Moving Beyond R and Exploratory Data Analysis

Matthew Renze

@matthewrenze | www.matthewrenze.com

Overview

Beyond Exploratory Data Analysis

Demo

Alternatives to R

Conclusion


This is just the tip of the iceberg!

Other Types of Data Analysis

Descriptive

Exploratory

Inferential

Predictive

Causal

Mechanistic

Advanced Programming Features

Procedural

Functional

Object-oriented

Outbound calls

Inbound calls

Distributed programming

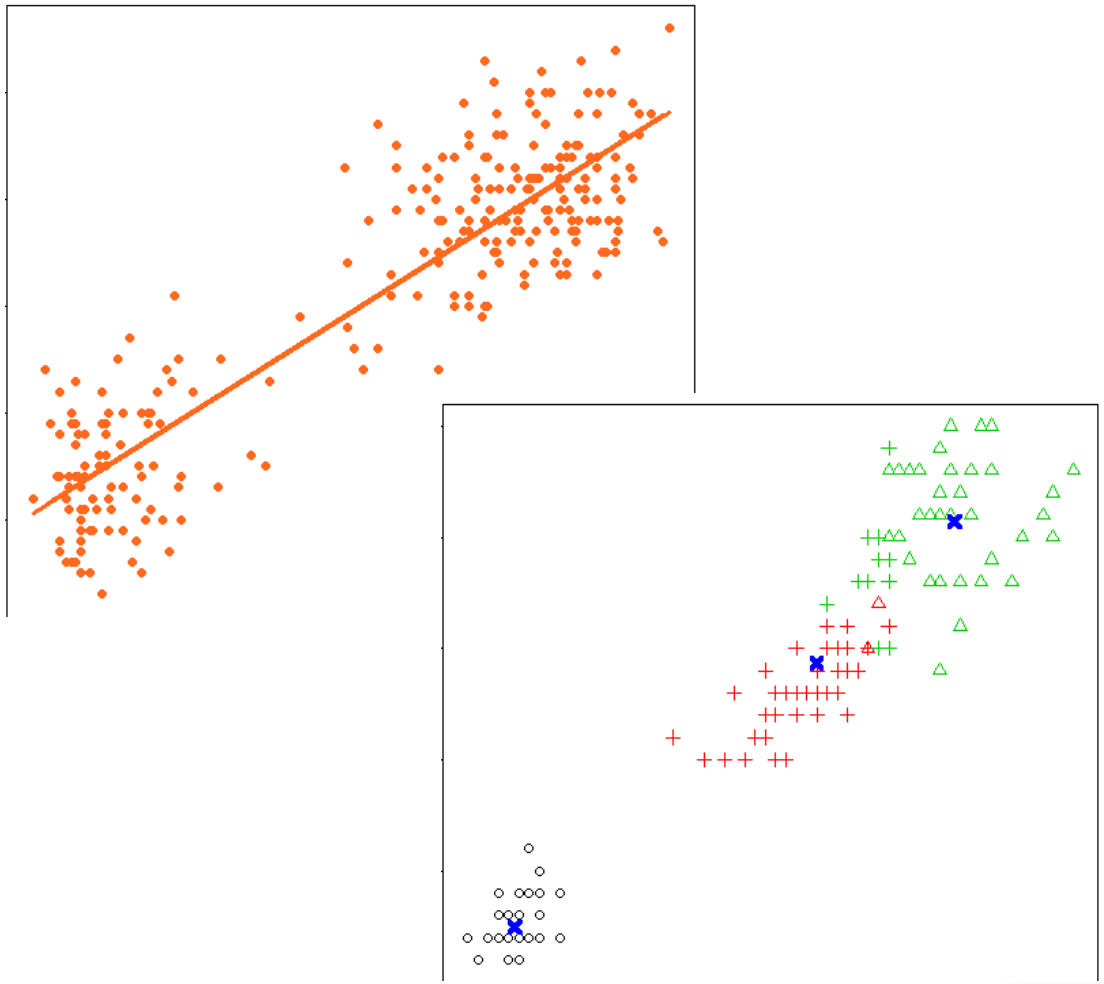
Advanced Statistical Analysis

Statistical modeling

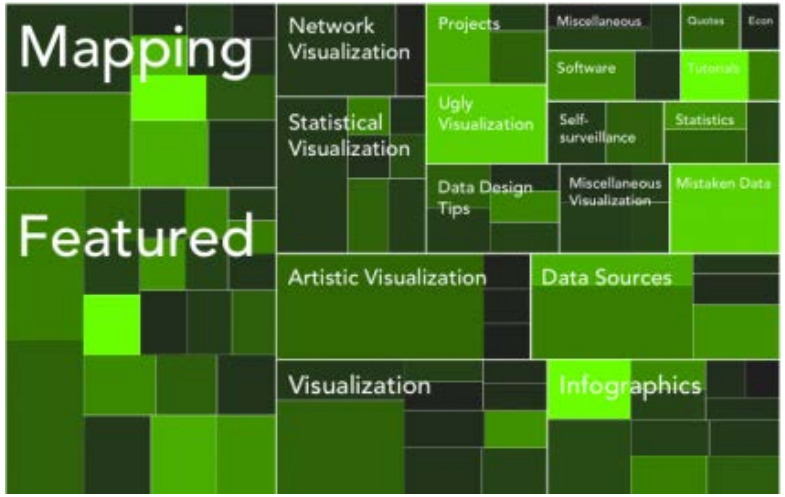
Cluster analysis

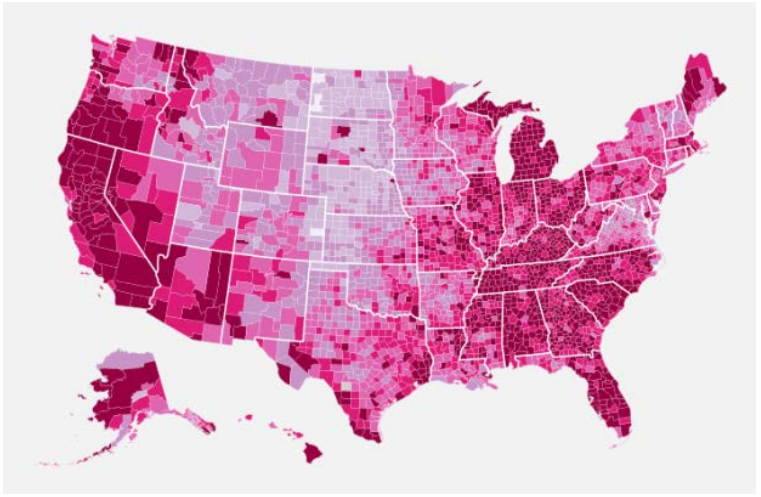
Dimensionality reduction

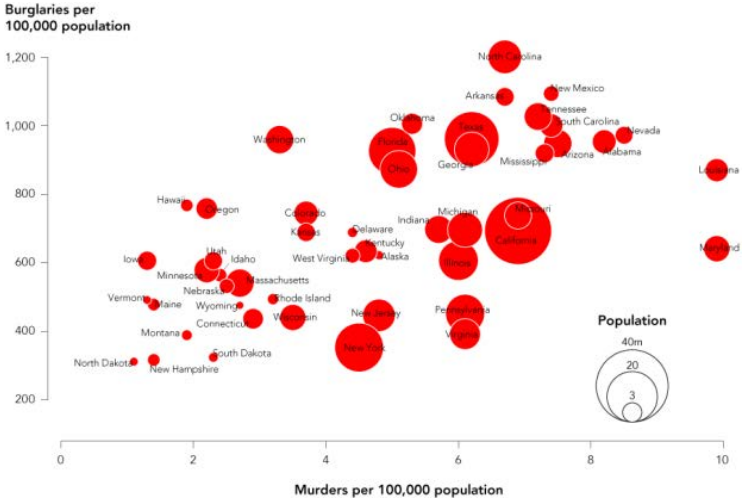
Analysis of Variance (ANOVA)

