Journal of
Information
Systems
Education

Volume 30
Issue 2
Spring 2019

Teaching Tip
Pedagogy for Business Analytics Courses

Anand Jeyaraj

Recommended Citation: Jeyaraj, A. (2019). Teaching Tip: Pedagogy for Business Analytics Courses. *Journal of Information Systems Education,* 30(2), 67-83.

Article Link: http://jise.org/Volume30/n2/JISEv30n2p67.html

Initial Submission: 20 July 2018
Accepted: 19 December 2018
Abstract Posted Online: 13 March 2019
Published: 5 June 2019

ISSN: 2574-3872 (Online) 1055-3096 (Print)

Teaching Tip
Pedagogy for Business Analytics Courses

Anand Jeyaraj
Information Systems and Supply Chain Management
Wright State University
Dayton, OH 45435, USA
anand.jeyaraj@wright.edu

ABSTRACT

Responding to the industry need for professionals to employ data-driven decision-making, educational institutions offer courses in business analytics (BA). Since BA professionals require a unique set of skills different from those found in specific business disciplines, a pedagogical framework to impart such knowledge and skills was developed. The framework encompasses multiple stages related to data – acquisition, preparation, analysis, visualization, and interpretation – and provides an end-to-end learning experience for students. It enables students to gain related knowledge and skills including Python scripting, data cleansing, statistical modeling, visualization, and interpretation, which provide a solid foundation for professional endeavors in BA.

Keywords: Business analytics, Pedagogy, Data acquisition, Data visualization

1. INTRODUCTION

As data-driven decision-making has gained prominence within organizations, the need to develop and equip business professionals with skills in business analytics (BA) has gained significant momentum (Provost and Fawcett, 2013). Educational institutions have risen to the challenge by developing new courses for undergraduate and graduate programs, concentrations or tracks, and certificates in BA.

BA courses that address the need to impart skills in acquiring, analyzing, and visualizing data are crucial for the success of business professionals (Waller and Fawcett, 2013; Asamoah, Doran, and Schiller, 2015). Following acquisition, it may be necessary to prepare the data for analysis and visualization and, ultimately, for interpretation. The combination of these skills, which includes analytical skills, information technology knowledge, and business domain knowledge (Chiang, Goes, and Stohr, 2012), is typically not within the scope of domain-specific undergraduate business degree programs (e.g., supply chain management, marketing). This is due to their idiosyncratic foci (i.e., supply chain management may focus on analyzing the efficiency of supply chains using existing data collections rather than acquiring new data from other sources) and their goals to build and graduate professionals in specific disciplines, which also indicates that business users may not be knowledgeable on BA skills (Kohavi, Rothleder, and Simoudis, 2002). Such skills, which are reasonably general and not constrained to specific domains, can apply to different disciplines in which data acquisition, analysis, and visualization are necessary.

After differentiating BA students from students in specific business disciplines (Wixom et al., 2011), a pedagogical framework was developed and implemented in an introductory course on BA within an undergraduate business degree program. The framework enables the students to gain skills in data acquisition, preparation, analysis, visualization, and interpretation. Such an approach seems to be consistent with industry trends and job prospects[1] in the BA domain and has the potential to offer significant benefits and marketability for students trained in the unique combination of BA skills. Students expressed considerable satisfaction with the learning process based on the framework and confidence in their ability to apply the skills across different business contexts.

The remainder of the paper is organized as follows: The next two sections introduce the pedagogical framework and its application in an undergraduate business degree course. The next section provides a discussion, reflection, and implications followed by a conclusion section.

2. PEDAGOGICAL FRAMEWORK

Taking the perspective that business professionals are more marketable and become more relevant for organizations when they possess skills within the broad spectrum of BA rather than specific domains, a pedagogical framework (shown in Figure 1) that can enable students to gain and enhance skills in acquiring, analyzing, and visualizing data was developed for the BA course. The framework takes into account the state of existing courses that focus on specialized skills, the end-to-end sequence of activities that may be needed to handle data, and the similarities in job requirements for BA professionals across different business fields.

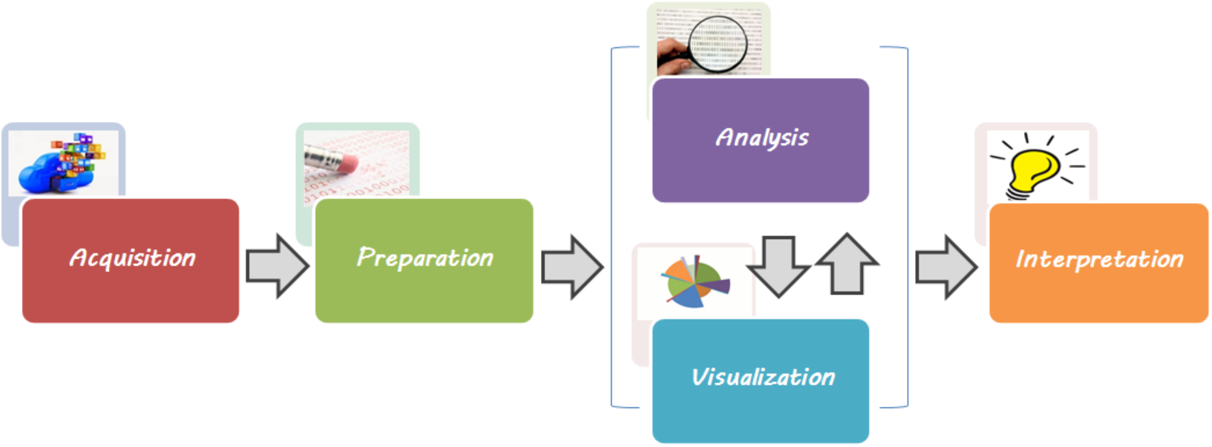

Figure 1. Pedagogical Framework

The framework synthesizes skills from disparate settings into a unified set of stages unlike other approaches – for instance, a statistical methods course may begin with an available data set and then deal with preparation and analysis; a database management course may begin with data modeling and then deal with populating and querying the database; and a marketing course may aim to acquire data from consumers. Due to the end-to-end focus of the framework, it has the potential to offer value to the aspiring BA professionals that may not be otherwise possible. The framework is relevant for BA professionals across different business fields.

