Code

Project 2: Modeling and Evaluation

CSE6242 - Data and Visual Analytics

Due: Friday, April 21, 2017 at 11:59 PM UTC-12:00 on T-Square

ID: clightsey3 (Hunter Lightsey)

Data

We will use the same dataset as Project 1: `movies_merged` (https://s3.amazonaws.com/content.udacity-data.com/courses/gt-cs6242/project/movies_merged).

Objective

Your goal in this project is to build a linear regression model that can predict the `Gross` revenue earned by a movie based on other variables. You may use R packages to fit and evaluate a regression model (no need to implement regression yourself). Please stick to linear regression, however.

Instructions

You should be familiar with using an RMarkdown (http://rmarkdown.rstudio.com) Notebook by now. Remember that you have to open it in RStudio, and you can run code chunks by pressing *Cmd+Shift+Enter*.

Please complete the tasks below and submit this R Markdown file (as **pr2.Rmd**) containing all completed code chunks and written responses, as well as a PDF export of it (as **pr2.pdf**) which should include all of that plus output, plots and written responses for each task.

*Note that **Setup** and **Data Preprocessing** steps do not carry any points, however, they need to be completed as instructed in order to get meaningful results.*

Setup

Hide

```
# Turn off warnings
options(warn=-1)
```

Same as Project 1, load the dataset into memory:

Hide

```
setwd("/Users/C3P0/Documents/Code/OMSCS/CSE6242_DVA/Projects/Project 2")
load('movies_merged')
```

This creates an object of the same name (`movies_merged`). For convenience, you can copy it to `df` and start using it:

Hide

```
df = movies_merged
cat("Dataset has", dim(df)[1], "rows and", dim(df)[2], "columns", end="\n", file="")
```

```
Dataset has 40789 rows and 39 columns
```

Hide

```
colnames(df)
```

```
 [1] "Title"             "Year"              "Rated"             "Released"          "Ru
ntime"
 [6] "Genre"             "Director"          "Writer"            "Actors"            "Pl
ot"
[11] "Language"          "Country"           "Awards"            "Poster"            "Me
tascore"
[16] "imdbRating"        "imdbVotes"         "imdbID"            "Type"              "to
matoMeter"
[21] "tomatoImage"       "tomatoRating"      "tomatoReviews"     "tomatoFresh"       "to
matoRotten"
[26] "tomatoConsensus"   "tomatoUserMeter"   "tomatoUserRating"  "tomatoUserReviews" "to
matoURL"
[31] "DVD"               "BoxOffice"         "Production"        "Website"           "Re
sponse"
[36] "Budget"            "Domestic_Gross"    "Gross"             "Date"
```

Load R packages

Load any R packages that you will need to use. You can come back to this chunk, edit it and re-run to load any additional packages later.

Hide

```
options(warn=-1)
library(ggplot2)
library(plyr)
library(hydroGOF)
library(GGally)
library(lattice)
library(mlbench)
library(car)
library(caret)
library(tm)
```

If you are using any non-standard packages (ones that have not been discussed in class or explicitly allowed for this project), please mention them below. Include any special instructions if they cannot be installed using the regular `install.packages('<pkg name>')` command.

Non-standard packages used: None

Data Preprocessing

Before we start building models, we should clean up the dataset and perform any preprocessing steps that may be necessary. Some of these steps can be copied in from your Project 1 solution. It may be helpful to print the dimensions of the resulting dataframe at each step.

1. Remove non-movie rows

Hide

```
# TODO: Remove all rows from df that do not correspond to movies
cat("Number of Rows before Removal:", dim(df)[1], "\n")
```

```
Number of Rows before Removal: 40789
```

Hide

```
df = df[df$Type == "movie",]
cat("Number of Rows after Removal:", dim(df)[1])
```

```
Number of Rows after Removal: 40000
```

2. Drop rows with missing `Gross` value

Since our goal is to model `Gross` revenue against other variables, rows that have missing `Gross` values are not useful to us.

Hide

```
# TODO: Remove rows with missing Gross value
cat("Number of Rows before Removal:", dim(df)[1], "\n")
```

```
Number of Rows before Removal: 40000
```

Hide

```
df = df[is.na(df$Gross) == FALSE,]
cat("Number of Rows after Removal:", dim(df)[1])
```

```
Number of Rows after Removal: 4558
```

3. Exclude movies released prior to 2000

Inflation and other global financial factors may affect the revenue earned by movies during certain periods of time. Taking that into account is out of scope for this project, so let's exclude all movies that were released prior to the year 2000 (you may use `Released`, `Date` or `Year` for this purpose).

Hide

```
# TODO: Exclude movies released prior to 2000
cat("Number of Rows before Removal:", dim(df)[1], "\n")
```

```
Number of Rows before Removal: 4558
```

Hide

```
df = df[df$Year >= 2000,]
cat("Number of Rows after Removal:", dim(df)[1])
```

```
Number of Rows after Removal: 3332
```

4. Eliminate mismatched rows

Note: You may compare the `Released` column (string representation of release date) with either `Year` or `Date` (numeric representation of the year) to find mismatches. The goal is to avoid removing more than 10% of the rows.

Hide

```
# TODO: Remove mismatched rows
# Convert Released to numeric year
df$Released_Year = as.numeric(gsub("-.*$", "", df$Released))
# Compare Year and Released columns in a new column named "Date_Compare"
df$Date_Compare = df$Year >= (df$Released_Year - 1) & df$Year <= (df$Released_Year + 1)

# Compare Year and Released columns in a new column named "Date_Compare"
#df$Date_Compare = df$Year >= df$Released_Year - 1 | df$Released_Year + 1 <= df$Year
# Subset the DataFrame to keep Matches or rows with NA
cat("Number of Rows before Removal:", dim(df)[1], "\n")
```

```
Number of Rows before Removal: 3332
```

Hide

```
df = subset(df, df$Date_Compare == TRUE | is.na(df$Date_Compare) == TRUE)
cat("Number of Rows after Removal:", dim(df)[1])
```

```
Number of Rows after Removal: 3257
```

Hide

```
# Dump the filtered df back into df and get rid of Date_Compare
df$Date_Compare = NULL
df$Released_Year = NULL
```

5. Drop `Domestic_Gross` column

`Domestic_Gross` is basically the amount of revenue a movie earned within the US. Understandably, it is very highly correlated with `Gross` and is in fact equal to it for movies that were not released globally. Hence, it should be removed for modeling purposes.

Hide

```
# TODO: Exclude the `Domestic_Gross` column
cat("Number of Columns before Removal:", dim(df)[2], "\n")
```

```
Number of Columns before Removal: 39
```

Hide

```
df$Domestic_Gross = NULL
cat("Number of Columns after Removal:", dim(df)[2])
```

```
Number of Columns after Removal: 38
```

6. Process `Runtime` column

Hide

```
# TODO: Replace df$Runtime with a numeric column containing the runtime in minutes
cat("Number of Rows before Removal:", dim(df)[1], "\n")
```

```
Number of Rows before Removal: 3257
```

Hide

```
df$Runtime = gsub("N/A", NA, df$Runtime)
df$Runtime = as.numeric(gsub(" min", "", df$Runtime)) # No "hr" in this dataset so the c
onversion is simpler
cat("Number of Rows after Removal:", dim(df)[1])
```

```
Number of Rows after Removal: 3257
```

Perform any additional preprocessing steps that you find necessary, such as dealing with missing values or highly correlated columns (feel free to add more code chunks, markdown blocks and plots here as necessary).

Hide

```
# TODO(optional): Additional preprocessing
# Convert N/A to NA for df$Metascore
df$Metascore = gsub("N/A", NA, df$Metascore)
# Setup functions to check for amount of or percent of NA for numeric columns
sum_na = function(x) {sum(is.na(x))}
perc_na = function(x) {sum(is.na(x))/length(x)*100}
# Check for NA pre
perc_pre_num_convert = apply(df[,c("Year", "Runtime", "imdbRating", "imdbVotes", "tomato
Rating", "tomatoUserMeter", "tomatoUserRating", "Date", "Gross", "Metascore", "tomatoMet
er", "tomatoReviews", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")], 2,
 perc_na)
sum_pre_num_convert = apply(df[,c("Year", "Runtime", "imdbRating", "imdbVotes", "tomatoR
ating", "tomatoUserMeter", "tomatoUserRating", "Date", "Gross", "Metascore", "tomatoMete
r", "tomatoReviews", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")], 2,
sum_na)
# Convert columns to numeric that should be numeric without converting any categorical v
ariables to binary yet
df$Metascore = as.numeric(df$Metascore)
df$tomatoMeter = as.numeric(df$tomatoMeter)
df$tomatoReviews = as.numeric(df$tomatoReviews)
df$tomatoFresh = as.numeric(df$tomatoFresh)
df$tomatoRotten = as.numeric(df$tomatoRotten)
df$tomatoUserReviews = as.numeric(df$tomatoUserReviews)
df$Budget = as.numeric(df$Budget)
# Check for NA pst
sum_post_num_convert = apply(df[,c("Year", "Runtime", "imdbRating", "imdbVotes", "tomato
Rating", "tomatoUserMeter", "tomatoUserRating", "Date", "Gross", "Metascore", "tomatoMet
er", "tomatoReviews", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")], 2,
 sum_na)
perc_pst_num_convert = apply(df[,c("Year", "Runtime", "imdbRating", "imdbVotes", "tomato
Rating", "tomatoUserMeter", "tomatoUserRating", "Date", "Gross", "Metascore", "tomatoMet
er", "tomatoReviews", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")], 2,
 perc_na)
# Build a data frame for summary data
df_na_check = data.frame(Sum_Pre=sum_pre_num_convert, Sum_Pst=sum_post_num_convert, Sum_
Diff=sum_post_num_convert-sum_pre_num_convert, Perc_Pre=perc_pre_num_convert, Perc_Pst=p
erc_pst_num_convert, Perc_Diff=perc_pst_num_convert-perc_pre_num_convert)  # Create data
 frame for PRE/PST NA count comparison
print(df_na_check)
```

	Sum_Pre <int>	Sum_Pst <int>	Sum_Diff <int>	Perc_Pre <dbl>	Perc_Pst <dbl>	Perc_Diff <dbl>
Year	0	0	0	0.000000	0.000000	0
Runtime	37	37	0	1.136015	1.136015	0
imdbRating	43	43	0	1.320233	1.320233	0
imdbVotes	43	43	0	1.320233	1.320233	0
tomatoRating	374	374	0	11.482960	11.482960	0
tomatoUserMeter	188	188	0	5.772183	5.772183	0

	Sum_Pre <int>	Sum_Pst <int>	Sum_Diff <int>	Perc_Pre <dbl>	Perc_Pst <dbl>	Perc_Diff <dbl>
tomatoUserRating	186	186	0	5.710777	5.710777	0
Date	0	0	0	0.000000	0.000000	0
Gross	0	0	0	0.000000	0.000000	0
Metascore	394	394	0	12.097022	12.097022	0

1-10 of 16 rows Previous **1** 2 Next

Hide

```
# TODO(optional): Additional preprocessing - continued
# Check for highly correlated columns using ggpairs
num_pairs_plot = ggpairs(df, columns = c("imdbRating", "imdbVotes", "tomatoRating", "tom
atoUserMeter", "tomatoUserRating", "Gross", "Metascore", "tomatoMeter", "tomatoReviews",
 "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget"), columnLabels = c("imdbRa
ting", "imdbVotes", "tomatoRating", "tomatoUserMeter", "tomatoUserRating", "Gross", "Met
ascore", "tomatoMeter", "tomatoReviews", "tomatoFresh", "tomatoRotten", "tomatoUserRevie
ws", "Budget")) + theme_grey(base_size = 4)
suppressWarnings(print(num_pairs_plot))
```

```
 plot: [1,1] [---------------------------------------------------------------------------
---]  1% est: 0s
 plot: [1,2] [=--------------------------------------------------------------------------
---]  1% est:13s
 plot: [1,3] [=--------------------------------------------------------------------------
---]  2% est:13s
 plot: [1,4] [==-------------------------------------------------------------------------
---]  2% est:15s
 plot: [1,5] [==-------------------------------------------------------------------------
---]  3% est:15s
 plot: [1,6] [===------------------------------------------------------------------------
---]  4% est:15s
 plot: [1,7] [===------------------------------------------------------------------------
---]  4% est:15s
 plot: [1,8] [====-----------------------------------------------------------------------
---]  5% est:15s
 plot: [1,9] [====-----------------------------------------------------------------------
---]  5% est:15s
 plot: [1,10] [====----------------------------------------------------------------------
---]  6% est:14s
 plot: [1,11] [=====---------------------------------------------------------------------
---]  7% est:15s
 plot: [1,12] [=====---------------------------------------------------------------------
---]  7% est:14s
 plot: [1,13] [======--------------------------------------------------------------------
---]  8% est:14s
 plot: [2,1] [======---------------------------------------------------------------------
---]  8% est:14s
 plot: [2,2] [=======--------------------------------------------------------------------
---]  9% est:14s
 plot: [2,3] [=======--------------------------------------------------------------------
---]  9% est:17s
 plot: [2,4] [========-------------------------------------------------------------------
---] 10% est:17s
 plot: [2,5] [========-------------------------------------------------------------------
---] 11% est:16s
 plot: [2,6] [=========------------------------------------------------------------------
---] 11% est:16s
 plot: [2,7] [=========------------------------------------------------------------------
---] 12% est:16s
 plot: [2,8] [==========-----------------------------------------------------------------
---] 12% est:15s
 plot: [2,9] [==========-----------------------------------------------------------------
---] 13% est:15s
 plot: [2,10] [==========----------------------------------------------------------------
---] 14% est:15s
 plot: [2,11] [===========---------------------------------------------------------------
---] 14% est:15s
 plot: [2,12] [===========---------------------------------------------------------------
---] 15% est:15s
 plot: [2,13] [============--------------------------------------------------------------
---] 15% est:15s
```

```
 plot: [3,1] [============--------------------------------------------------------------
---] 16% est:14s
 plot: [3,2] [=============-------------------------------------------------------------
---] 17% est:14s
 plot: [3,3] [=============-------------------------------------------------------------
---] 17% est:14s
 plot: [3,4] [==============------------------------------------------------------------
---] 18% est:14s
 plot: [3,5] [==============------------------------------------------------------------
---] 18% est:14s
 plot: [3,6] [===============-----------------------------------------------------------
---] 19% est:14s
 plot: [3,7] [===============-----------------------------------------------------------
---] 20% est:14s
 plot: [3,8] [===============-----------------------------------------------------------
---] 20% est:13s
 plot: [3,9] [================----------------------------------------------------------
---] 21% est:13s
 plot: [3,10] [================---------------------------------------------------------
---] 21% est:13s
 plot: [3,11] [=================--------------------------------------------------------
---] 22% est:13s
 plot: [3,12] [=================--------------------------------------------------------
---] 22% est:13s
 plot: [3,13] [==================-------------------------------------------------------
---] 23% est:13s
 plot: [4,1] [==================--------------------------------------------------------
---] 24% est:13s
 plot: [4,2] [===================-------------------------------------------------------
---] 24% est:13s
 plot: [4,3] [===================-------------------------------------------------------
---] 25% est:12s
 plot: [4,4] [====================------------------------------------------------------
---] 25% est:12s
 plot: [4,5] [====================------------------------------------------------------
---] 26% est:12s
 plot: [4,6] [=====================-----------------------------------------------------
---] 27% est:12s
 plot: [4,7] [=====================-----------------------------------------------------
---] 27% est:12s
 plot: [4,8] [=====================-----------------------------------------------------
---] 28% est:12s
 plot: [4,9] [======================----------------------------------------------------
---] 28% est:12s
 plot: [4,10] [======================---------------------------------------------------
---] 29% est:11s
 plot: [4,11] [======================---------------------------------------------------
---] 30% est:11s
 plot: [4,12] [=======================--------------------------------------------------
---] 30% est:11s
 plot: [4,13] [=======================--------------------------------------------------
---] 31% est:11s
 plot: [5,1] [========================--------------------------------------------------
---] 31% est:11s
```

