The Use of Wikipedia in Universities

Introduction

The Wiki4HE data set contains research information from two different universities, Universitat Oberta de Catalunya (UOC) and Universitat Pompeu Fabra (UPF). Research was conducted on the perceptions and use by professors and other faculty. The goal of this analysis is to accurately predict the use of Wikipedia among faculty members.

Data

The data set contains 53 variables

AGE	Numeric
GENDER	0=Male 1=Female
DOMAIN	1=Arts & Humanities, 2= Sciences, 3= Health Sciences, 4= Engineering & Architecture, 5= Law & Politics, 6=Social Sciences
PhD	0=No, 1=Yes
YEARSEXP	Years of university teaching experience
UNIVERSITY	1= UOC, 2=UPF
UOC_POSITION	1=Professor, 2=Associate, 3=Assistant, 4=Lecturer, 5=Instructor, 6=Adjunct
OTHER	Main job in another university for part time . 1=Yes, 0=No
OTHER_POSITION	Work as part-time in another university and UPF members. 1=Professor, 2=Associate, 3=Assistant, 4= Lecturer, 5=Instructor, 6=Adjunct
USERWIKI	Wikipedia registered user. 0=No, 1=Yes

Perceived Usefulness

PU1	The use of Wikipedia make is easier for students to develop new skills
PU2	The use of Wikipedia improves students' learning.
PU3	Wikipedia is useful for teaching

Perceived Ease of Use

PEU1	Wikipedia is user-friendly
PEU2	It is easy to find in Wikipedia the information you seek
PEU3	It is easy to add or edit information in Wikipedia

Perceived Enjoyment

ENJ1	The use of Wikipedia stimulates curiosity
ENJ2	The use of Wikipedia is entertaining

Quality

QU1	Articles in Wikipedia are reliable
QU2	Articles in Wikipedia are updated
QU3	Articles in Wikipedia are comprehensive
QU4	In my area of expertise, Wikipedia has lower quality than other educational resources
QU5	I trust in the editing system of Wikipedia

Visibility

VIS1	Wikipedia improves visibility of students' work
Vis2	It is easy to have a record of the contributions made in Wikipedia
Vis3	I cite Wikipedia in my academic papers

Social Image

IM1	The use of Wikipedia is well considered among colleagues
IM2	In academia, sharing open educational resources is appreciated
IM3	My colleagues use Wikipedia

Sharing attitude

SA1	It is important to share academic content in open platforms

SA2	It is important to publish research results in other media than academic journals or books
SA3	It is important that students become familiar with online collaborative environments

Use Behavior

Use1	I use Wikipedia to develop my teaching
Use2	I use Wikipedia as a platform to develop educational activities with students.
Use3	I recommend my students to use Wikipedia
Use4	I recommend my colleagues to use Wikipedia
Use5	I agree my students use Wikipedia in my course

Profile 2.0

PF1	I contribute to blogs
PF2	I actively participate in social networks
PF3	I publish academic content in open platforms

Job Relevance

JR1	My university promotes the use of open collaborative environments in the Internet
JR2	My university considers the use of open collaborative environments in the Internet as a teaching merit

Behavior Intention

BI1	In the future I will recommend the use of Wikipedia to my colleagues and students
BI2	In the future, I will use Wikipedia in my teaching activity

Incentives

INC1	To design educational activities using Wikipedia, it would be helpful: a best practice guide
INC2	To design educational activities using Wikipedia, it would be helpful: getting instruction from a colleague
INC3	To design educational activities using Wikipedia, it would be helpful: getting specific training
INC4	To design educational activities using Wikipedia, it would be helpful: greater institutional recognition.