2.1 Acquisition

The acquisition stage represents activities in obtaining the data necessary for the analysis from existing internal and external sources. The internal data sources include departmental systems (Accounting, Marketing, Finance), enterprise systems (Human Resources, Enterprise Resource Planning), and boundary-spanning systems (Customer Relationship Management, Supplier Relationship Management). The external data sources include third-party archives (e.g., Compustat), websites (e.g., Yahoo! Finance), and industry reports (e.g., Standish Group). The data in these repositories may be structured, semi-structured, or unstructured and available in different formats (e.g., XLSX, TXT, XML, HTML, PDF) (Chen, Chiang, and Storey, 2012).

2.2 Preparation

The preparation stage comprises activities that enable the data to be made ready for analysis. These activities include, but are not restricted to, cleansing, formatting, and organizing the data such that the data are consistent, accurate, and sufficient. Cleansing may involve the recoding of data values when data are combined from multiple sources, the substitution of data values for missing data cells, and the recasting of data into categorical or ordinal types. It may be necessary to employ Extract-Transform-Load (ETL) processes to mold the data from multiple sources (Vassiliadis, 2009). The data may also be organized into required formats (e.g., XLSX, SAV, DTA) and configurations (e.g., wide-form vs. long-form).

2.3 Analysis

The analysis stage involves activities to examine the data gathered and prepared in response to the problem to be addressed. A basic level of analysis involves descriptive statistics such as frequencies, central tendency, and cross tabs. Advanced statistical methods such as ANOVA, correlation, decision trees, cluster analysis, and discriminant analysis may be applied to the data to uncover patterns and associations. Data mining techniques for classification, clustering, and predication may also be relevant for analysis (Han, Kamber, and Pei, 2012). Techniques such as linear regression and logistic regression may be appropriate when prediction becomes necessary for the given problem.

2.4 Visualization

The visualization stage represents activities that enable effective visual presentation of the data and relationships such that the decision makers can better understand the data with minimal cognitive effort (Zheng, 2017). Such visuals are frequently created using software tools such as Tableau or dedicated websites like Plotly. Visualizations may be static (tables or charts), interactive (drill-down), or dynamic (animation over time) (Wang, Wang, and Alexander, 2015). They may be used to explore and explain the data and can present descriptive statistics such as frequencies and means and inferential statistics such as regression lines or trend lines.

2.5 Interpretation

The interpretation stage includes activities to make sense of the visuals and interpret the results obtained from the analyses. Depending on the statistical methods employed, the interpretations may lead to insights that can be helpful in describing, diagnosing, predicting, or prescribing (Porter and Heppelmann, 2015), which enable organizations to understand and plan their activities.

2.6 The Framework

The framework typically assumes a forward progression from the acquisition stage through the interpretation stage. However, iteration between the stages is possible and may be necessary depending on the specific situations. For instance: a) the BA user may discover during the analysis phase that additional data acquisition is needed for a more insightful analysis; b) the BA user may determine during the visualization phase that an additional ETL process is needed to develop a more effective visual; or c) the BA user may discover during the interpretation stage that additional visuals may be needed as evidence to support the findings. Similarly, the BA user may iterate between the analysis and visualization stages. For instance, the BA user may have developed an exploratory visualization and desire to conduct a statistical analysis, or the BA user may have found a significant difference using a statistical test and desire to show it using a visualization.

3. APPLICATION OF THE PEDAGOGICAL FRAMEWORK

The pedagogical framework was implemented at a large university in the midwestern United States. The setting was an introductory course on business analytics for undergraduate students in a business school. Students entering the course have basic knowledge of IS through a prerequisite introductory survey course on IS for business students, which introduces the various types of IS and a working knowledge of spreadsheets and databases (but not any programming). Some students may have taken a business statistics course that introduced descriptive statistics and basic statistical techniques such as correlation and linear regression. Stated differently, students may have gained insights into certain aspects of the framework in isolation but have not seen an integrated view of the various capabilities enabling BA.

The application of the pedagogical framework had two parts: a) the in-class instruction based on the framework over a semester and b) the independent work completion by students consistent with the in-class instruction and the framework. Based on the availability of 14 weeks for the course, the framework was applied using a timeline of 6 weeks for the acquisition and preparation stages, 6 weeks for the analysis and visualization stages, and 2 weeks for the interpretation stage.

3.1 In-Class Instruction

Students taking the course were introduced to various software tools including Microsoft Excel, Python, and SPSS, consistent with the stages of the pedagogical framework. Students learned various techniques in Excel and Python during the acquisition and preparation stages and in SPSS, Excel, and Python during the analysis and visualization stages. The interpretation stage largely dealt with the ways in which students can extract meaning from the results and visualizations. At appropriate times, in-class discussions also focused on heuristics to identify the best possible tools, analysis techniques, and visualizations.

Using Microsoft Excel, students learned advanced techniques related to text manipulation (e.g., LEFT, RIGHT, MID, FIND), data sourcing (e.g., CSV, TXT, XML), data formatting (e.g., type conversion, conditional formats), and data substitution (e.g., VLOOKUP). Using Python and its many libraries, students learned the basics of designing scripts to: visit different websites and fetch desired web pages (e.g., Requests), extract the desired content from the HTML pages (e.g., BeautifulSoup, JSON, ElementTree), create data files (e.g., TXT, CSV), conduct analysis of numeric and textual data (e.g., Pandas, NLTK), and create visualizations (e.g., MatPlot, WordCloud). Using SPSS, students learned various techniques for data analysis including correlation, regression (linear, logistic, and ordinal), decision trees, cluster analysis, and discriminant analysis. Students were also introduced to different software tools that can enable the creation of visualizations. These include Microsoft Excel (e.g., Charts, PivotCharts), Tableau (e.g., Bubble charts, Tree maps), and Python MatPlot.

3.2 Independent Work Completion

Students were placed on small teams and tasked with acquiring, preparing, analyzing, visualizing, and interpreting data. The broad problem chosen for these activities pertained to the success of movies. Specifically, the student teams were required to take a data-centric approach to explain and predict the success of animated movies by Disney. To get started, a list of 60 animated movies was given to students in an Excel workbook. The list included movies by three studios: Walt Disney Animation (beginning with *Beauty and the Beast* in 1991), Pixar Animation (beginning with *Toy Story* in 1995), and Disney Toons (beginning with *A Goofy Movie* in 1995). See Appendix A for additional details about the movies.

The initial briefing session included a discussion on “movie success” – students suggested box office earnings and Academy awards received as potential indicators of success. The initial briefing session also included a discussion on the potential measures that may be used to explain movie success. Several measures, such as the genre, rating, run time, reviews received from movie critics, votes received from moviegoers, number of weeks as a top-10 movie, and popularity of directors and actors, were considered.