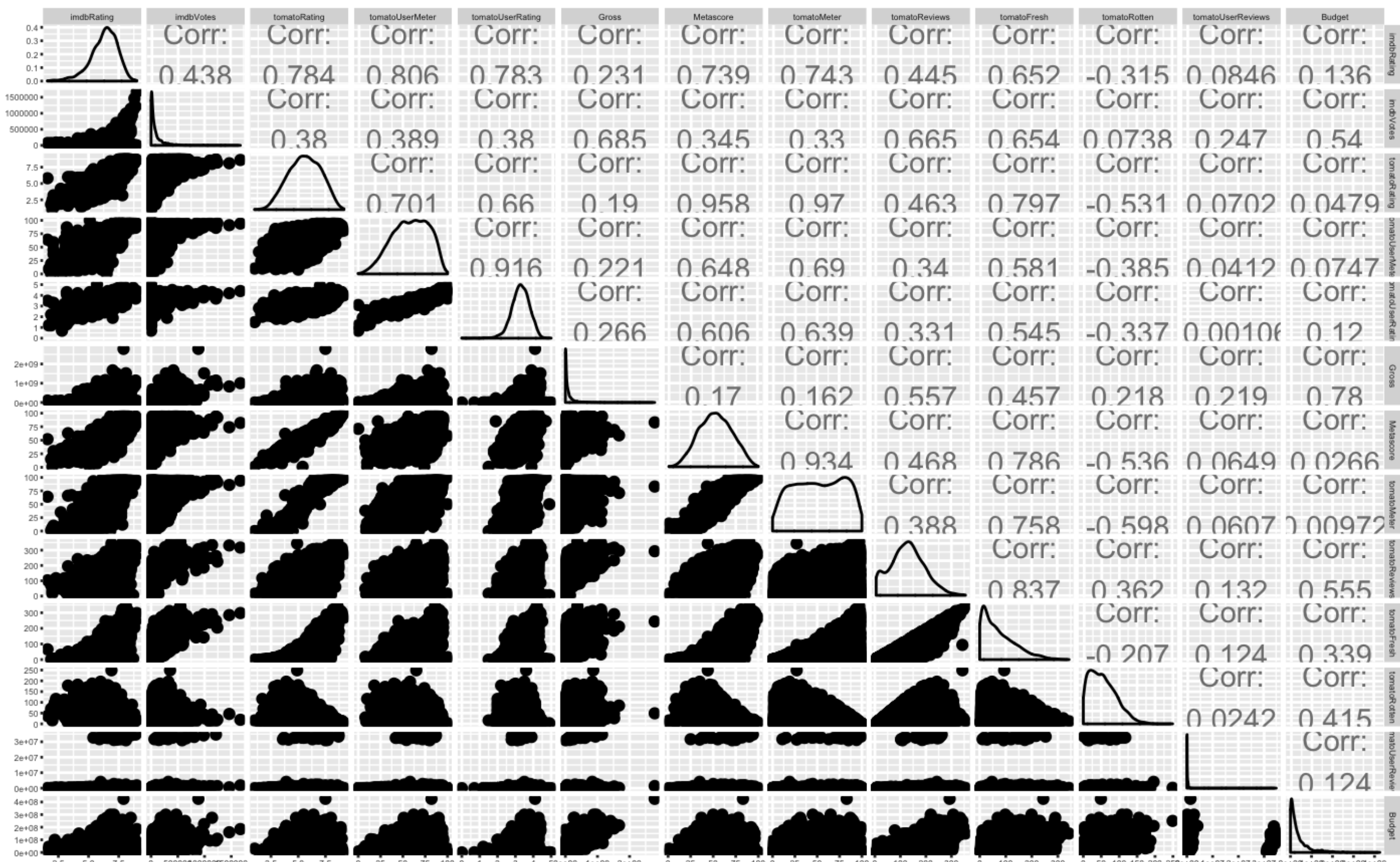
```
 plot: [5,2] [========================-------------------------------------------------
---] 32% est:11s
 plot: [5,3] [========================-------------------------------------------------
---] 33% est:11s
 plot: [5,4] [=========================------------------------------------------------
---] 33% est:11s
 plot: [5,5] [=========================------------------------------------------------
---] 34% est:11s
 plot: [5,6] [=========================------------------------------------------------
---] 34% est:10s
 plot: [5,7] [==========================-----------------------------------------------
---] 35% est:10s
 plot: [5,8] [==========================-----------------------------------------------
---] 36% est:10s
 plot: [5,9] [===========================----------------------------------------------
---] 36% est:10s
 plot: [5,10] [===========================---------------------------------------------
---] 37% est:10s
 plot: [5,11] [===========================---------------------------------------------
---] 37% est:10s
 plot: [5,12] [============================--------------------------------------------
---] 38% est:10s
 plot: [5,13] [============================--------------------------------------------
---] 38% est:10s
 plot: [6,1] [=============================--------------------------------------------
---] 39% est:10s
 plot: [6,2] [==============================-------------------------------------------
---] 40% est:10s
 plot: [6,3] [==============================-------------------------------------------
---] 40% est: 9s
 plot: [6,4] [==============================-------------------------------------------
---] 41% est: 9s
 plot: [6,5] [===============================------------------------------------------
---] 41% est: 9s
 plot: [6,6] [===============================------------------------------------------
---] 42% est: 9s
 plot: [6,7] [================================-----------------------------------------
---] 43% est: 9s
 plot: [6,8] [================================-----------------------------------------
---] 43% est: 9s
 plot: [6,9] [=================================----------------------------------------
---] 44% est: 9s
 plot: [6,10] [=================================---------------------------------------
---] 44% est: 9s
 plot: [6,11] [=================================---------------------------------------
---] 45% est: 9s
 plot: [6,12] [==================================--------------------------------------
---] 46% est: 9s
 plot: [6,13] [==================================--------------------------------------
---] 46% est: 9s
 plot: [7,1] [===================================--------------------------------------
---] 47% est: 8s
 plot: [7,2] [===================================--------------------------------------
---] 47% est: 8s
```

```
 plot: [7,3] [=====================================--------------------------------------
---] 48% est: 8s
 plot: [7,4] [=====================================--------------------------------------
---] 49% est: 8s
 plot: [7,5] [======================================-------------------------------------
---] 49% est: 8s
 plot: [7,6] [======================================-------------------------------------
---] 50% est: 8s
 plot: [7,7] [=======================================------------------------------------
---] 50% est: 8s
 plot: [7,8] [=======================================------------------------------------
---] 51% est: 8s
 plot: [7,9] [========================================-----------------------------------
---] 51% est: 8s
 plot: [7,10] [========================================----------------------------------
---] 52% est: 8s
 plot: [7,11] [========================================----------------------------------
---] 53% est: 7s
 plot: [7,12] [========================================----------------------------------
---] 53% est: 7s
 plot: [7,13] [=========================================---------------------------------
---] 54% est: 7s
 plot: [8,1] [==========================================---------------------------------
---] 54% est: 7s
 plot: [8,2] [==========================================---------------------------------
---] 55% est: 7s
 plot: [8,3] [===========================================--------------------------------
---] 56% est: 7s
 plot: [8,4] [===========================================--------------------------------
---] 56% est: 7s
 plot: [8,5] [============================================-------------------------------
---] 57% est: 7s
 plot: [8,6] [============================================-------------------------------
---] 57% est: 7s
 plot: [8,7] [=============================================------------------------------
---] 58% est: 7s
 plot: [8,8] [=============================================------------------------------
---] 59% est: 6s
 plot: [8,9] [==============================================-----------------------------
---] 59% est: 6s
 plot: [8,10] [=============================================-----------------------------
---] 60% est: 6s
 plot: [8,11] [==============================================----------------------------
---] 60% est: 6s
 plot: [8,12] [==============================================----------------------------
---] 61% est: 6s
 plot: [8,13] [===============================================---------------------------
---] 62% est: 6s
 plot: [9,1] [================================================---------------------------
---] 62% est: 6s
 plot: [9,2] [================================================---------------------------
---] 63% est: 6s
 plot: [9,3] [=================================================--------------------------
---] 63% est: 6s
```

```
 plot: [9,4] [==================================================-------------------------
---] 64% est: 6s
 plot: [9,5] [===================================================------------------------
---] 64% est: 6s
 plot: [9,6] [===================================================------------------------
---] 65% est: 5s
 plot: [9,7] [====================================================-----------------------
---] 66% est: 5s
 plot: [9,8] [====================================================-----------------------
---] 66% est: 5s
 plot: [9,9] [====================================================-----------------------
---] 67% est: 5s
 plot: [9,10] [====================================================----------------------
---] 67% est: 5s
 plot: [9,11] [=====================================================---------------------
---] 68% est: 5s
 plot: [9,12] [=====================================================---------------------
---] 69% est: 5s
 plot: [9,13] [======================================================--------------------
---] 69% est: 5s
 plot: [10,1] [======================================================--------------------
---] 70% est: 5s
 plot: [10,2] [=======================================================-------------------
---] 70% est: 5s
 plot: [10,3] [=======================================================-------------------
---] 71% est: 4s
 plot: [10,4] [=======================================================-------------------
---] 72% est: 4s
 plot: [10,5] [========================================================------------------
---] 72% est: 4s
 plot: [10,6] [========================================================------------------
---] 73% est: 4s
 plot: [10,7] [=========================================================-----------------
---] 73% est: 4s
 plot: [10,8] [=========================================================-----------------
---] 74% est: 4s
 plot: [10,9] [==========================================================----------------
---] 75% est: 4s
 plot: [10,10] [=========================================================----------------
---] 75% est: 4s
 plot: [10,11] [==========================================================---------------
---] 76% est: 4s
 plot: [10,12] [==========================================================---------------
---] 76% est: 4s
 plot: [10,13] [===========================================================--------------
---] 77% est: 4s
 plot: [11,1] [============================================================--------------
---] 78% est: 3s
 plot: [11,2] [============================================================--------------
---] 78% est: 3s
 plot: [11,3] [=============================================================-------------
---] 79% est: 3s
 plot: [11,4] [=============================================================-------------
---] 79% est: 3s
```

```
 plot: [11,5] [=============================================================------------
---] 80% est: 3s
 plot: [11,6] [=============================================================------------
---] 80% est: 3s
 plot: [11,7] [==============================================================-----------
---] 81% est: 3s
 plot: [11,8] [==============================================================-----------
---] 82% est: 3s
 plot: [11,9] [===============================================================----------
---] 82% est: 3s
 plot: [11,10] [==============================================================----------
---] 83% est: 3s
 plot: [11,11] [===============================================================---------
---] 83% est: 3s
 plot: [11,12] [===============================================================---------
---] 84% est: 2s
 plot: [11,13] [===============================================================---------
---] 85% est: 2s
 plot: [12,1] [=================================================================--------
---] 85% est: 2s
 plot: [12,2] [=================================================================--------
---] 86% est: 2s
 plot: [12,3] [==================================================================-------
---] 86% est: 2s
 plot: [12,4] [==================================================================-------
---] 87% est: 2s
 plot: [12,5] [===================================================================------
---] 88% est: 2s
 plot: [12,6] [===================================================================------
---] 88% est: 2s
 plot: [12,7] [===================================================================------
---] 89% est: 2s
 plot: [12,8] [====================================================================-----
---] 89% est: 2s
 plot: [12,9] [====================================================================-----
---] 90% est: 2s
 plot: [12,10] [====================================================================----
---] 91% est: 1s
 plot: [12,11] [====================================================================----
---] 91% est: 1s
 plot: [12,12] [=====================================================================---
---] 92% est: 1s
 plot: [12,13] [=====================================================================---
---] 92% est: 1s
 plot: [13,1] [=======================================================================--
---] 93% est: 1s
 plot: [13,2] [=======================================================================--
---] 93% est: 1s
 plot: [13,3] [========================================================================-
---] 94% est: 1s
 plot: [13,4] [========================================================================-
---] 95% est: 1s
 plot: [13,5] [========================================================================-
---] 95% est: 1s
```

```
 plot: [13,6] [=============================================================================
---] 96% est: 1s
 plot: [13,7] [=============================================================================
---] 96% est: 1s
 plot: [13,8] [=============================================================================
=--] 97% est: 0s
 plot: [13,9] [=============================================================================
=--] 98% est: 0s
 plot: [13,10] [============================================================================
==-] 98% est: 0s
 plot: [13,11] [============================================================================
==-] 99% est: 0s
 plot: [13,12] [============================================================================
===] 99% est: 0s
 plot: [13,13] [============================================================================
===]100% est: 0s
```



_**Note**: The following observations were made from the above ggpairs plot and the df_na_check data frame: - tomatoUserRating (5.7% NA) and tomatoUserMeter (5.7% NA) are very highly correlated with a coefficient of 0.91. - tomatoMeter (11.5% NA) and tomatoRating (11.5% NA) are very highly correlated with a coefficient of 0.97. - Metascore (12.1% NA) and tomatoRating (11.5% NA) are very highly correlated with a coefficient of 0.96. - tomatoMeter (11.5% NA) and Metascore (12.1% NA) are very highly correlated with a coefficient of 0.93. - tomatoUserRating will be used for modeling and tomatoUserMeter will be dropped. - tomatoMeter will be used for modeling, and tomatoRating will be dropped. - Metascore will not be dropped as well due to high correlation and high percent NA. - Date is highly correlated to Year, so Year will be used for modeling and Date will be dropped. - tomatoReviews is actually just tomatoFresh + tomatoRotten, so tomatoReviews will also be dropped.

Hide

```
# TODO(optional): Additional preprocessing - continued
# Drop highly correlated columns
df$tomatoUserMeter = NULL
df$tomatoRating = NULL
df$Metascore = NULL
df$Date = NULL
df$tomatoReviews = NULL
# Check NA count and percent again
sum_num_convert = apply(df[,c("Gross", "Year", "Runtime", "imdbRating", "imdbVotes", "tomatoUserRating", "tomatoMeter", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")], 2, sum_na)
perc_num_convert = apply(df[,c("Gross", "Year", "Runtime", "imdbRating", "imdbVotes", "tomatoUserRating", "tomatoMeter", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")], 2, perc_na)
df_na_check = data.frame(Sum=sum_num_convert, Perc=perc_num_convert)  # Create data frame for PRE/PST NA count comparison
print(df_na_check)
```

	Sum <int>	Perc <dbl>
Gross	0	0.000000
Year	0	0.000000
Runtime	37	1.136015
imdbRating	43	1.320233
imdbVotes	43	1.320233
tomatoUserRating	186	5.710777
tomatoMeter	374	11.482960
tomatoFresh	374	11.482960
tomatoRotten	374	11.482960
tomatoUserReviews	84	2.579060

1-10 of 11 rows Previous 1 2 Next

Hide

```
# TODO(optional): Additional preprocessing - continued
# Create a data frame of useable numeric variables as defined in the code above
df_num_init = df[,c("Gross", "Year", "Runtime", "imdbRating", "imdbVotes", "tomatoUserRa
ting", "tomatoMeter", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")]
df_num = df[,c("Gross", "Year", "Runtime", "imdbRating", "imdbVotes", "tomatoUserRating"
, "tomatoMeter", "tomatoFresh", "tomatoRotten", "tomatoUserReviews", "Budget")]
# Deal with NA values via imputation based on column medians (assume data is MCAR). Gros
s, Year, and Budget have no na values.
medians = apply(df_num, 2, median, na.rm=TRUE)
df_num$Runtime[is.na(df_num$Runtime) == TRUE] = medians[["Runtime"]]
df_num$imdbRating[is.na(df_num$imdbRating) == TRUE] = medians[["imdbRating"]]
df_num$imdbVotes[is.na(df_num$imdbVotes) == TRUE] = medians[["imdbVotes"]]
df_num$tomatoUserRating[is.na(df_num$tomatoUserRating) == TRUE] = medians[["tomatoUserRa
ting"]]
df_num$tomatoMeter[is.na(df_num$tomatoMeter) == TRUE] = medians[["tomatoMeter"]]
df_num$tomatoFresh[is.na(df_num$tomatoFresh) == TRUE] = medians[["tomatoFresh"]]
df_num$tomatoRotten[is.na(df_num$tomatoRotten) == TRUE] = medians[["tomatoRotten"]]
df_num$tomatoUserReviews[is.na(df_num$tomatoUserReviews) == TRUE] = medians[["tomatoUser
Reviews"]]
df_num$Runtime[is.na(df_num$Runtime) == TRUE] = medians[["Runtime"]]
# Check NA count and percent again
sum_num_convert = apply(df_num[,c("Gross", "Year", "Runtime", "imdbRating", "imdbVotes",
 "tomatoUserRating", "tomatoMeter", "tomatoFresh", "tomatoRotten", "tomatoUserReviews",
"Budget")], 2, sum_na)
perc_num_convert = apply(df_num[,c("Gross", "Year", "Runtime", "imdbRating", "imdbVotes"
, "tomatoUserRating", "tomatoMeter", "tomatoFresh", "tomatoRotten", "tomatoUserReviews",
 "Budget")], 2, perc_na)
df_na_check = data.frame(Sum=sum_num_convert, Perc=perc_num_convert)  # Create data fram
e for PRE/PST NA count comparison
print(df_na_check)
```

	Sum <int>	Perc <dbl>
Gross	0	0
Year	0	0
Runtime	0	0
imdbRating	0	0
imdbVotes	0	0
tomatoUserRating	0	0
tomatoMeter	0	0
tomatoFresh	0	0
tomatoRotten	0	0
tomatoUserReviews	0	0

1-10 of 11 rows Previous **1** 2 Next

***Note**: Do NOT convert categorical variables (like `Genre`) into binary columns yet. You will do that later as part of a model improvement task.*

Final preprocessed dataset

Report the dimensions of the preprocessed dataset you will be using for modeling and evaluation, and print all the final column names. (Again, `Domestic_Gross` should not be in this list!)

Hide

```
# TODO: Print the dimensions of the final preprocessed dataset and column names
cat("Final number of rows/movies in data frame: ", dim(df)[1], "\n")
```

```
Final number of rows/movies in data frame:  3257
```

Hide

```
cat("Final number of columns/parameters in data frame: ", dim(df)[2], "\n")
```

```
Final number of columns/parameters in data frame:  33
```

Hide

```
colnames(df)
```

```
 [1] "Title"              "Year"               "Rated"              "Released"           "Ru
ntime"
 [6] "Genre"              "Director"           "Writer"             "Actors"             "Pl
ot"
[11] "Language"           "Country"            "Awards"             "Poster"             "im
dbRating"
[16] "imdbVotes"          "imdbID"             "Type"               "tomatoMeter"        "to
matoImage"
[21] "tomatoFresh"        "tomatoRotten"       "tomatoConsensus"    "tomatoUserRating"   "to
matoUserReviews"
[26] "tomatoURL"          "DVD"                "BoxOffice"          "Production"         "We
bsite"
[31] "Response"           "Budget"             "Gross"
```

Evaluation Strategy

In each of the tasks described in the next section, you will build a regression model. In order to compare their performance, use the following evaluation procedure every time:

1. Randomly divide the rows into two sets of sizes 5% and 95%.
2. Use the first set for training and the second for testing.
3. Compute the Root Mean Squared Error (RMSE) on the train and test sets.
4. Repeat the above data partition and model training and evaluation 10 times and average the RMSE results so the results stabilize.

5. Repeat the above steps for different proportions of train and test sizes: 10%-90%, 15%-85%, …, 95%-5% (total 19 splits including the initial 5%-95%).
6. Generate a graph of the averaged train and test RMSE as a function of the train set size (%).

You can define a helper function that applies this procedure to a given model and reuse it.

Hide

```
# Build helper functions
# Function for splitting into train and test samples 1x
random.split = function(df, train_perc = 0.05) {
  random_sample = sample(1:nrow(df), train_perc*nrow(df))
  train = df[random_sample,]
  test = df[-random_sample,]
  df_list = list(train, test)
  return(df_list)
}
calc.RMSE = function(train, test) {
  train_rmse = vector()
  for (i in 1:nrow(train)) {
    train_rmse = append(train_rmse, rmse(train$Gross_Prediction[i], train$Gross[i]))
  }
  test_rmse = vector()
  for (i in 1:nrow(test)) {
    test_rmse = append(test_rmse, rmse(test$Gross_Prediction[i], test$Gross[i]))
  }
  rmse_list = list(train_rmse, test_rmse)
  return(rmse_list)
}
run.lm = function(df, train_perc = 0.05, model) {
  # Parts 1-3 for a single train/test dataset
  # Part 1 - Randomly divide the rows into two sets of sizes 5% and 95%
  tt = random.split(df, train_perc)
  train = tt[[1]]
  test = tt[[2]]
  # Part 2 - Use the first set for training and the second for testing
  M = lm(formula(model), train)
  train_pred = predict(M, train[,-which(names(train) == "Gross")])  # Predict train valu
es
  test_pred = predict(M, test[,-which(names(test) == "Gross")])  # Predict test values
  train$Gross_Prediction = as.vector(train_pred)
  test$Gross_Prediction = as.vector(test_pred)
  # Part 3 - Compute the Root Mean Squared Error (RMSE) on the train and test sets
  tt_rmse = calc.RMSE(train, test)
  train$Train_RMSE = tt_rmse[[1]]
  test$Test_RMSE = tt_rmse[[2]]

  final_list = list(train, test, M)
  return(final_list)
}
rep.run.lm = function(df, train_perc, model, rep_times=10) {

  # Initialize Train and Test df from run.lm
  tt_init = run.lm(df, train_perc, model)
  train = tt_init[[1]]
  train$r_squared = summary(tt_init[[3]])$r.squared
  train$adj_r_squared = summary(tt_init[[3]])$adj.r.squared
  test = tt_init[[2]]
  test$r_squared = summary(tt_init[[3]])$r.squared
  test$adj_r_squared = summary(tt_init[[3]])$adj.r.squared
```

```
  # Initialize model list
  m_list = list()
  m_list = append(m_list, tt_init[3])

  # Rerun run.lm by rep_times and add to Train and Test df (also add each model to m_lis
t)
  for (i in 1:(rep_times-1)) {
    tt_temp = run.lm(df, train_perc, model)
    tt_temp[[1]]$r_squared = summary(tt_temp[[3]])$r.squared
    tt_temp[[1]]$adj_r_squared = summary(tt_temp[[3]])$adj.r.squared
    tt_temp[[2]]$r_squared = summary(tt_temp[[3]])$r.squared
    tt_temp[[2]]$adj_r_squared = summary(tt_temp[[3]])$adj.r.squared
    train = rbind(train, tt_temp[[1]])
    test = rbind(test, tt_temp[[2]])
    m_list = append(m_list, tt_temp[3])
  }

  # Create new columns for avg r-squared, avg adjusted r-squared and avg rmse for Train
 and Test df
  train$R_Squared_Avg = mean(train$r_squared)
  test$R_Squared_Avg = mean(test$r_squared)
  train$Adj_R_Squared_Avg = mean(train$adj_r_squared)
  test$Adj_R_Squared_Avg = mean(test$adj_r_squared)
  train$RMSE_Avg = mean(train$Train_RMSE, na.rm = TRUE)
  test$RMSE_Avg = mean(test$Test_RMSE, na.rm = TRUE)

  final_list = list(train, test, m_list)
}
build.plot.df = function(tt_list, n_mod) {
  # Build a data frame to be used for plotting all models on one plot
  train_seq = seq(1,37,2)
  test_seq = seq(2,38,2)

  avg_train_rsquared = vector()
  for (b in train_seq) {avg_train_rsquared = append(avg_train_rsquared, tt_list[[b]]$R_S
quared_Avg[1])}

  avg_test_rsquared = vector()
  for (c in test_seq) {avg_test_rsquared = append(avg_test_rsquared, tt_list[[c]]$R_Squa
red_Avg[1])}

  avg_train_adjrsquared = vector()
  for (d in train_seq) {avg_train_adjrsquared = append(avg_train_adjrsquared, tt_list[[d
]]$Adj_R_Squared_Avg[1])}

  avg_test_adjrsquared = vector()
  for (e in test_seq) {avg_test_adjrsquared = append(avg_test_adjrsquared, tt_list[[e]]$
Adj_R_Squared_Avg[1])}

  avg_train_rmse = vector()
  for (f in train_seq) {avg_train_rmse = append(avg_train_rmse, tt_list[[f]]$RMSE_Avg[1]
)}

  avg_test_rmse = vector()
```

```
  for (g in test_seq) {avg_test_rmse = append(avg_test_rmse, tt_list[[g]]$RMSE_Avg[1])}

  train_partition_size = vector()
  for (h in train_seq) {train_partition_size = append(train_partition_size, dim(tt_list[
[h]])[1] / 10)}

  plot_df_train = data.frame(Partition_Train_Group=as.character(seq(0.05,0.95,0.05)), Pa
rtition_Type=rep("Train", times=19), Partition_Type_M=rep("Train - Model 1", times=19),
Model_Number=rep("Model 1", times=19), Train_Partition_Size=train_partition_size, Avg_R_
Squared=avg_train_rsquared, Avg_Adj_R_Squared=avg_train_adjrsquared, Avg_RMSE=avg_train_
rmse)

  plot_df_test = data.frame(Partition_Train_Group=as.character(seq(0.05,0.95,0.05)), Par
tition_Type=rep("Test", times=19), Partition_Type_M=rep("Test - Model 1", times=19), Mod
el_Number=rep("Model 1", times=19), Train_Partition_Size=train_partition_size, Avg_R_Squ
ared=avg_test_rsquared, Avg_Adj_R_Squared=avg_test_adjrsquared, Avg_RMSE=avg_test_rmse)

  plot_df = rbind(plot_df_train, plot_df_test)

  n = 39

  if (n_mod == 1) {return(plot_df)} else {

    for (i in 2:n_mod) {
      tt_list_temp = list()
      tt_list_temp = append(tt_list_temp, tt_list[n:(i*38)])
      n = 1 + i*38
      avg_train_rsquared = vector()
      for (j in train_seq) {avg_train_rsquared = append(avg_train_rsquared, tt_list_temp
[[j]]$R_Squared_Avg[1])}

      avg_test_rsquared = vector()
      for (k in test_seq) {avg_test_rsquared = append(avg_test_rsquared, tt_list_temp[[k
]]$R_Squared_Avg[1])}

      avg_train_adjrsquared = vector()
      for (l in train_seq) {avg_train_adjrsquared = append(avg_train_adjrsquared, tt_lis
t_temp[[l]]$Adj_R_Squared_Avg[1])}

      avg_test_adjrsquared = vector()
      for (m in test_seq) {avg_test_adjrsquared = append(avg_test_adjrsquared, tt_list_t
emp[[m]]$Adj_R_Squared_Avg[1])}

      avg_train_rmse = vector()
      for (o in train_seq) {avg_train_rmse = append(avg_train_rmse, tt_list_temp[[o]]$RM
SE_Avg[1])}
      avg_test_rmse = vector()
      for (p in test_seq) {avg_test_rmse = append(avg_test_rmse, tt_list_temp[[p]]$RMSE_
Avg[1])}
      train_partition_size = vector()
      for (q in train_seq) {train_partition_size = append(train_partition_size, dim(tt_l
ist_temp[[q]])[1] / 10)}
      plot_df_train = data.frame(Partition_Train_Group=as.character(seq(0.05,0.95,0.05))
, Partition_Type=rep("Train", times=19), Partition_Type_M=rep(paste("Train - Model", i),
```

```
 times=19), Model_Number=rep(paste("Model", i), times=19), Train_Partition_Size=train_pa
rtition_size, Avg_R_Squared=avg_train_rsquared, Avg_Adj_R_Squared=avg_train_adjrsquared,
 Avg_RMSE=avg_train_rmse)
      plot_df_test = data.frame(Partition_Train_Group=as.character(seq(0.05,0.95,0.05)),
 Partition_Type=rep("Test", times=19), Partition_Type_M=rep(paste("Test - Model", i), ti
mes=19), Model_Number=rep(paste("Model", i), times=19), Train_Partition_Size=train_parti
tion_size, Avg_R_Squared=avg_test_rsquared, Avg_Adj_R_Squared=avg_test_adjrsquared, Avg_
RMSE=avg_test_rmse)

      plot_df = rbind(plot_df, plot_df_train)
      plot_df = rbind(plot_df, plot_df_test)
    }
    return(plot_df)
  }
}
plot.lm = function(plot_df, rmse_x_breaks, rmse_plot_title, adjrsquared_x_breaks, adjrsq
uared_plot_title, n_mod = 3) {

  # plot all models on one plot
  if (n_mod < 4) {
    rmse_plot = ggplot(plot_df, aes(x=Train_Partition_Size, y=Avg_RMSE, color=Model_Numb
er, shape=Partition_Type)) + geom_point(size=4) + geom_line() + labs(x="Train Set Size",
 y="Average RMSE") + scale_x_continuous(breaks=rmse_x_breaks) + labs(title=rmse_plot_tit
le)
    adjrsquared_plot = ggplot(plot_df, aes(x=Train_Partition_Size, y=Avg_Adj_R_Squared,
color=Model_Number, shape=Partition_Type)) + geom_point(size=4) + geom_line() + labs(x=
"Train Set Size", y="Average Adjusted R-Squared") + scale_x_continuous(breaks=adjrsquare
d_x_breaks) + labs(title=adjrsquared_plot_title)
  } else {
    rmse_plot = ggplot(plot_df, aes(x=Train_Partition_Size, y=Avg_RMSE, color=Model_Numb
er, shape=Partition_Type)) + geom_point(size=4) + geom_line() + labs(x="Train Set Size",
 y="Average RMSE") + scale_x_continuous(breaks=rmse_x_breaks) + labs(title=rmse_plot_tit
le)
    adjrsquared_plot = ggplot(plot_df, aes(x=Train_Partition_Size, y=Avg_Adj_R_Squared,
color=Model_Number)) + geom_point(size=4) + geom_line() + labs(x="Train Set Size", y="Av
erage Adjusted R-Squared") + scale_x_continuous(breaks=adjrsquared_x_breaks) + labs(titl
e=adjrsquared_plot_title)
  }

  print(rmse_plot)
  print(adjrsquared_plot)

  plot_list = list(rmse_plot, adjrsquared_plot)

  return(plot_list)
}
```