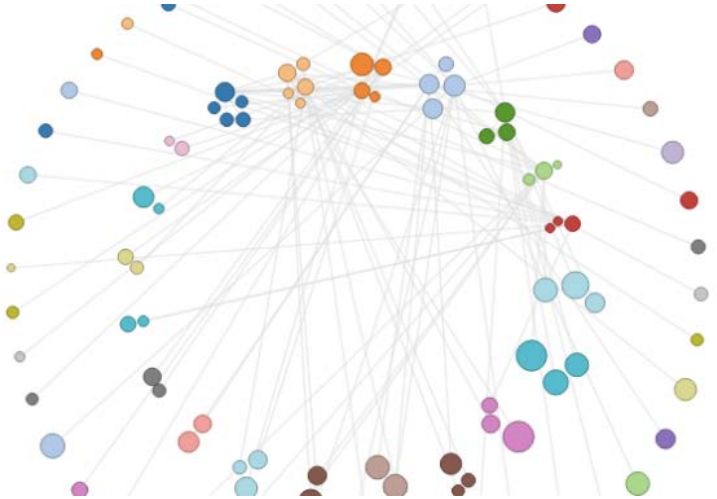


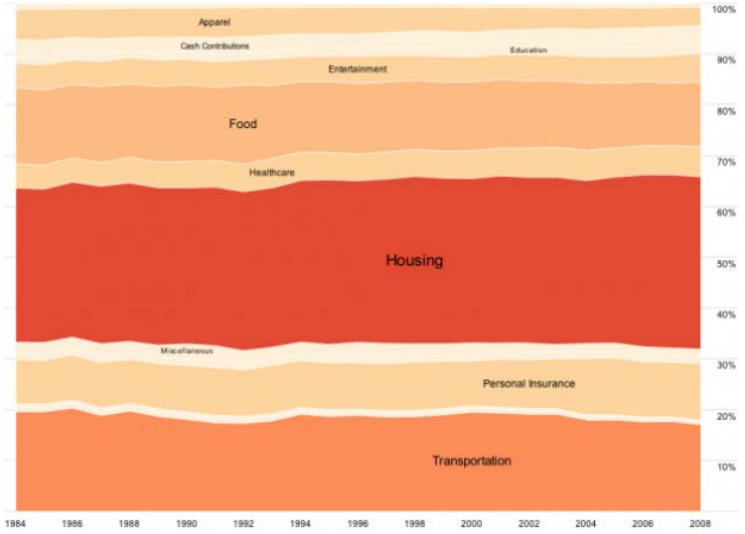
Advanced Data Visualization

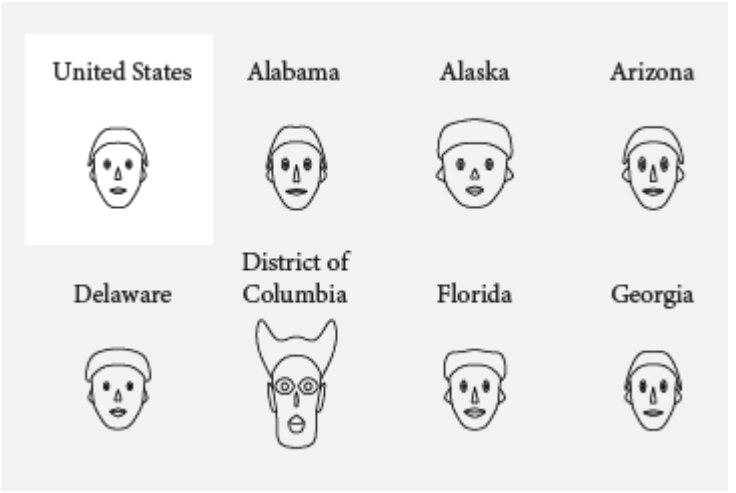






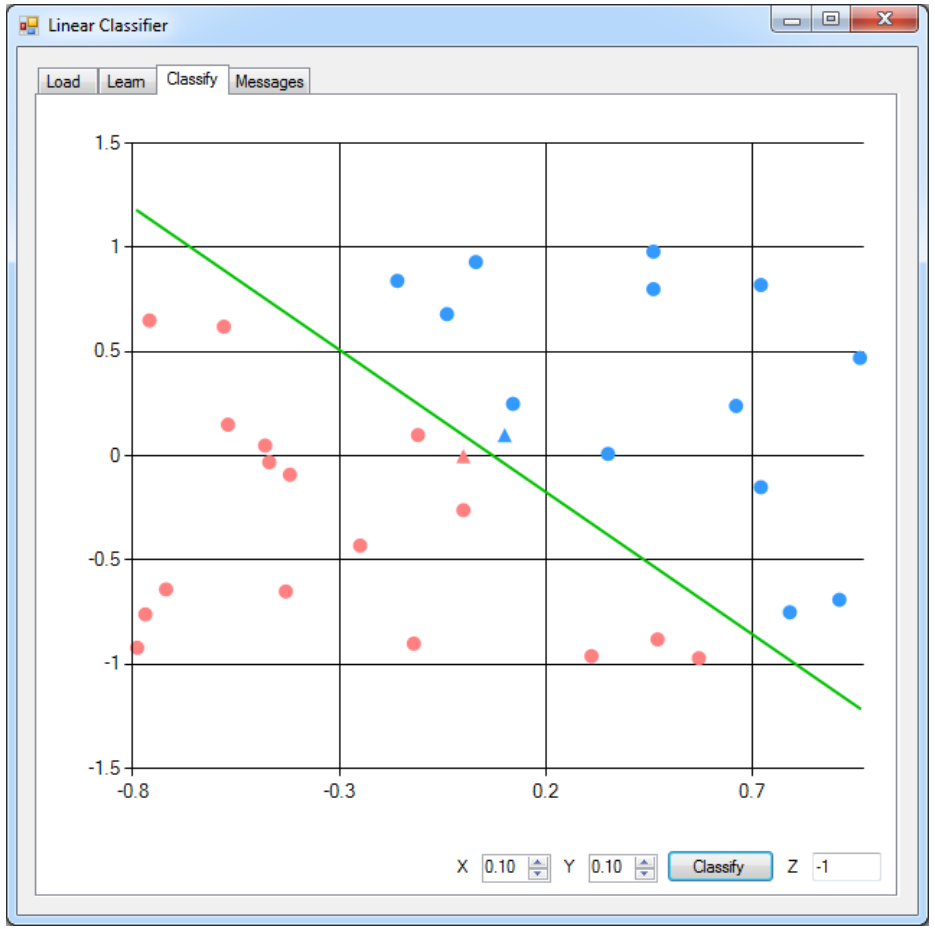






Source: Nathan Yau (www.flowingdata.com)