During the acquisition stage, students had the flexibility to gather as much relevant data as possible using any or a combination of manual methods (copy and paste from external websites), semi-automated methods (table lookup using Excel), and automated methods (Python scripts to parse HTML pages and save data into CSV files). The student teams employed different methods in acquiring the data. One team found a third-party website that contained data on several variables for all Disney movies in table format. Those students copied and pasted the entire data table into an Excel worksheet and then used the VLOOKUP function to gather the data for the given list of animated movies. Another team found that the IMDB (Internet Movie Data Base) website could provide useful data and also allowed users to create their own custom lists of movies for tracking purposes. Those students created an IMDB list with the given list of animated movies and then developed Python scripts to fetch and parse the HTML pages and create a TXT file with the data provided by IMDB. Appendix B shows the Python script developed for this purpose. Another team used movie reviews on the variety.com website and gathered the hyperlinks for the reviews of the animated movies. Those students developed Python scripts to fetch and parse the HTML pages and extract the unstructured textual data into TXT files. Appendix C shows the Python script developed for this purpose.

The data gathered (which were submitted as Excel workbooks by the student teams) were consolidated into a master dataset by the instructor. The resulting dataset contained various data attributes: movie title, studio, MPAA rating, genre, summary, movie review, directors, actors, budget, and gross earnings. Some attributes such as directors and actors shared a one-to-many relationship with the movie such that they were organized in long format while the remaining attributes such as gross earnings and movie review shared a one-to-one relationship with the movie and were organized in wide format. The consolidated dataset, essentially "crowd sourced," was then shared with all students such that they can proceed with the other stages of the framework. Thus, all students benefited from the acquisition stage by having access to the large consolidated dataset that may not have been available otherwise.

During the preparation stage, students were required to examine the consolidated dataset and implement any transformations on the data as may be needed for the analysis and the overall goal of explaining and predicting movie success. Several transformations were completed by the student teams on the dataset:

i) recode the "genre" attribute (which included multiple values such as action, drama, and family for a single movie) into multiple categorical variables (e.g., Drama = 0 or Drama = 1 and Family = 0 or Family = 1);
ii) recode the "gross earnings" attribute to have only numeric values so that the value of $26.2M representing 26 million dollars was transformed to 26.2 or 26200000 depending on the perspective taken by the student team;
iii) a new "adjusted gross earnings" attribute was created to account for inflation;
iv) the "year of Oscar" attribute was recoded into a single categorical variable (Oscar award = 0 or Oscar award = 1) representing whether or not the movie received an Oscar award;
v) a new attribute was created to differentiate the MPAA rating of "G" ("general rating" = 1) with other ratings such as "PG" and "PG 13" ("general rating" = 0);
vi) the "directors" attribute which listed the names of directors was transformed into a "directors count" attribute containing the number of directors for the movie and a categorical variable representing whether or not the movie had a single director or multiple directors; and
vii) the "review" attribute which contained the movie review obtained from variety.com was transformed into a numeric score based on sentiment analysis, for which students developed a Python script that employed the APIs provided by a third-party web service. Appendix D shows the Python script for this purpose.

During the analysis stage, students had the option to use any statistical method to uncover the relationships based on the research questions of interest. Student teams applied different methods such as logistic regression, cluster analysis, and linear regression on the consolidated dataset after preparation (Han, Kamber, and Pei, 2012). The results of a logistic regression, a cluster analysis, and an ordinary least squares (OLS) regression done by students are shown in Appendix E, Appendix F, and Appendix G, respectively. Since the consolidated dataset had a small sample size, the findings should be used with caution. During the visualization stage, students had the flexibility to develop the necessary, relevant, useful, and usable visuals using any combination of software packages. Appendix H documents a few visuals developed by student teams.

During the interpretation stage, students used the results of the statistical tests as well as the visuals to make sense of the data, including patterns and relationships. These activities enable the presentation of findings to the decision-makers so that informed decisions can be made for the organization. The Appendices show illustrations of such interpretations. For instance, the logistic regression (Appendix E) shows that "G" rated movies in the Comedy genre produced by Pixar Studio that run for a longer time have the potential to earn an Oscar award, and the OLS regression (Appendix G) shows that movies by Pixar Studio generate greater gross earnings. Such results provide insights for organizations (Disney in this case) for their next endeavors.

4. REFLECTION

Through the application of the pedagogical framework introduced earlier, students aspiring to be trained as BA professionals learned the intricacies of data acquisition, preparation, analysis, visualization, and interpretation that can be applied in different settings. Students learned a variety of skills related to scripting (in Python), data cleansing and ETL techniques, statistical modeling and data mining methods (in SPSS), and visualization, the combination of which could be gainfully applied in several disciplines.

Students liked the practical, hands-on instruction in class, the real-world application of tools and techniques, and the potential for data-driven decision-making that they experienced in the course. Comments such as "I appreciate the hands on nature of work and activities," "course used real-world examples," and "materials seem to be immediately useful to a business workplace setting" were given by students. Since the course was fairly new and open to business students from different disciplines, it was helpful for the instructor to be "very open to class input as to what should be covered" and offer a course that "had structure but remained flexible based on student skills" while "communicating why and how topic areas are useful and the purpose of a particular piece of knowledge." While not explicitly stated by students, it was also interesting to see them engage in out-of-the-box thinking to solve problems, at times employing techniques more advanced than those introduced during in-class instruction.

Students were appreciative of the unique nature of the course as they shared during the debriefing session. One student stated:

> I had previously taken a course in which we were given a data set for analysis. We had to clean data and handle missing data and apply regression methods. We never had to deal with data acquisition or visualization but I see how valuable those may be for business decision-making.

Another student mentioned:

> I had taken a course in which we learned how we can gather data using primary data collection methods such as surveys but the use of Python scripts to gather data from web sites was fascinating. I can see how it can be immensely helpful for projects in different business domains.

Such comments by students offer indirect evidence that the pedagogical framework offered a valuable learning experience as they holistically dealt with the various stages in BA.

The timeline adopted for the framework seemed to work reasonably well for student learning. Students learned the essentials of Python in about six weeks, which helped with the data acquisition and preparation stages. Over the next six weeks, students learned the different analysis and visualization techniques using SPSS and Excel. While this timeline offers a reasonable start, it is possible that it can be tweaked further depending on the focus of the course planned (is the course data acquisition or data analysis intensive?), the skills of the students (are students already trained in statistical methods?, are students interested in specific disciplines such as marketing or finance?), and the availability of time (is the course offered over a typical 14-week semester or a compressed 6-week schedule?). The framework can thus be customized to be heavy on the acquisition and preparation stages or analysis and visualization stages depending on certain contingencies.