Tasks

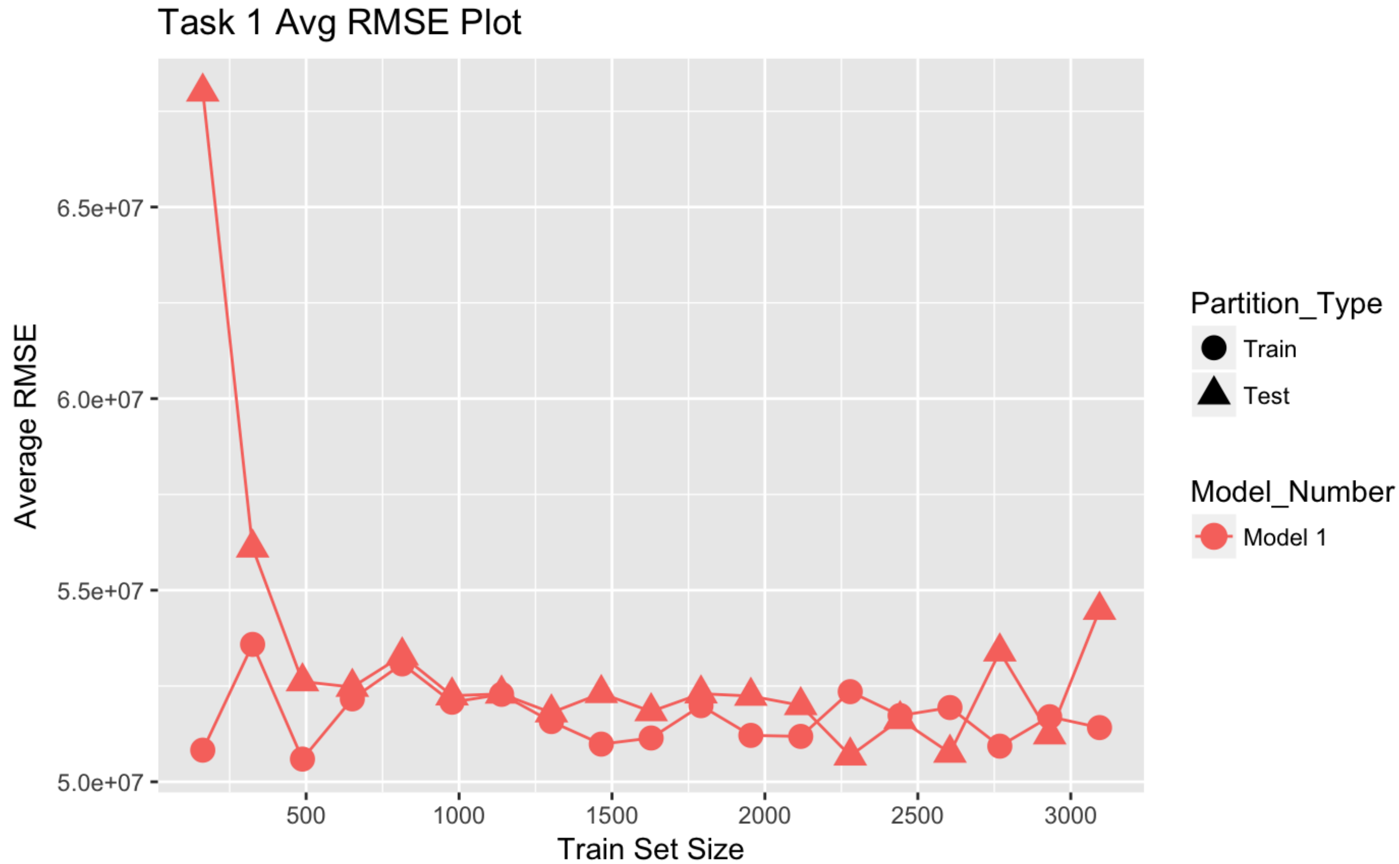
Each of the following tasks is worth 20 points. Remember to build each model as specified, evaluate it using the strategy outlined above, and plot the training and test errors by training set size (%).

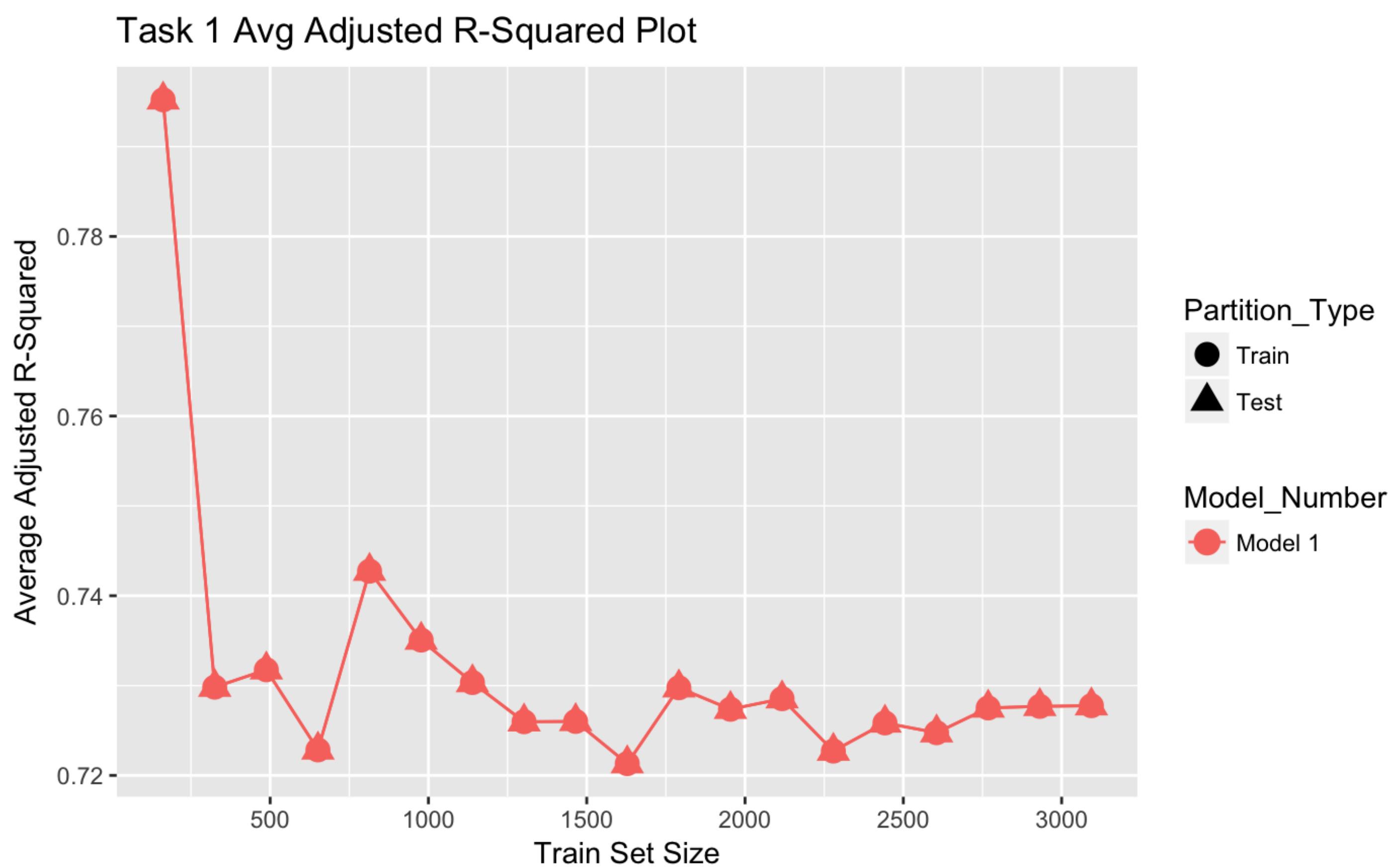
1. Numeric variables

Use linear regression to predict `Gross` based on all available *numeric* variables.

Hide

```
# TODO: Build & evaluate model 1 (numeric variables only)
### Pre-processing for task #1
df1 = df_num  # Keep only columns with numeric data for this task
### Evaluation Part 1 - Part 5 - Build a list of train and test dataframes for each of t
he 19 partitions, where each partition is repeated 10x and the average RMSE is obtained
partition_list = list(0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.7
5,0.8,0.85,0.9,0.95)  # List to define train partitions
# Task 1, Model 1 - rank deficiency warnings obtained in model 1 (t1_m1)
t1_m1 = Gross~.
t1_m1_tt_list = list()
t1_m1_model_list = list()
for (i in partition_list) {
  t1_tt_temp = rep.run.lm(df1, i, t1_m1)
  t1_m1_tt_list = append(t1_m1_tt_list, t1_tt_temp[1:2])
  t1_m1_model_list = append(t1_m1_model_list, t1_tt_temp[3])
}
# Task 1, Model 2 - tomatoReviews = tomatoFresh + tomatoRotten, so tomatoReviews was dro
pped in model 2 (t1_m2) and rank deficiency warnings were not obtained using this model
#t1_m2 =
#t1_m2_tt_list = list()
#t1_m2_model_list = list()
#for (i in partition_list) {
#  t1_tt_temp = rep.run.lm(df1, i, t1_m2)
#  t1_m2_tt_list = append(t1_m2_tt_list, t1_tt_temp[1:2])
#  t1_m2_model_list = append(t1_m2_model_list, t1_tt_temp[3])
#}
### Evauation Part 6 - Plot all models on one plot
t1_models = list(t1_m1_tt_list)  # List of models for this task
t1_n_models = length(t1_models)
# Create a large list of all partitions of train and test for all models
t1_mall_tt_list = list()
for (i in t1_models) {t1_mall_tt_list = append(t1_mall_tt_list, i)}
# Create the final plot df
t1_plot_df = build.plot.df(t1_mall_tt_list, t1_n_models)
# Plot all the models
t1_plot = plot.lm(t1_plot_df, seq(0,3000,500), "Task 1 Avg RMSE Plot", seq(0,3000,500),
"Task 1 Avg Adjusted R-Squared Plot", t1_n_models)
```





Q: List all the numeric variables you used.

A: “Year”, “Runtime”, “Budget”, “imdbRating”, “imdbVotes”, “tomatoUserRating”, “tomatoMeter”, “tomatoFresh”, “tomatoRotten”, “tomatoUserReviews”

2. Feature transformations

Try to improve the prediction quality from **Task 1** as much as possible by adding feature transformations of the numeric variables. Explore both numeric transformations such as power transforms and non-numeric transformations of the numeric variables like binning (e.g. `is_budget_greater_than_3M`).

Hide

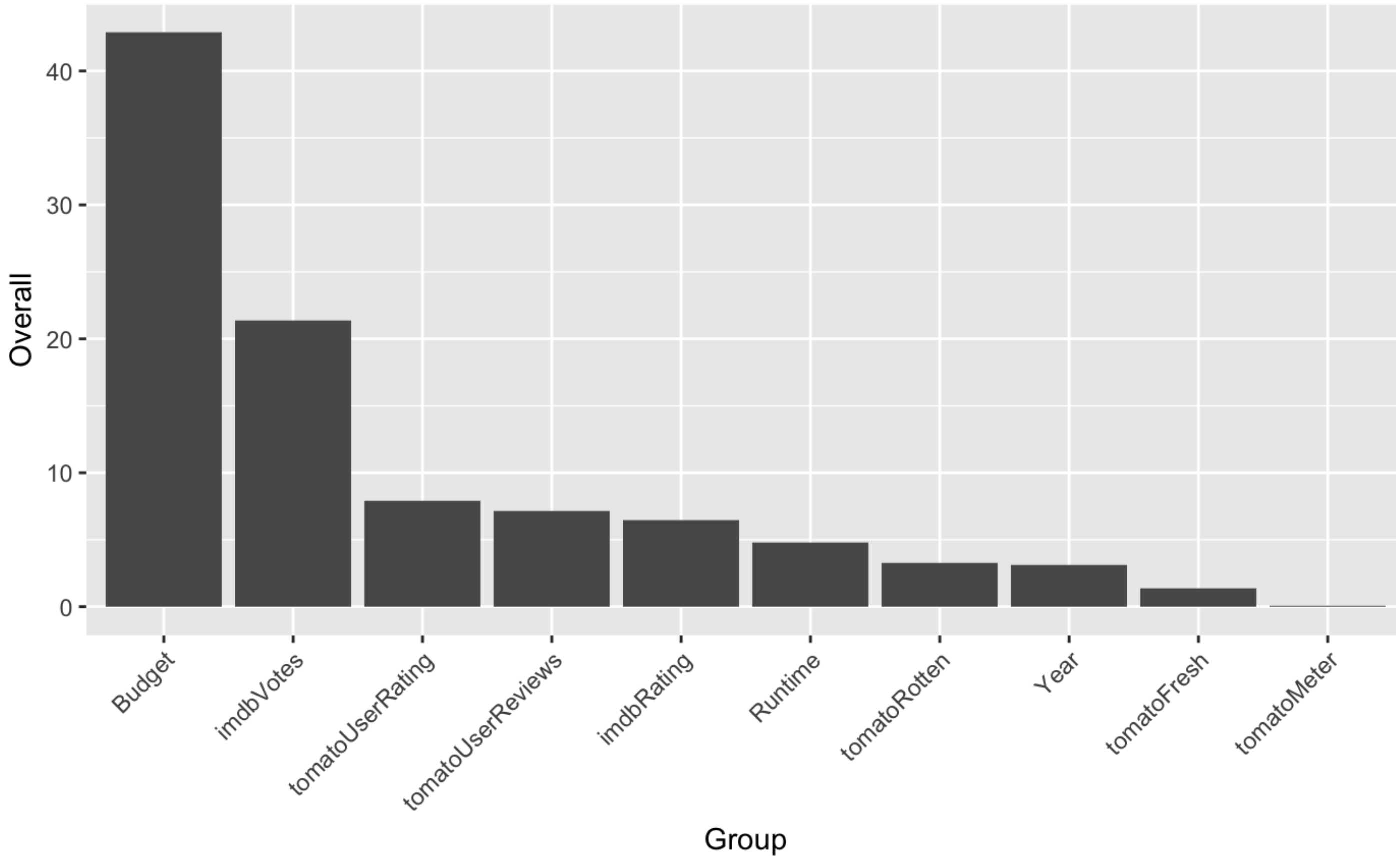
```
### TODO: Build & evaluate model 2 (transformed numeric variables only)
options(warn=-1)
### Pre-processing for Task #2
df2 = df1
# Check variable importance of Task 1, Model 2 of partition 0.75 (as this partition uses
 enough train data to give accurate predictions) using the Caret R package
t1_importance = varImp(t1_m1_model_list[[15]][[1]], useModel = "lm", scale = FALSE)
t1_importance$Group = rownames(t1_importance)
t1_importance = t1_importance[order(t1_importance$Overall, decreasing = TRUE),]
print(t1_importance)
```

	Overall <dbl>	Group <chr>
Budget	42.90363532	Budget
imdbVotes	21.40618978	imdbVotes
tomatoUserRating	7.92065860	tomatoUserRating
tomatoUserReviews	7.12722051	tomatoUserReviews
imdbRating	6.49065797	imdbRating
Runtime	4.77361039	Runtime
tomatoRotten	3.29257912	tomatoRotten
Year	3.14477778	Year
tomatoFresh	1.39083352	tomatoFresh
tomatoMeter	0.08847067	tomatoMeter

1-10 of 10 rows

Hide

```
t1_importance$Group <- factor(t1_importance$Group, levels=unique(as.character(t1_importa
nce$Group)) )
ggplot(t1_importance, aes(y=Overall, x=Group)) + geom_bar(stat="identity") + theme(axis.
text.x = element_text(angle = 45, hjust = 1))
```



_**Note**: Budget and imdbVotes have much higher importance values, as given by varImp in the Caret package.

Hide

```
### TODO: Build & evaluate model 2 (transformed numeric variables only) - Continued
options(warn=-1)
### Pre-processing for Task #2
## Use BoxTidwell approach/function to find the correct power transformation for most im
portant numeric variables
# boxTidwell function cannot handle 0's, so those rows need to be ignored/dropped for th
is approach
df2_temp = df2
cat("Number of Columns and Rows before Removal:", dim(df2), "\n", "\n")
```

```
Number of Columns and Rows before Removal: 3257 11
```

Hide

```
df2_temp = df2_temp[df2_temp$tomatoUserRating != 0,]
cat("Number of Columns and Rows after Removal:", dim(df2_temp), "\n", "\n")
```

```
Number of Columns and Rows after Removal: 3256 11
```

Hide

```
df2_temp = df2_temp[df2_temp$tomatoMeter != 0,]
cat("Number of Columns and Rows after Removal:", dim(df2_temp), "\n", "\n")
```

```
Number of Columns and Rows after Removal: 3234 11
```

Hide

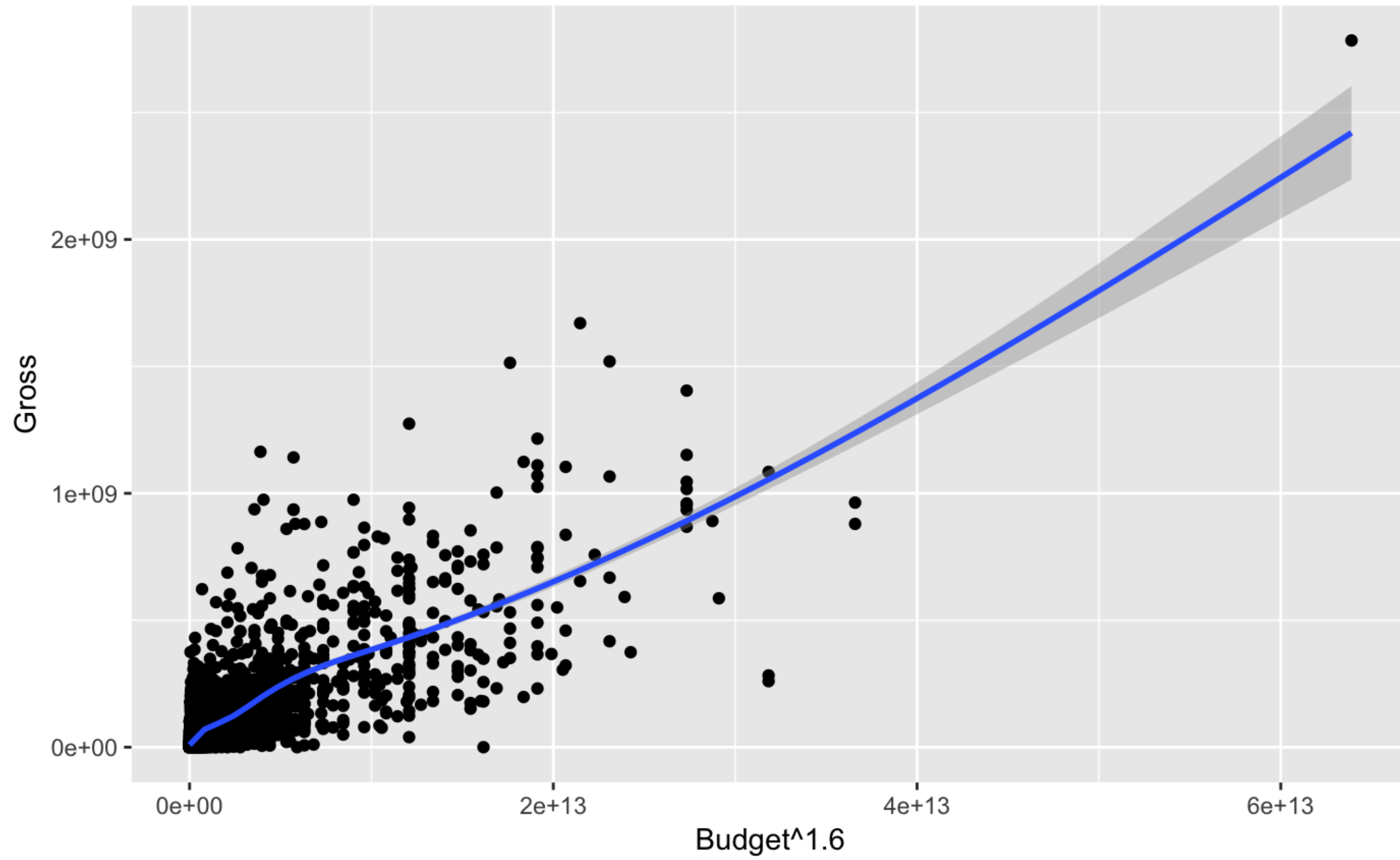
```
# Calculate the Box-Tidwell transformation of the Explanatory Variables
bt_transform = boxTidwell(Gross~Budget+imdbVotes+tomatoUserRating+imdbRating, data = df2
_temp)
print(bt_transform)
```

```
                 Score Statistic   p-value MLE of lambda
Budget                 13.998541 0.0000000      1.593217
imdbVotes              -4.430462 0.0000094      0.882543
tomatoUserRating        3.319568 0.0009016      3.290609
imdbRating             -5.804545 0.0000000      5.537583

iterations =  5
```