Experience

Exp1	I consult Wikipedia for issues related to my field of expertise
Exp2	I consult Wikipedia for other academic related issues
Exp3	I consult Wikipedia for personal issues
Exp4	I contribute to Wikipedia (editions, revisions, articles improvement)
Exp5	I use wikis to work with my students

The variables in this data set surveying faculty's view of Wikipedia are all completed on an ordinal scale from 1(strongly disagree/never) to 5 (strongly agree/always). The summary table below helps demonstrate which variables are rated more agreed upon by faculty. PEU1 and PEU2 show an overall higher agreement. SA1, SA2, SA3 also show a higher rate of agreement. USE2 and EXP4 show low disagree/never scores. Many of the other variables have an overall mean of approximately neutral.

```
      AGE            GENDER          DOMAIN            PhD         
 Min.   :23.00   Min.   :0.000   Min.   :1.000   Min.   :0.0000  
 1st Qu.:36.00   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:0.0000  
 Median :42.00   Median :0.000   Median :5.000   Median :0.0000  
 Mean   :42.25   Mean   :0.425   Mean   :4.098   Mean   :0.4644  
 3rd Qu.:47.00   3rd Qu.:1.000   3rd Qu.:6.000   3rd Qu.:1.0000  
 Max.   :69.00   Max.   :1.000   Max.   :6.000   Max.   :1.0000  
                                 NA's   :2                        
    YEARSEXP       UNIVERSITY     UOC_POSITION   OTHER_POSITION 
 Min.   : 0.00   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.: 5.00   1st Qu.:1.000   1st Qu.:6.000   1st Qu.:1.000  
 Median :10.00   Median :1.000   Median :6.000   Median :2.000  
 Mean   :10.87   Mean   :1.124   Mean   :5.406   Mean   :1.589  
 3rd Qu.:15.00   3rd Qu.:1.000   3rd Qu.:6.000   3rd Qu.:2.000  
 Max.   :43.00   Max.   :2.000   Max.   :6.000   Max.   :2.000  
 NA's   :23                      NA's   :113     NA's   :261    
  OTHERSTATUS       USERWIKI           PU1             PU2            PU3      
 Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.00   Min.   :1.00  
 1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:3.00  
 Median :4.000   Median :0.0000   Median :3.000   Median :3.00   Median :3.00  
 Mean   :4.209   Mean   :0.1375   Mean   :3.138   Mean   :3.15   Mean   :3.45  
 3rd Qu.:7.000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:4.00  
 Max.   :7.000   Max.   :1.0000   Max.   :5.000   Max.   :5.00   Max.   :5.00  
 NA's   :540     NA's   :4        NA's   :7       NA's   :11     NA's   :5     
      PEU1            PEU2            PEU3            ENJ1      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:4.000   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:3.000  
 Median :5.000   Median :4.000   Median :3.000   Median :4.000  
 Mean   :4.356   Mean   :4.046   Mean   :3.384   Mean   :3.795  
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
 NA's   :4       NA's   :14      NA's   :97      NA's   :7      
```

```
      ENJ2            Qu1             Qu2             Qu3       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000  
 Median :4.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :3.821   Mean   :3.195   Mean   :3.422   Mean   :2.981  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
 NA's   :17      NA's   :7       NA's   :10      NA's   :15     
      Qu4             Qu5             Vis1            Vis2      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :3.238   Mean   :3.042   Mean   :2.945   Mean   :3.069  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
 NA's   :22      NA's   :29      NA's   :72      NA's   :117    
      Vis3            Im1             Im2             Im3       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000  
 Median :2.000   Median :2.000   Median :3.000   Median :3.000  
 Mean   :2.027   Mean   :2.478   Mean   :3.295   Mean   :2.888  
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
 NA's   :8       NA's   :22      NA's   :20      NA's   :57     
      SA1             SA2            SA3             Use1            Use2      
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:4.000   1st Qu.:4.00   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :4.000   Median :4.00   Median :5.000   Median :2.000   Median :1.000  
 Mean   :4.191   Mean   :4.13   Mean   :4.384   Mean   :2.116   Mean   :1.831  
 3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:2.000  
 Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000  
 NA's   :11      NA's   :12     NA's   :11      NA's   :14      NA's   :17     
```

```
     Use3            Use4            Use5             Pf1
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.000
 Median :3.000   Median :3.000   Median :3.000   Median :2.000
 Mean   :2.662   Mean   :2.554   Mean   :3.305   Mean   :2.274
 3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
 NA's   :9       NA's   :23      NA's   :15      NA's   :11
     Pf2             Pf3             JR1             JR2
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:2.000
 Median :3.000   Median :2.000   Median :4.000   Median :3.000
 Mean   :2.861   Mean   :2.551   Mean   :3.699   Mean   :3.108
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
 NA's   :6       NA's   :14      NA's   :27      NA's   :53
     BI1             BI2            Inc1            Inc2            Inc3
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.00   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000
 Median :3.000   Median :3.00   Median :4.000   Median :4.000   Median :3.000
 Mean   :2.952   Mean   :2.99   Mean   :3.746   Mean   :3.461   Mean   :3.442
 3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000
 NA's   :32      NA's   :43     NA's   :35      NA's   :35      NA's   :37
     Inc4           Exp1            Exp2            Exp3            Exp4
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:3.00   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:1.000
 Median :4.00   Median :3.000   Median :4.000   Median :4.000   Median :1.000
 Mean   :3.49   Mean   :3.001   Mean   :3.492   Mean   :3.651   Mean   :1.588
 3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
 Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
 NA's   :42     NA's   :13      NA's   :11      NA's   :13      NA's   :14
```

```
     Exp5
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.487
 3rd Qu.:4.000
 Max.   :5.000
 NA's   :13
```