While the framework was useful in student learning, there may be opportunities for tweaking the future offerings of the course. First, the context for the project could be different from animated movies by Disney. Several industries such as finance, hospitality, and insurance serve as potential candidates. Second, regardless of the context chosen, the sample size could be much larger. Data on stock market performance may easily afford hundreds of ticker symbols that may be used to acquire data (within the limits imposed by the websites for robots). Third, it may be possible to incorporate natural language processing to a greater extent. Packages such as NLTK in Python or NLP in R would engender additional skills for students in dealing with unstructured data. Fourth, the repertoire of statistical modeling and data mining methods may be adapted or extended for a more in-depth or expanded learning. Students may benefit from applying multiple methods such as cluster analysis and discriminant analysis in tandem to develop finer insights on the problem. Finally, alternate or additional visualization techniques may be used to enhance the learning experience for students.

5. CONCLUSION

A pedagogical framework was developed considering BA professionals as a distinct group of students who require distinctive skills. The framework enables students to experience end-to-end learning on the related activities of data acquisition, preparation, analysis, visualization, and interpretation for business decision-making. Using the framework, students gained a unique combination of knowledge and skills on various techniques (Python coding, ETL, logistic regression) and software tools (Python, Excel, SPSS), and they expressed satisfaction with the learning process. These students are expected to be highly competitive and marketable in the emerging domain of BA that draws from a variety of disciplines (e.g., statistics, information systems, engineering) and applicable to various fields (e.g., marketing, actuarial science, insurance).

6. ENDNOTES

[1] According to jobs posted on websites such as monster.com, glassdoor.com, and careerbuilder.com, job titles such as business data analyst, actuarial analyst, data mining analyst, healthcare data analyst, and marketing analyst are common. Such positions typically call for completed degrees in business administration, information systems, statistics, economics, finance, applied math, operations research, industrial engineering, data analytics, or a related field and experience in data visualization, predictive modeling, data cleansing including ETL, and programming (in R, SAS, or Python).

7. REFERENCES

Asamoah, D. A., Doran, D., & Schiller, S. (2015). Teaching the Foundations of Data Science: An Interdisciplinary Approach. *Proceedings of the Pre-ICIS SIGDSA Workshop*, Fort Worth, Texas.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), 1165-1188.

Chiang, R. H. L., Goes, P., & Stohr, E. A. (2012). Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline. *ACM Transactions on Management Information Systems*, 3(3), 1-13.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. Waltham, MA: Morgan Kauffman (Elsevier).

Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging Trends in Business Analytics. *Communications of the ACM*, 45(8), 45-48.

Porter, M. E. & Heppelmann, J. E. (2015). How Smart, Connected Products are Transforming Companies. *Harvard Business Review*, 93(10), 96-114.

Provost, F. & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-driven Decision Making. *Big Data*, 1(1), 51-59.

Vassiliadis, P. (2009). A Survey of Extract-Transform-Load Technology. *International Journal of Data Warehousing & Mining*, 5(3), 1-27.

Waller, M. A. & Fawcett, E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution that will Transform Supply Chain Design and Management. *Journal of Business Logistics*, 34(2), 77-84.

Wang, L., Wang, G., & Alexander, C. A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, 1(1), 33-38.

Wixom, B., Ariyachandra, T., Goul, M., Gray, P., Kulkarni, U., & Phillips-Wren, G. (2011). The Current State of Business Intelligence in Academia. *Communications of the Association for Information Systems*, 29(16), 299-312.

Zheng, J. G. (2017). Data Visualization for Business Intelligence. In *Global Business Intelligence*, Chapter 6, J. M. Munoz (ed.), New York, NY: Routledge, 67-82.

AUTHOR BIOGRAPHY

Anand Jeyaraj is a professor of information systems at the Raj Soin College of Business at Wright State University. He has been an instructor for over 20 years and taught a variety of courses related to information systems including programming, geographic information systems, supply chain information management, and business analytics. He has employed learner-centered teaching and active learning techniques in his classes and received multiple teaching awards.

APPENDIX A
Description of the Movies in the Dataset

https://en.wikipedia.org/wiki/List_of_Disney_theatrical_animated_features shows the titles of the movies (last accessed April 5, 2018). Of the 60 movies included in the dataset, Walt Disney Animation had 27 movies, Pixar Animation had 19 movies, and Disney Toons had 14 movies. There were 12 movies which had received Oscar awards. At least one movie was released every year between 1991 and 2017 except 1993 with four movies each in 2000 and 2003 (see Figure A1). Table A1 contains the descriptive statistics of the movies.

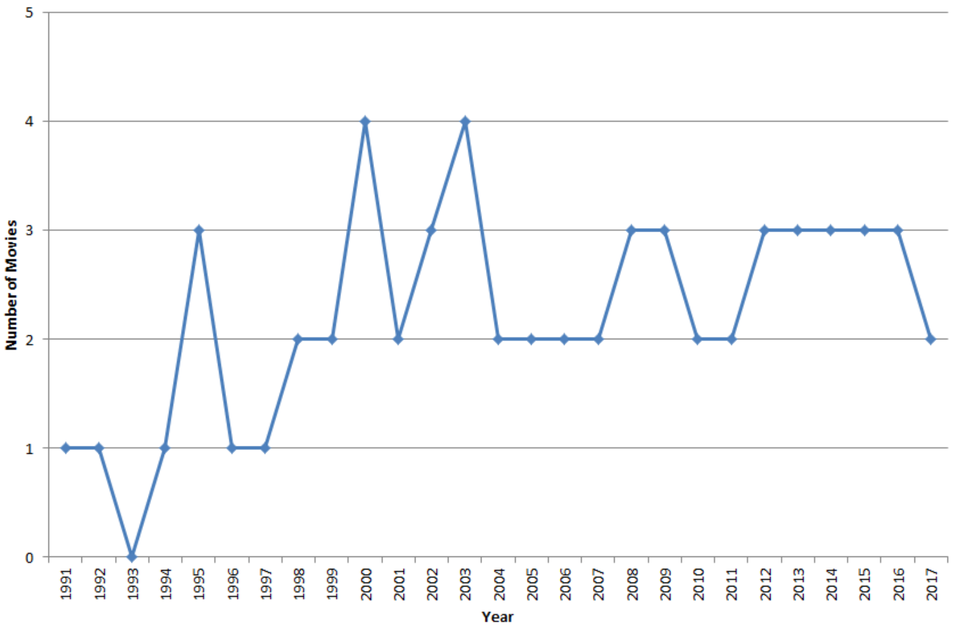

Figure A1. Number of Movies by Year of Release

Variable	Mean	SD
Oscar Award	0.20	0.40
Adjusted Gross Earnings ($)	132523363.30	81078872.02
Pixar Studio	0.31	0.46
G Rating	0.51	0.50
Movie Run Time (minutes)	90	12.23
Comedy Genre	0.67	0.47
Family Genre	0.30	0.46
Action Genre	0.05	0.22
Positive Sentiment Rating	0.73	0.13
Number of Fan Votes	228093.88	241167.08