Data Mining and Machine Learning



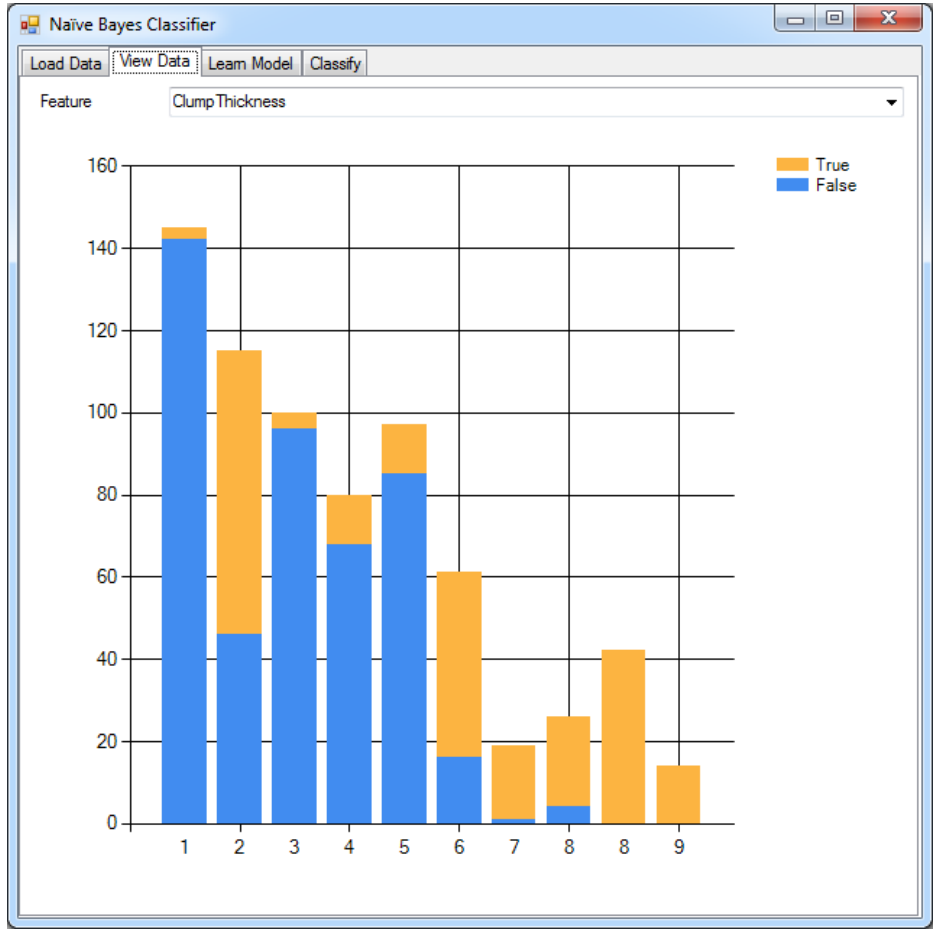


Photo by Danielle Langlois

Load the Data

Scatterplot Matrix

Scatterplot

Linear Regression Model

Plot Linear Model

Correlation Coefficient

Linear Model Summary

Linear Model Prediction

Caveats

Photos by Radomił Binek,
Danielle Langlois, and Frank Mayfield

Load the Data

Scatterplot Matrix

Scatterplot

Create k-Means Clusters

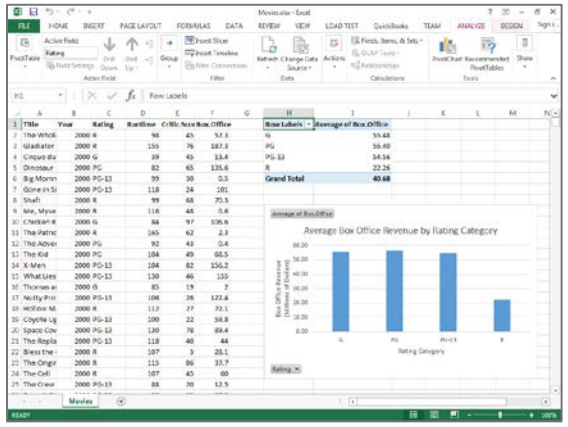
Plot Clusters as Shapes

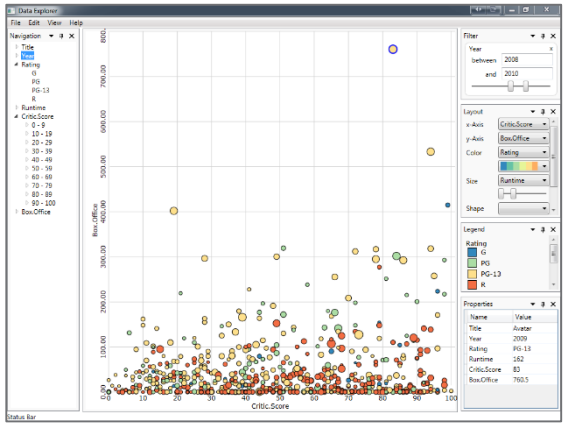
Plot Centers of Clusters

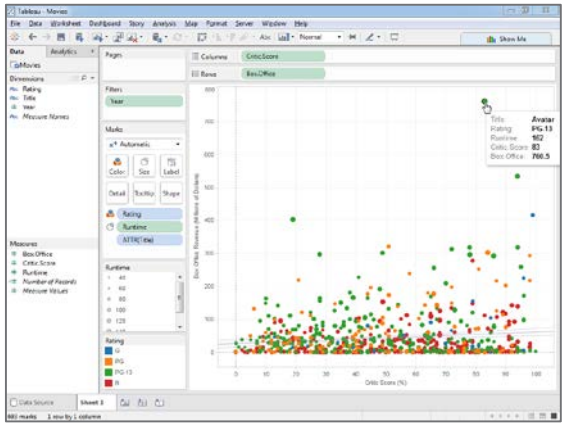
Table of Clusters

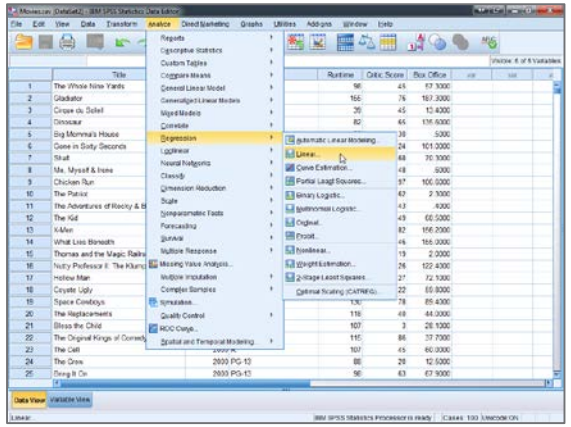
Caveats

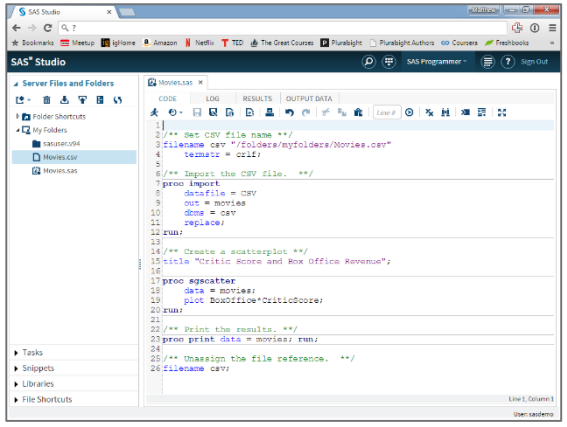
Alternatives to R

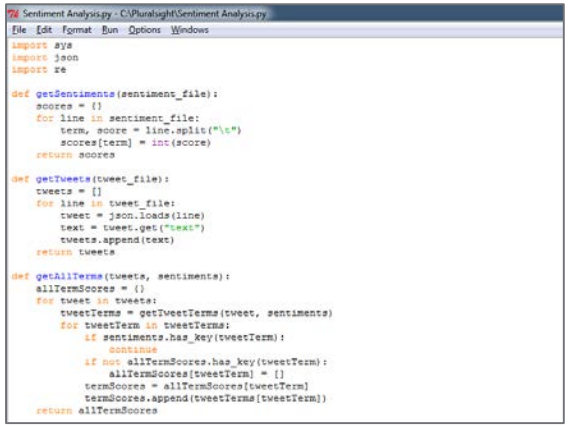












Movies.xlsx - Excel

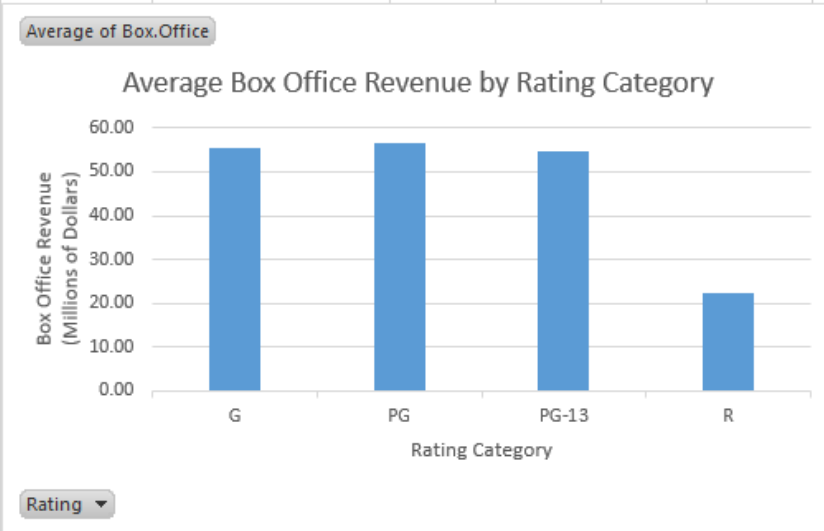
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW LOAD TEST QuickBooks TEAM ANALYZE DESIGN Sign i

PivotTable | Active Field: Rating | Field Settings | Drill Down | Drill Up | Group | Insert Slicer | Insert Timeline | Filter Connections | Refresh | Change Data Source | Actions | Fields, Items, & Sets | OLAP Tools | Relationships | PivotChart | Recommended PivotTables | Show

