DSO102 L10 workshop

Meredith Dodd <meredith.dodd@woz-u.com>

Thu 4/30/2020 3:40 PM

To: Meredith Dodd <meredith.dodd@woz-u.com>

```
# Loading in Libraries

library("ggplot2")
library("dplyr")

# Create a boxplot of the energy rating by genre

ggplot(top10s, aes(x = top.genre, y = nrgy)) +
  geom_boxplot()

# Which genres have the highest energy ratings, and which have the lowest?

top10s %>% group_by(top.genre) %>% summarise(Mean = mean(nrgy)) %>% arrange(Mean)

#Alaska Indie is the slowest.

top10s %>% group_by(top.genre) %>% summarise(Mean = mean(nrgy)) %>% arrange(desc(Mean))

#French Indie pop is the fastest.

#Find the median energy rating for all the genres for each year.

MedianDF <- top10s %>% group_by(year) %>% summarise(Median = median(nrgy))

#Plot this median value for the years 2010-2019

ggplot(MedianDF, aes(x=year, y=Median)) + geom_point()

#Create a scatter plot of energy level versus danceability for all genres and all years.

ggplot(top10s, aes(x=nrgy, y=dnce)) + geom_point() + geom_smooth(method=lm)

# Is the energy level and danceability correlated?

cor.test(top10s$nrgy, top10s$dnce, method="pearson", use="complete.obs")

#It looks like the higher the energy, the more it's danceable.  This looks significantly but slightly correlated, r=.17,
p < .05.

#Change the scatter plot to show the points for each year in a different color.

ggplot(top10s, aes(x=nrgy, y=dnce, color=year)) + geom_point() + geom_smooth(method=lm)

#Does the relationship between the two variables change over time?

# No, it does not seem to.
```

```
# Do a linear regression for those two variables.

regression <- lm(dnce ~ nrgy, top10s)
summary(regression)

# How much variability does the line explain?
#It explains 3% of the variability.

# The duration of the song is in seconds, which is hard for people to understand! Convert it to minutes and then graph it to see how the number of minutes is related to the amount of speech in the song.

DurMin <- top10s %>% mutate(Minutes = dur / 60)

ggplot(DurMin, aes(x=Minutes, y=spch)) + geom_point()

# Create a data frame with just the rows from 2010.

Songs2010 <- top10s %>% filter(year == 2010)
Songs2010Subset <- Songs2010[1:31,]

# Create a second data frame with just the rows from 2019

Songs2019 <- top10s %>% filter(year == 2019)

#Use a paired t-test to see if the amount of speech has changed over time.

t.obj <- t.test(Songs2010Subset$spch, Songs2019$spch, paired=TRUE)
t.obj
```

Meredith Dodd, Ph.D. | Data Science Program Chair and Instructor
meredith.dodd@woz-u.com
o: 480-291-8068
powered by WOZ U
https://woz-u.com