Table A1. Descriptive Statistics

APPENDIX B
Python Script for Fetching and Parsing IMDB Web Page and Extracting Data

https://www.the-numbers.com/movies/distributor/Walt-Disney provided data such as release data, genre, MPAA rating, and total gross amount for all movies by Walt Disney.
https://www.imdb.com/ provided data such as MPAA rating, user star rating, genre, summary, meta score, genre, run time, votes received, and total gross. https://www.imdb.com/list/ls026028045/ contains the custom IMDB list created by the student team (all last accessed April 5, 2018).

```
import requests
from bs4 import BeautifulSoup

url = 'https://www.imdb.com/list/ls026028045'
page = requests.get(url)
soup = BeautifulSoup(page, 'html.parser')

#get title
title = soup.findAll('a',href=lambda x: x and x.startswith('/title/'))
title = list(map(lambda x: x.text, title))

#get ratings
rating = soup.findAll('span',attrs={"class":"certificate"})
rating = list(map(lambda x: x.text, rating))

#get runtime
runtime = soup.findAll('span',attrs={"class":"runtime"})
runtime = list(map(lambda x: x.text.split(' ')[0], runtime))

#get genre
genre = soup.findAll('span',attrs={"class":"genre"})
genre = list(map(lambda x: x.text, genre))

#get vote
vote = soup.findAll('span',attrs={"name":"nv"})
vote = list(map(lambda x: x.text, vote))

data = list(zip(title, rating, runtime, genre, vote))
with open('IMDB.txt','w') as outfile:
   for d in data:
      record = '\t'.join(d)+'\n'
      outfile.write(record)
```

APPENDIX C
Python Script for Fetching and Parsing variety.com Reviews and Extracting Text

https://variety.com/v/film/reviews/ shows the film review site.
Examples of reviews obtained: http://variety.com/2017/film/reviews/coco-review-pixar-1202595605/ and http://variety.com/1990/film/reviews/beauty-and-the-beast-3-1200428997/ (all last accessed June 20, 2018).

```
import requests
from bs4 import BeautifulSoup
import time
import glob

# Read URLs from file
data = []
with open('urls.txt','r') as infile:
    data = infile.read().splitlines()

# Get Pages for URLs
for url in urls:
    page = requests.get(url)
    soup = BeautifulSoup(page.text,'html.parser')
    # Write HTML page to file
    with open(movies[urls.index(url)]+'.html','w') as outfile:
        outfile.write(str(soup))
    time.sleep(5)

# Function to drop tags with attributes
def fix(x):
    if len(x.attrs)==0:
        return x
    else:
        return ''

# Retrieve file names
files = glob.glob('*.html')
movies = list(map(lambda x: x.split('.',1)[0], files))

# Extract text from HTML files
for f in files:
    print(f)
    soup = BeautifulSoup(open(f),'html.parser')
    review = soup.findAll('p')
    review = list(set(review)) #eliminate duplicates
    review = list(map(lambda x: fix(x), review)) #drop <p> with attr
    review = [i for i in review if
              not ('</div>' in str(i)) and
              not ('</section>' in str(i)) and
              not ('</strong>' in str(i))]
    review = list(filter(None, review)) #drop empty items
    review = list(map(lambda x: x.text, review)) #drop <p> </p> tags
    with open(movies[files.index(f)]+'.txt','w') as outfile:
        outfile.write(' '.join(review))
```

APPENDIX D
Python Script for Obtaining Sentiment Scores using Third-Party API Service

https://www.census.gov/topics/income-poverty/income/guidance/current-vs-constant-dollars.html provides the CPI values for each year and https://www.usinflationcalculator.com/frequently-asked-questions-faqs/ provides the formula for conversion. All gross earnings were converted to 2017 rates. A sentiment analysis score was provided by http://text-processing.com/api/sentiment/ indicating whether the text was positive, negative, or neutral along with a numeric score (all last accessed June 20, 2018).

```
import requests
import glob
import random
import time
import json
import csv

# Retrieve file names
files = list(set(glob.glob('*.txt')))
reviews = []
results = []

# Use third-party API to get sentiment scores
for f in files:
    with open(f,'r') as infile:
        review = infile.read()
    revdata = [('text', review),]
    response = requests.post('http://text-processing.com/api/sentiment/',
     data=revdata)
    results.append(response.text)
    time.sleep(2)

#Format results for use
fresult = []
for r in results: #flatten
    jobj = json.loads(r)
    jobj['movie'] = files[results.index(r)].split('.')[0] #add movie name
    for k in jobj['probability'].keys():
        jobj[k] = jobj['probability'][k]
    jobj.pop('probability')
    fresult.append(jobj)
print(fresult)

# Save results to file
fields = ['movie','label','neutral','pos','neg']
with open('sentresult-all.csv','w',newline='') as outfile:
    csvwriter = csv.DictWriter(outfile,fields)
    csvwriter.writeheader()
    for obj in fresult:
        csvwriter.writerow(obj)
```

APPENDIX E
Results of Logistic Regression

To determine the predictors of a movie receiving an Oscar award, students employed a logistic regression method. The dependent variable was a categorical variable (i.e., Oscar award, coded as 1 or 0 depending on whether or not the movie received an Oscar award). The attributes for the analysis included: Pixar studio (coded as 1 or 0 depending on whether or not the movie was produced by Pixar Animation), general rating (coded as 1 or 0 based on whether the movie received an MPAA rating of "G" or others such as "PG"), movie run time (in minutes), comedy genre (whether or not the movie is listed as a "comedy" movie), family genre (whether the movie is listed as a "family" movie), and action genre (whether or not the movie is listed as an "action" movie). Table E1 contains the results of the logistic regression.

