

CS573: Homework 2 Solution

Compiled from submissions of Balamurugan Anandan, David Haye
Wahbeh Qardaji, John Ross Wallrabenstein, Jin Yu

1 Identifying Hypotheses (15 pts)

1. Identify five potentially controversial hypotheses, claims, or conjectures in news articles that can be analyzed with data. See some examples claims below. Summarize the articles and the claims.

Sample solution

- (a) *Study Reveals Why Women Apologize So Much.* The article details the findings of two studies, both conducted exclusively on college students who ranged in age from 18 to 44. The first study found that both men and women apologize about 81% of the time when they feel that have done something wrong that might warrant an apology. However, the second study found that men are less likely to deem a certain action (such as calling a friend late at night and causing them to perform poorly at an interview the next day) as warranting an apology, and therefore apologize often than women.
- (b) *Diesel vs. Gas: Fuel Economy vs. Emissions.* The article claims that while diesel fuel has higher greenhouse gas emissions, it has a higher fuel economy that compensates for this increase in greenhouse gas. The article states that diesel engines get 20% to 40% better fuel economy and produce 15% more greenhouse gases. No indication of the scope of these claims was explicitly made.
- (c) *Obese higher-status women might suffer from a particular income disadvantage in their jobs.* The authors studied the connection between weight and individual income among gainfully employed Finnish women and men at different levels of socioeconomic status. They used a population-based survey including 2068 women and 2314 men with linked income data from a taxation register. Regression analysis was used to calculate mean income levels within educational and occupational groups. The study revealed a clear income disadvantage among obese women in higher socioeconomic status groups, whereas a similar wage-depressant association with obesity was not found among women in lower socioeconomic groups. It also showed that excess body weight was not associated with income disadvantages in men.
- (d) *Alcohol consumption may protect against risk of Alzheimer's disease(AD), particularly in female nonsmokers.* A new study suggests that the consumption of alcohol reduces the risk of Alzheimer's disease. It suggests the tobacco consumption does not affect disease but there is a considerable effect in alcohol consumers. Moreover, the effect was more in nonsmokers and hence it raises an interesting question of interactions between tobacco and alcohol.

- (e) *A study published Monday drew on a host of metrics to conclude something we all already knew: The media covers Apple more than any other company.* The study analyzed the percent of technology related news headlines where Apple Inc. was the primary subject. The results show that Apple Inc. accounts for 15.1 percent of the coverage, with Google a close second at 11.4 percent. Microsoft accounted for 3 percent of the total coverage.

2. For each of your identified claims, match it to one of the following tasks:

- (a) Classification
- (b) Regression
- (c) Pattern discovery
- (d) Clustering
- (e) Anomaly detection

Restate the claim and describe, in one or more paragraphs, how the claim could be analyzed as an example of the task to which you have matched it. Consider carefully whether you could obtain the needed data from web-accessible sources or elsewhere, since you will need to find at least one data source to use in question 2 below.

Sample solution

- (a) *Study Reveals Why Women Apologize So Much.* This claim could best be described as pattern discovery. Because the study focused exclusively on college aged students, it cannot be assumed to correlate to the population on the whole. Instead the study sought to determine how often a college student would likely apologize given gender, how often a college student would perceive fault warranting an apology given gender, and how often a college student would apologize given that they have perceived fault and their gender. The findings of the study were simple rules that could be applied to the given population that would reflect how often a person meeting those criteria would likely apologize and how often they would likely perceive fault.
- (b) *Diesel vs. Gas: Fuel Economy vs. Emissions.* This claim could best be described as regression. The claim states that given the entire population of cars, the section of cars which run on diesel fuel are 20% to 40% more fuel efficient and have 15% more emissions. Some undisclosed subset of cars were tested and the results of those tests were extrapolated to apply the entire population of cars.
- (c) *Obese higher-status women might suffer from a particular income disadvantage in their jobs.* This hypothesis can be matched to pattern discovery. We can extract information from the samples on certain attributes, i.e. self-reported income, weight and socioeconomic status, such as education level and occupation status. Then we can compare the income of women at different weight levels with the income of women in same educational and occupational group. The result of such comparison will lead to the result of the study.