Active Field | Filter | Data | Calculations | Tools

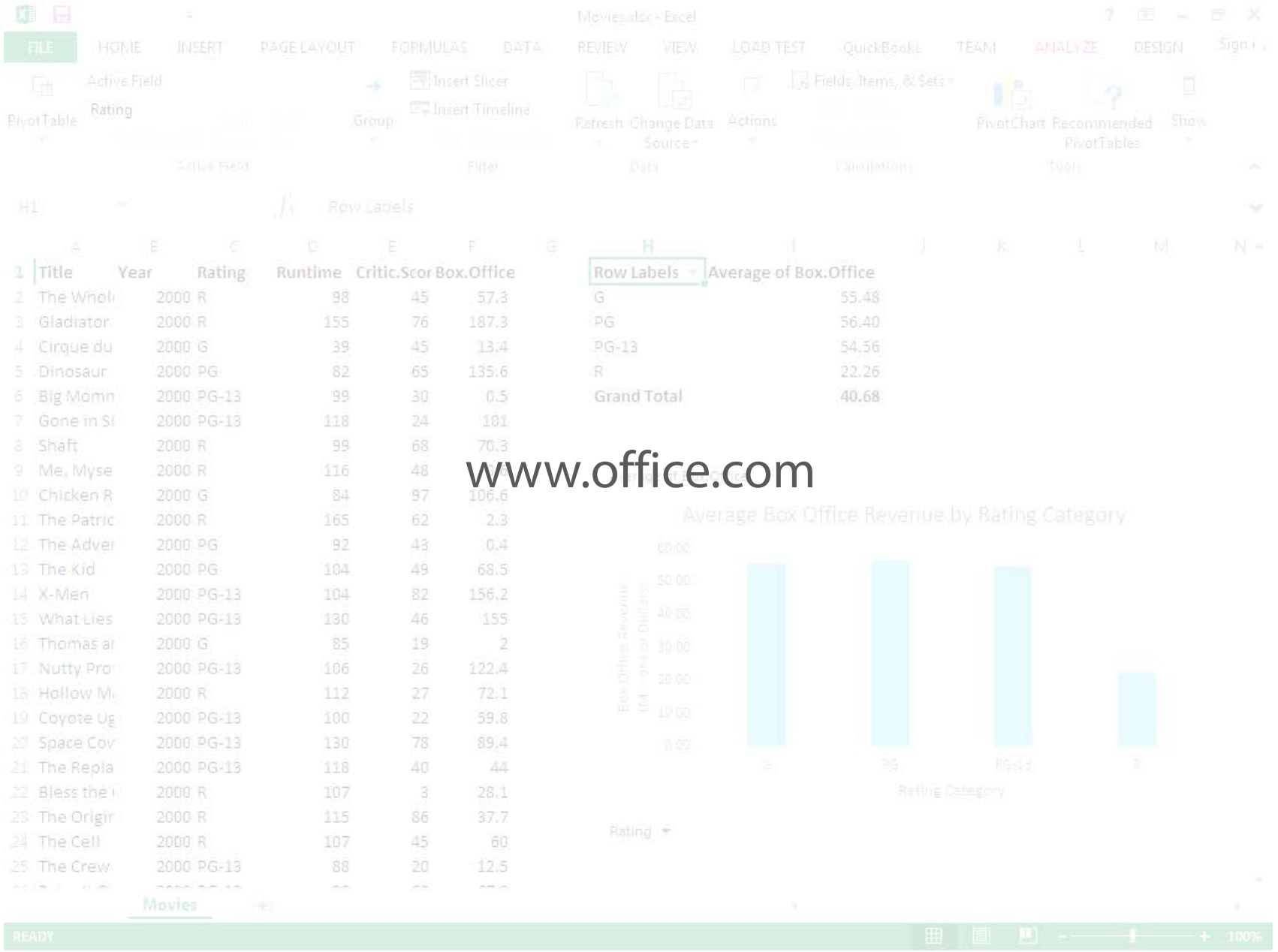
H1 | Row Labels

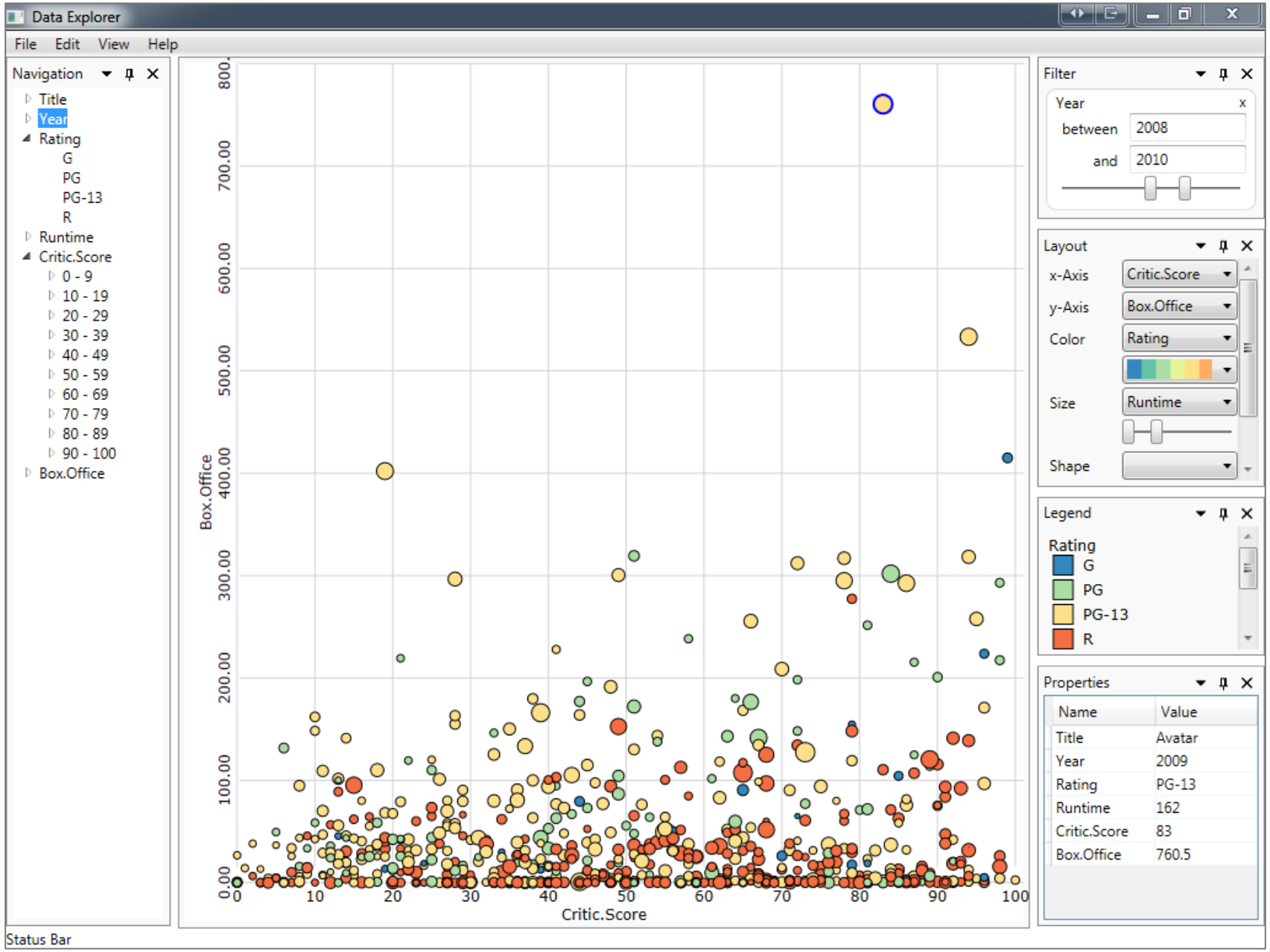
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Title	Year	Rating	Runtime	Critic.Scor	Box.Office		Row Labels	Average of Box.Office					
2	The Whol	2000	R	98	45	57.3		G	55.48					
3	Gladiator	2000	R	155	76	187.3		PG	56.40					
4	Cirque du	2000	G	39	45	13.4		PG-13	54.56					
5	Dinosaur	2000	PG	82	65	135.6		R	22.26					
6	Big Momn	2000	PG-13	99	30	0.5		Grand Total	40.68					
7	Gone in Si	2000	PG-13	118	24	101								
8	Shaft	2000	R	99	68	70.3								
9	Me, Myse	2000	R	116	48	0.6								
10	Chicken R	2000	G	84	97	106.6								
11	The Patric	2000	R	165	62	2.3								
12	The Adver	2000	PG	92	43	0.4								
13	The Kid	2000	PG	104	49	68.5								
14	X-Men	2000	PG-13	104	82	156.2								
15	What Lies	2000	PG-13	130	46	155								
16	Thomas ar	2000	G	85	19	2								
17	Nutty Prof	2000	PG-13	106	26	122.4								
18	Hollow Ma	2000	R	112	27	72.1								
19	Coyote Ug	2000	PG-13	100	22	59.8								
20	Space Cov	2000	PG-13	130	78	89.4								
21	The Repla	2000	PG-13	118	40	44								
22	Bless the	2000	R	107	3	28.1								
23	The Origir	2000	R	115	86	37.7								
24	The Cell	2000	R	107	45	60								
25	The Crew	2000	PG-13	88	20	12.5								



Movies

READY 100%


Movies.xlsx - Excel
FILE
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
LOAD TEST
QuickBooks
TEAM
ANALYZE
DESIGN
PivotTable
Active Field
Rating
Insert Slicer
Insert Timeline
Group
Refresh
Change Data Source
Actions
Fields, Items, & Sets
PivotChart
Recommended PivotTables
Show
Active Field
Filter
Data
Calculations
Tools
H1
Row Labels
Title
Year
Rating
Runtime
Critic.Scor
Box.Office
Row Labels
Average of Box.Office
G
55.48
PG
56.40
PG-13
54.56
R
22.26
Grand Total
40.68
www.office.com
Average Box Office Revenue by Rating Category
Box Office Revenue (in Tens of Dollars)
Rating Category
Rating
Movies
READY
100%


Data Explorer
File
Edit
View
Help
Navigation
Title
Year
Rating
G
PG
PG-13
R
Runtime
Critic.Score
0 - 9
10 - 19
20 - 29
30 - 39
40 - 49
50 - 59
60 - 69
70 - 79
80 - 89
90 - 100
Box.Office
Box.Office
0.00
100.00
200.00
300.00
400.00
500.00
600.00
700.00
800.00
0
10
20
30
40
50
60
70
80
90
100
Critic.Score
Filter
Year
between
2008
and
2010
Layout
x-Axis
Critic.Score
y-Axis
Box.Office
Color
Rating
Size
Runtime
Shape
Legend
Rating
G
PG
PG-13
R
Properties
Name
Value
Title
Avatar
Year
2009
Rating
PG-13
Runtime
162
Critic.Score
83
Box.Office
760.5
Status Bar