Variable	**B**	**Exp(B)**
Constant	-43.303**	0.000
Pixar Studio	-4.362**	0.013
G Rating	3.585**	36.059
Movie Run Time	0.313**	1.367
Comedy Genre	7.810**	2464.029
Family Genre	5.471*	237.705
Action Genre	5.226	186.071

Dependent variable: Oscar Award (=1 if movie received an Oscar award and =0 otherwise)
N = 60, Model X^2 = 33.192***, Cox & Snell R^2 = 0.425, Nagelkarke R^2 = 0.672, 88.3% of cases correctly classified
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table E1. Logistic Regression

The logistic regression model was significant and explained 42.5% of the variance. The results showed that variables for Pixar Studio, "G" Rating, movie run time, and comedy genre were significant at $p < 0.05$ and the variable for family genre was marginally significant at $p < 0.10$. Therefore, it can be surmised that "G" rated movies in the Comedy genre produced by Pixar Studio that run for a longer time have the potential to earn an Oscar award.

APPENDIX F
Results of Cluster Analysis

To understand the similarities and differences between the animated movies, students applied a cluster analysis on a selected set of attributes given the small sample size. The attributes for the analysis included: movie run time (in minutes), number of votes received from fans, positive sentiment rating for the movie review on variety.com (range: 0 to 1), Oscar award (coded as 1 or 0 depending on whether or not the movie received an Oscar award), family genre (whether the movie is listed as a “family” movie), and comedy genre (whether or not the movie is listed as a “comedy” movie). After excluding movies with missing data, the cluster analysis was applied on 54 movies using SPSS. Based on the agglomeration schedule (pictured in Figure F1), a two-cluster solution was chosen due to the significant jump in the error value indicating a significant typology.

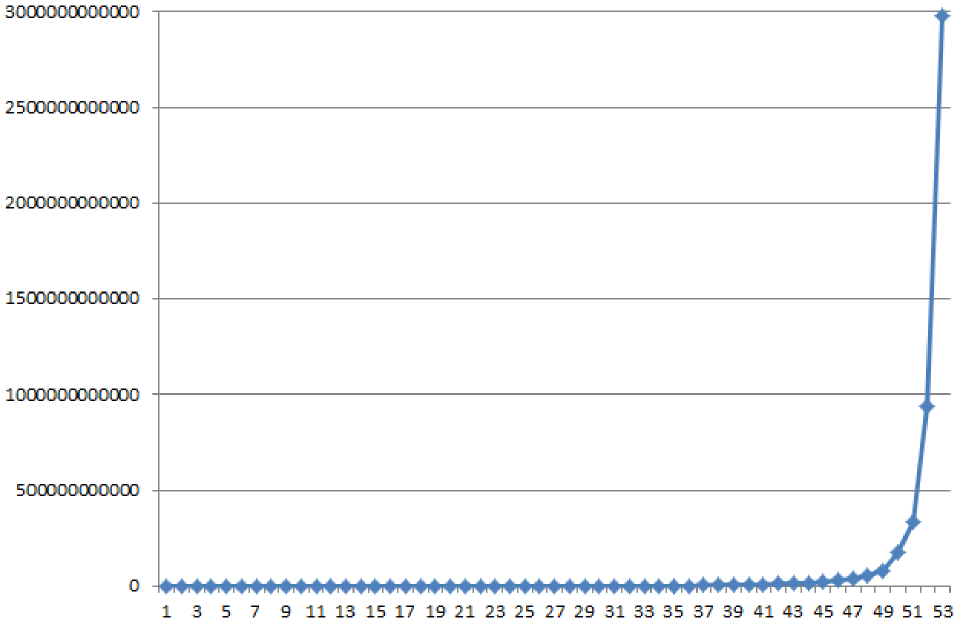

Figure F1. Agglomeration Schedule Visualization

The two-cluster solution showed 21 movies in the first cluster and 33 movies in the second cluster. The differences between the two clusters in terms of means and standard deviations as well as an ANOVA comparison of the means are captured in Table F1. Barring the two genre attributes for Comedy and Family, the other four attributes significantly differ between the clusters.

	Cluster 1	Cluster 2	F (ANOVA)
Oscar Award	0.52 (0.51)	0.03 (0.17)	26.18***
Positive Sentiment Rating	0.80 (0.13)	0.70 (0.13)	8.46***
Movie Run Time	98.81 (9.35)	86.42 (11.23)	17.71***
Number of Fan Votes	483379.67 (200489.92)	84905.94 (64775.62)	112.94***
Comedy Genre	0.76 (0.44)	0.67 (0.48)	0.54
Family Genre	0.14 (0.36)	0.33 (0.48)	2.44

N = 54, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table F1. Cluster Analysis and ANOVA

Table F2 shows the movies in the two clusters. The first cluster contains movies produced by Walt Disney Animation and Pixar Animation including 11 of the 12 movies which had received Oscar awards. The second cluster contains movies produced by all three studios with the fewest representation from Pixar Animation.

Cluster 1		Cluster 2	
A Bug's Life	Pixar	A Goofy Movie	Disney Toons
Aladdin	Walt Disney	Atlantis: The Lost Empire	Walt Disney
Beauty and the Beast	Walt Disney	Bolt	Walt Disney
Big Hero 6	Walt Disney	Brother Bear	Walt Disney
Brave	Pixar	Cars 2	Pixar
Cars	Pixar	Cars 3	Pixar
Finding Nemo	Pixar	Chicken Little	Walt Disney
Frozen	Walt Disney	Coco	Pixar
Inside Out	Pixar	Dinosaur	Walt Disney
Monsters University	Pixar	Fantasia 2000	Walt Disney
Ratatouille	Pixar	Finding Dory	Pixar
Tangled	Walt Disney	Hercules	Walt Disney
The Incredibles	Pixar	Home on the Range	Walt Disney
The Lion King	Walt Disney	Lilo & Stitch	Walt Disney
Toy Story	Pixar	Meet the Robinsons	Walt Disney
Toy Story 2	Pixar	Moana	Walt Disney
Toy Story 3	Pixar	Mulan	Walt Disney
Up	Pixar	Planes	Disney Toons
WALL-E	Pixar	Planes: Fire & Rescue	Disney Toons
Wreck-It Ralph	Walt Disney	Pocahontas	Walt Disney
Zootopia	Walt Disney	Pooh's Heffalump Movie	Disney Toons
		Secret of the Wings	Disney Toons
		Tarzan	Walt Disney
		The Emperor's New Groove	Walt Disney
		The Good Dinosaur	Pixar
		The Hunchback of Notre Dame	Walt Disney
		The Jungle Book 2	Disney Toons
		The Princess and the Frog	Walt Disney
		The Tigger Movie	Disney Toons
		Tinker Bell	Disney Toons
		Tinker Bell and the Legend of the NeverBeast	Disney Toons
		Treasure Planet	Walt Disney
		Winnie the Pooh	Walt Disney