- (d) *Alcohol consumption may protect against risk of Alzheimer’s disease(AD), particularly in female nonsmokers.* The claim can be matched to the classification task. A dataset containing healthy and AD patients can be collected along with their age, sex, alcohol-consumption (Nil, Low, Moderate, High), smoking (Nil, Social, Chain), Alzheimer’s disease (yes , no). Then, Naive bayes classifier can be used to determine the conditional probability of the class variable(alzheimers disease) given other attributes. The conditional probabilities($P(AD|female, non-smoker, alcohol-consumption)$) can verify if there the alcohol consumption has a positive or negative effect on female non smokers.
- (e) *A study published Monday drew on a host of metrics to conclude something we all already knew: The media covers Apple more than any other company.* This article is an example of Anomaly Detection, in that the observed frequency of technology articles where Apple Inc. is the subject is noticeably greater than similar technology companies. In order to analyze this claim, we need a frequency count representing the number of articles each major technology company has been featured in during some common timespan. A histogram plotting the respective frequencies would allow the statistician to visualize the relationship between the company and the number of news articles that feature that company.

2 Working with Data (10 pts)

1. Identify, obtain, and prepare a data set to support one of the claims you identified in question 1. Data.gov is one possible place to look for public data that relates to claims found in news articles.

Your data must have $D > 500$, where $D = NA - M$.

- N is the number of instances (rows)
- A is the number of attributes (columns)
- M is the total number of missing data points

Make sure that $N > 40$, and $A > 4$ and do not count identifiers (e.g., name, unique code number, etc) toward A . In addition, make sure that the data have at least three continuous attributes and two discrete attributes.

For example, if you analyzed the House races in 2008 by looking at the party of the winner, the margin of victory, the party split in the district, whether there was an opponent, and campaign spending, then $N = 435$, $A = 5$, and M might be as large as 100 (because of unavailable data). Thus, $D = 2075$.

Sample solution

The data we use to support the claim from the news article is from The Association of Religion Data Archives. The data set is titled “*Religion and America’s Role in the World*”. We use a subset of the full data set for our analysis. Our data set has

16 meaningful attributes, and 1400 rows with 2850 missing values. This, our value of D is:

$$D = NA - M = (1400)(16) - 2850 = 19550 > 500$$

As $N = 1400 > 40$ and $A = 16 > 4$, and we have at least three continuous and two discrete attributes, the data set is valid for analysis.

2. Write a description of the data that includes:

(a) Information about the data, including at least the following:

- Data representation
- Population, sample, and sampling mechanism
- Attribute description that includes semantics (i.e., what they record) , type (e.g., nominal, ordinal), and range
- Which of your claims from question 1 the data is appropriate for

Sample solution

Data Representation The data representation of our data set is *tabular*, with 16 non-identifier attributes and 1400 instances.

Population, Sample, Sampling Mechanism The *population* under consideration is the U.S. population, which was sampled. The *sample* is composed of three sets: 1000 adult respondents chosen randomly, 100 young evangelical christians selected randomly, and 300 young evangelical christians drawn from an opt-in web panel designed to be demographically representative at the national level. The *sampling mechanism* for the first two sets is a random digit dial process that included both listed and unlisted phones.

Attribute Description

- **CaseID**; *ordinal*; 1-1400: The unique identifier for the individual.
- **GENDER**; *nominal*; {Male=1, Female=2}: The gender of the individual.
- **AGE**; *interval*; 18-100: The age of the individual.

- **RELIG1A**; *nominal*;

<i>Protestant</i>	=	1
<i>RomanCatholic</i>	=	2
<i>Jewish</i>	=	3
<i>Muslim/Islam</i>	=	4
<i>Mormon</i>	=	5
<i>Orthodox</i>	=	6
<i>OtherChristian</i>	=	7
<i>Other</i>	=	8
<i>NoPref/Atheist/Agnostic</i>	=	9
<i>Refused</i>	=	10

The religious preference of the individual.

- **RELIGSID**; *nominal*;

<i>Fundamentalist</i>	=	1
<i>Evangelical</i>	=	2
<i>Charismatic</i>	=	3
<i>Pentecostal</i>	=	4
<i>Mainline</i>	=	5
<i>Liberal</i>	=	6
<i>None</i>	=	7
<i>Other</i>	=	8
<i>Refused</i>	=	9

This identifier is specific *only* to individuals who have Protestant or other christian preference. The modifier most closely associated by the individual with their religion.