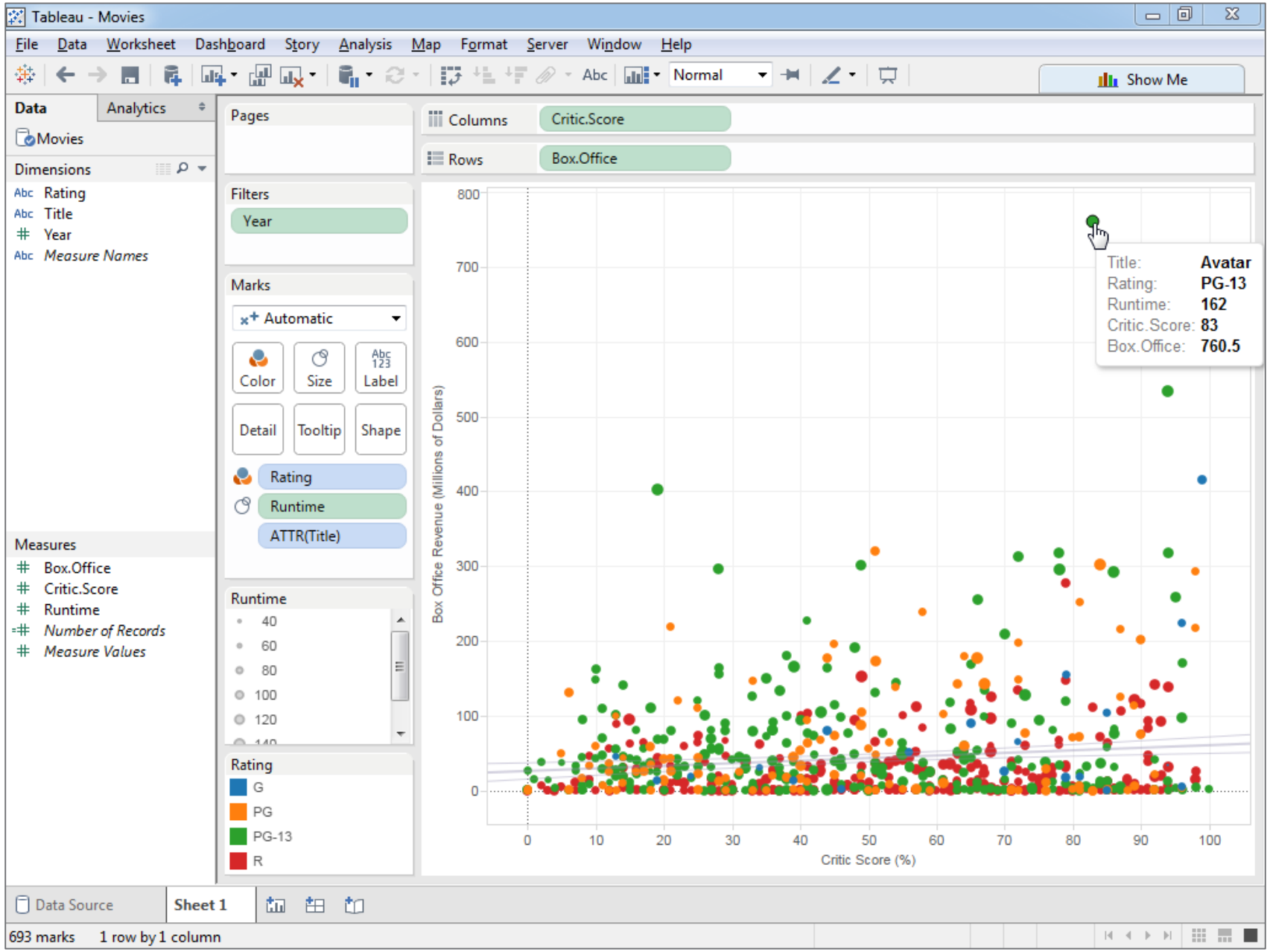
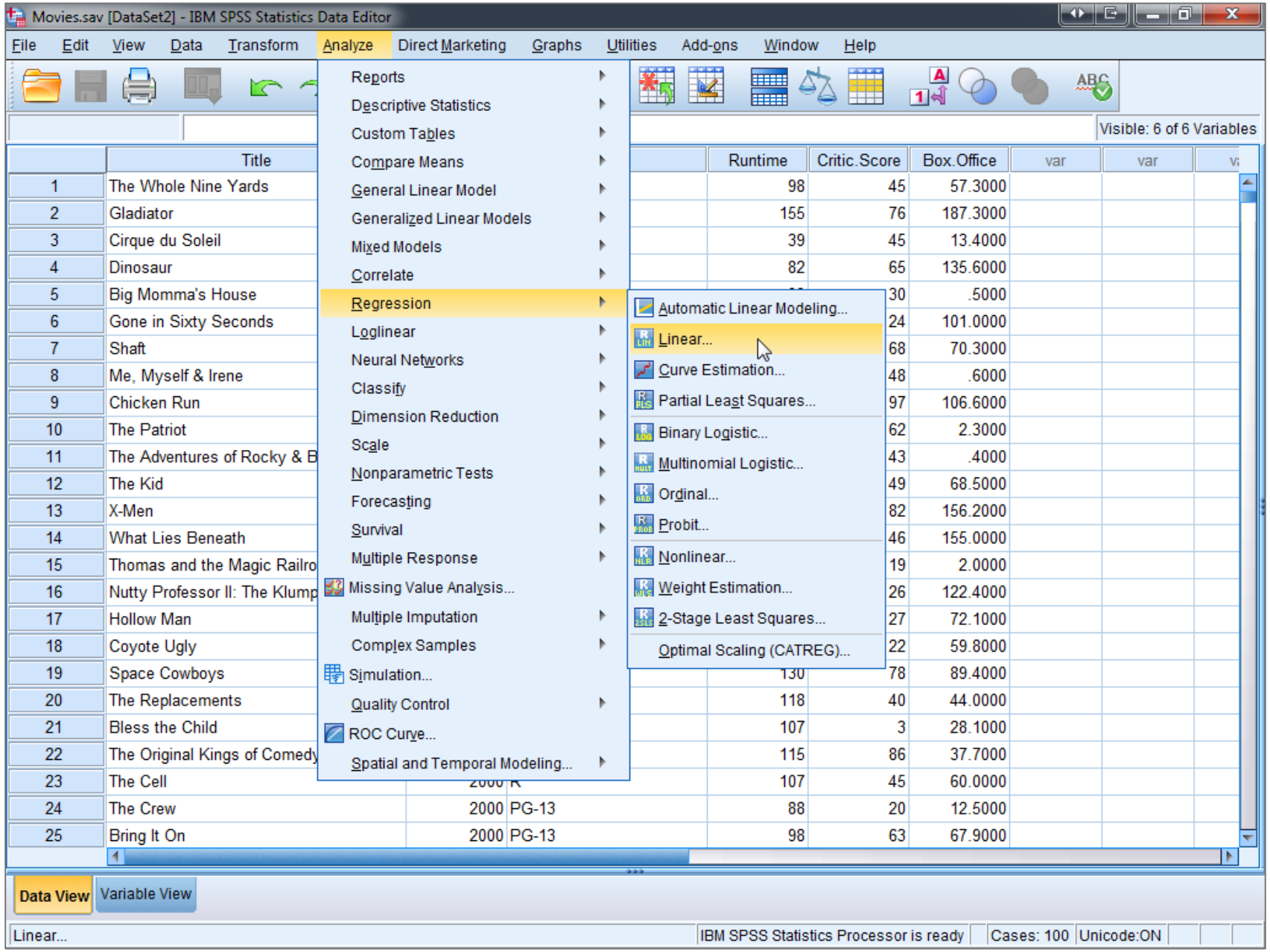
