R Notebook

Basic Idea

This project tries to analyze the car brands mentioned in lyrics.

Import Packages

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## 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
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

Import Data

```
load(file="../data/lyrics.RData")
lrc <- dt_lyrics
```

Data Cleaning

```
# keep songs' year from 1968-2016
lrc <- lrc%>%filter(year>=1968 & year <=2016)
# remove duplicated rows
lrc <- unique(lrc)
# remove songs with no lyrics or too many lyrics
lrc <- lrc%>%mutate(word_ct=str_count(lrc$lyrics, '\\w+'))
wd_outliers = boxplot(lrc$word_ct, plot=FALSE)$out
'%ni%' <- Negate('%in%')
lrc <- lrc%>%filter(lrc$word_ct %ni% wd_outliers)
```

Data Analysis

Let's have a look at how many songs in each genre after data cleaning.

```
lrc_genre <- lrc%>%group_by(genre)%>%summarise(genre_ct=n())%>%arrange(desc(genre_ct))
lrc_genre
```

```
## # A tibble: 12 x 2
##    genre         genre_ct
##    <chr>            <int>
##  1 Rock             64099
##  2 Pop              17619
##  3 Metal            11132
##  4 Country           7494
##  5 Jazz              4081
##  6 Hip-Hop           3858
##  7 Not Available     3258
##  8 Electronic        2736
##  9 R&B               2114
## 10 Indie             1314
## 11 Folk               497
## 12 Other              132
```

It turns out "Rock" has the most songs in this dataset.

How many cars are mentioned in lyrics?

```
# lexus
lexus <- sum(grepl("\\lexus\\b", lrc$lyrics, ignore.case = TRUE))
lexus
```

```
## [1] 10
```

```
# ferari
ferari <- sum(grepl("rari", lrc$lyrics, ignore.case = TRUE)) + sum(grepl("ferari", lrc$lyrics, ignore.ca
ferari
```

```
## [1] 95
```

```
# bentley
bentley <- sum(grepl("bentley", lrc$lyrics, ignore.case = TRUE))
bentley
```

```
## [1] 56
```

```
# bmw
bmw <- sum(grepl("bmw", lrc$lyrics, ignore.case = TRUE)) + sum(grepl("beamer", lrc$lyrics, ignore.case =
bmw
```

```
## [1] 39
```

```
# lamborghini
lambo <- sum(grepl("lambo", lrc$lyrics, ignore.case = TRUE))
lambo
```

```
## [1] 47
```

```
# maserati
maserati <- sum(grepl("maserati", lrc$lyrics, ignore.case = TRUE))
maserati
```

```
## [1] 20
```

```
# mcLaren
mclaren <- sum(grepl("mclaren", lrc$lyrics, ignore.case = TRUE))
mclaren
```

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

```
# benz
benz <- sum(grepl("\\benz\\b", lrc$lyrics, ignore.case = TRUE)) + sum(grepl("\\mercedes\\b", lrc$lyrics
benz
```

```
## [1] 6
```

```
# porsche
porsche <- sum(grepl("porsche", lrc$lyrics, ignore.case = TRUE))
porsche
```

```
## [1] 33
```

```r
# amg
amg <- sum(grepl("\\amg\\b", lrc$lyrics, ignore.case = TRUE))
amg
```

```
## [1] 0
```

```r
# cadillac
caddy <- sum(grepl("cadillac", lrc$lyrics, ignore.case = TRUE)) + sum(grepl("caddy", lrc$lyrics, ignore
caddy
```

```
## [1] 377
```

```r
# ford
ford <- sum(grepl("\\ford\\b", lrc$lyrics, ignore.case = TRUE))
ford
```

```
## [1] 0
```

```r
# honda
honda <- sum(grepl("honda", lrc$lyrics, ignore.case = TRUE))
honda
```

```
## [1] 39
```

```r
#toyota
toyota <- sum(grepl("toyota", lrc$lyrics, ignore.case = TRUE))
toyota
```

```
## [1] 7
```

```r
# nissan
nissan <- sum(grepl("nissan", lrc$lyrics, ignore.case = TRUE))
nissan
```

```
## [1] 0
```

```r
# volvo
volvo <- sum(grepl("volvo", lrc$lyrics, ignore.case = TRUE))
volvo
```

```
## [1] 12
```

```r
# chevrolet
chevy <- sum(grepl("chevrolet", lrc$lyrics, ignore.case = TRUE)) + sum(grepl("chevy", lrc$lyrics, ignor

# jeep
jeep <- sum(grepl("jeep", lrc$lyrics, ignore.case = TRUE))
jeep
```

```
## [1] 73
```

```r
# buick
buick <- sum(grepl("buick", lrc$lyrics, ignore.case = TRUE))
buick
```

```
## [1] 18
```

```r
# jaguar
jaguar <- sum(grepl("jaguar", lrc$lyrics, ignore.case = TRUE))
jaguar
```

```
## [1] 22
```

```r
# land rover
rover <- sum(grepl("\\rover\\b", lrc$lyrics, ignore.case = TRUE))
rover
```

```
## [1] 0
```

```r
# lexus
audi <- sum(grepl("\\audi\\b", lrc$lyrics, ignore.case = TRUE)) + sum(grepl("\\audis\\b", lrc$lyrics, ig
audi
```

```
## [1] 0
```

```r
# tesla
tesla <- sum(grepl("tesla", lrc$lyrics, ignore.case = TRUE))
tesla
```

```
## [1] 2
```

```r
cars <- tibble('lexus'=lexus, 'ferari'=ferari, 'bentley'=bentley, 'bmw'=bmw, 'lambo'=lambo, 'maserati'=
cars <- cars%>%pivot_longer(everything(), names_to = "car brand", values_to = "count")%>%arrange(desc(c
cars <- cars%>%mutate(percentage=count/sum(count)*100)
cars
```

```
## # A tibble: 23 x 3
##    `car brand` count percentage
##    <chr>       <int>      <dbl>
##  1 caddy         377      36.4
##  2 chevy         178      17.2
##  3 ferari         95       9.18
##  4 jeep           73       7.05
##  5 bentley        56       5.41
##  6 lambo          47       4.54
##  7 bmw            39       3.77
##  8 honda          39       3.77
##  9 porsche        33       3.19
## 10 jaguar         22       2.13
## # ... with 13 more rows
```