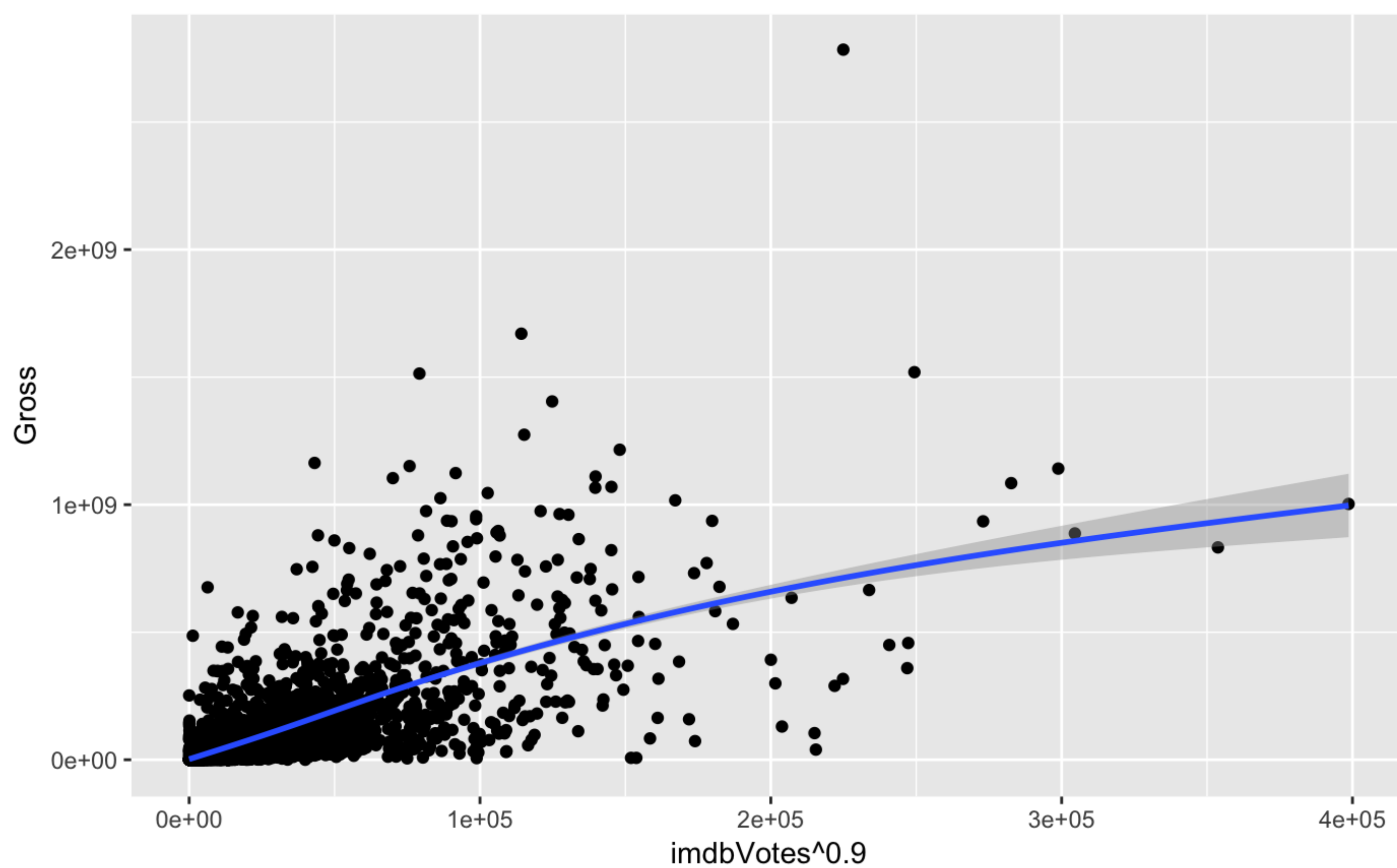
Hide

```
### TODO: Build & evaluate model 2 (transformed numeric variables only) - Continued
options(warn=-1)
### Pre-processing for Task #2 - Continued
# Plot the Box-Tidwell transformed, explanatory variables vs the response variable
ggplot(df2, aes(x=Budget^1.6, y=Gross)) + geom_point() + stat_smooth(method="auto")
```



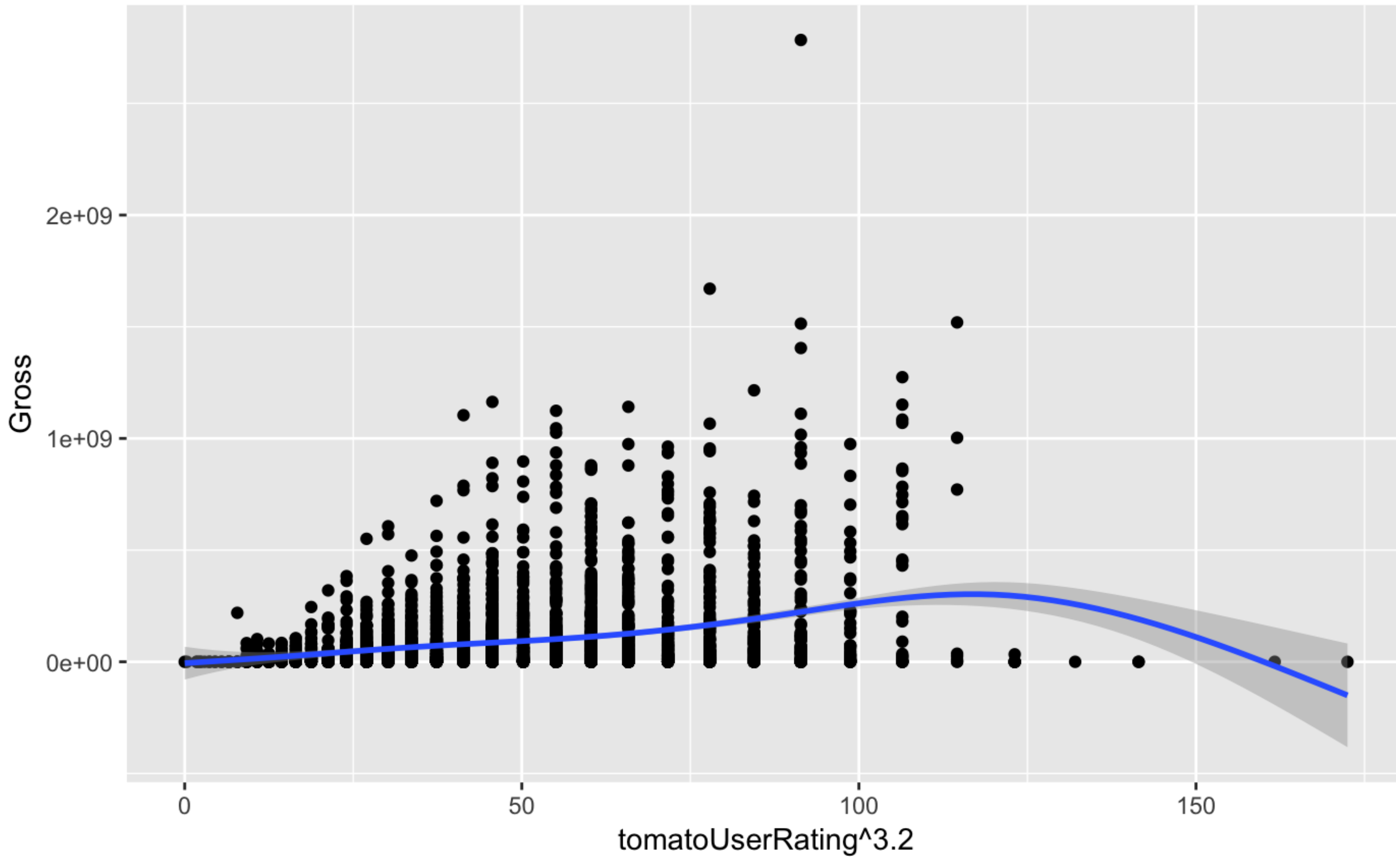
Hide

```
ggplot(df2, aes(x=imdbVotes^0.9, y=Gross)) + geom_point() + stat_smooth(method="auto")
```



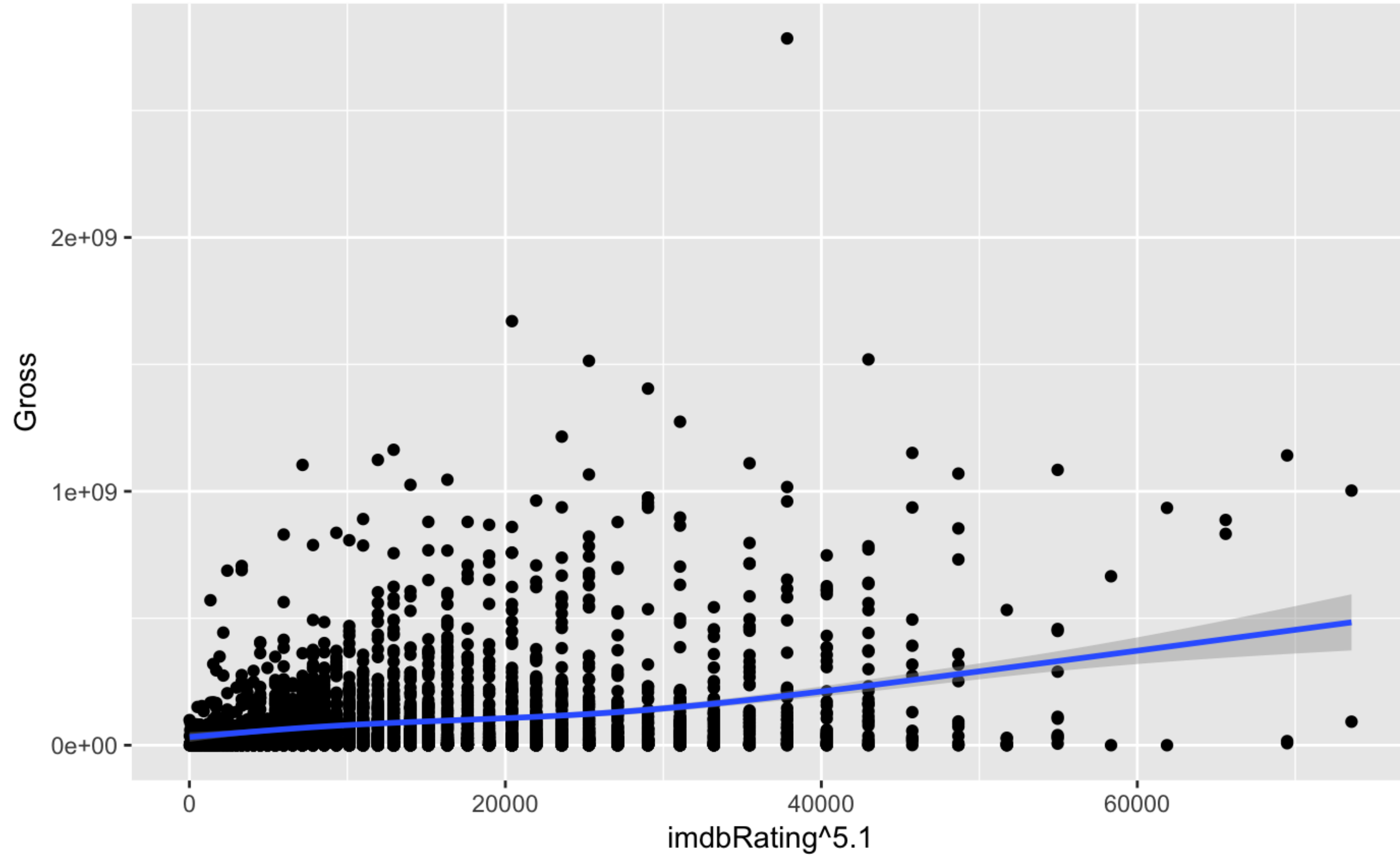
Hide

```
ggplot(df2, aes(x=tomatoUserRating^3.2, y=Gross)) + geom_point() + stat_smooth(method="a
uto")
```



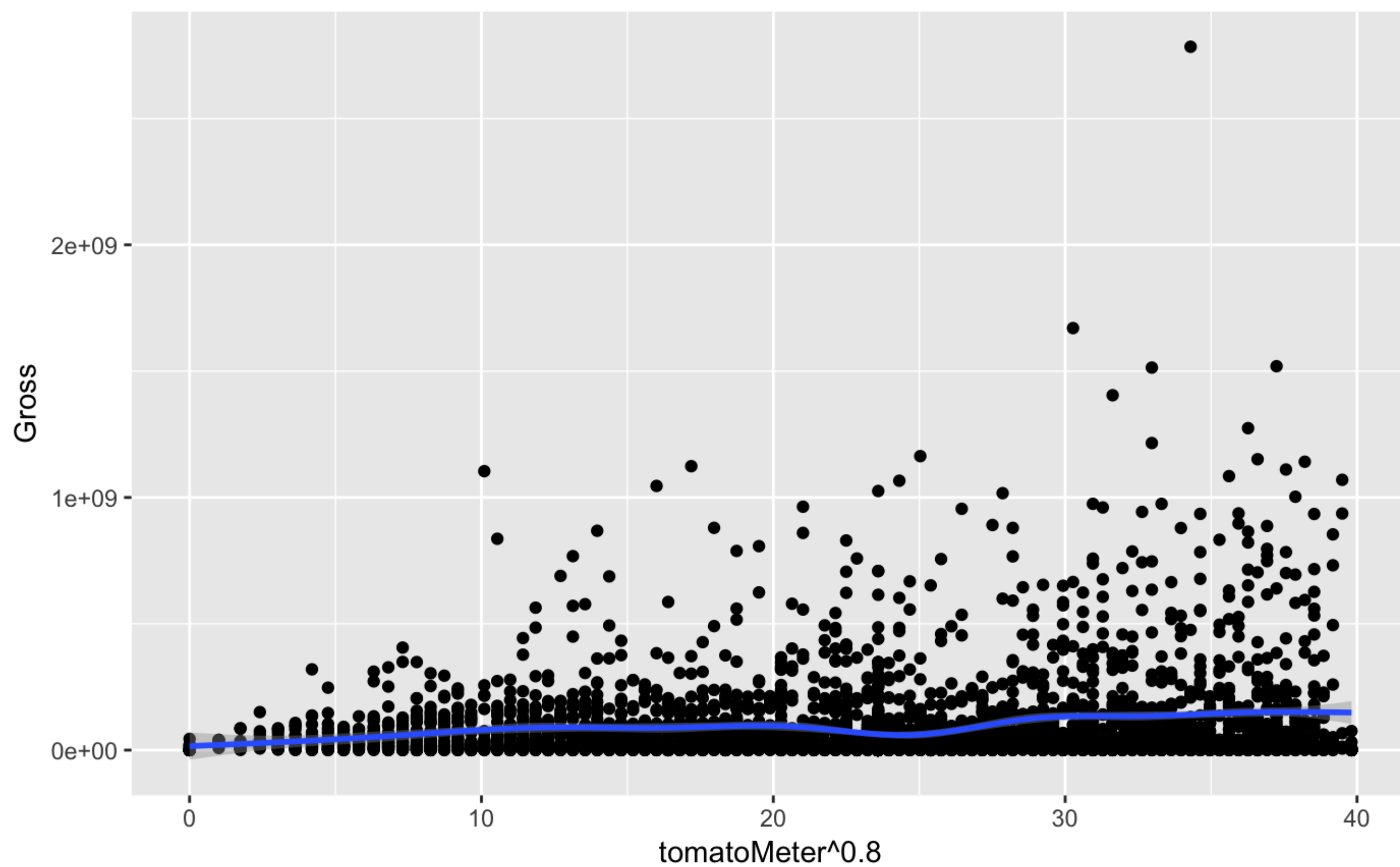
Hide

```
ggplot(df2, aes(x=imdbRating^5.1, y=Gross)) + geom_point() + stat_smooth(method="auto")
```