Getting the Data Ready

Before the data is used for analysis, there is concern about missing values. There are 1,995 missing values in the entire data set. Removing all of the observations that contain missing values would greatly reduce the data set. The variable UOC_POSITION contains 113 missing values, this is due to the fact that 113 of these observations are from UPF and do not hold a position at UOC. The variable OTHER_POSITION contains 261 missing values, and OTHERSTATUS contains 540 missing values. This number of missing values is significant for one variable to contain and can result in numerous questions about the usefulness of these two variables. The missing values were replaced with the column means in order to properly conduct the following analysis. There are some disadvantages of replacing the missing values with the column means. For example, in the Use3 column, there were 9 missing values, when they were replaced with the mean for the column, there were 9 more observations in the Yes category. In a survey situation, it is not likely that all of these observations would in fact go into one category. There are other methods for predicting missing values (packages missForest), but for this analysis, a simpler approach was used. (A summary table of the replaced data was examined, but it was not output)

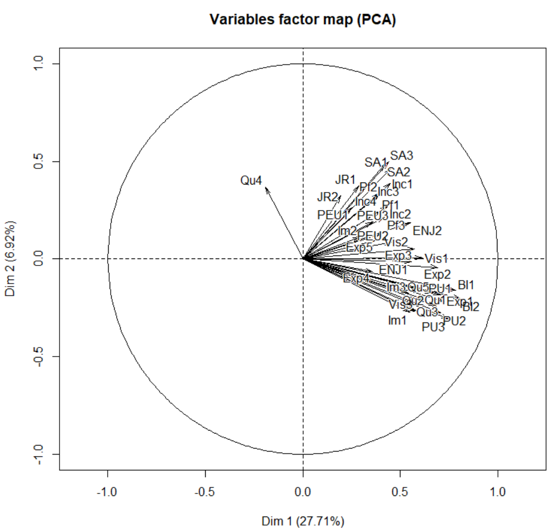
The data set was split into training and a testing subsets. They were split 50% in each of the subsets. This split allowed a large amount of observations in each category to help with the model building and model testing. A split could also have been done with 1/3 of the data in training and 2/3 of the data in testing, but ½ of the data in the training data set allowed for a better model building process.

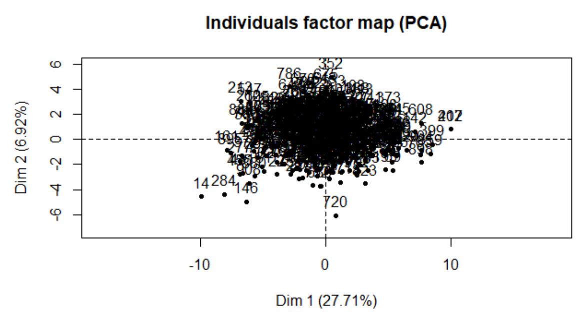
The split was stratified so that there was an equal representation of the two Universities in the training and testing set as well as an equal representation of the domain variables in each the training and testing data. Faculty from UOC make up 800 observations, while faculty from UPF only 113 observations. If the data was not properly subset, the training or the testing data could contain observations from only one university.

