

Code (from RStudio)

(Note: The following is not cleaned of unused code; code is in no specific order; please excuse any repetition)

```
library(tidyverse)
library(ggplot2)

getOption("max.print")
options(max.print=999999)
songs <- read.csv(file = "songs.csv")
songs$decade <- cut(songs$year,
                    c(1919,1929,1939,1949,1959,1969,
                      1979,1989,1999,2009,2019,2029),
                    c("20s", "30s", "40s", "50s", "60s",
                      "70s", "80s", "90s", "00s", "10s", "2020-21"))

as.vector(songs$key)

trim<-songs %>% filter(year >= 1950)
plot(aggregate( danceability ~ year, trim, mean ))

songs %>% filter(keys == "F#") %>%
  filter(decade == "50s") %>%
  nrow()/nrow(songs[songs$decade == "50s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "60s") %>%
  nrow()/nrow(songs[songs$decade == "60s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "70s") %>%
  nrow()/nrow(songs[songs$decade == "70s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "80s") %>%
  nrow()/nrow(songs[songs$decade == "80s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "90s") %>%
  nrow()/nrow(songs[songs$decade == "90s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "00s") %>%
  nrow()/nrow(songs[songs$decade == "00s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "10s") %>%
  nrow()/nrow(songs[songs$decade == "10s",])
songs %>% filter(keys == "F#") %>%
  filter(decade == "2020-21") %>%
  nrow()/nrow(songs[songs$decade == "2020-21",])

decades <- c("50s", "60s", "70s", "80s", "90s", "00s", "10s", "2020-21")
prop.fsharp <- c(0.03705, 0.03227945, 0.03678298, 0.04847066, 0.05893349,
                 0.06319837, 0.07285682, 0.07743724)

df <- data.frame(decades, prop.fsharp)
as.numeric(df$prop.fsharp)
df$decades <- factor(df$decades, levels = df$decades)
ggplot(df, aes(x=decades, y=prop.fsharp))+
  geom_bar(stat = "identity", col = "blue")+
  ggtitle("Sample Proportion of Songs in F# per Decade") +
  labs(y = "Proportion of Songs in F#", x = "Decade")
```

```

nrow(songs[songs$decade == "30s",])
nrow(songs[songs$decade == "40s",])
nrow(songs[songs$decade == "50s",])
nrow(songs[songs$decade == "60s",])
nrow(songs[songs$decade == "80s",])
nrow(songs[songs$decade == "90s",])
nrow(songs[songs$decade == "00s",])
nrow(songs[songs$decade == "10s",])
nrow(songs[songs$decade == "2020-21",])

nrow(songs[songs$keys == "A",])/174389
nrow(songs[songs$keys == "A#",])/174389
nrow(songs[songs$keys == "B",])/174389
nrow(songs[songs$keys == "C",])/174389
nrow(songs[songs$keys == "C#",])/174389
nrow(songs[songs$keys == "D",])/174389
nrow(songs[songs$keys == "D#",])/174389
nrow(songs[songs$keys == "E",])/174389
nrow(songs[songs$keys == "F",])/174389
nrow(songs[songs$keys == "F#",])/174389
nrow(songs[songs$keys == "G",])/174389
nrow(songs[songs$keys == "G#",])/174389

hist(index$danceability)

index<-songs[sample(nrow(songs), 5000), ]

songstrim <- songs %>%
  select(2,3,11,13,16,17,21)

mean(songstrim[["duration_ms"]])

songs$length <- as.numeric(as.character(songs$duration_ms)) / 60000
head(songs)

m<-mean(songs$tempo);std<-sqrt(var(songs$tempo))
hist(songs$tempo,prob=T, main = "Distribution of Song Tempo", xlab = "Tempo (bpm)", col =
"darkblue")
curve(dnorm(x, mean=m, sd=std), col="red", lwd=2, add=TRUE)
abline(v=mean(songs$tempo),col="red",lwd=2)

m<-mean(songs$danceability);std<-sqrt(var(songs$danceability))
hist(songs$danceability,prob=T, main = "Distribution of Song Danceability", xlab =
"Danceability", col = "yellow")
curve(dnorm(x, mean=m, sd=std), col="red", lwd=2, add=TRUE)
abline(v=mean(songs$danceability),col="red",lwd=2)

m<-mean(songs$loudness);std<-sqrt(var(songs$loudness))
hist(songs$loudness,prob=T, main = "Distribution of Song Loudness", xlab = "Tempo (dB)", col
= "green")
curve(dnorm(x, mean=m, sd=std), col="red", lwd=2, add=TRUE)
abline(v=mean(songs$loudness),col="red",lwd=2)

plot(songs$danceability, songs$popularity)

trimmed<-songs%>%select(3, 11, 17)

summary(trimmed)

songs %>% summarise(min = min(danceability))