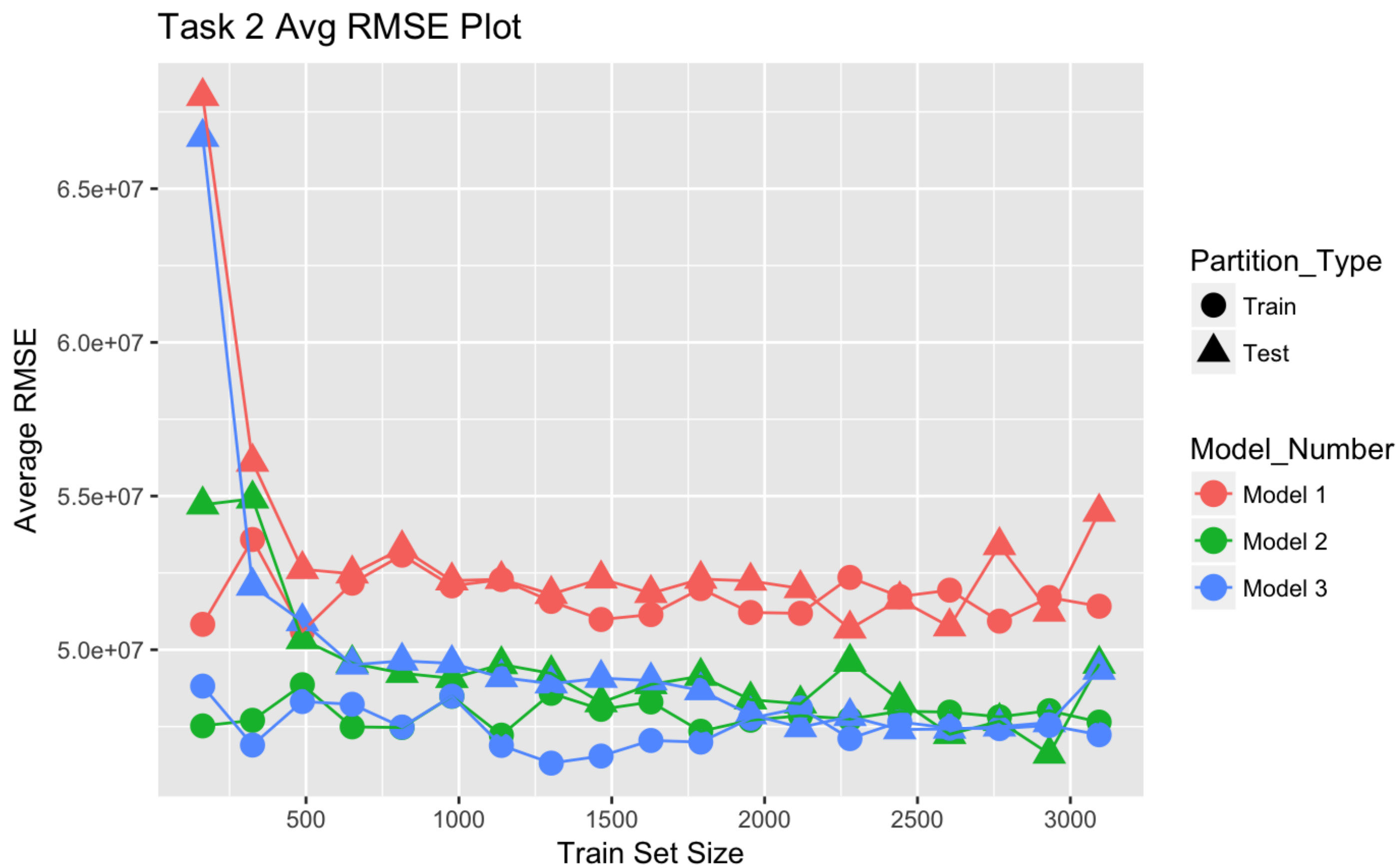
Hide

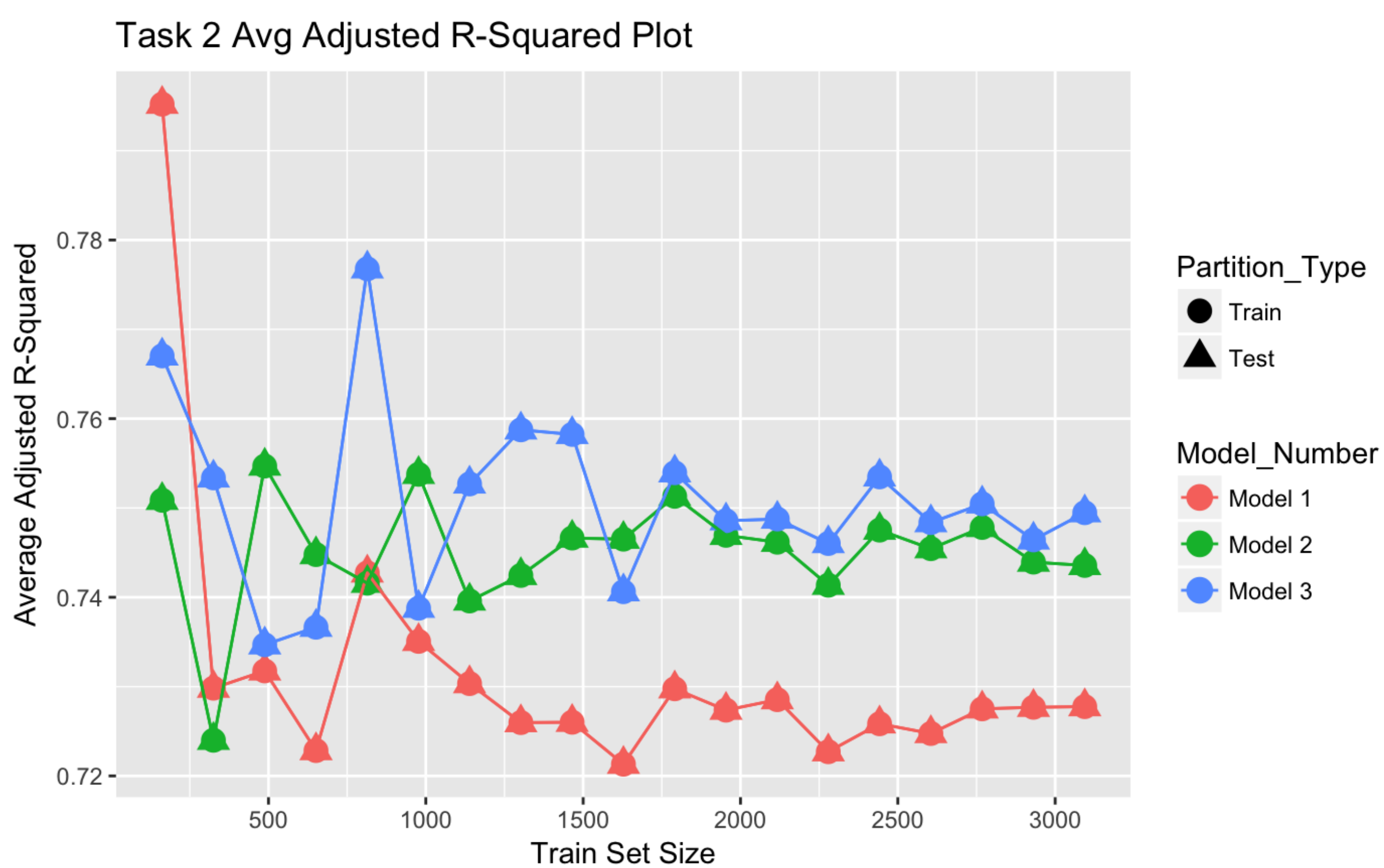
```
ggplot(df2, aes(x=tomatoMeter^0.8, y=Gross)) + geom_point() + stat_smooth(method="auto")
```



Hide

```
# TODO: Build & evaluate model 2 (transformed numeric variables only) - Continued
options(warn=-1)
### Evaluation Part 1 - Part 5
# Task 2, Model 1 - Same as Task 1, Model 1 (Best Task 1 Model), and this model will be
 used for comparison
# t2_m1 = Gross~Year+Runtime+Budget+imdbRating+imdbVotes+tomatoUserRating+tomatoMeter+to
matoFresh+tomatoRotten+tomatoUserReviews
# Task 2, Model 2
t2_m2 = Gross~.+I(Budget^1.6)+I(imdbVotes^0.9)
t2_m2_tt_list = list()
t2_m2_model_list = list()
for (i in partition_list) {
  t2_tt_temp = rep.run.lm(df2, i, t2_m2)
  t2_m2_tt_list = append(t2_m2_tt_list, t2_tt_temp[1:2])
  t2_m2_model_list = append(t2_m2_model_list, t2_tt_temp[3])
}
# Task 2, Model 3
t2_m3 = Gross~.+I(Budget^2)+I(imdbVotes^2)+I(tomatoUserRating^3.2)+I(imdbRating^5.1)+I(t
omatoMeter^0.8)
t2_m3_tt_list = list()
t2_m3_model_list = list()
for (i in partition_list) {
  t2_tt_temp = rep.run.lm(df2, i, t2_m3)
  t2_m3_tt_list = append(t2_m3_tt_list, t2_tt_temp[1:2])
  t2_m3_model_list = append(t2_m3_model_list, t2_tt_temp[3])
}
### Evauation Part 6 - Plot all models on one plot
t2_models = list(t1_m1_tt_list, t2_m2_tt_list, t2_m3_tt_list)  # List of models for this
 task
t2_n_models = length(t2_models)
# Create a large list of all partitions of train and test for all models
t2_mall_tt_list = list()
for (i in t2_models) {t2_mall_tt_list = append(t2_mall_tt_list, i)}
# Create the final plot df
t2_plot_df = build.plot.df(t2_mall_tt_list, t2_n_models)
# Plot all the models
t2_plot = plot.lm(t2_plot_df, seq(0,3000,500), "Task 2 Avg RMSE Plot", seq(0,3000,500),
"Task 2 Avg Adjusted R-Squared Plot", t2_n_models)
```


Task 2 Avg RMSE Plot
Average RMSE
6.5e+07
6.0e+07
5.5e+07
5.0e+07
500
1000
1500
2000
2500
3000
Train Set Size
Partition_Type
Train
Test
Model_Number
Model 1
Model 2
Model 3


Task 2 Avg Adjusted R-Squared Plot
Average Adjusted R-Squared
0.78
0.76
0.74
0.72
500
1000
1500
2000
2500
3000
Train Set Size
Partition_Type
Train
Test
Model_Number
Model 1
Model 2
Model 3

Hide

```
options(warn=-1)
t1_perc_diff_models = 100 * (t2_plot_df$Avg_RMSE[34] - t2_plot_df$Avg_RMSE[72]) / t2_plo
t_df$Avg_RMSE[34]
cat("Percent Difference in RMSE from Task 1, Model 1 and Task 2, Model 2:   ", round(t1_
perc_diff_models, 0), "%")
```

```
Percent Difference in RMSE from Task 1, Model 1 and Task 2, Model 2:    6 %
```

Hide

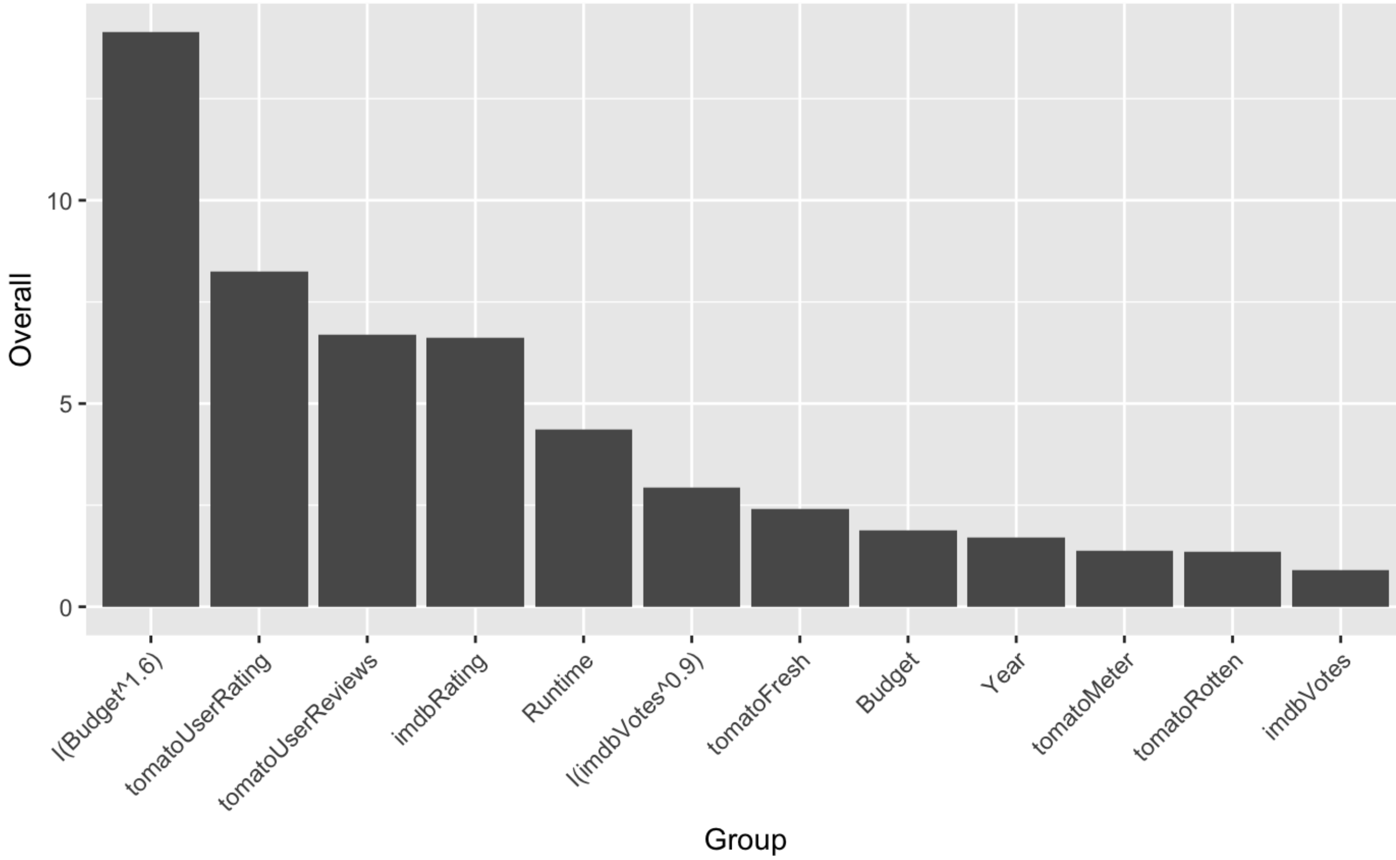
```
options(warn=-1)
# Check variable importance of best Task 2 model (i.e. Model 2 and 3 performed equally w
ell in terms of RMSE and adjusted R-squared) of partition 0.5 (as this partition uses en
ough train data to give accurate predictions) using the Caret R package
t2_importance = varImp(t2_m2_model_list[[15]][[1]], useModel = "lm", scale = FALSE)
t2_importance$Group = rownames(t2_importance)
t2_importance = t2_importance[order(t2_importance$Overall, decreasing = TRUE),]
print(t2_importance)
```

	Overall <dbl>	Group <chr>
I(Budget^1.6)	14.1443612	I(Budget^1.6)
tomatoUserRating	8.2504323	tomatoUserRating
tomatoUserReviews	6.7010497	tomatoUserReviews
imdbRating	6.6256779	imdbRating
Runtime	4.3722615	Runtime
I(imdbVotes^0.9)	2.9260775	I(imdbVotes^0.9)
tomatoFresh	2.3970808	tomatoFresh
Budget	1.8905773	Budget
Year	1.7121106	Year
tomatoMeter	1.3701408	tomatoMeter

1-10 of 12 rows Previous **1** 2 Next

Hide

```
t2_importance$Group <- factor(t2_importance$Group, levels=unique(as.character(t2_importa
nce$Group)) )
ggplot(t2_importance, aes(y=Overall, x=Group)) + geom_bar(stat="identity") + theme(axis.
text.x = element_text(angle = 45, hjust = 1))
```



Q: Explain which transformations you used and why you chose them.

A: I transformed the two most important explanatory variables (i.e. Budget and imdbVotes) along with a few other important explanatory variables, where importance was defined by the varImp() function in the caret package. Budget and imdbVotes had a much higher ranking than all other explanatory variables, so those variables alone were transformed in Task 2, Model 2. I used the BoxTidwell() function in the car package to find the appropriate power transformation (MLE of lambda) for the most important variables. After finding the MLE of lambda for Budget was 1.6 and imdbVotes was 0.9, I plotted Gross vs Budget^1.6 and Gross vs imdbVotes^0.9 (and plotted a smoothing function) to ensure the lambda values provided a good linear transformation. I did the same plotting checks for the other important explanatory variables. Using the power transformations for Budget and imdbVotes, I was able to build a model (Task 2, Model 2) which reduced the RMSE of the predicted Gross value in the Test dataset by about 5-10%, depending on the run, compared to the best model from Task 1 (Task 1, Model 1 which is the same as Task 2, Model 1). Task 3, Model 3 used more transformed explanatory variables and did slightly improve the adjusted R-squared value, but this model did not improve on the predicted Gross RMSE. Thus, Model 2 was the best model from Task 2.

3. Non-numeric variables

Write code that converts genre, actors, directors, and other categorical variables to columns that can be used for regression (e.g. binary columns as you did in Project 1). Also process variables such as awards into more useful columns (again, like you did in Project 1). Now use these converted columns only to build your next model.

Hide

```
### TODO: Build & evaluate model 3 (converted non-numeric variables only)
options(warn=-1)
### Pre-processing for Task #2
df3 = df[,c("Gross", "Director", "Rated", "Actors", "Genre", "Awards")]
### Evaluation Part 1 - Part 5
## Task 3, Model 1 - same as Task 2, Model 2 (current best model for comparison)
# t3_m1 = t2_m2 = Gross~.+I(Budget^1.6)+I(imdbVotes^0.9)
## Pre-Processing for Task 3, Model 2
# Convert the Awards column into a numeric column which can be used for regression
df3$Awards_Wins = regmatches(df3$Awards, gregexpr("\\d+(?= (win))", df3$Awards, perl=TRU
E))
df3$Awards_Wins = as.numeric(df3$Awards_Wins)
df3$Awards_Wins[is.na(df3$Awards_Wins)] = 0
df3$Awards_Wins1 = regmatches(df3$Awards, gregexpr("(?<=\\b(Won)\\s)[0-9]", df3$Awards,
perl=TRUE))
df3$Awards_Wins1 = as.numeric(df3$Awards_Wins1)
df3$Awards_Wins1[is.na(df3$Awards_Wins1)] = 0
df3$Awards_Nomination = regmatches(df3$Awards, gregexpr("\\d+(?= (nomination))", df3$Awa
rds, perl=TRUE))
df3$Awards_Nomination = as.numeric(df3$Awards_Nomination)
df3$Awards_Nomination[is.na(df3$Awards_Nomination)] = 0
df3$Awards_Nomination1 = regmatches(df3$Awards, gregexpr("(?<=\\b(for)\\s)[0-9]", df3$Aw
ards, perl=TRUE))
df3$Awards_Nomination1 = as.numeric(df3$Awards_Nomination1)
df3$Awards_Nomination1[is.na(df3$Awards_Nomination1)] = 0
df3$Awards_NA = ifelse(df3$Awards == "N/A", NA, 0)
df3$Award_Wins = df3$Awards_Wins + df3$Awards_Wins1 + df3$Awards_NA
df3$Award_Nominations = df3$Awards_Nomination + df3$Awards_Nomination1 + df3$Awards_NA
# Remove unused columns
df3["Awards"] = NULL
df3["Awards_Wins"] = NULL
df3["Awards_Wins1"] = NULL
df3["Awards_Nomination"] = NULL
df3["Awards_Nomination1"] = NULL
df3["Awards_NA"] = NULL
# Replace Genre with a collection of binary columns
# Create a vector of the df$Genre column
genre = df3$Genre
# Create a list of genres for each movie
genres_l = strsplit(genre, ", ")
# Build the Corpus
genre_corp = Corpus(VectorSource(genres_l))
# Build the DocumentTermMatrix
genre_dtm = DocumentTermMatrix(genre_corp)
# Convert DocumentTermMatrix to a DataFrame
genre_dtm_matrix = as.matrix(genre_dtm)
genre_dtm_df = as.data.frame(genre_dtm_matrix)
# Add a join key to both DataFrames
genre_dtm_df$JoinKey = seq(1, dim(genre_dtm_df)[1], by=1)
df3$JoinKey = seq(1, dim(df3)[1], by=1)
# Merge the two DataFrames
df3 = merge(x=df3, y=genre_dtm_df, by.x="JoinKey", by.y= "JoinKey")
# Drop JoinKey and Genre
```

```
df3$JoinKey = NULL
df3$Genre = NULL
# Rename Sci-fi column
df3$scifi = df3$`sci-fi`
df3$`sci-fi` = NULL
## Task 3, Model 2 - Build a model using Awards and Genre
t3_m2 = Gross~Award_Wins+Award_Nominations+action+adventure+animation+biography+comedy+c
rime+documentary+drama+family+fantasy+history+horror+music+musical+mystery+news+romance+
scifi+short+sport+thriller+war+western
t3_m2_tt_list = list()
t3_m2_model_list = list()
for (i in partition_list) {
  t3_tt_temp = rep.run.lm(df3, i, t3_m2)
  t3_m2_tt_list = append(t3_m2_tt_list, t3_tt_temp[1:2])
  t3_m2_model_list = append(t3_m2_model_list, t3_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
```

```
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleading
```

Hide

```
## Pre-processing for Task 3, Model 3
# Replace Actors with a collection of binary columns
# Create a vector of the Actors column
actor = df3$Actors
# Create a tokenizer for DTM to use
commasplit_tokenizer = function(x) {unlist(strsplit(as.character(x), ", "))}
# This code is used to investigate counts of actors to find the cutoff for using actors
 as columns and build an actors dictionary
actors = unlist(strsplit(actor, ", "))
actors_df = as.data.frame(count(actors))
actors_df = actors_df[order(actors_df$freq, decreasing = TRUE),]
actors_df_f = actors_df[actors_df$freq >= 10,]
actors_dict = as.character(actors_df_f$x)
# Build the Corpus
actors_corp = VCorpus(VectorSource(actor))
# Build the DocumentTermMatrix
actors_dtm = DocumentTermMatrix(actors_corp, control=list(tokenize=commasplit_tokenizer,
 dictionary=actors_dict, tolower=FALSE))
# Convert DocumentTermMatrix to a DataFrame
actors_dtm_matrix = as.matrix(actors_dtm)
actors_dtm_df = as.data.frame(actors_dtm_matrix)
# Add a join key to both DataFrames
actors_dtm_df$JoinKey = seq(1, dim(actors_dtm_df)[1], by=1)
df3$JoinKey = seq(1, dim(df3)[1], by=1)
# Merge the two DataFrames
df3 = merge(x=df3, y=actors_dtm_df, by.x="JoinKey", by.y= "JoinKey")
# Drop JoinKey and Actors
df3$JoinKey = NULL
df3$Actors = NULL
# Keep only columns used in Model 3
df3_m3 = df3[,c(-2,-3)]
## Task 3, Model 3 - Build a model using Awards and Genre and Actors in >= 10 movies
t3_m3 = Gross~.
t3_m3_tt_list = list()
t3_m3_model_list = list()
for (i in partition_list) {
  t3_tt_temp = rep.run.lm(df3_m3, i, t3_m3)
  t3_m3_tt_list = append(t3_m3_tt_list, t3_tt_temp[1:2])
  t3_m3_model_list = append(t3_m3_model_list, t3_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
```

```
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
 a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
 from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
 a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
 from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
 a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
 from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
```

```
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
```

```
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleading
```

Hide

```
## Pre-processing for Task 3, Model 4
## Replace Rated with a collection of binary columns
# Create a vector of the Actors column
rated = df3$Rated
rated = gsub(" ", "", rated)
rated = gsub("/", "", rated)
rated = gsub("-", "", rated)
# Create tokenizer
rated_tokenizer = function(x) {x}
# Build the Corpus
rated_corp = VCorpus(VectorSource(rated))
# Build the DocumentTermMatrix
rated_dtm = DocumentTermMatrix(rated_corp, control=list(tokenize=rated_tokenizer, wordLe
ngths=c(1,Inf)))
# Convert DocumentTermMatrix to a DataFrame
rated_dtm_matrix = as.matrix(rated_dtm)
rated_dtm_df = as.data.frame(rated_dtm_matrix)
# Add a join key to both DataFrames
rated_dtm_df$JoinKey = seq(1, dim(rated_dtm_df)[1], by=1)
df3$JoinKey = seq(1, dim(df3)[1], by=1)
# Merge the two DataFrames
df3 = merge(x=df3, y=rated_dtm_df, by.x="JoinKey", by.y= "JoinKey")
# Drop JoinKey and Actors
df3$JoinKey = NULL
df3$Rated = NULL
# Keep only columns used in Model 3
df3_m4 = df3[,-2]
## Task 3, Model 4 -  Build a model using Awards and Genre and Actors in >= 10 movies an
d Rated
t3_m4 = Gross~.
t3_m4_tt_list = list()
t3_m4_model_list = list()
for (i in partition_list) {
  t3_tt_temp = rep.run.lm(df3_m4, i, t3_m4)
  t3_m4_tt_list = append(t3_m4_tt_list, t3_tt_temp[1:2])
  t3_m4_model_list = append(t3_m4_model_list, t3_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
```

```
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
 a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
 from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
 a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
 from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
 a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
 from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
```

```
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
```

```
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleading
```