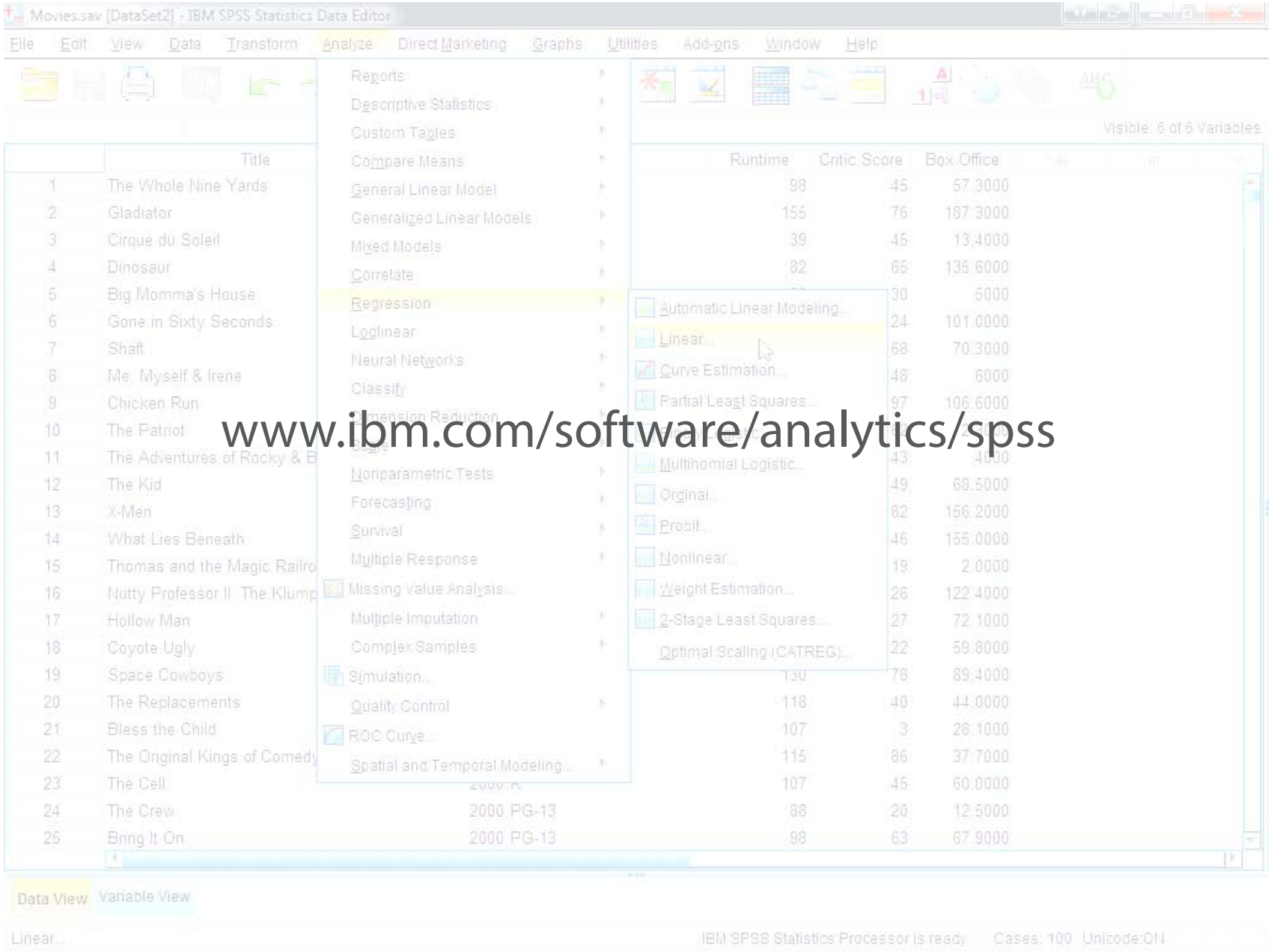