```

```

    ,max = max(danceability)
    ,mean = mean(danceability)
    ,sd = sd(danceability)
    ,count = n()
    ,q25 = quantile(danceability, .25)
    ,q75 = quantile(danceability, .75))

songs %>% summarise(min = min(loudness)
    ,max = max(loudness)
    ,mean = mean(loudness)
    ,sd = sd(loudness)
    ,n = n()
    ,q25 = quantile(loudness, .25)
    ,q75 = quantile(loudness, .75))

songs %>% summarise(min = min(tempo)
    ,max = max(tempo)
    ,mean = mean(tempo)
    ,sd = sd(tempo)
    ,n = n()
    ,q25 = quantile(tempo, .25)
    ,q75 = quantile(tempo, .75))

ggplot(songs, aes(x=danceability, fill=factor(keys))) +
  geom_boxplot() +
  ggtitle("Danceability of Songs by Their Key") +
  labs(fill = "Key", x = "Danceability")+
  annotate(geom="text", x=0.75, y=.4, label="sample median = 0.548",
    color="blue")+
  geom_vline(xintercept = 0.548, linetype = "dashed", color = "blue") +
  theme_linedraw()

ggplot(songs, aes(danceability, fill=explicit)) +
  geom_boxplot() +
  ggtitle("Danceability of Songs by Their Explicitness")
theme_linedraw()

ggplot(songs, aes(valence, fill=factor(key))) +
  geom_boxplot() +
  ggtitle("Valence of Songs by Their Mode")

songs$explicitness[songs$explicit == 1] <- "explicit"
songs$explicitness[songs$explicit == 0] <- "clean"

nrow(songs%>%filter(explicitness == "explicit"))
nrow(songs%>%filter(explicitness == "clean"))

songs$modes[songs$mode == 1] <- "major"
songs$modes[songs$mode == 0] <- "minor"

songs$keys[songs$key == 0] <- "C"
songs$keys[songs$key == 1] <- "C#"
songs$keys[songs$key == 2] <- "D"
songs$keys[songs$key == 3] <- "D#"
songs$keys[songs$key == 4] <- "E"
songs$keys[songs$key == 5] <- "F"
songs$keys[songs$key == 6] <- "F#"
songs$keys[songs$key == 7] <- "G"
songs$keys[songs$key == 8] <- "G#"
songs$keys[songs$key == 9] <- "A"
songs$keys[songs$key == 10] <- "A#"