Hide

```
## Pre-processing for Task 3, Model 5
## Replace Director with a collection of binary columns
# Create a vector of the Directors column
director = df3$Director
# Create a tokenizer for DTM to use
commasplit_tokenizer = function(x) {unlist(strsplit(as.character(x), ", "))}
# This code is used to investigate counts of actors to find the cutoff for using actors
 as columns and build an actors dictionary
directors = unlist(strsplit(director, ", "))
directors_df = as.data.frame(count(directors))
directors_df = directors_df[order(directors_df$freq, decreasing = TRUE),]
directors_df_f = directors_df[directors_df$freq >= 7,]
directors_dict = as.character(directors_df_f$x)
# Build the Corpus
directors_corp = VCorpus(VectorSource(director))
# Build the DocumentTermMatrix
directors_dtm = DocumentTermMatrix(directors_corp, control=list(tokenize=commasplit_toke
nizer, dictionary=directors_dict, tolower=FALSE))
# Convert DocumentTermMatrix to a DataFrame
directors_dtm_matrix = as.matrix(directors_dtm)
directors_dtm_df = as.data.frame(directors_dtm_matrix)
# Add a join key to both DataFrames
directors_dtm_df$JoinKey = seq(1, dim(directors_dtm_df)[1], by=1)
df3$JoinKey = seq(1, dim(df3)[1], by=1)
# Merge the two DataFrames
df3 = merge(x=df3, y=directors_dtm_df, by.x="JoinKey", by.y= "JoinKey")
# Drop JoinKey and directors
df3$JoinKey = NULL
df3$Director = NULL
## Task 3, Model 5 - Build a model using Awards and Genre and Actors in >= 10 movies and
 Rated and Director of >= 7 movies
t3_m5 = Gross~.
t3_m5_tt_list = list()
t3_m5_model_list = list()
for (i in partition_list) {
  t3_tt_temp = rep.run.lm(df3, i, t3_m5)
  t3_m5_tt_list = append(t3_m5_tt_list, t3_tt_temp[1:2])
  t3_m5_model_list = append(t3_m5_model_list, t3_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
```

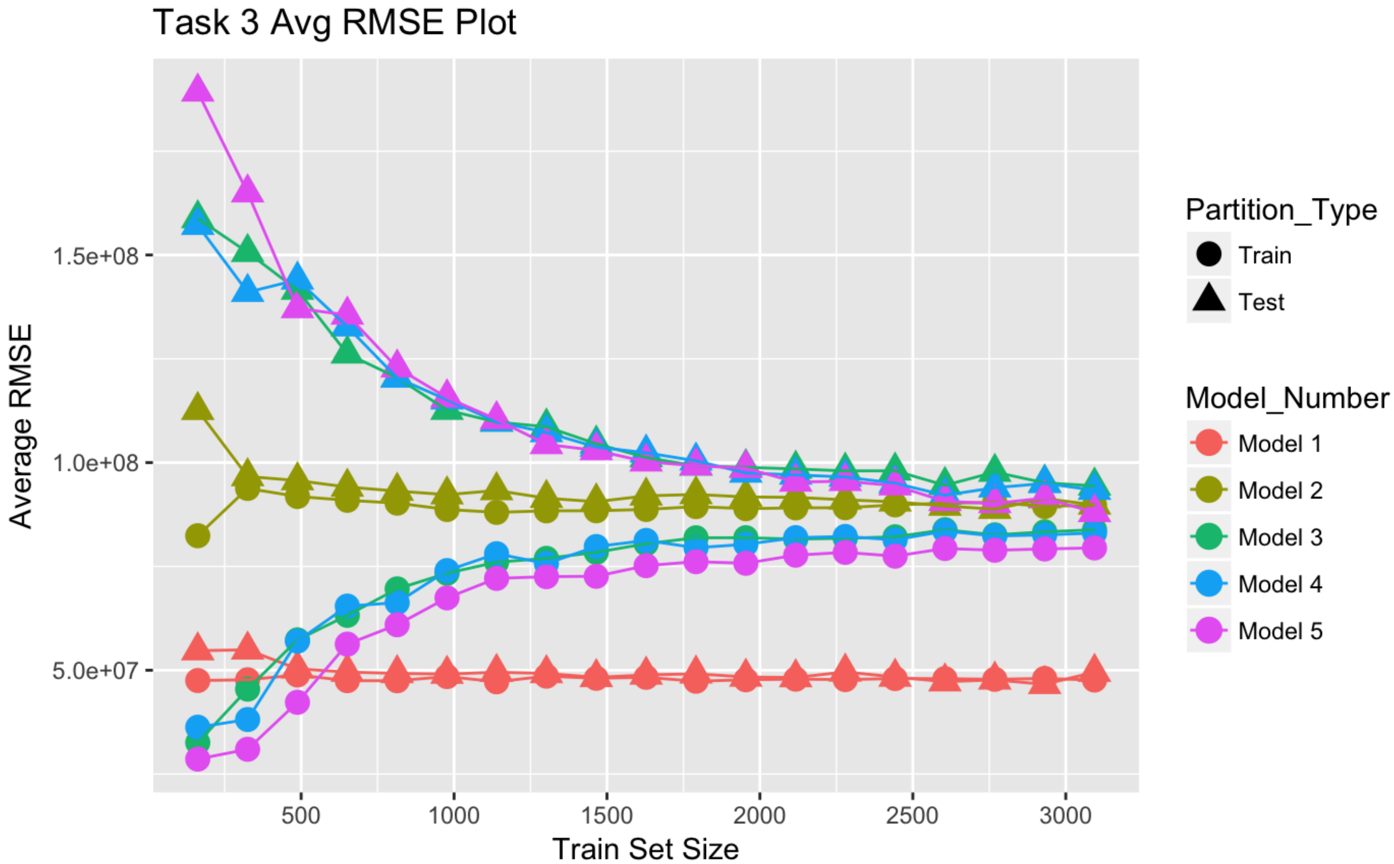
```
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
```

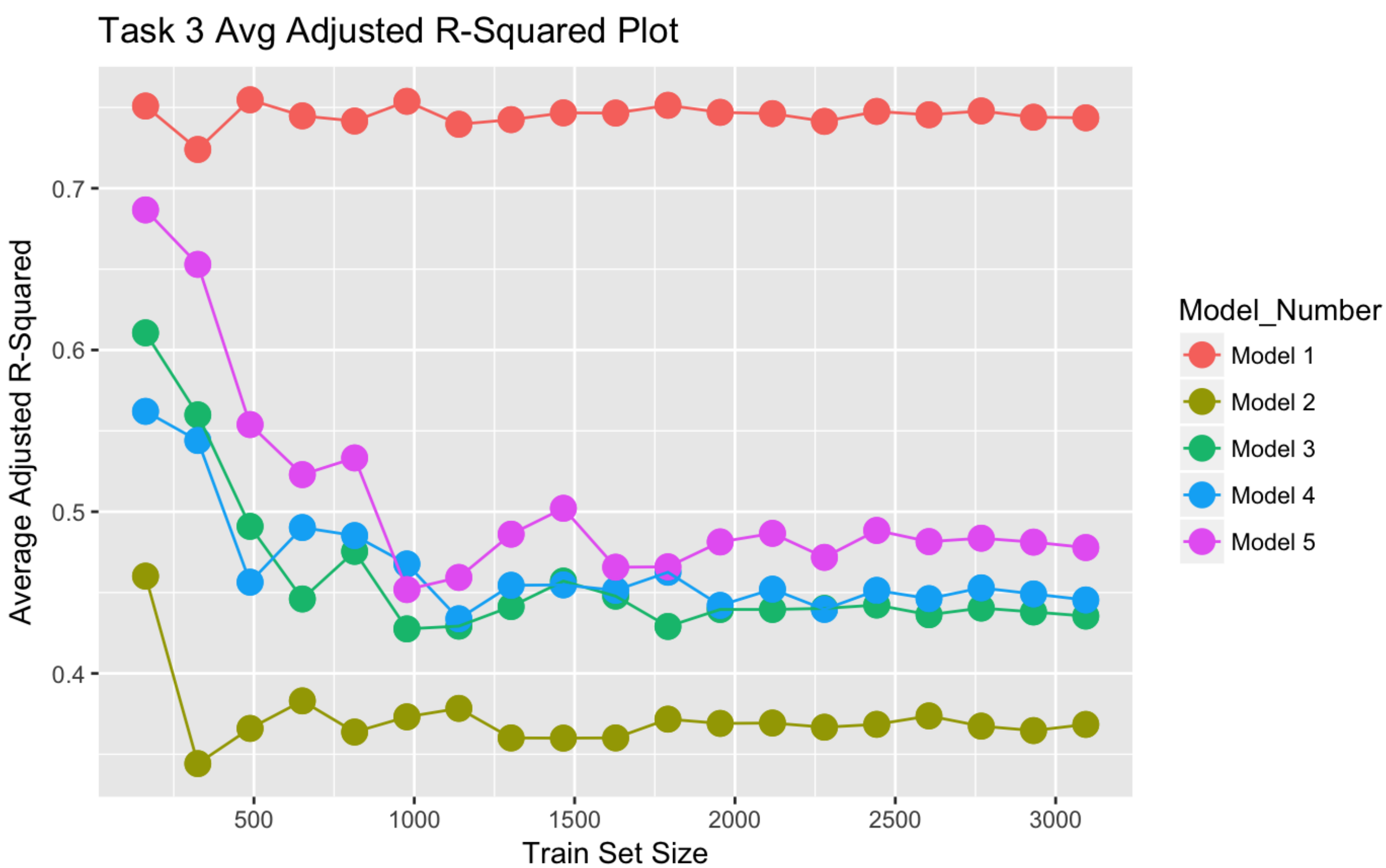
```
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
```

```
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient 
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may 
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleading
```

Hide

```
### Evauation Part 6 - Plot all models on one plot
t3_models = list(t2_m2_tt_list, t3_m2_tt_list, t3_m3_tt_list, t3_m4_tt_list, t3_m5_tt_li
st)  # List of models for this task
t3_n_models = length(t3_models)
# Create a large list of all partitions of train and test for all models
t3_mall_tt_list = list()
for (i in t3_models) {t3_mall_tt_list = append(t3_mall_tt_list, i)}
# Create the final plot df
t3_plot_df = build.plot.df(t3_mall_tt_list, t3_n_models)
# Plot all the models
t3_plot = plot.lm(t3_plot_df, seq(0,3000,500), "Task 3 Avg RMSE Plot", seq(0,3000,500),
"Task 3 Avg Adjusted R-Squared Plot", n_mod = t3_n_models)
```


Task 3 Avg RMSE Plot
Average RMSE
1.5e+08
1.0e+08
5.0e+07
500
1000
1500
2000
2500
3000
Train Set Size
Partition_Type
Train
Test
Model_Number
Model 1
Model 2
Model 3
Model 4
Model 5


Task 3 Avg Adjusted R-Squared Plot
Average Adjusted R-Squared
0.7
0.6
0.5
0.4
500
1000
1500
2000
2500
3000
Train Set Size
Model_Number
Model 1
Model 2
Model 3
Model 4
Model 5

Hide

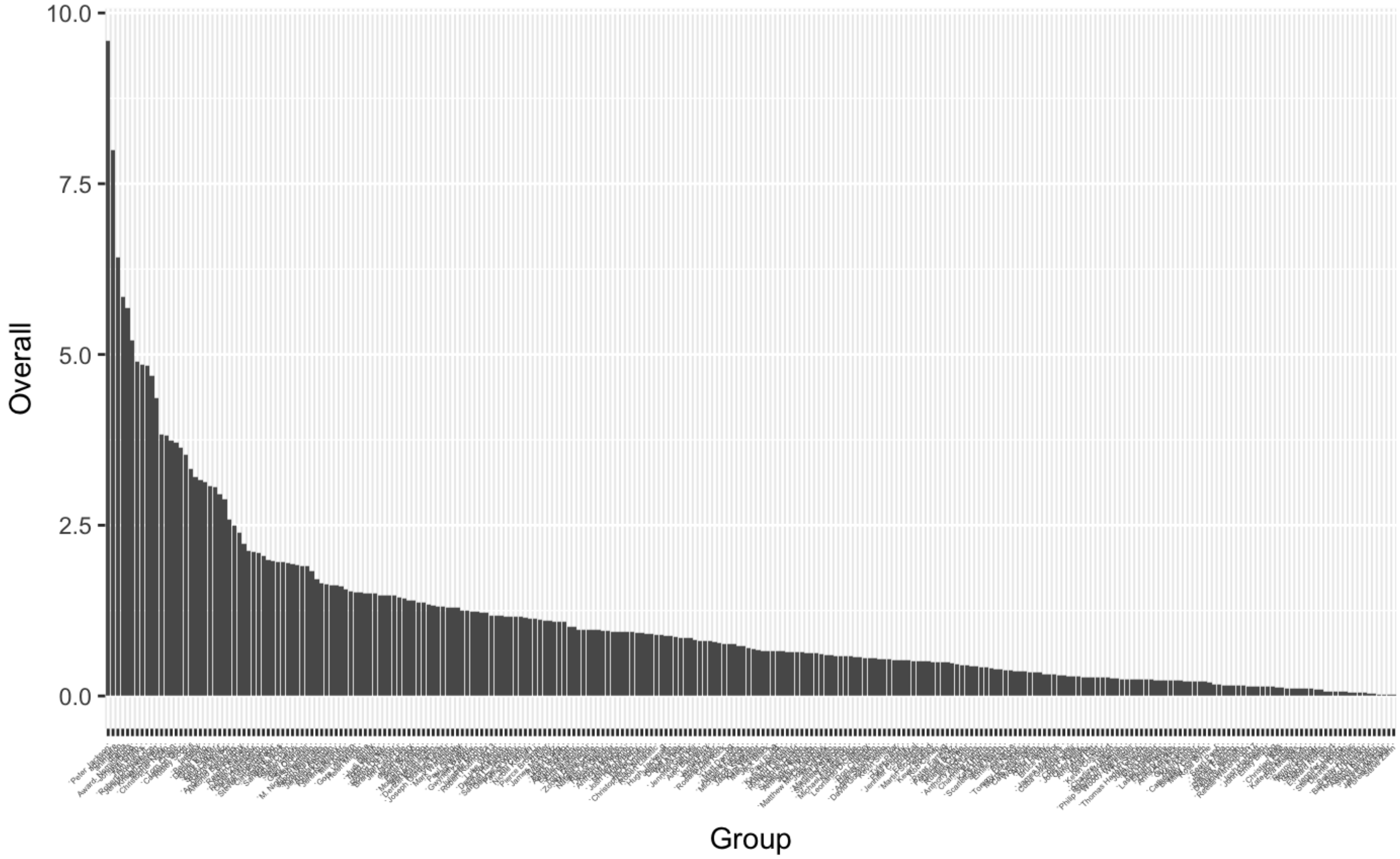
```
options(warn=-1)
# Check variable importance of Task 3, Model 5 of partition 0.75 (as this partition uses
 enough train data to give accurate predictions) using the Caret R package
t3_importance = varImp(t3_m5_model_list[[15]][[1]], useModel = "lm", scale = FALSE)
t3_importance$Group = rownames(t3_importance)
t3_importance = t3_importance[order(t3_importance$Overall, decreasing = TRUE),]
print(t3_importance)
```

	Overall <dbl>	**Group** <chr>
\`Peter Jackson\`	9.59208823	\`Peter Jackson\`
adventure	7.99718414	adventure
animation	6.42041252	animation
drama	5.83931075	drama
Award_Nominations	5.67542063	Award_Nominations
\`Johnny Depp\`	5.21283032	\`Johnny Depp\`
\`Will Smith\`	4.89498839	\`Will Smith\`
\`Tom Hanks\`	4.85407575	\`Tom Hanks\`
\`Robert Downey Jr.\`	4.83744479	\`Robert Downey Jr.\`
\`Jennifer Lawrence\`	4.68381700	\`Jennifer Lawrence\`

1-10 of 266 rows Previous **1** 2 3 4 5 6 ... 27 Next

Hide

```
t3_importance$Group = factor(t3_importance$Group, levels=unique(as.character(t3_importan
ce$Group)) )
ggplot(t3_importance, aes(y=Overall, x=Group)) + geom_bar(stat="identity") + theme(axis.
text.x = element_text(angle = 45, hjust = 1, size=3,))
```



Hide

```
options(warn=-1)
## Pre-processing for Task 3, Model 6
t3_m6_imp = as.character(t3_importance$Group[1:15])
t3_m6_imp = gsub("`", "", t3_m6_imp)
df3_m6 = df3[,c("Gross", t3_m6_imp)]
## Task 3, Model 6 - Build a model using the 15 most important variables as defined by v
arImp from the caret package
t3_m6 = Gross~.
t3_m6_tt_list = list()
t3_m6_model_list = list()
for (i in partition_list) {
  t3_tt_temp = rep.run.lm(df3_m6, i, t3_m6)
  t3_m6_tt_list = append(t3_m6_tt_list, t3_tt_temp[1:2])
  t3_m6_model_list = append(t3_m6_model_list, t3_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleading
```

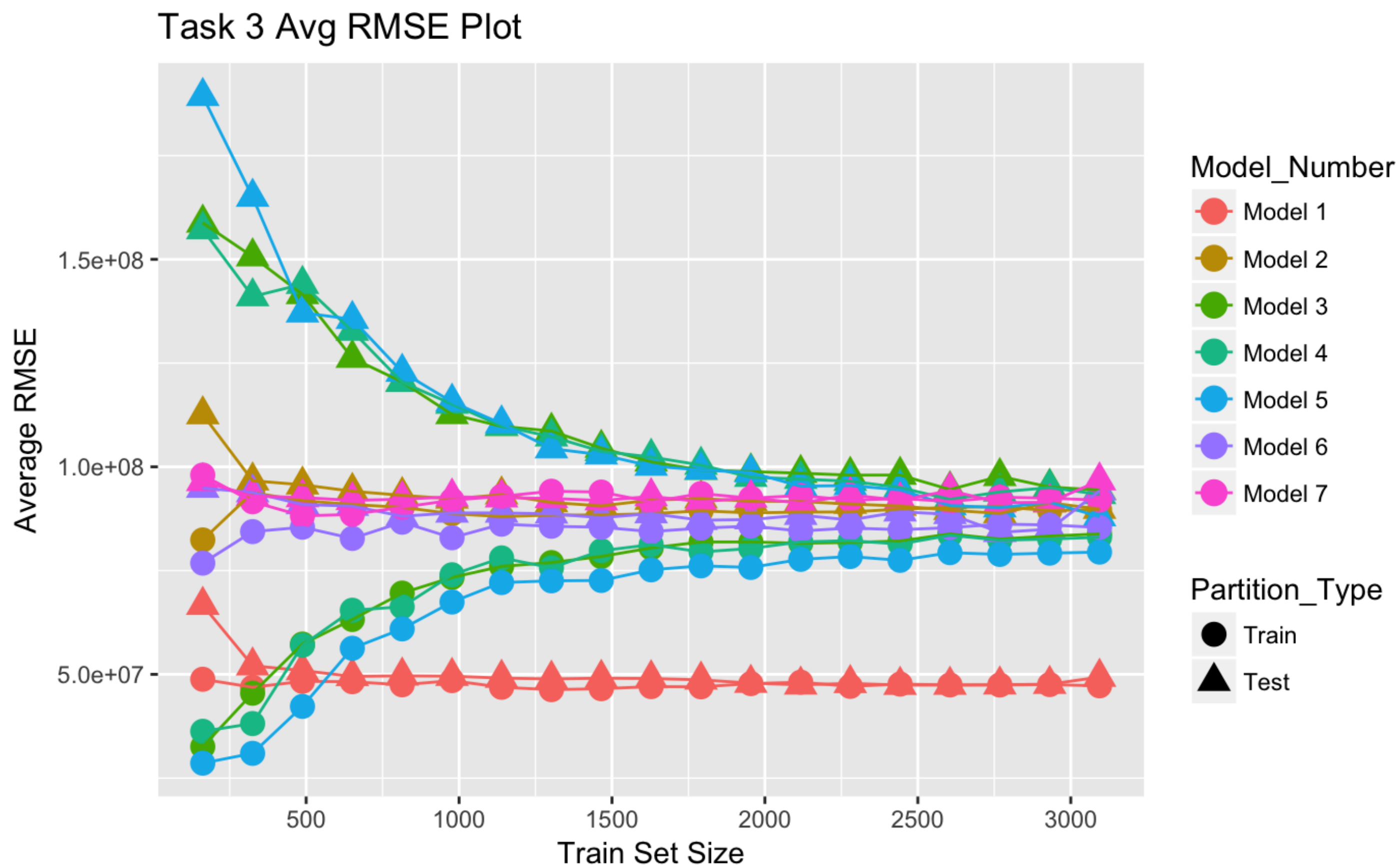
Hide

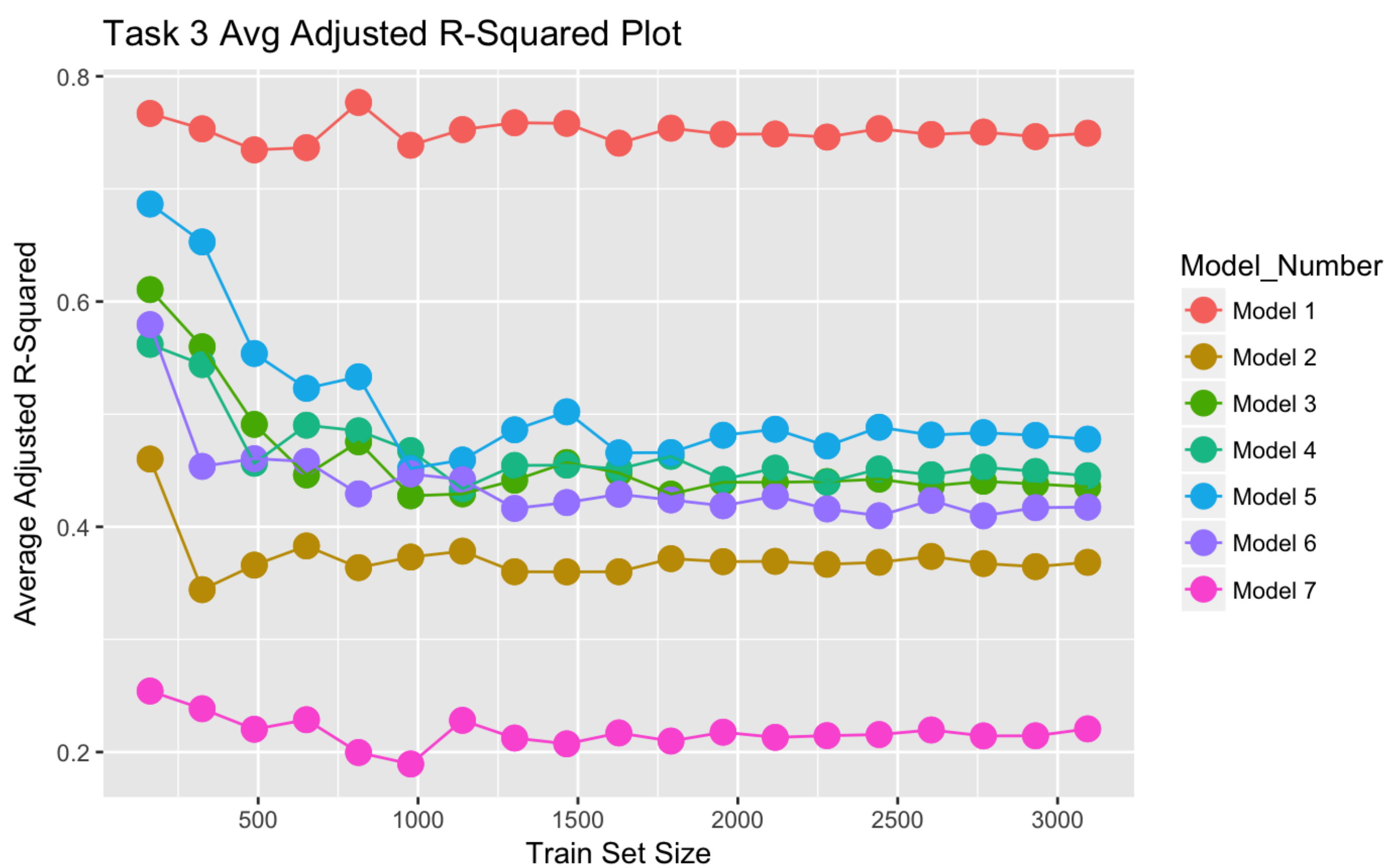
```
## Pre-processing for Task 3, Model 7
# Create a data frame with only a few of the most important parameters defined by the va
rImp function
t3_m7_imp = as.character(t3_importance$Group[1:2])
t3_m7_imp = gsub("`", "", t3_m7_imp)
df3_m7 = df3[,c("Gross", t3_m7_imp)]
## Task 3, Model 7
t3_m7 = Gross~.
t3_m7_tt_list = list()
t3_m7_model_list = list()
for (i in partition_list) {
  t3_tt_temp = rep.run.lm(df3_m7, i, t3_m7)
  t3_m7_tt_list = append(t3_m7_tt_list, t3_tt_temp[1:2])
  t3_m7_model_list = append(t3_m7_model_list, t3_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleading
```

