DATA 606 Data Project Proposal

Alice Ding

Data Preparation

```r
library(dplyr)
library(ggplot2)
library(psych)

# load data
full_data <- read.csv("https://raw.githubusercontent.com/addsding/data606/main/project/Spotify-2000.csv

head(full_data)
```

```
##   Index                                 Title            Artist
## 1     1                               Sunrise       Norah Jones
## 2     2                           Black Night       Deep Purple
## 3     3                        Clint Eastwood          Gorillaz
## 4     4                         The Pretender      Foo Fighters
## 5     5                Waitin’ On A Sunny Day Bruce Springsteen
## 6     6 The Road Ahead (Miles Of The Unknown)      City To City
##              Top.Genre Year Beats.Per.Minute..BPM. Energy Danceability
## 1      adult standards 2004                    157     30           53
## 2           album rock 2000                    135     79           50
## 3  alternative hip hop 2001                    168     69           66
## 4    alternative metal 2007                    173     96           43
## 5         classic rock 2002                    106     82           58
## 6 alternative pop rock 2004                     99     46           54
##   Loudness..dB. Liveness Valence Length..Duration. Acousticness Speechiness
## 1           -14       11      68               201           94           3
## 2           -11       17      81               207           17           7
## 3            -9        7      52               341            2          17
## 4            -4        3      37               269            0           4
## 5            -5       10      87               256            1           3
## 6            -9       14      14               247            0           2
##   Popularity
## 1         71
## 2         39
## 3         69
## 4         76
## 5         59
## 6         45
```

Research question

Is there a correlation between tempo (BPM) and popularity tracks on Spotify?

Cases

What are the cases, and how many are there?

The cases are each songs; the data is comprised of the top 2000 top tracks on Spotify released from 1956 to 2019.

Data collection

Describe the method of data collection.

This data was retrieved from Kaggle by Sumat Singh. They retrieved it from Spotify's API specifically.

Type of study

Observational.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

https://www.kaggle.com/datasets/iamsumat/spotify-top-2000s-mega-dataset

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variable would be popularity and it is quantitative. This is the description taken from Spotify:

```
The popularity of the track. The value will be between 0 and 100, with 100 being the
most popular. The popularity of a track is a value between 0 and 100, with 100 being
the most popular. The popularity is calculated by algorithm and is based, in the most
part, on the total number of plays the track has had and how recent those plays are.
Generally speaking, songs that are being played a lot now will have a higher popularity
than songs that were played a lot in the past. Duplicate tracks (e.g. the same track
from a single and an album) are rated independently. Artist and album popularity is
derived mathematically from track popularity. Note: the popularity value may lag actual
popularity by a few days: the value is not updated in real time.
```

Independent Variable(s)

The explanatory variable would be tempo (BPM) and it is quantitative. This is the description taken from Spotify:

```
The overall estimated tempo of a track in beats per minute (BPM). In musical terminology,
tempo is the speed or pace of a given piece and derives directly from the average beat
duration.
```

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
describe(full_data$Popularity)
```

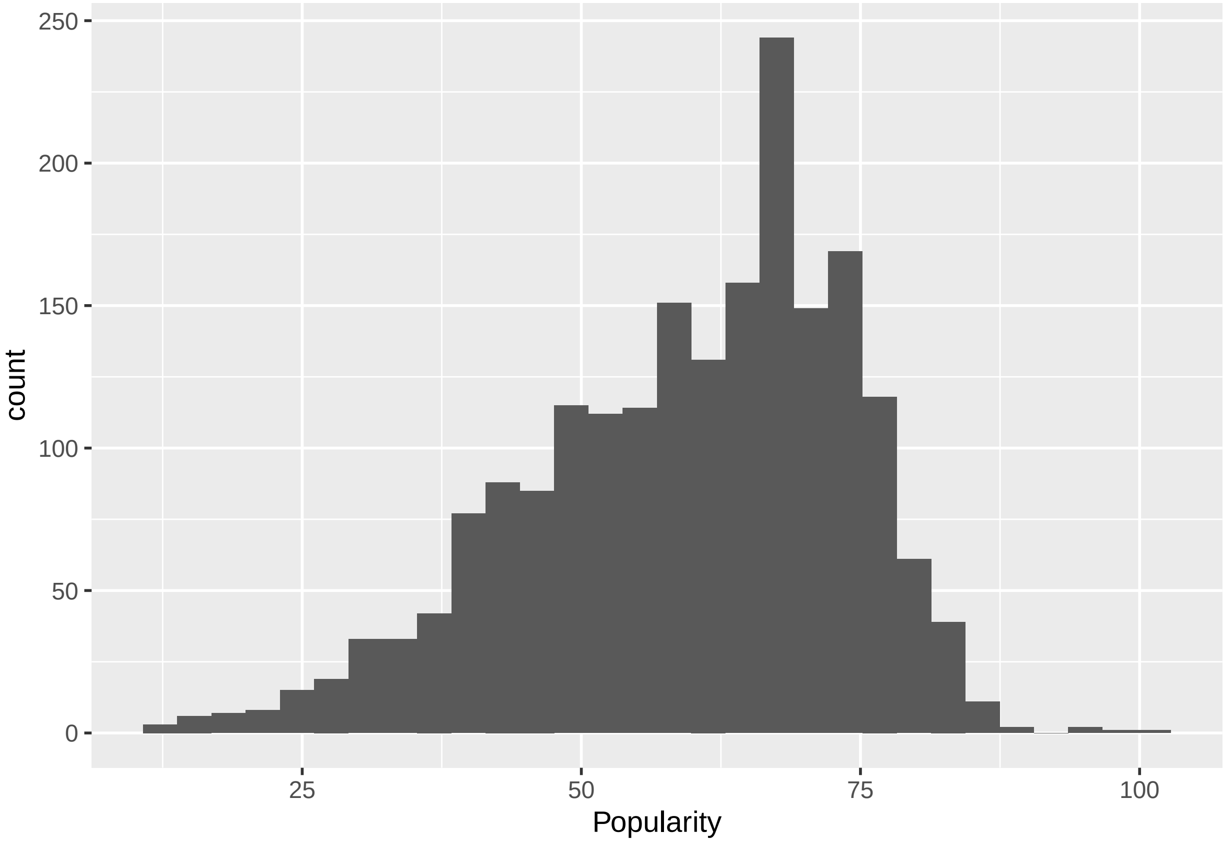
```
##    vars    n  mean    sd median trimmed   mad min max range  skew kurtosis   se
## X1    1 1994 59.53 14.35     62    60.4 14.83  11 100    89 -0.53    -0.12 0.32
```