```

```

songs$keys[songs$key == 11] <- "B"

songs %>% group_by(year) %>%
ggplot(mapping = aes(x = year, y = mean(danceability)), alpha = 0.0001, color=decade) +
  geom_point()
trim<-songs %>% filter(year>=1950) %>%
plot(aggregate(danceability ~ year, trim, mean))

dat<-aggregate(danceability ~ year, trim, mean)

dataset<-aggregate(loudness ~ year, trim, mean)

resid(dataset$loudness)

ggplot(data = dataset, mapping = aes(x = year, y = loudness)) +
  geom_point() +
  stat_smooth(method = lm) +
  annotate(geom="text", x=2008, y=-10.5, label="y = 0.106x - 221.359",
    color="black")+
  annotate(geom="text", x=2008, y=-11, label="r = 0.9357078",
    color="black") +
  ggtitle("Linear Regression of Loudness Over the Years") +
  theme_linedraw()

table(songs$keys)
table(songs$explicitness)
table(songs$decade)

abline()
songs %>%
  ggplot(mapping = aes(x = tempo)) +
  geom_density(aes(fill = decade)) +
  facet_wrap(~ decade)

songs %>%
  ggplot(mapping = aes(x = danceability)) +
  geom_density(aes(fill = decade)) +
  facet_wrap(~ decade) +
  theme_linedraw()+
  ggtitle("Density Plots of Danceability Over the Years")

chisq.test(table(songs$decade, songs$keys))

deckey<-as.data.frame(table(songs$decade, songs$keys))

ggplot(deckey, aes(x=Var1, y=Freq, fill=factor(Var2)))+
  geom_bar(stat = "identity", position = "fill")+
  ggtitle("Distribution of Song Keys Over Past Ten Decades") +
  labs(fill = "Key", y = "Proportion", x = "Decade")

plot(table(songs$decade, songs$key))

ggplot(songs, aes(x=decade, y= key, fill=factor(key))) +
  geom_bar(stat="identity", position="fill")

table(songs$decade, songs$key)

summary(songs)

zdata_percentage <- apply(songs, 2, function(songs){x*100/sum(x,na.rm=T)})

```

```

songs$id<-songs$id[!duplicated(songs$id)]

ggplot(songs, aes(fill=key, y=nrow(songs[key]), x=decade)) +
  geom_bar(position="fill", stat="identity")

prop.table(songs$key)

ggplot(songs, aes(x = decade, fill = factor(key))) +
  geom_bar(position = "fill", stat = "identity") +
  scale_y_continuous(labels = scales::percent_format())+
  theme_linedraw()+
  ggtitle("Proportional Stacked Bars of Keys Every Decade")

library(ggpubr)

shapiro.test(index$danceability)

library(ggthemes)
library(ggthemr)
ggthemr_reset()
ggthemr::ggthemr_reset()

ggplot(data = dat, mapping = aes(x = year, y = danceability)) +
  geom_point() +
  stat_smooth(method = lm) +
  annotate(geom="text", x=1995, y=.52, label="y = 0.001865x - 3.163114",
    color="black")+
  annotate(geom="text", x=1995, y=.505, label="r = 0.9084426",
    color="black") +
  ggtitle("Linear Regression of Danceability Over the Years") +
  xlab("Average Danceability")
  theme_linedraw()

table(songs$decade)

t.test(x, y, alternative = "two.sided", var.equal = FALSE)

dance.lm<-lm(danceability ~ year, data = dat)
loud.lm<-lm(loudness ~ year, data = dataset)
dance.res<-resid(dance.lm)
loud.res<-resid(loud.lm)

plot(dat$year, dance.res,
  ylab="Residuals", xlab="Danceability",
  main="Residuals of Danceability Over the Years")
abline(0, 0)

plot(dataset$year, loud.res,
  ylab="Residuals", xlab="Loudness",
  main="Residuals of Loudness Over the Years")
abline(0, 0)

cor(dat$year, dat$danceability)
cor(dataset$year, dataset$loudness)

var<-songs %>% group_by(decade) %>%
  summarize(mean_tempo = mean(tempo), stdev = sqrt(var(tempo)))

barplot(var)

```

```
seventiestempo <- (songs %>% filter(decade == "70s"))$tempo
tenstempo <- (songs %>% filter(decade == "10s"))$tempo
t.test(tenstempo, seventiestempo, alternative = "two.sided", conf.level = .99)

explicitdanceability <- (songs %>% filter(explicit == "1"))$danceability
cleandanceability <- (songs %>% filter(explicit == "0"))$danceability
t.test(explicitdanceability, cleandanceability, alternative = "greater", conf.level = .99)

fsharpdance <- (songs %>% filter(keys == "F#"))$danceability
cdance <- (songs %>% filter(keys == "C"))$danceability
t.test(fsharpdance, cdance, alternative = "greater", conf.level = NULL)

median(songs$danceability)

songs %>% filter(decade == c("10s", "70s")) %>%
ggplot(data = songs, mapping = aes(x = tempo)) +
  geom_density(aes(col = decade))

songs %>%
ggplot(aes(tempo, fill=factor(decade))) +
  geom_boxplot()
```