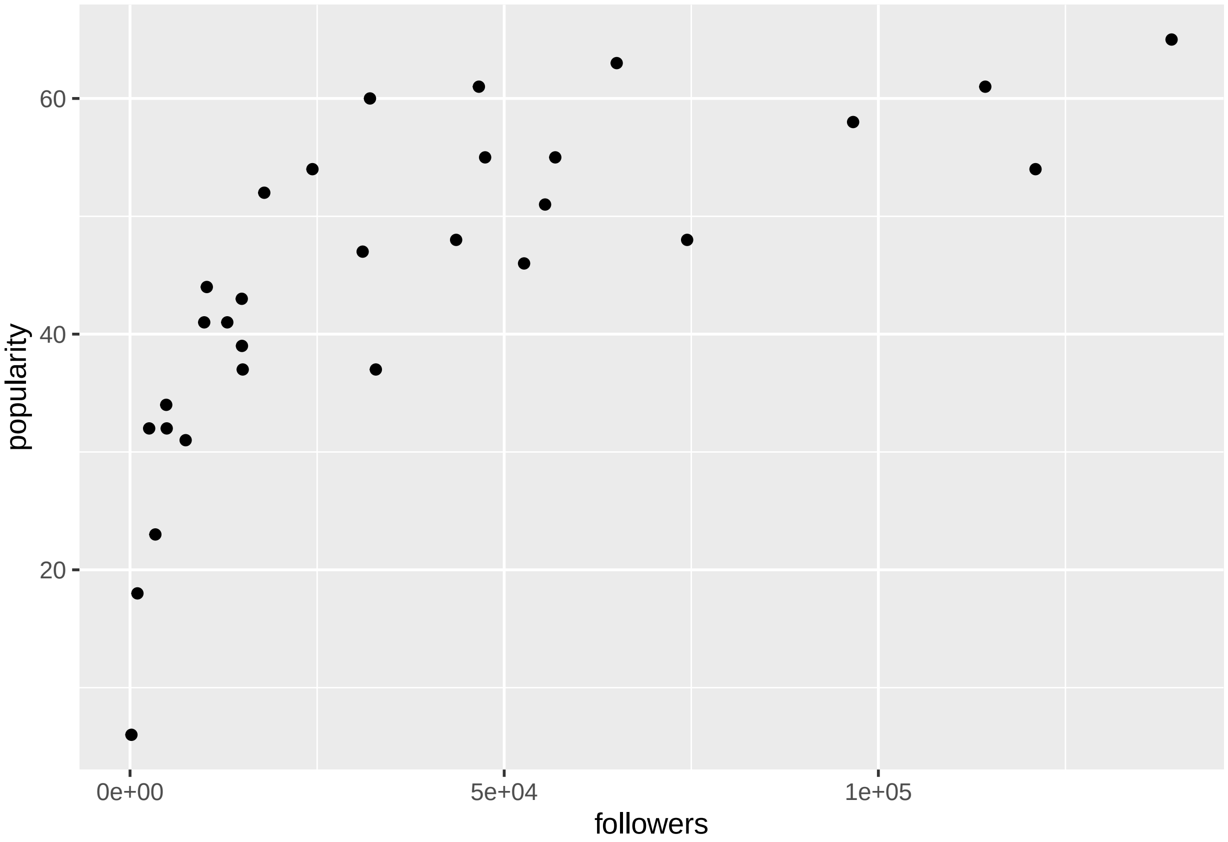


```
ggplot(artists_small, aes(followers)) +
  geom_histogram()
```

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

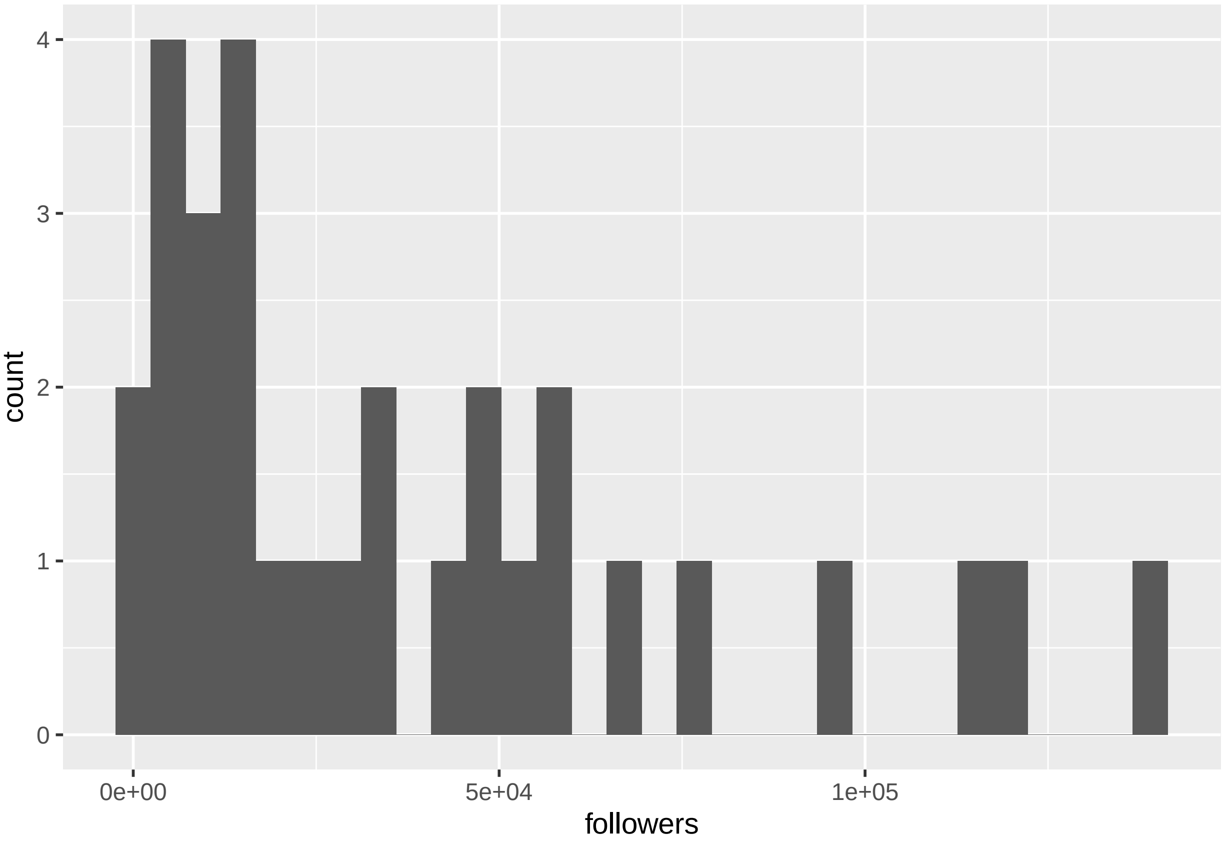


```r
set.seed(123)

artists_small_sample_i <- sample(nrow(artists_small), size = 30)

artists_small_sample <- artists_small[artists_small_sample_i, ]

ggplot(artists_small_sample, aes(x = followers, y = popularity)) +
  geom_point()
```



```r
ggplot(artists_small_sample, aes(followers)) +
  geom_histogram()
```

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



```
#write.csv(artists_small_sample, file = "/Users/Anton/Desktop/initial_small_artists.csv", row.names = F
```