Table F2. Movies in Clusters

When the clusters were visualized using a bivariate relationship involving adjusted gross earnings and one of the six attributes employed in cluster analysis, interesting patterns emerged from the analysis. For instance, the movies in the first cluster reported higher adjusted gross earnings with longer movie run times, higher number of fan votes, and higher positive sentiment ratings of movie reviews. Figure F2 shows the relevant charts.

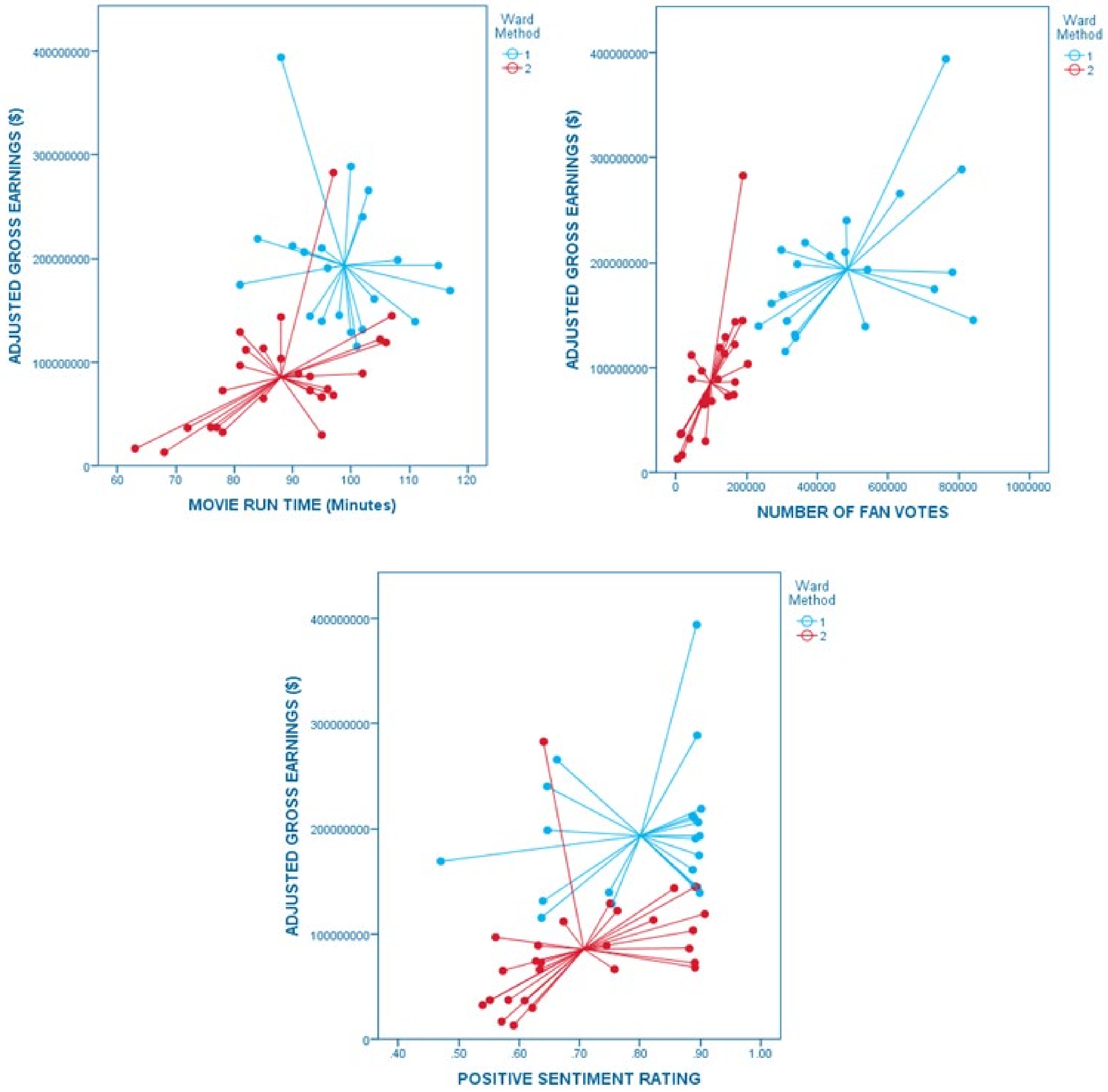

Figure F2. Cluster Charts involving Adjusted Gross Earnings

APPENDIX G
Results of OLS Regression

To determine the predictors of the adjusted gross earnings for a movie, students employed an ordinary least squares (OLS) regression method. The dependent variable was a continuous variable. The attributes for the analysis included: Pixar studio (coded as 1 or 0 based on whether or not the movie was produced by Pixar Animation), general rating (coded as 1 or 0 depending on whether the movie received an MPAA rating of "G" or others such as "PG"), movie run time (in minutes), comedy genre (whether or not the movie is listed as a "comedy" movie), and family genre (whether the movie is listed as a "family" movie). Table G1 contains the results of the OLS regression.

Variable	Beta	T
Pixar Studio	0.307	2.023**
G Rating	0.095	0.698
Movie Run Time	0.305	1.968*
Comedy Genre	-0.314	-1.911*
Family Genre	-0.268	-1.639

Dependent variable: Adjusted Gross Earnings
N = 50, $R^2 = 0.318$, Adjusted $R^2 = 0.240$, Model F (ANOVA) = 4.101***
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table G1. OLS Regression

The OLS regression model was significant and explained 31.8% of the variance. The results showed that the variable for Pixar Studio was significant at $p < 0.05$, and the variables for movie run time and comedy genre were marginally significant at $p < 0.10$. These findings demonstrate that movies by Pixar Studio have the potential to generate greater gross earnings.