Hide

```
### Evauation Part 6 - Plot all models on one plot
t3_models = list(t2_m3_tt_list, t3_m2_tt_list, t3_m3_tt_list, t3_m4_tt_list, t3_m5_tt_li
st, t3_m6_tt_list, t3_m7_tt_list)  # List of models for this task
t3_n_models = length(t3_models)
# Create a large list of all partitions of train and test for all models
t3_mall_tt_list = list()
for (i in t3_models) {t3_mall_tt_list = append(t3_mall_tt_list, i)}
# Create the final plot df
t3_plot_df = build.plot.df(t3_mall_tt_list, t3_n_models)
# Plot all the models
t3_plot = plot.lm(t3_plot_df, seq(0,3000,500), "Task 3 Avg RMSE Plot", seq(0,3000,500),
"Task 3 Avg Adjusted R-Squared Plot", n_mod = t3_n_models)
```





Q: Explain which categorical variables you used, and how you encoded them into features.

A: I converted "Awards", "Genre", "Actors", "Rated", and "Directors" to either numeric features or multiple features of binary that can be used in lm(). I converted each categorical, explanatory variable as follows:

- Awards - I searched for and removed numbers preceeding “win” or “nomination” and numbers after “nominations for” (actually just “for”) and “Won”. I put all of these numbers into 4 columns (2 columns for nominations and 2 columns for wins). I also created a column for rows that had N/A. After creating these 5 new rows, I added the win columns and NA column and created a new column for wins, and I added the nomination columns and NA column and created a new column for nominations. I spot checked my two new columns for Wins (Award_Wins) and Nominations (Award_Nominations) vs the old Awards column for errors in groups of 50 rows across the entire length of the data frame.

- Genre - I used the DocumentTermMatrix() function from the tm package to build a DocumentTermMatrix (a matrix where the columns are all available genre categories and where each column contains a 1 or 0 depending on whether or not the movie was in that genre) from the Genre column. I used all genres in the models (i.e. did not filter any genres out). After converting the DocumentTermMatrix to a data frame, I merged that data frame to my original data frame.

- Actors - I used the DocumentTermMatrix() function from the tm package to build a DocumentTermMatrix (a matrix where the columns are actors and where each column contains a 1 or 0 depending on whether or not the actor was in that movie) from the Actor column. In the DocumentTermMatrix() function, I used a subset of all actors in all movies, defined as a dictionary, where only actors in >10 movies were included. This dictionary limited the amount of actor variables in my final data frame. Also, I defined a new tokenizer function which did not split actors names by whitespace. The tokenizer function split actors on actors full name.

- Rated - I used the DocumentTermMatrix() function from the tm package to build a DocumentTermMatrix (a matrix where the columns are Ratings and where each column contains a 1 or 0 depending on whether or not movie received that rating or not) from the Rated column. Since the DocumentTermMatrix() function ignores word matches with 2 or less characters in a string and the Rated column has ratings like “R” and “G”, I had to update the “control” argument to set wordLengths to include strings with 1 or more characters.

- Directors - I processed this categorical, explanatory variable in the exact same way I processed the Actors column, except that I only used Directors which directed >= 7 movies.

4. Numeric and categorical variables

Try to improve the prediction quality as much as possible by using both numeric and non-numeric variables from **Tasks 2 & 3**.

Hide

```
### TODO: Build & evaluate model 4 (numeric & converted non-numeric variables)
options(warn=-1)
### Evaluation Part 1 - Part 5
## Task 4, Model 1 - Same as Task 2, Model 2 (current best model for comparison)
# t4_m1 = t2_m2 = Gross~.+I(Budget^1.6)+I(imdbVotes^0.9)
## Pre-processing for Task 4, Model 2 - Build a data frame with the most important numer
ic variables (t2_m2) and most important categorical (t3_m6) variables
df4_m2 = df1
# Add a join key to both DataFrames
df4_m2$JoinKey = seq(1, dim(df4_m2)[1], by=1)
df3_m6$JoinKey = seq(1, dim(df3_m6)[1], by=1)
# Merge the two DataFrames
df4_m2 = merge(x=df4_m2, y=df3_m6, by.x="JoinKey", by.y= "JoinKey")
# Remove JoinKeys and correct Gross columns
df4_m2$JoinKey = NULL
df3_m6$JoinKey = NULL
df4_m2$Gross.y = NULL
df4_m2$Gross = df4_m2$Gross.x
df4_m2$Gross.x = NULL
## Task 4, Model 2 - Build a model using the most important numeric (t2_m2) and categori
cal (t3_m6) variables
t4_m2 = Gross~.+I(Budget^1.6)+I(imdbVotes^0.9)
t4_m2_tt_list = list()
t4_m2_model_list = list()
for (i in partition_list) {
  t4_tt_temp = rep.run.lm(df4_m2, i, t4_m2)
  t4_m2_tt_list = append(t4_m2_tt_list, t4_tt_temp[1:2])
  t4_m2_model_list = append(t4_m2_model_list, t4_tt_temp[3])
}
```

```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleading
```

Hide

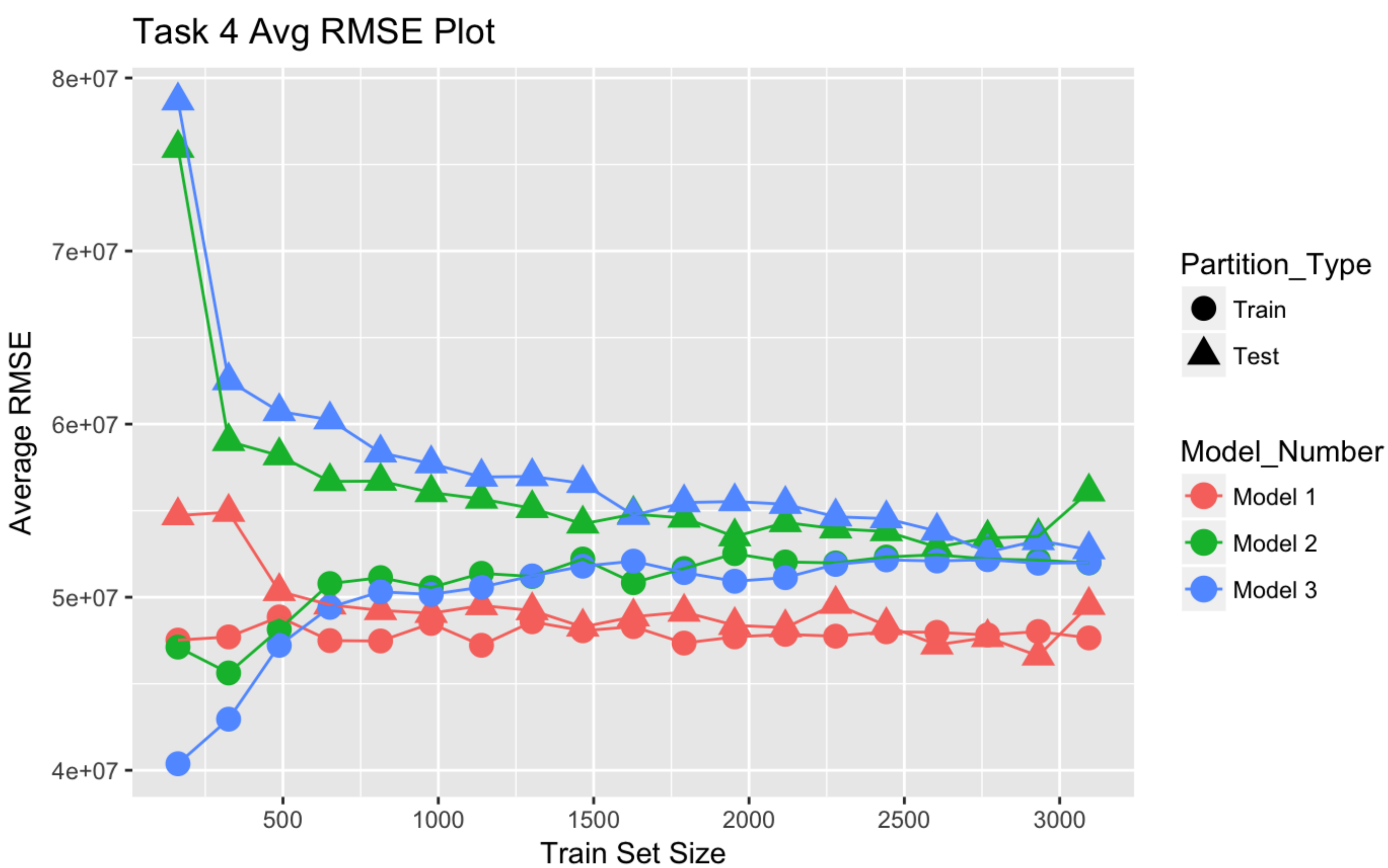
```
## Pre-processing for Task 4, Model 3 - Build a data frame with the most important numer
ic (t2_m2) and even more categorical variables
t3_imp_temp = as.character(t3_importance$Group[1:40])
t3_imp_temp = gsub("`", "", t3_imp_temp)
df3_temp = df3[,c("Gross", t3_imp_temp)]
df4_m3 = df1[,-8]
# Add a join key to both DataFrames
df4_m3$JoinKey = seq(1, dim(df4_m3)[1], by=1)
df3_temp$JoinKey = seq(1, dim(df3_temp)[1], by=1)
# Merge the two DataFrames
df4_m3 = merge(x=df4_m3, y=df3_temp, by.x="JoinKey", by.y= "JoinKey")
# Remove JoinKeys and correct Gross columns
df4_m3$JoinKey = NULL
df3_temp$JoinKey = NULL
df4_m3$Gross.y = NULL
df4_m3$Gross = df4_m3$Gross.x
df4_m3$Gross.x = NULL
## Task 4, Model 3 - Build a model using the most important numeric (t2_m2) and even mor
e categorical variables
t4_m3 = Gross~.+I(Budget^1.6)+I(imdbVotes^0.9)
t4_m3_tt_list = list()
t4_m3_model_list = list()
for (i in partition_list) {
  t4_tt_temp = rep.run.lm(df4_m3, i, t4_m3)
  t4_m3_tt_list = append(t4_m3_tt_list, t4_tt_temp[1:2])
  t4_m3_model_list = append(t4_m3_model_list, t4_tt_temp[3])
}
```

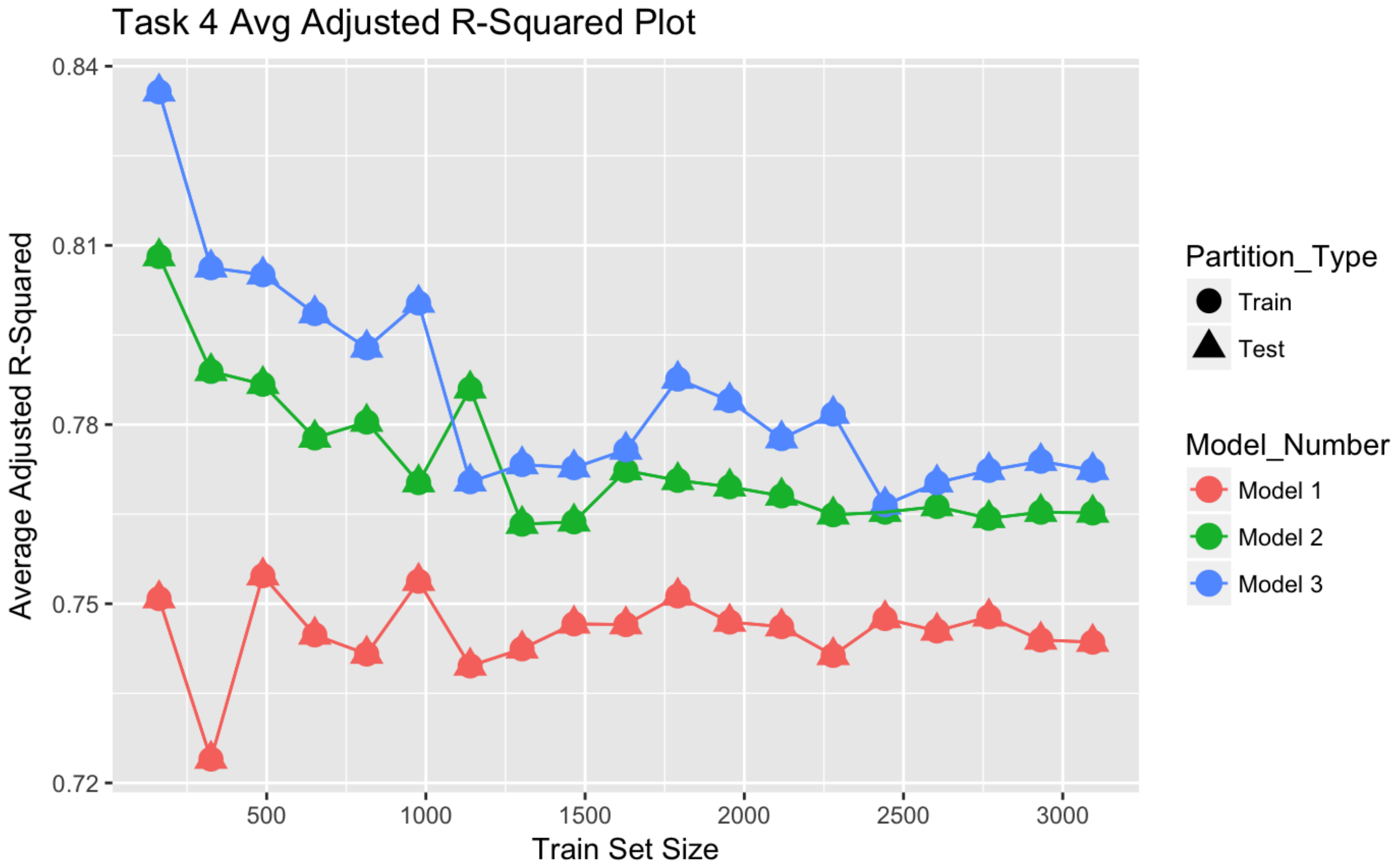
```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleading
```

Hide

```
### Evauation Part 6 - Plot all models on one plot
t4_models = list(t2_m2_tt_list, t4_m2_tt_list, t4_m3_tt_list)  # List of models for this
 task
t4_n_models = length(t4_models)
# Create a large list of all partitions of train and test for all models
t4_mall_tt_list = list()
for (i in t4_models) {t4_mall_tt_list = append(t4_mall_tt_list, i)}
# Create the final plot df
t4_plot_df = build.plot.df(t4_mall_tt_list, t4_n_models)
# Plot all the models
t4_plot = plot.lm(t4_plot_df, seq(0,3000,500), "Task 4 Avg RMSE Plot", seq(0,3000,500),
"Task 4 Avg Adjusted R-Squared Plot", n_mod = t4_n_models)
```





Hide

```
options(warn=-1)
# Check variable importance of Task 4, Model 2 of partition 0.75 (as this partition uses
 enough train data to give accurate predictions) using the Caret R package
t4_importance = varImp(t4_m3_model_list[[15]][[1]], scale = FALSE)
t4_importance$Group = rownames(t4_importance)
t4_importance = t4_importance[order(t4_importance$Overall, decreasing = TRUE),]
print(t4_importance)
```

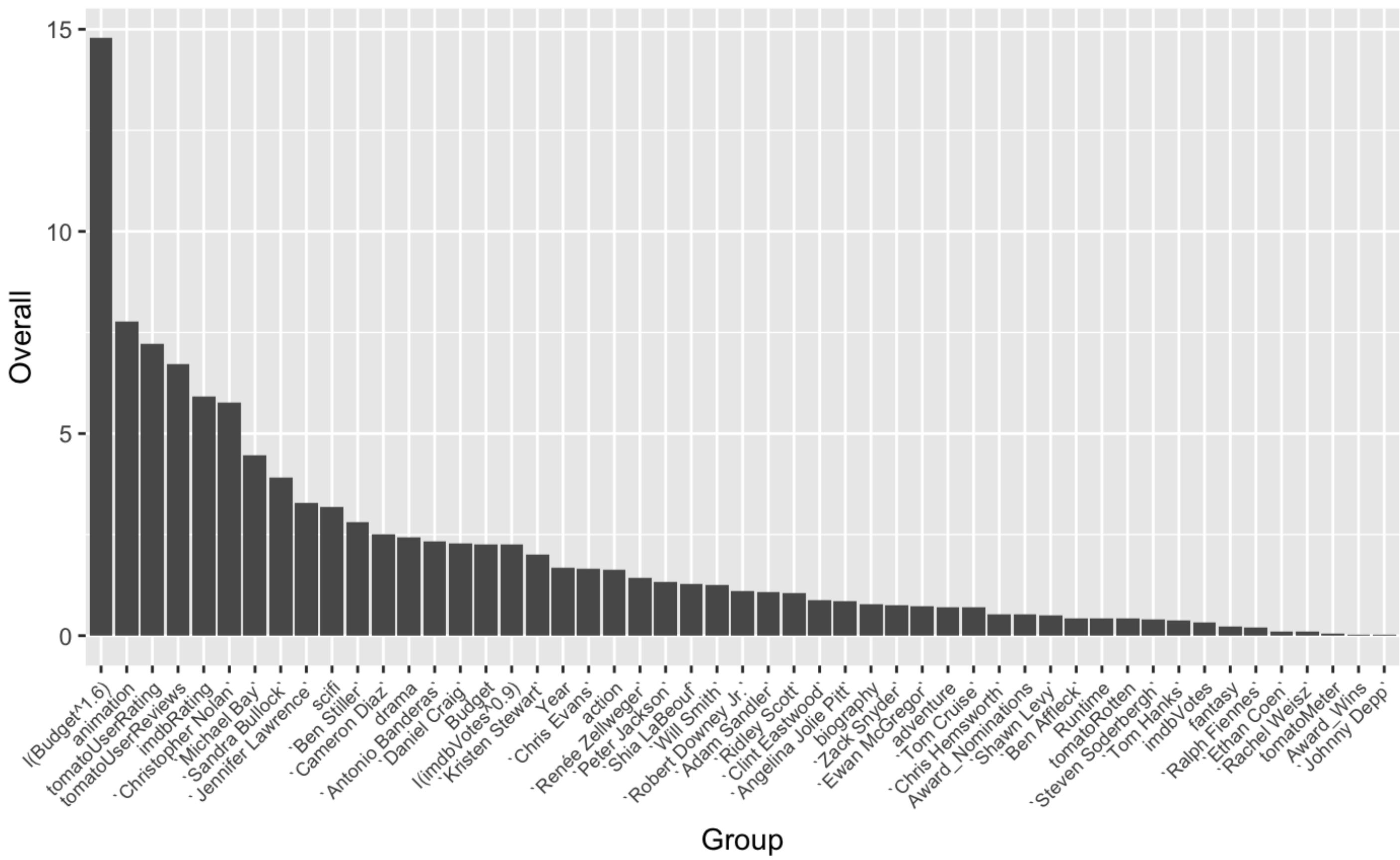
	Overall <dbl>	Group <chr>
I(Budget^1.6)	14.78540777	I(Budget^1.6)
animation	7.76833398	animation
tomatoUserRating	7.20942507	tomatoUserRating
tomatoUserReviews	6.72564188	tomatoUserReviews
imdbRating	5.91802796	imdbRating
`Christopher Nolan`	5.76089074	`Christopher Nolan`
`Michael Bay`	4.47394477	`Michael Bay`
`Sandra Bullock`	3.92059673	`Sandra Bullock`
`Jennifer Lawrence`	3.28791041	`Jennifer Lawrence`

	Overall	Group
	<dbl>	<chr>
scifi	3.18063189	scifi

1-10 of 51 rows Previous **1** 2 3 4 5 6 Next

Hide

```
t4_importance$Group <- factor(t4_importance$Group, levels=unique(as.character(t4_importa
nce$Group)) )
ggplot(t4_importance, aes(y=Overall, x=Group)) + geom_bar(stat="identity") + theme(axis.
text.x = element_text(angle = 45, hjust = 1, size=7,))
```



When adding any amount of categorical varialbles, which were converted to columns of binary, or columns that bin to numerical columns, I see an increase in the adjusted r-squared value and an increase in the RMSE value.