PCA on Training Dataset.

PCA was used to reduce the dimensionality of this large data set and identify variables that will be useful predictors in later analysis. The first principal component explains 27.71% of the variance found in the data, while the second principal component explains 6.92% of the variance. Qu4 (Wikipedia has a lower quality in my area of expertise) reveals a principal component vector that is going in the opposite direction. As observations move to the right, the Qu4 value is decreasing while the other variable values are increasing.

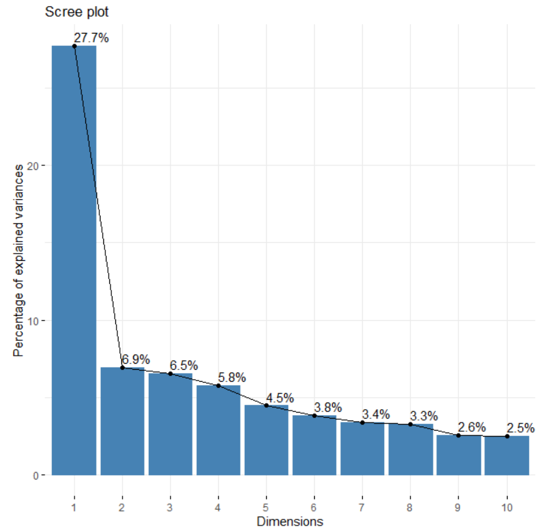
The survey variables are being used as continuous data in PCA. The program is thus assuming that the variables can take on values between their agreed upon scale (1,2,3,4,5). This can change the results of the best model output.





The eigenvalues and the scree plot reveal that there is an elbow after the first principal component. After the first pc, the remaining principal components only explain a small portion of the data, 7% or less per component. For this particular data set, PCA revels that the first prinipal component is enough for explaining the variance in this training subset, even though 27.71% is not a large amount of variance explained.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	10.5293650	27.7088552	27.70886
comp 2	2.6312007	6.9242124	34.63307
comp 3	2.4822387	6.5322071	41.16527
comp 4	2.1892099	5.7610788	46.92635
comp 5	1.6962596	4.4638410	51.39019
comp 6	1.4567650	3.8335922	55.22379
comp 7	1.2895704	3.3936062	58.61739
comp 8	1.2474355	3.2827251	61.90012
comp 9	0.9722695	2.5586041	64.45872
comp 10	0.9451078	2.4871258	66.94585
comp 11	0.9081152	2.3897769	69.33562
comp 12	0.8654178	2.2774152	71.61304
comp 13	0.7439930	1.9578763	73.57092
comp 14	0.7125959	1.8752524	75.44617
comp 15	0.6721296	1.7687621	77.21493
comp 16	0.6343186	1.6692595	78.88419
comp 17	0.5628653	1.4812245	80.36541
comp 18	0.5402718	1.4217678	81.78718
comp 19	0.5199489	1.3682865	83.15547
comp 20	0.5144848	1.3539072	84.50938
comp 21	0.5064217	1.3326886	85.84206
comp 22	0.4880679	1.2843892	87.12645
comp 23	0.4611783	1.2136270	88.34008
comp 24	0.4329599	1.1393681	89.47945
comp 25	0.4056864	1.0675958	90.54704
comp 26	0.3796969	0.9992024	91.54625
comp 27	0.3687622	0.9704269	92.51667
comp 28	0.3577378	0.9414153	93.45809
comp 29	0.3272891	0.8612870	94.31938
comp 30	0.3137955	0.8257777	95.14515
comp 31	0.3031524	0.7977696	95.94292
comp 32	0.2994788	0.7881022	96.73103
comp 33	0.2633729	0.6930866	97.42411
comp 34	0.2513929	0.6615604	98.08567
comp 35	0.2397184	0.6308378	98.71651
comp 36	0.1979409	0.5208971	99.23741
comp 37	0.1773939	0.4668261	99.70423
comp 38	0.1123912	0.2957662	100.00000



The loading vectors and $contrib were used to identify which variables would be most useful in Logisitics regression, LDA and QDA.