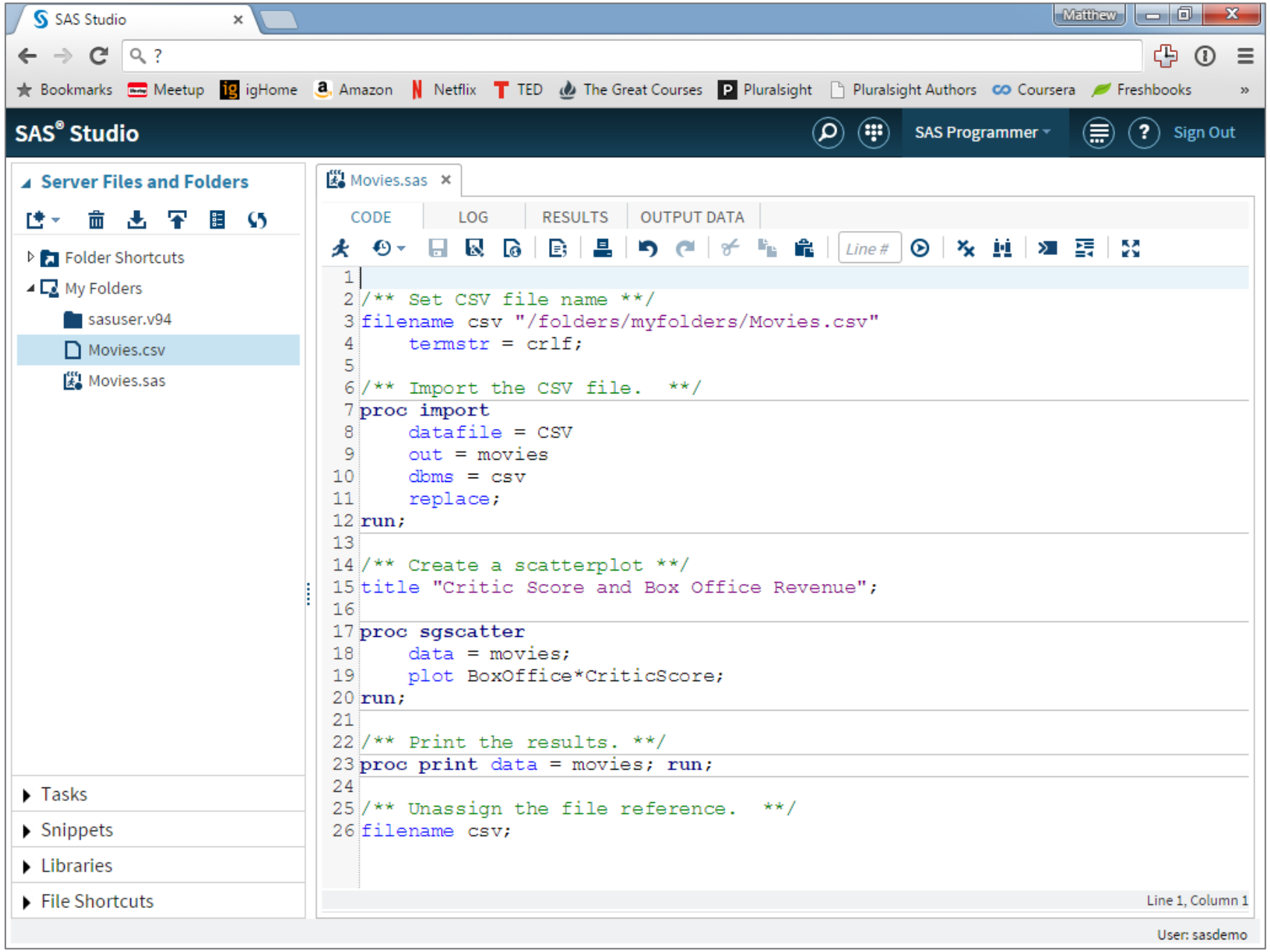
Tableau - Movies
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Normal
Show Me
Data
Analytics
Movies
Dimensions
Abc Rating
Abc Title
Year
Abc Measure Names
Measures
Box.Office
Critic.Score
Runtime
Number of Records
Measure Values
Pages
Filters
Year
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
Rating
Runtime
ATTR(Title)
Runtime
40
60
80
100
120
Rating
G
PG
PG-13
R
Columns
Critic.Score
Rows
Box.Office
Box Office Revenue (Millions of Dollars)
Critic Score (%)
Title: Avatar
Rating: PG-13
Runtime: 162
Critic.Score: 83
Box.Office: 760.5
Data Source
Sheet 1
693 marks
1 row by 1 column