5. Additional features

Now try creating additional features such as interactions (e.g. `is_genre_comedy x is_budget_greater_than_3M`) or deeper analysis of complex variables (e.g. text analysis of full-text columns like `Plot`).

Hide

```
### TODO: Build & evaluate model 5 (numeric, non-numeric and additional features)
options(warn=-1)
### General Pre-processing for Task 5
df5 = df4_m3
df5$Plot = df$Plot
### Evaluation Part 1 - Part 5
partition_list = list(0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.7
5,0.8,0.85,0.9,0.95)  # List to define train partitions
## Task 5, Model 1 - Same as Task 2, Model 2 (current best model for comparison)
# t4_m1 = t2_m2 = Gross~.+I(Budget^1.6)+I(imdbVotes^0.9)
## Pre-processing for Task 5, Model 2
# Replace Plot with a collection of binary columns
df5_m2 = df5
# Create a vector of the Plot column
plot = df5_m2$Plot
# Build the Corpus
plots_corp = VCorpus(VectorSource(plot))
# Remove punctuation, stop words, etc
plots_corp = tm_map(plots_corp, removeNumbers)
plots_corp = tm_map(plots_corp, removePunctuation)
plots_corp = tm_map(plots_corp, content_transformer(tolower))
plots_corp = tm_map(plots_corp, removeWords, stopwords("english"))
plots_corp = tm_map(plots_corp, stripWhitespace)
plots_corp = tm_map(plots_corp, stemDocument)
# Create a tokenizer for DTM to use
spacesplit_tokenizer = function(x) {unlist(strsplit(as.character(x), " "))}
# Build the DocumentTermMatrix
plots_dtm = DocumentTermMatrix(plots_corp, control=list(tokenize=commasplit_tokenizer, t
olower=FALSE))
# Build a dictionary of the most frequently occurring terms
plots_dict = findFreqTerms(plots_dtm, 200)
# Rebuild the DocumentTermMatrix using the plots_dict
plots_dtm = DocumentTermMatrix(plots_corp, control=list(tokenize=commasplit_tokenizer, d
ictionary=plots_dict, tolower=FALSE))
# Convert DocumentTermMatrix to a DataFrame
plots_dtm_matrix = as.matrix(plots_dtm)
plots_dtm_df = as.data.frame(plots_dtm_matrix)
# Add a join key to both DataFrames
plots_dtm_df$JoinKey = seq(1, dim(plots_dtm_df)[1], by=1)
df5_m2$JoinKey = seq(1, dim(df5_m2)[1], by=1)
# Merge the two DataFrames
df5_m2 = merge(x=df5_m2, y=plots_dtm_df, by.x="JoinKey", by.y= "JoinKey")
# Drop JoinKey and plots
df5_m2$JoinKey = NULL
df5_m2$Plot = NULL
## Task 5, Model 2
t5_m2 = Gross~I(Budget^1.6)+I(imdbVotes^0.9)+.
t5_m2_tt_list = list()
t5_m2_model_list = list()
for (i in partition_list) {
  t5_tt_temp = rep.run.lm(df5_m2, i, t5_m2)
  t5_m2_tt_list = append(t5_m2_tt_list, t5_tt_temp[1:2])
```

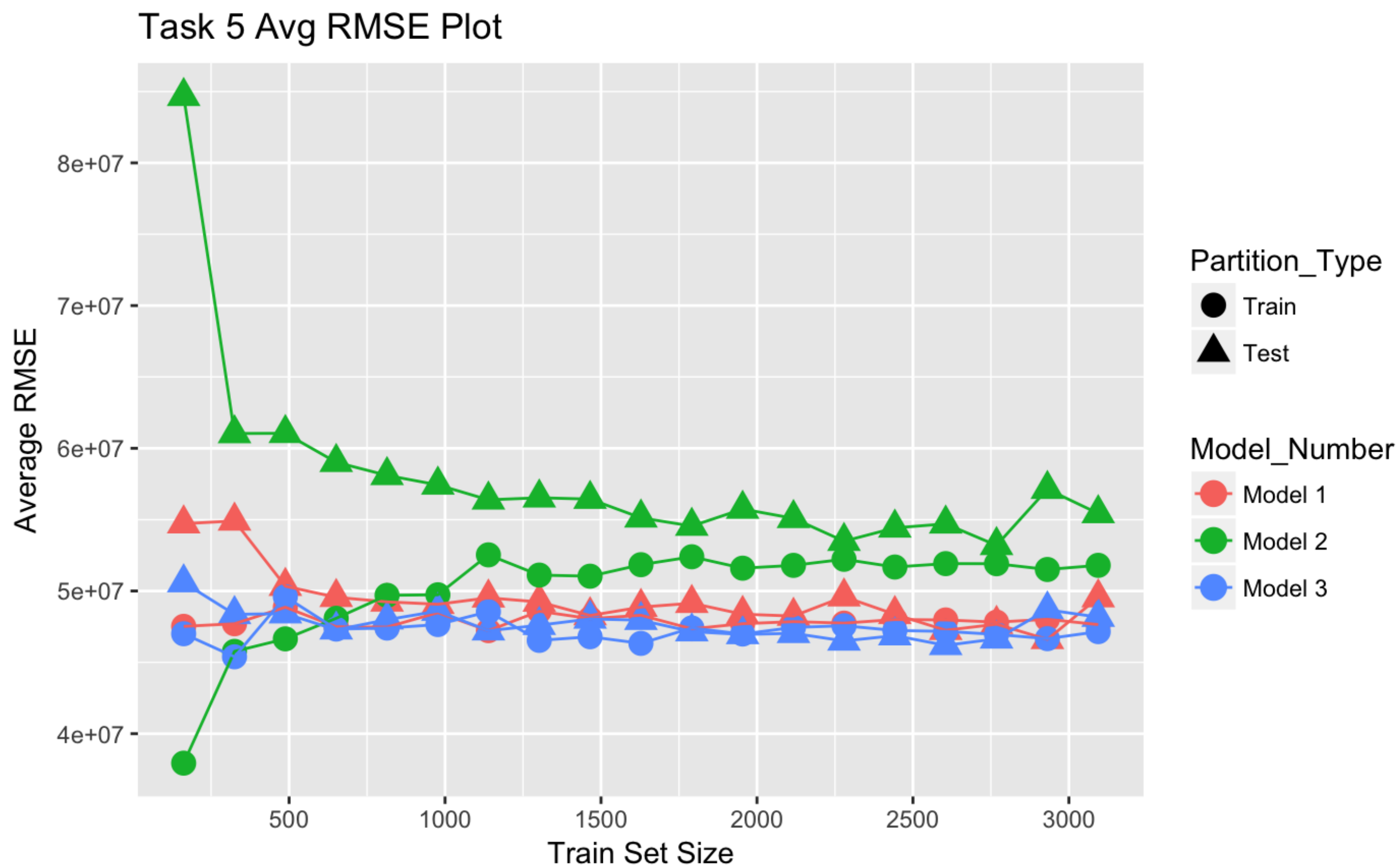
```
  t5_m2_model_list = append(t5_m2_model_list, t5_tt_temp[3])
}
```

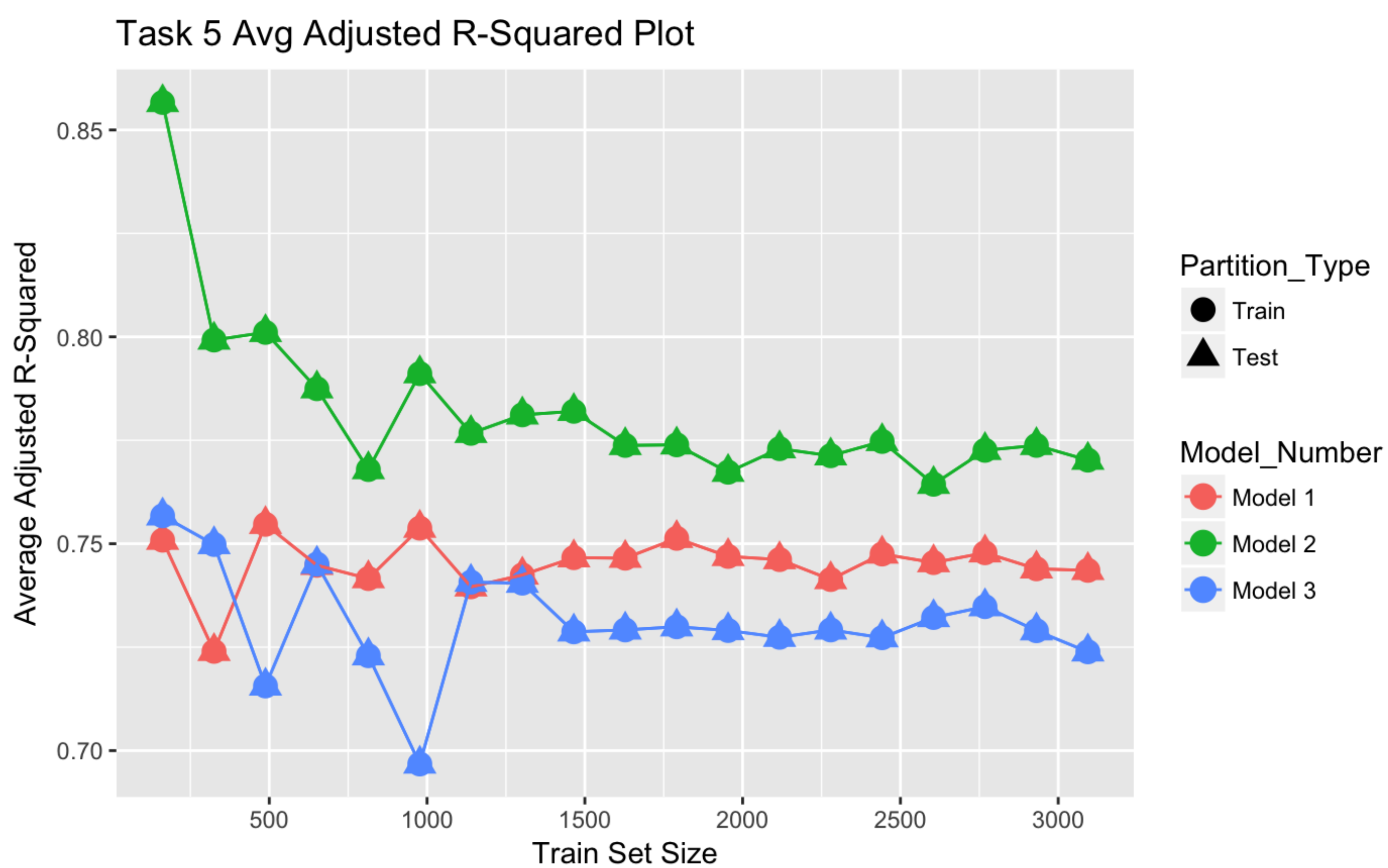
```
prediction from a rank-deficient fit may be misleadingprediction from a rank-deficient f
it may be misleadingprediction from a rank-deficient fit may be misleadingprediction fro
m a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misl
eadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defic
ient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredicti
on from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may b
e misleadingprediction from a rank-deficient fit may be misleadingprediction from a rank
-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpr
ediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit
may be misleadingprediction from a rank-deficient fit may be misleadingprediction from a
rank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadi
ngprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leadingprediction from a rank-deficient fit may be misleadingprediction from a rank-defi
```

```
cient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredict
ion from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may
be misleadingprediction from a rank-deficient fit may be misleadingprediction from a ran
k-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadingp
rediction from a rank-deficient fit may be misleadingprediction from a rank-deficient fi
t may be misleadingprediction from a rank-deficient fit may be misleadingprediction from
a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mislea
dingprediction from a rank-deficient fit may be misleadingprediction from a rank-deficie
nt fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction
from a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be m
isleadingprediction from a rank-deficient fit may be misleadingprediction from a rank-de
ficient fit may be misleadingprediction from a rank-deficient fit may be misleadingpredi
ction from a rank-deficient fit may be misleadingprediction from a rank-deficient fit ma
y be misleadingprediction from a rank-deficient fit may be misleadingprediction from a r
ank-deficient fit may be misleadingprediction from a rank-deficient fit may be misleadin
gprediction from a rank-deficient fit may be misleadingprediction from a rank-deficient
fit may be misleadingprediction from a rank-deficient fit may be misleadingprediction fr
om a rank-deficient fit may be misleadingprediction from a rank-deficient fit may be mis
leading
```

Hide

```
## Pre-processing for Task 5, Model 3
df5_m3 = df2
df5_m3$Plot = df$Plot
df5_m3 = df5_m3[,c("Gross","imdbVotes","Budget")]
# Add a join key to both DataFrames
plots_dtm_df$JoinKey = seq(1, dim(plots_dtm_df)[1], by=1)
df5_m3$JoinKey = seq(1, dim(df5_m3)[1], by=1)
# Merge the two DataFrames
df5_m3 = merge(x=df5_m3, y=plots_dtm_df, by.x="JoinKey", by.y= "JoinKey")
# Drop JoinKey and plots
df5_m3$JoinKey = NULL
df5_m3$Plot = NULL
## Task 5, Model 3
t5_m3 = Gross~I(Budget^1.6)+I(imdbVotes^0.9)+.
t5_m3_tt_list = list()
t5_m3_model_list = list()
for (i in partition_list) {
  t5_tt_temp = rep.run.lm(df5_m3, i, t5_m3)
  t5_m3_tt_list = append(t5_m3_tt_list, t5_tt_temp[1:2])
  t5_m3_model_list = append(t5_m3_model_list, t5_tt_temp[3])
}
### Evauation Part 6 - Plot all models on one plot
t5_models = list(t2_m2_tt_list, t5_m2_tt_list, t5_m3_tt_list)  # List of models for this
 task
t5_n_models = length(t5_models)
# Create a large list of all partitions of train and test for all models
t5_mall_tt_list = list()
for (i in t5_models) {t5_mall_tt_list = append(t5_mall_tt_list, i)}
# Create the final plot df
t5_plot_df = build.plot.df(t5_mall_tt_list, t5_n_models)
# Plot all the models
t5_plot = plot.lm(t5_plot_df, seq(0,3000,500), "Task 5 Avg RMSE Plot", seq(0,3000,500),
"Task 5 Avg Adjusted R-Squared Plot", n_mod = t5_n_models)
```


Task 5 Avg RMSE Plot
Average RMSE
8e+07
7e+07
6e+07
5e+07
4e+07
500
1000
1500
2000
2500
3000
Train Set Size
Partition_Type
Train
Test
Model_Number
Model 1
Model 2
Model 3


Task 5 Avg Adjusted R-Squared Plot
Average Adjusted R-Squared
0.85
0.80
0.75
0.70
500
1000
1500
2000
2500
3000
Train Set Size
Partition_Type
Train
Test
Model_Number
Model 1
Model 2
Model 3

Hide

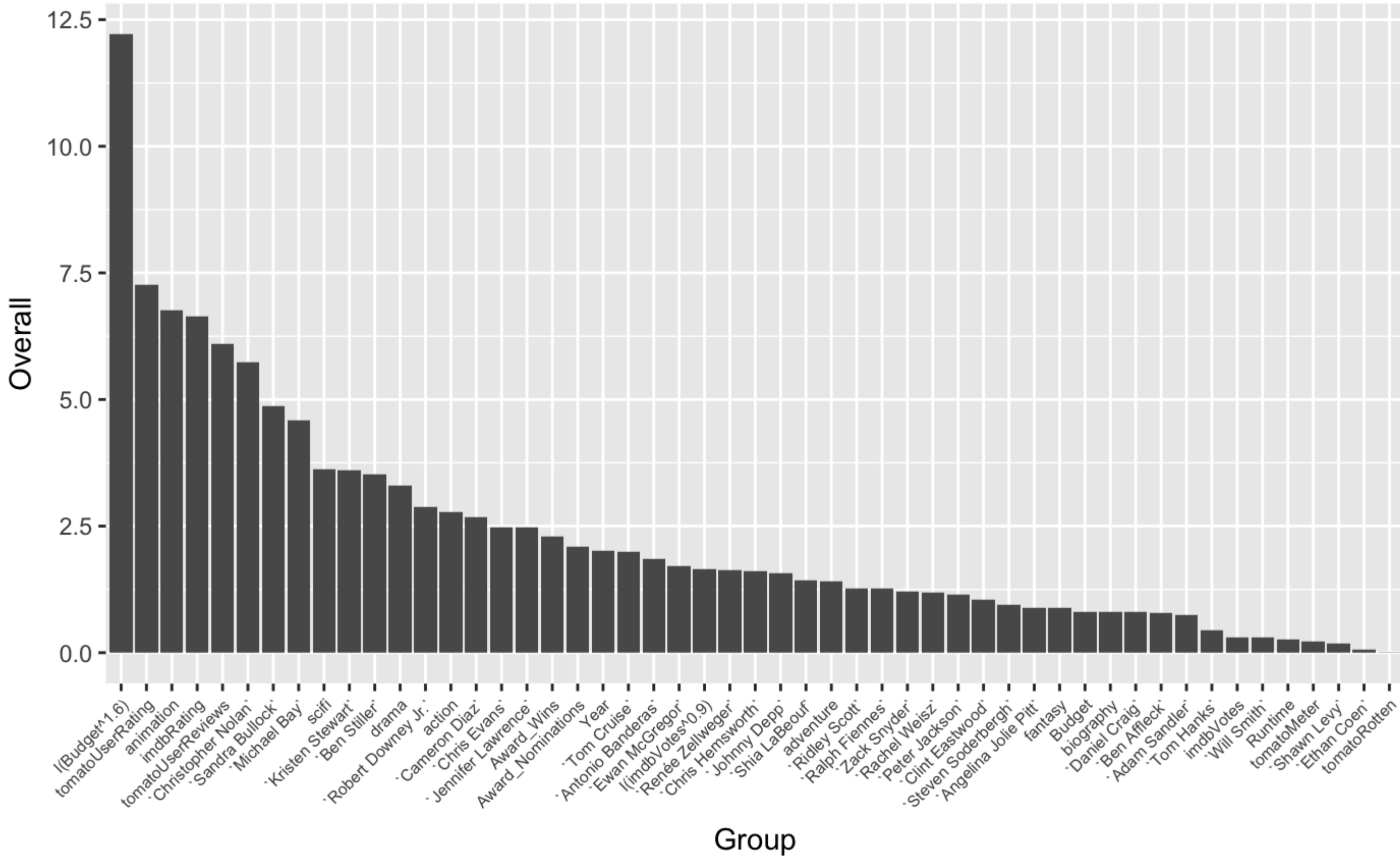
```
options(warn=-1)
# Check variable importance of the best model from Task 5 at partition 0.75 (as this par
tition uses enough train data to give accurate predictions) using the Caret R package
t5_importance = varImp(t5_m2_model_list[[15]][[1]], useModel = "lm", scale = FALSE)
t5_importance$Group = rownames(t5_importance)
t5_importance = t5_importance[order(t5_importance$Overall, decreasing = TRUE),]
print(t5_importance)
```

	Overall <dbl>	Group <chr>
I(Budget^1.6)	12.216228157	I(Budget^1.6)
tomatoUserRating	7.268002910	tomatoUserRating
animation	6.752686173	animation
imdbRating	6.638588168	imdbRating
tomatoUserReviews	6.108089454	tomatoUserReviews
`Christopher Nolan`	5.729211411	`Christopher Nolan`
`Sandra Bullock`	4.862546585	`Sandra Bullock`
`Michael Bay`	4.594580775	`Michael Bay`
scifi	3.618267137	scifi
`Kristen Stewart`	3.596396753	`Kristen Stewart`

1-10 of 51 rows Previous **1** 2 3 4 5 6 Next

Hide

```
t5_importance$Group <- factor(t5_importance$Group, levels=unique(as.character(t5_importa
nce$Group)) )
ggplot(t5_importance, aes(y=Overall, x=Group)) + geom_bar(stat="identity") + theme(axis.
text.x = element_text(angle = 45, hjust = 1, size=6,))
```



Q: Explain what new features you designed and why you chose them.

A: I chose to convert the plot column to multiple columns of binary. By doing this, I was able to use common words in the Plot column to predict the Gross value of a movie. I chose plot because there is a lot of text and information in this column that could be used to produce a better model. However, these attempts tended to increase the adjusted R-squared value and increase the RMSE compared to Task 2, Model 2. Task 5, Model 2, where I used numerical features (from Task 2), plot features (from Task 5), and other categorical features converted to columns of binary (from Task 3), performed almost as well as Task 2, Model 2 for RMSE.