APPENDIX H
Visualization Gallery

Different visuals were developed by students. Figure H1 shows a bar chart of adjusted gross earnings by movie studio (Disney Toons vs. Pixar vs. Walt Disney), MPAA rating ("G" or "PG"), and review rating (=1 if positive and =0 otherwise). Disney Toons did not have a movie with a PG rating, but in all other categories, movies that had positive review ratings also had higher adjusted gross earnings.

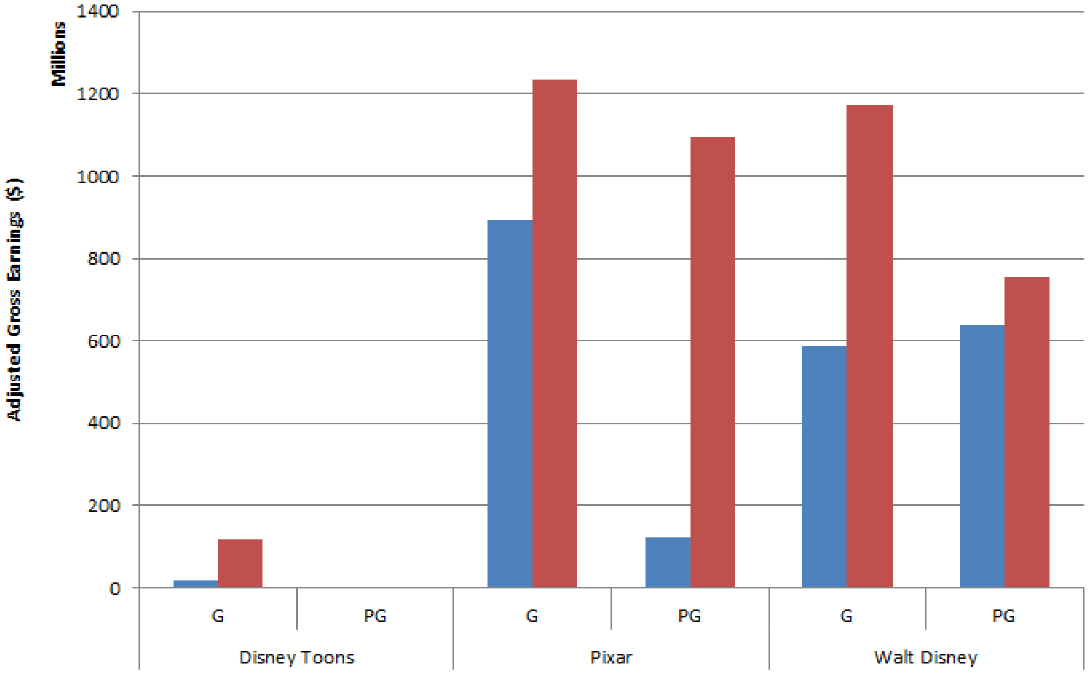

Figure H1. Adjusted Gross Earnings by Movie Studio, MPAA Rating, and Review Rating

Figure H2 shows a scatter graph of adjusted gross earnings and number of fan votes organized into two clusters: movies by Pixar Animation and others. Across all movies, fan votes seems to have a positive association with adjusted gross earnings, i.e., the greater the number of fan votes, the higher the adjusted gross earnings. Based on the two trend lines for the clusters, it seems that the association between fan votes and adjusted gross earnings is stronger for movies by other studios (Disney Toons and Walt Disney) relative to Pixar Animation. However, the scatter graph also shows that movies by Pixar generally seem to garner a higher number of fan votes relative to the movies produced by the other studios.

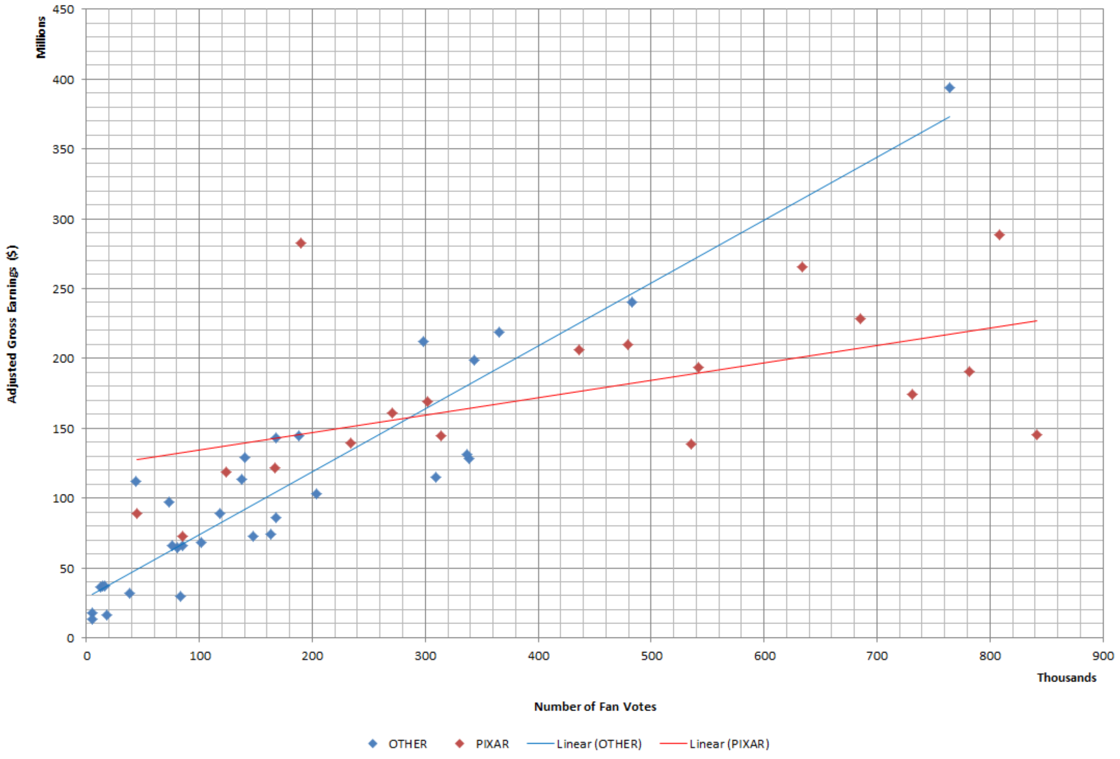

Figure H2. Adjusted Gross Earnings by Fan Votes and Movie Studio

iscap
Information Systems & Computing
Academic Professionals

EDSIG
Serving Information Systems Educators

STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the Journal of Information Systems Education have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN 2574-3872