- **BORNAGIN**; *nominal*; {Yes=1, No=2, Refused=3}: This identifier is specific *only* to individuals who are not Jewish, Muslim or Atheist/Agnostic. Whether or not the individual considers themselves to be a born-again christian.
- **FEELOBAM**; *interval*; 0-100: The individual's feelings towards Obama, where 0 represents very unfavorable and 100 represents very favorable.
- **FEELREP**; *interval*; 0-100: The individual's feelings towards Republicans, where 0 represents very unfavorable and 100 represents very favorable.
- **FEELDEM**; *interval*; 0-100: The individual's feelings towards Democrats, where 0 represents very unfavorable and 100 represents very favorable.

- **PRES08A**; *nominal*;

DemocratBarackObama = 1
RepublicanJohnMcCain = 2
Undecided = 3
OtherCandidate = 4
Refused = 5

If the election were today, the candidate that the individual would vote for.

- **RELIGCAN**; *nominal*;

DemocratBarackObama = 1
RepublicanJohnMcCain = 2
EquallyReligious = 3
Refused = 4

The candidate the individual thinks is more religious.

- **EDUC**; *ordinal*;

1 – 11th grade = 1
HighSchoolGraduate = 2
Non – CollegePostHighSchool = 3
SomeCollege = 4
CollegeGraduate = 5
GraduateSchool = 6
Refused = 7

The highest level of education the individual has completed.

- **EMPLOY**; *nominal*;

FullTimeEmployment = 1
PartTimeEmployment = 2
Unemployed = 3
Retired = 4
Student = 5
Homemaker = 6
Other = 7
Refused = 8

The job status of the individual.

- **MARITAL2**; *nominal*;

<i>Married</i>	=	1
<i>Single</i>	=	2
<i>Separated</i>	=	3
<i>Divorced</i>	=	4
<i>Widowed</i>	=	5
<i>Refused</i>	=	6
<i>DomesticPartnership</i>	=	7

The relationship status of the individual.

- **RELIG2**; *ordinal*;

<i>MoreThan1/week</i>	=	1
<i>1/week</i>	=	2
<i>1/month</i>	=	3
<i>SeveralTimes/year</i>	=	4
<i>HardlyEver</i>	=	5
<i>Never</i>	=	6
<i>Refused</i>	=	7

How often the individual attends religious services.

- **RACE**; *nominal*;

<i>White</i>	=	1
<i>AfricanAmerican</i>	=	2
<i>Hispanic/Latino</i>	=	3
<i>Other</i>	=	4
<i>Refused</i>	=	5

The race of the individual.

- **INCOME**; *ordinal*;

< \$10,000	=	1
\$10,000 – \$20,000	=	2
\$20,000 – \$30,000	=	3
\$30,000 – \$50,000	=	4
\$50,000 – \$75,000	=	5
\$75,000 – \$100,000	=	6
\$100,000+	=	7

The total family income from all sources, before taxes, of the individual.

Claims Addressed This data is appropriate to address <a specific claim in the five claims from Question 1>.

- (b) A discussion of possible data quality issues and an outline as to what might be an appropriate response to address the issues.

Sample solution

The fact that the web panel sample used non-probability based sampling methods should be addressed when considering data quality. That is, the users that visited the opt-in website were not chosen at random and may not be representative of the underlying population. This effect of this possible *convenience sample* is likely negligible, given that only 300 of the 1400 participant responses were obtained through the website.

3 Data Exploration (10 pts)

Using your dataset from question 2, load your data into R¹, and do the following analysis:

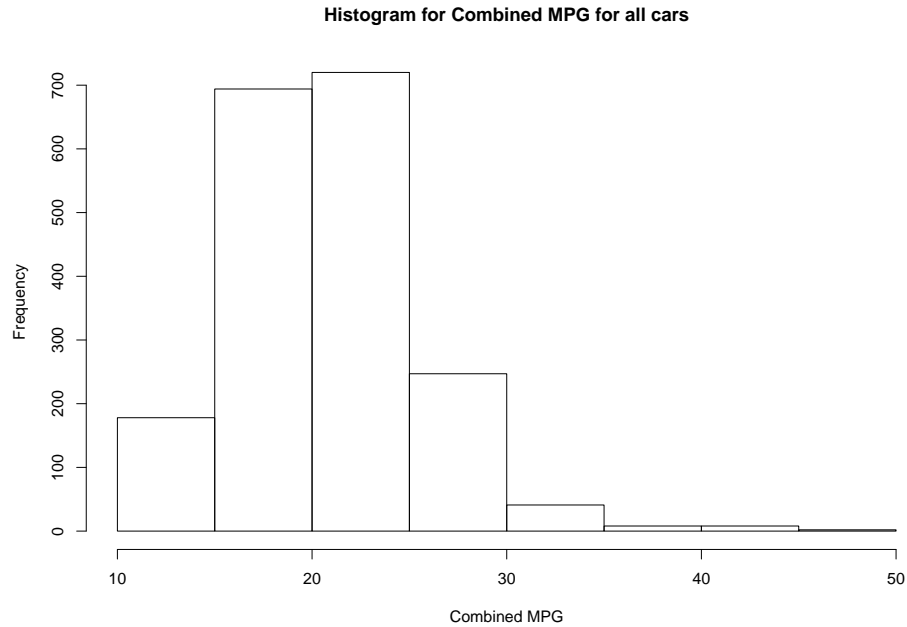
1. Find two attributes with interesting histograms. Plot them and describe briefly why they are interesting.

Sample solution

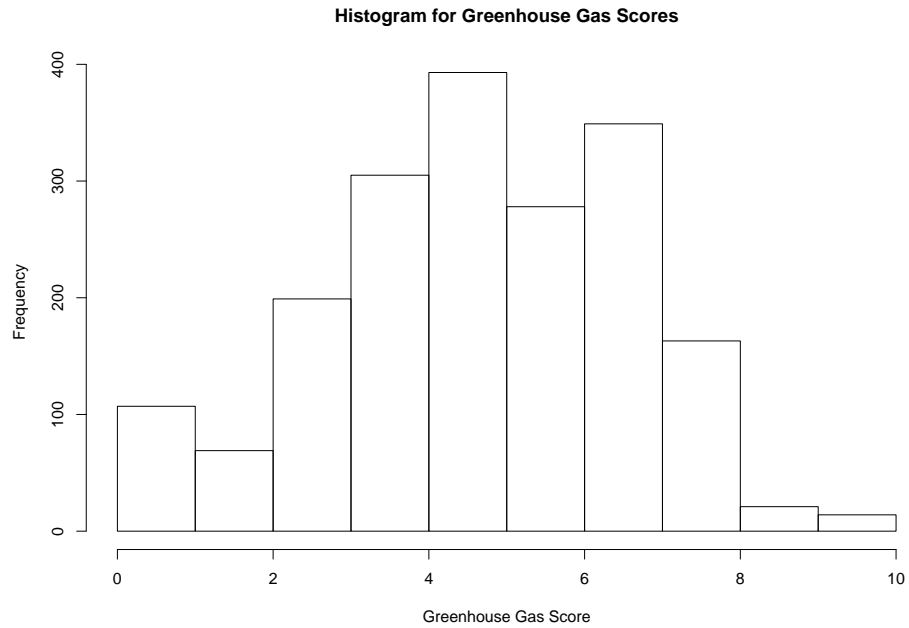
- (a) Histogram for combined MPG. This is interesting for car buyers to compare the fuel economy of the vehicle they are considering to other vehicles on the market. If this is an aim of the car buyer, then knowing how the MPG of one car compares to other is important. An interesting feature shown in the

¹The class resource page has links to the R software, a tutorial, and a cheatsheet with basic command. If you prefer to use Matlab or some other other statistical package, that's fine—but the TA will not be able to support you for packages other than R.

histogram is that the distribution is skewed to the right, which indicates that there are very few cars with high MPGs. Hence, if one is considering an car with a high MPG (i.e. > 30), then one would know that the car is likely to be one of the few very efficient vehicles.



