Vanessa Caldera

Final Project Update 3

R-Code Used for Looking at Popularity with Runtime

#Librarying the Packages

library(ggplot2)

library(tidyverse)

library(imputeTS)

library(readr)

#Reading the file

df <- read.csv("/Users/vanessacaldera/Desktop/Syracuse University/Intro to Data Science/Final Project/Final Project Dataset.csv", header=TRUE,stringsAsFactors = FALSE)

df$runtime <- na\_interpolation(df$runtime)

df$popularity <- na\_interpolation(df$popularity)

#Put it all together

df <- df %>% arrange(df$runtime)

#Take away the 0

df <- df[-1:-35, ]

#Mean of the original df after cleaning

mean(df$runtime)

* 107.6543

mean(df$popularity)

* 2.15165

#Creating short and long runtime subsets

shortestaruntimes <- df[df$runtime > 150, ]

longestruntimes <- df[df$runtime <90, ]

#How many movies fall in each

length(shortetruntimes$popularity)

* 671

length(longestruntimes$popularity)

* 17

#Mean of the shortest variables

mean(shortestruntimes$runtime)

* 83.77124

mean(shortestruntimes$popularity)

* 12.15499

#Mean of the longest variables

mean(longestruntimes$runtime)

* 173.6725

mean(longestruntimes$popularity)

* 42.35088

#Linear models

shortmodel <- lm(formula= popularity ~ runtime, data= shortestruntimes)

longmodel <- lm(formula=popularity ~ runtime, data=longestruntimes)

originalmodel <- lm(formula=popularity ~ runtime, data=df)

#Summary of linear models

summary(shortmodel)

* R-Squared: 0.01491

summary(longmodel)

* R-Squared: 0.009518

summary(originalmodel)

* R-Squared: 0.04889

#Plotting

ggplot(df) + aes(x=runtime, y=popularity) + geom\_point()

ggplot(shortestruntimes) + aes(x=runtime, y=popularity) + geom\_point(aes(color=popularity))

ggplot(longestruntimes) + aes(x=runtime, y=popularity) + geom\_point(aes(color=popularity))

ggplot(middleruntimes,aes(x=runtime, y=popularity)) + geom\_point(aes(color=popularity))

#Looking at the top10 most popular movies

df2 <- df %>% arrange(desc(df$popularity))

top10 <- head(df2, 10)

mean(top10$runtime)

* 123.1

Tim Tieng

Final Project Update 3

R-Code Used for Looking at How Budget Affects Popularity

# Load Packages

library(tidyverse)

library(ggplot2)

library(jsonlite)

library(readr)

library(lattice)

library(caret)

library(measures)

library(rpart)

library(rpart.plot)

library(kernlab)

library(quanteda)

library(quanteda.textplots)

# Create a dataframe from the movies.csv file

setwd("/Users/timtieng/Library/CloudStorage/OneDrive-Personal/Desktop/Masters in Applied Data Science/IST-687 Intro to Data Science/Project/TMDB 5000 Movie/Datasets")

movieFile <- "tmdb\_5000\_movies.csv"

moviesDF <- data.frame(read\_csv(movieFile))

# Inspect to see if moviesDF was successfully created

str(moviesDF) # 4803 Rows, 20 Columns

# Columns: budget, genres, homepage, id, keywords, orginal\_language, original\_title, overview, popularity, production\_companies, production\_countries, release\_date, revenue, runtime,

# spoken\_languages, status, tagline, title, vote\_average, vote\_count,

# Task 1 - Analyse how budget correlates with movie popularity

# Dependent Variable - popularity

# Independent Variable - budget

# Determine if there are any NA values in the two columns -- Expecting 0's

numNABudgetColumn <- sum(is.na(moviesDF$budget))

numNABudgetColumn

numNAPopularityColumn <- sum(is.na(moviesDF$popularity))

numNAPopularityColumn

numNABudgetColumn <- sum(is.na(moviesDF$budget))

numNABudgetColumn

# Visualize using Scatterplot

budgetVSPopularityPlot <- ggplot(moviesDF) +

geom\_point(aes(x=budget, y=popularity)) +

ggtitle("Budget Versus Popularity Rating")

budgetVSPopularityPlot

# Linear Regression Model

budgetLM <- lm(formula = popularity ~ budget, data=moviesDF)

summary(budgetLM)

# Visualize Populartiy vs additional ind. variables

genresVSPopularityPlot <- ggplot(moviesDF) +

geom\_point(aes(x=genres, y=popularity)) +

ggtitle("Genres Versus Popularity Rating")

genresVSPopularityPlot

voteAvgVSPopularityPlot <- ggplot(moviesDF) +

geom\_point(aes(x=vote\_average, y=popularity)) +

ggtitle("Vote Average Versus Popularity Rating")

voteAvgVSPopularityPlot

voteCountVSPopularityPlot <- ggplot(moviesDF) +

geom\_point(aes(x=vote\_count, y=popularity)) +

ggtitle("Vote Count Versus Popularity Rating")

voteCountVSPopularityPlot

# Linear Regrssion Model - Multi Variable analysis (budget + genres + runtime)

budgetMultiVariable <- lm(formula = popularity ~ budget + genres + runtime + vote\_average + vote\_count, data = moviesDF)

summary(budgetMultiVariable)

# Task 2 - Use another technique to determine which attributes have the most influence on the dependent variable

# Create a predictable sample

set.seed(111)

# Create the indices to use to create two additional data sets

trainList <- createDataPartition(y = moviesDF$popularity, p = .80, list = FALSE)

#Create test and train data sets

trainSet <- moviesDF[trainList, ]

testSet <- moviesDF[-trainList, ]

# Inspect

str(trainSet)

str(testSet)

dim(trainSet)

dim(testSet)

head(trainSet)

# Build treebag model

fit1 <- train(popularity ~ budget, genres, runtime, vote\_average, vote\_count, data = trainSet, method = "treebag",preProc = c("center", "scale"))

cartTree <- rpart(popularity ~ ., data = trainSet, method = "class")

prp(cartTree, faclen = 0, cex = 0.8, extra = 1)