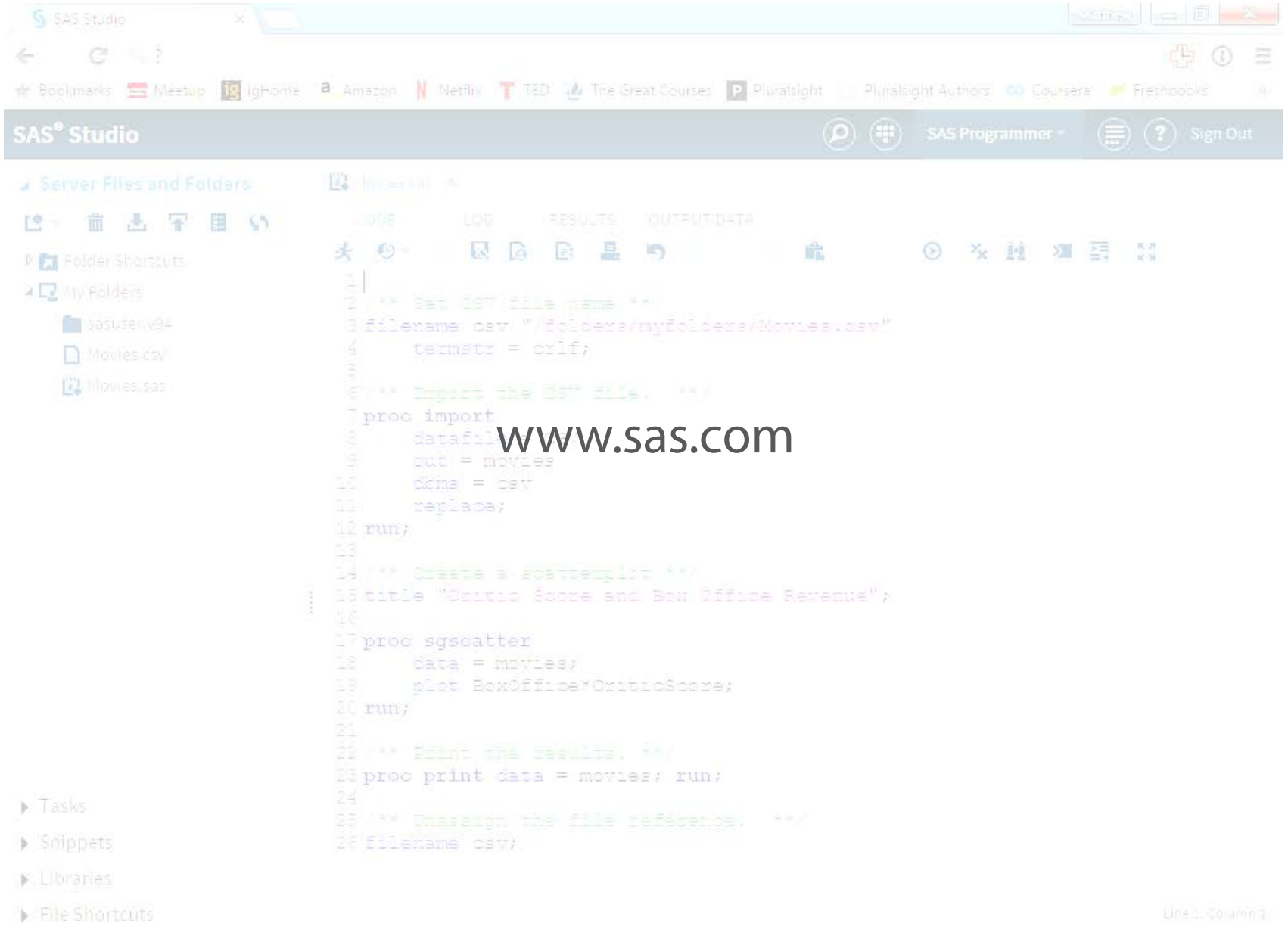

Tableau - Movies
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Normal
Show Me
Data
Analytics
Movies
Dimensions
Rating
Title
Year
Measure Names
Pages
Columns
Critic.Score
Rows
Box.Office
Filters
Year
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
Rating
Runtime
ATTR(Title)
Measures
Box.Office
Critic.Score
Runtime
Number of Records
Measure Values
Runtime
Rating
G
PG
PG-13
R
Avatar
PG-13
162
83
760.5
www.tableau.com
Data Source
Sheet 1
693 marks
1 row by 1 column


Movies.sav [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Spatial and Temporal Modeling...
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...
Visible: 6 of 6 Variables
Title
Runtime
Critic.Score
Box.Office
var
var
1 The Whole Nine Yards 98 45 57.3000
2 Gladiator 155 76 187.3000
3 Cirque du Soleil 39 45 13.4000
4 Dinosaur 82 65 135.6000
5 Big Momma's House 30 .5000
6 Gone in Sixty Seconds 24 101.0000
7 Shaft 68 70.3000
8 Me, Myself & Irene 48 .6000
9 Chicken Run 97 106.6000
10 The Patriot 62 2.3000
11 The Adventures of Rocky & B 43 .4000
12 The Kid 49 68.5000
13 X-Men 82 156.2000
14 What Lies Beneath 46 155.0000
15 Thomas and the Magic Railro 19 2.0000
16 Nutty Professor II: The Klump 26 122.4000
17 Hollow Man 27 72.1000
18 Coyote Ugly 22 59.8000
19 Space Cowboys 130 78 89.4000
20 The Replacements 118 40 44.0000
21 Bless the Child 107 3 28.1000
22 The Original Kings of Comedy 115 86 37.7000
23 The Cell 107 45 60.0000
24 The Crew 2000 PG-13 88 20 12.5000
25 Bring It On 2000 PG-13 98 63 67.9000
Data View
Variable View
Linear...
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON



www.ibm.com/software/analytics/spss


SAS Studio
Matthew
Bookmarks
Meetup
igHome
Amazon
Netflix
TED
The Great Courses
Pluralsight
Pluralsight Authors
Coursera
Freshbooks
SAS® Studio
SAS Programmer
Sign Out
Server Files and Folders
Folder Shortcuts
My Folders
sasuser.v94
Movies.csv
Movies.sas
Tasks
Snippets
Libraries
File Shortcuts
Movies.sas
CODE
LOG
RESULTS
OUTPUT DATA
Line #
1
2 /** Set CSV file name **/
3 filename csv "/folders/myfolders/Movies.csv"
4 termstr = crlf;
5
6 /** Import the CSV file. **/
7 proc import
8 datafile = CSV
9 out = movies
10 dbms = csv
11 replace;
12 run;
13
14 /** Create a scatterplot **/
15 title "Critic Score and Box Office Revenue";
16
17 proc sgscatter
18 data = movies;
19 plot BoxOffice*CriticScore;
20 run;
21
22 /** Print the results. **/
23 proc print data = movies; run;
24
25 /** Unassign the file reference. **/
26 filename csv;
Line 1, Column 1
User: sasdemo


SAS Studio
SAS® Studio
SAS Programmer
Sign Out
Server Files and Folders
Folder Shortcuts
My Folders
sasuser.v94
Movies.csv
Movies.sas
www.sas.com
Tasks
Snippets
Libraries
File Shortcuts

Sentiment Analysis.py - C:\Pluralsight\Sentiment Analysis.py

File Edit Format Run Options Windows

```
import sys
import json
import re

def getSentiments(sentiment_file):
    scores = {}
    for line in sentiment_file:
        term, score = line.split("\t")
        scores[term] = int(score)
    return scores

def getTweets(tweet_file):
    tweets = []
    for line in tweet_file:
        tweet = json.loads(line)
        text = tweet.get("text")
        tweets.append(text)
    return tweets

def getAllTerms(tweets, sentiments):
    allTermScores = {}
    for tweet in tweets:
        tweetTerms = getTweetTerms(tweet, sentiments)
        for tweetTerm in tweetTerms:
            if sentiments.has_key(tweetTerm):
                continue
            if not allTermScores.has_key(tweetTerm):
                allTermScores[tweetTerm] = []
            termScores = allTermScores[tweetTerm]
            termScores.append(tweetTerms[tweetTerm])
    return allTermScores
```

Summary

Beyond Exploratory Data Analysis

Demo

Alternatives to R

Conclusion

Where to Go Next

R website: http://www.cran.r-project.org

R Studio: http://www.rstudio.com

Pluralsight: http://www.pluralsight.com

Coursera: https://www.coursera.org

Flowing Data: http://flowingdata.com

Revolutions: http://blog.revolutionanalytics.com

R-Blogger: http://www.r-bloggers.com

Feedback

Rate the course

Ask questions

Leave comments

Course Summary

Introduction to R

Transforming and Cleaning Data

Calculating Descriptive Statistics

Visualizing Data

Moving Beyond R and EDA

Special Thanks

Nathan Yau

Jeff Leek

Roger Peng

Karen Bittner

Anne Herlache

Cory House

Thank You!