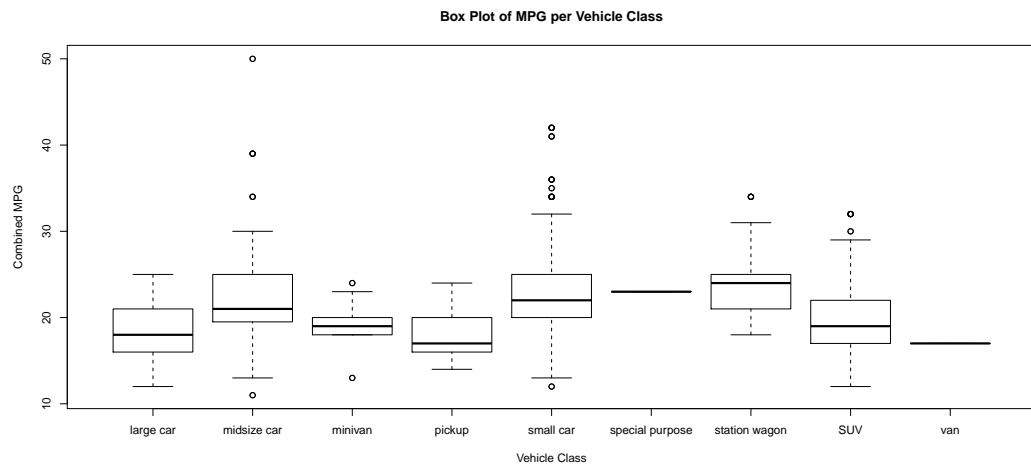
- (b) Histogram for greenhouse gas scores. This is interesting because it shows that the scores are normally distributed over the entire range with the peak of the distribution around 5. This indicates that very few cars release a large amount of greenhouse gasses and a very few release a small amount. Hence, this will help car buyers compare the scores of a given car to the whole population. While one would expect the distribution of greenhouse gas scores to mimic the fuel economy (i.e. combined MPG) distribution, the following histogram shows that the distributions differ in their skewness. This might prompt further investigation into the correlation between the two variables.



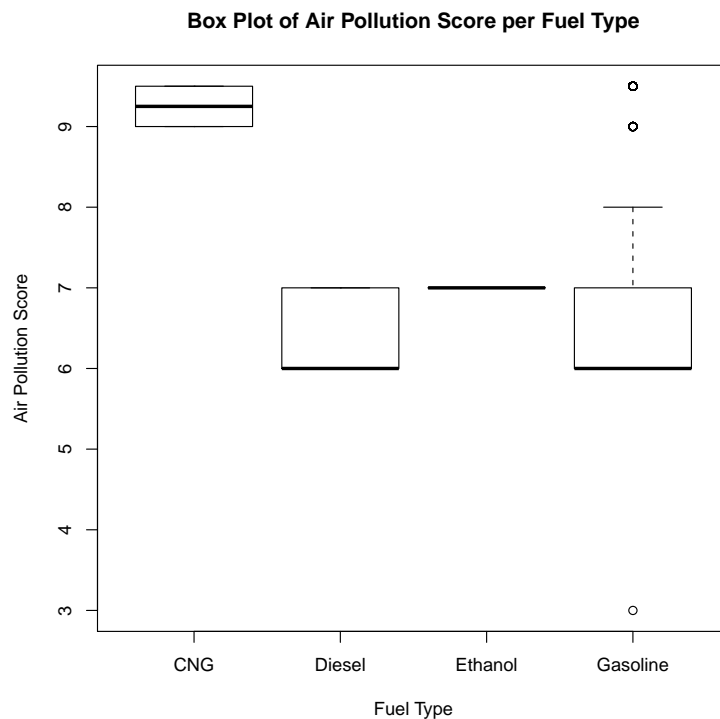
2. Find two sets of two attributes that have interesting scatter plots or box plots. Plot them and describe briefly why they are interesting.

Sample solution

- (a) Box plot for combined MPG and vehicle class. This is interesting because it shows whether fuel economy is a feature of vehicle class/size. This plot helps answer the claim in question 1. It shows that medium cars can have similar fuel efficiencies to smaller cars. It also shows that a medium-sized car has the most fuel efficiency (albeit, it is an outlier). Furthermore, it shows some separation between the different classes of cars (for instance by comparing the fuel efficiency of large cars to small cars):



- (b) Box plot for air pollution score and fuel type. This plot is interesting for several reasons. Firstly, it shows a clear separation between CNG and other types of fuel. Cars that run on CNG tend to have much higher air pollution scores. Furthermore, it shows that for all other types of fuel, the air pollution scores are more or less the same. This might be used to disprove the hypothesis that cars that run on diesel have worst air pollution scores than those that run on gasoline. Furthermore, it shows that very few cars that run on gasoline have high air pollution scores. This might indicate that the air pollution scores are determined by factors other than the type of fuel itself if considering cars that run on gasoline. Hence, this plot clearly shows the relationship between fuel type and the air pollution it causes.