```
describe(full_data$Beats.Per.Minute..BPM.)
```

```
##    vars    n   mean    sd median trimmed   mad min max range skew kurtosis   se
## X1    1 1994 120.22 28.03    119  118.71 28.17  37 206   169 0.42    -0.15 0.63
```

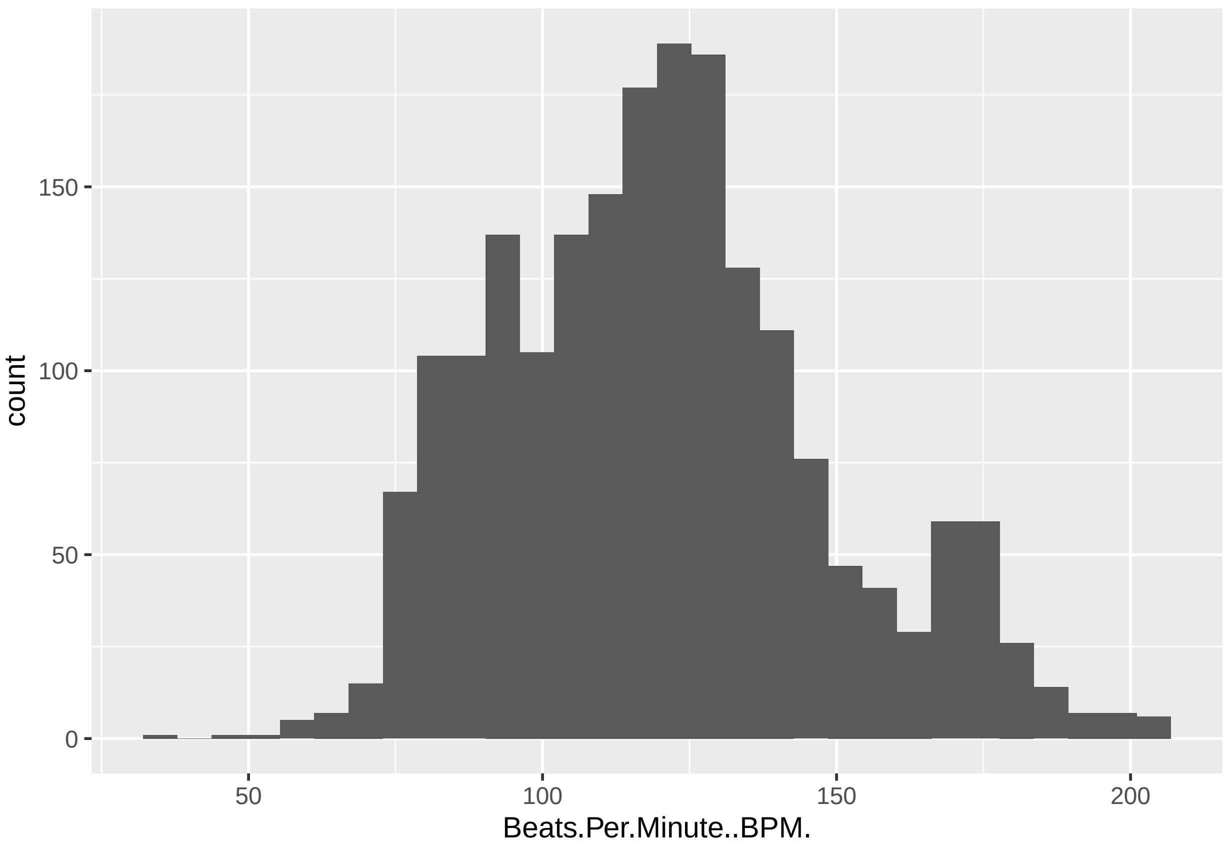
```
ggplot(full_data, aes(x=Popularity)) + geom_histogram()
```

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



```
ggplot(full_data, aes(x=Beats.Per.Minute..BPM.)) + geom_histogram()
```

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


count
150
100
50
0
50
100
150
200
Beats.Per.Minute..BPM.