It turns out “Cadillac” is the most mentioned car brand among all songs.

So my next question is: What genre mentions cars most? It can be seen from the above tibble that top 4 car brands count for nearly 70% among all brands. So I work on these four brands:

```
# cadilac
lrc_caddy <- lrc%>%mutate(caddy=(grepl("cadillac", lrc$lyrics, ignore.case = TRUE)) | grepl("caddy", lrc
lrc_caddy <- lrc_caddy%>%group_by(genre)%>%summarise(caddy_count=n())%>%arrange(desc(caddy_count))
lrc_car <- lrc_genre%>%left_join(lrc_caddy)%>%mutate(caddy_percentage=caddy_count/genre_ct*100)
```

```
## Joining, by = "genre"
```

```
# chevrolet
lrc_chevy <- lrc%>%mutate(chevy=(grepl("chevrolet", lrc$lyrics, ignore.case = TRUE)) | grepl("chevy", li
lrc_chevy <- lrc_chevy%>%group_by(genre)%>%summarise(chevy_count=n())%>%arrange(desc(chevy_count))
lrc_car <- lrc_car%>%left_join(lrc_chevy)%>%mutate(chevy_percentage=chevy_count/genre_ct*100)
```

```
## Joining, by = "genre"
```

```
# ferari
lrc_ferari <- lrc%>%mutate(ferari=(grepl("rari", lrc$lyrics, ignore.case = TRUE)) | grepl("ferari", lrc$
lrc_ferari <- lrc_ferari%>%group_by(genre)%>%summarise(ferari_count=n())%>%arrange(desc(ferari_count))
lrc_car <- lrc_car%>%left_join(lrc_ferari)%>%mutate(ferari_percentage=ferari_count/genre_ct*100)
```

```
## Joining, by = "genre"
```

```
# jeep
lrc_jeep <- lrc%>%mutate(jeep=grepl("jeep", lrc$lyrics, ignore.case = TRUE))%>%filter(jeep==TRUE)
lrc_jeep <- lrc_jeep%>%group_by(genre)%>%summarise(jeep_count=n())%>%arrange(desc(jeep_count))
lrc_car <- lrc_car%>%left_join(lrc_jeep)%>%mutate(jeep_percentage=jeep_count/genre_ct*100)
```

```
## Joining, by = "genre"
```

```
# Add up all the percentage w/i genre
lrc_car <- lrc_car%>%select(genre, caddy_percentage, chevy_percentage, ferari_percentage, jeep_percentag
lrc_car <- lrc_car%>%mutate(sum=caddy_percentage+chevy_percentage+ferari_percentage+jeep_percentage)%>%a
lrc_car
```

```
## # A tibble: 12 x 2
##    genre             sum
##    <chr>           <dbl>
##  1 Other           3.79
##  2 Hip-Hop         2.90
##  3 Country         1.31
##  4 Rock            0.587
##  5 Not Available   0.552
##  6 Pop             0.369
##  7 Metal          NA
##  8 Jazz           NA
##  9 Electronic     NA
## 10 R&B            NA
## 11 Indie          NA
## 12 Folk           NA
```

Conclusion

It seems like “Other” genre has the most cars mentioned. But have a look at the singers: “a-boogie-wit-da-hoodie”, “asap twelvyy”, “g-herbo”... are rappers so their songs should belong to “Hip-Hop”. Thus the conclusion would be “Hip-Hop” songs mention cars most in our dataset. And among all car brands, “Cadillac” is singers’ favourite choice.