Pu1, Pu2, Pu3,Qu1, BI1, BI2, Exp1, and Exp2 were identified as having the biggest contribution to the training subset. Each one explaining about 5% of the variance in the first principal component.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PU1	0.21786192	-0.112825299	0.003552012	-0.016447166	-0.014396442
PU2	0.21832910	-0.170598032	-0.036279950	-0.019868965	-0.006256616
PU3	0.22990236	-0.188674294	0.001762005	0.001693490	-0.043153347
PEU1	0.07715523	0.163520896	-0.238969354	0.273990276	-0.043869984
PEU2	0.14016171	0.053112442	-0.248157426	0.164528626	-0.097298322
PEU3	0.11026393	0.116779544	0.158986617	0.156346011	-0.111481631
ENJ1	0.17163222	-0.008033593	-0.138462030	0.067128080	-0.052667891
ENJ2	0.17065822	0.114796216	-0.149447198	0.175937620	-0.066064032
Qu1	0.21013372	-0.148386704	-0.172860946	0.077688602	-0.026643629
Qu2	0.19569025	-0.129022159	-0.134442017	0.084629486	-0.069553873
Qu3	0.17649044	-0.164775644	-0.147158886	0.081298510	0.067132871
Qu4	-0.05936295	0.227707186	0.039949346	0.061860688	0.003654756
Qu5	0.20432436	-0.109364277	-0.003573449	0.008852651	-0.047092030
Vis1	0.18891988	0.003896215	0.066156871	-0.080508332	-0.023317613
Vis2	0.12428218	0.079918773	0.125950437	0.039397337	0.015989586
Vis3	0.17818733	-0.163096095	0.194226099	-0.025556588	0.054636499
Im1	0.16936229	-0.167448811	0.014601827	0.004561667	0.325652827
Im2	0.09028459	0.070316738	-0.056071369	0.106051428	0.370033243
Im3	0.16700282	-0.107290693	-0.034939363	0.040695288	0.277923673
SA1	0.13906373	0.285722881	-0.046633785	0.203032770	-0.082963817
SA2	0.13041540	0.300585845	-0.050276577	0.232974200	-0.035320186
SA3	0.13521345	0.307490810	-0.086058512	0.187104417	-0.051706651
Pf1	0.12175317	0.150662853	0.386604220	-0.012705327	-0.072861050
Pf2	0.12202886	0.208172604	0.322288419	-0.007938610	-0.053932710
Pf3	0.12916936	0.132099926	0.338484535	0.063204493	-0.126043354
JR1	0.08842933	0.232219353	0.003676490	0.002410265	0.466564375
JR2	0.05974947	0.199882411	0.006893505	0.003587997	0.524687767
BI1	0.24178326	-0.099805396	0.062005167	-0.121605468	0.067441763
BI2	0.24877347	-0.124348029	0.059096570	-0.110955392	0.034988450
Inc1	0.13831521	0.239496195	-0.119161597	-0.328782854	-0.033959985
Inc2	0.13407893	0.166210615	-0.157632578	-0.395591725	-0.015923725
Inc3	0.11472961	0.193040750	-0.165703633	-0.456358862	0.012403233
Inc4	0.12025584	0.206914293	-0.091576418	-0.375428744	-0.184130749
Exp1	0.22357036	-0.133718310	0.039960259	-0.048173084	-0.077497398
Exp2	0.21293958	-0.026423298	-0.007383797	0.052214740	-0.158404694
Exp3	0.17649418	0.032017039	-0.099322730	0.104494760	-0.164426874
Exp4	0.11015387	-0.039070719	0.364355572	-0.007517229	-0.021367894
Exp5	0.11437159	0.059015585	0.269927443	0.017209921	0.054141262