3. Consider three pairs of continuous attributes and three pairs of discrete attributes. (If you don't have enough discrete attributes for this, then first discretize a continuous attribute to pair up with a discrete attribute.) Determine whether there are any pairs with strong associations. For pairs of continuous attributes, calculate correlation; for pairs of discrete attributes calculate χ^2 scores from contingency tables. Include assessments of significance.

Sample solution

- (a) Pairs of discrete attributes:
- Vehicle Class and Drive: $\chi^2 = 274.3561$, degrees of freedom = 8, p -value $< 2.2e - 16$. By looking at the degrees of freedom in addition to the χ^2 value

we can deduce that such a combination is unlikely. This is given by the p -value which indicates the significance of the test. It is less than 0.05 which indicates insignificance. Thus, we reject the null hypothesis that the joint distribution of the cell counts in the contingency table is the product of the row and column marginals; i.e. we reject the hypothesis that the variables are independent. Thus, there is a dependence (i.e. high association) between vehicle class and its drive.

- Vehicle Class and Fuel: $\chi^2 = 146.8878$, degrees of freedom = 24, p -value $< 2.2e - 16$. Since the probability for the combination of χ^2 value and the given degrees of freedom is small (< 0.05), then we reject the null hypothesis that the variables are independent. Thus, there is a dependence between vehicle class and the type of fuel it runs on.
- Vehicle transmission type and Fuel: $\chi^2 = 74.0647$, degrees of freedom = 45, p -value = 0.004091. Since the probability for the combination of χ^2 value and the given degrees of freedom is small (< 0.05), then we reject the null hypothesis that the variables are independent. Thus, there is a dependence.

(b) Pairs of continuous attributes:

- Car Displacement and Combined MPG: correlation = -0.7706756. Significance: $t = -52.6619$, $df = 1896$, p -value $< 2.2e - 16$. The correlation value indicates slightly negative correlation. In addition, the significance test gave a p -value < 0.05 . This indicates that the null hypothesis stating that the correlation is 0 should be rejected. Thus, there exists a negative correlation between the variables.
- City MPG and Highway MPG: correlation = 0.9018159. Significance: $t = 90.8724$, $df = 1896$, p -value $< 2.2e - 16$. The correlation value indicates strong positive correlation. In addition, the significance test gave a p -value < 0.05 . This indicates that the null hypothesis stating that the correlation is 0 should be rejected. Thus, there exists a strong positive correlation between the variables.
- Car Displacement and Number of cylinders: Correlation = 0.9079518. Significance: $t = 94.3389$, $df = 1896$, p -value $< 2.2e - 16$. The correlation value indicates strong positive correlation. In addition, the significance test gave a p -value < 0.05 . This indicates that the null hypothesis stating that the correlation is 0 should be rejected. Thus, there exists a strong positive correlation between the variables.

4. Transform all discrete attributes into numerical attributes, center the data to have zero means, and run PCA on the data. Plot the scree plot and determine whether the data can be reduced to a smaller number of dimensions. Discuss how to choose an appropriate number of reduced dimensions. Report the details of the first two principal components (i.e., the component attributes and their weights), as well as the proportion of variance that they explain. Discuss whether this gives any insight into the claim that has been made about the data in question 1.

Table 1: Principal Components

	Comp1	Comp2
Aid	-0.485	0.814
Rank	-0.621	-0.562
Faculty Ratio	-0.248	
SAT	0.518	
Cost	0.223	

Table 2: Principal Components

	Comp1	Comp2
Standard deviation	2.7735846	1.5755042
Proportion of Variance	0.6262929	0.2020849
Cumulative Proportion	0.6262929	0.8283778

Sample solution

The dimension can be chosen by selecting only the principal components which captures the variance of at least a threshold θ (say 3%). The scree plot shown in the following figure determines that the data can be reduced to smaller dimension. When a threshold of 3% was used to select the principal components the dimension got reduced to 4 from 6.

Table 1 shows the component attributes and its weights for the first two principal component. Table 2 shows the proportion of variance of the principal components.

PCA supports the claim done in the article. The weights of the component attributes shows that the SAT plays some role in predicting success (graduation rate) but it is not the sole criteria. Other things like Aid also plays a major role in graduation rate.

pl.pca