$contrib

	Dim.1
PU1	4.7463816
PU2	4.7667597
PU3	5.2855093
PEU1	0.5952929
PEU2	1.9645305
PEU3	1.2158134
ENJ1	2.9457619
ENJ2	2.9124227
Qu1	4.4156181
Qu2	3.8294674
Qu3	3.1148874
Qu4	0.3523960
Qu5	4.1748445
Vis1	3.5690721
Vis2	1.5446059
Vis3	3.1750723
Im1	2.8683584
Im2	0.8151306
Im3	2.7889943
SA1	1.9338721
SA2	1.7008175
SA3	1.8282678
Pf1	1.4823834
Pf2	1.4891042
Pf3	1.6684723
JR1	0.7819746
JR2	0.3570000
BI1	5.8459143
BI2	6.1888238
Inc1	1.9131098
Inc2	1.7977159
Inc3	1.3162883
Inc4	1.4461467
Exp1	4.9983705
Exp2	4.5343265
Exp3	3.1150196
Exp4	1.2133875
Exp5	1.3080861

Logistic Regression

The variables above identified from the PCA, were used in addition to the Age, Gender, Domain, PhD, YearsExp, University, and Userwiki. The faculty variables UOC Position, Other Status, and Other Position were not added to the model because of the significant number of missing values in these variables, even with the mean of the variables inserted, the data could be skewed higher in the Yes or No categories as the data has been reduced to binary. Before logistic regression could be conducted, the Likert 5 point scale was converted to a binary scale, a score of 1 or 2 were changed to "No" and 3,4 & 5 were changed to "Yes". Three was converted to "Yes" because it indicates that the Professor doesn't have strong feelings for or against the use of Wikipedia, and it seems to indicate that the Professor is not against the use of Wikipedia. The Professors may not actively use Wikipedia or tell their students to use it, but they are not against students using it. Converting it to "Yes" on the scale also preserves more of the data.

In addition to the survey results being converted to a binary scale, the variables were also converted to factors so that they could be entered into the data, and their levels preserved.

The Use3 variable (I recommend my students to use Wikipedia) was used as the response variable to determine a faculty member's use of Wikipedia. This variable was chosen because it encompasses how professors and faculty truly feel about Wikipedia. If a faculty member is recommending Wikipedia, that means that they are in favor of Wikipedia. Use1 and Use2 seemed too narrow to use because many faculty members lecture from real life experiences or in alignment with a textbook.

```
Call:
glm(formula = Use3 ~ PU1 + PU2 + PU3 + Qu1 + BI1 + BI2 + Exp1 +
    Exp2 + AGE + GENDER + DOMAIN + PhD + YEARSEXP + UNIVERSITY +
    USERWIKI, family = binomial, data = wiki, subset = trainwiki)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5375  -0.4163   0.2592   0.5996   2.6444

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.34903    1.21019  -3.594 0.000326 ***
PU11         0.37852    0.38570   0.981 0.326404
PU21         0.86821    0.40766   2.130 0.033191 *
PU31         0.94329    0.50943   1.852 0.064078 .
Qu11         0.86123    0.40955   2.103 0.035475 *
BI11         0.68182    0.45828   1.488 0.136806
BI21         1.48641    0.44805   3.317 0.000908 ***
Exp11        1.11416    0.36097   3.087 0.002025 **
Exp21        1.05041    0.47036   2.233 0.025534 *
AGE         -0.01451    0.02259  -0.642 0.520625
GENDER1     -0.41427    0.29832  -1.389 0.164929
DOMAIN2      0.78698    0.71775   1.096 0.272880
DOMAIN3     -0.64029    0.59084  -1.084 0.278495
DOMAIN4     -0.81754    0.51683  -1.582 0.113687
DOMAIN5     -0.39087    0.54107  -0.722 0.470053
DOMAIN6     -0.85458    0.41837  -2.043 0.041088 *
PhD1        -0.23623    0.32543  -0.726 0.467892
YEARSEXP     0.02140    0.02652   0.807 0.419713
UNIVERSITY2 -0.62841    0.45516  -1.381 0.167391
USERWIKI1    1.57815    0.53016   2.977 0.002913 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 632.99  on 458  degrees of freedom
Residual deviance: 343.16  on 439  degrees of freedom
AIC: 383.16

Number of Fisher Scoring iterations: 6
```

```
        Use3test
glm.pred   0   1
     No  162  29
     Yes  48 215
```

(Yes is referred to as the positive class)

Sensitivity	Specificity	Accuracy
88.11%	**77.14%**	**83.04%**

The first run of logistic regression using the training and testing data was able to accurately predict a faculty members feelings towards the Use3 variable 83.04% of the time. This is a high prediction rate, but there are many coefficients in the data that are not significant at a 0.05 significance level.

Variables were removed from the model, and a second round of logistics regression was completed, but additional variables still needed to be removed from the model. (See additional output). To ensure that all of the correct variables were identified by the PCA, a full model logistic regression was also run, revealing that VIS3 was a significant predictor.

Logistic Regression Model 3

A final logistic regression model was run containing the variables PU2, BI2, Exp1, Exp2, Vis3, Domain, and UserWiki.

Domain 4 (Engineering & Architecture) and Domain 6 (Social Science) were the only two significant domain responses, and contained a negative coefficient value. This indicates that that these two departments are not in favor of the use of Wikipedia in the academic setting.

The remaining variables indicated that a value of 1 (Yes) was a significant predictor of the Use3 variable.

```
Call:
glm(formula = Use3 ~ PU2 + BI2 + Exp1 + Exp2 + Vis3 + DOMAIN +
    USERWIKI, family = binomial, data = wiki, subset = trainwiki)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7565  -0.4749   0.2128   0.5516   2.5635

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8150     0.5694  -6.701 2.08e-11 ***
PU21          1.5627     0.3516   4.444 8.81e-06 ***
BI21          1.7705     0.3241   5.462 4.71e-08 ***
Exp11         1.1657     0.3368   3.461 0.000538 ***
Exp21         1.1221     0.4584   2.448 0.014378 *
Vis31         1.4531     0.3635   3.997 6.41e-05 ***
DOMAIN2       0.4227     0.7121   0.594 0.552805
DOMAIN3      -0.9881     0.5818  -1.698 0.089443 .
DOMAIN4      -0.9352     0.4904  -1.907 0.056487 .
DOMAIN5      -0.3887     0.5299  -0.734 0.463228
DOMAIN6      -0.9953     0.4006  -2.485 0.012966 *
USERWIKI1     1.4525     0.5076   2.862 0.004215 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 632.99  on 458  degrees of freedom
Residual deviance: 344.23  on 447  degrees of freedom
AIC: 368.23

Number of Fisher Scoring iterations: 5
```

The confusion matrix was able to accurately predict the Yes/ No value of the Use3 variable 376 out of 454 times or 82.82%. This is lower than the percentage for the larger model fit above, but there are fewer variables in the model with makes it easier to understand, and the difference is the earlier model had one more correct prediction. The AIC of 368.23 is also lower for the final model indicating that it is a better fit for the data.

```
          Use3test
glm.pred3   0   1
      No  162  30
      Yes  48 214
```

(Yes is referred to as the Positive class)

Sensitivity	Specificity	Accuracy
87.7%	**77.14%**	**82.82%**

Linear Discriminant Analysis

LDA was conducted on the same variables from the logistic regression analysis that was found to be the same model.

```
Call:
lda(Use3 ~ PU2 + BI2 + Exp1 + Exp2 + Vis3 + DOMAIN + USERWIKI,
    data = wiki, subset = trainwiki)

Prior probabilities of groups:
        0         1 
0.4575163 0.5424837 

Group means:
       PU21      BI21     Exp11     Exp21      Vis31    DOMAIN2    DOMAIN3   DOMAIN4
0 0.5047619 0.3571429 0.3952381 0.6523810 0.07142857 0.03333333 0.09523810 0.1142857
1 0.9076305 0.9196787 0.8835341 0.9518072 0.49799197 0.08433735 0.06827309 0.1847390
     DOMAIN5   DOMAIN6  USERWIKI1
0 0.13809524 0.4476190 0.05238095
1 0.08835341 0.3493976 0.22489960

Coefficients of linear discriminants:
                 LD1
PU21       0.7699849
BI21       1.2591342
Exp11      0.8251365
Exp21      0.4059278
Vis31      0.7763169
DOMAIN2    0.1970657
DOMAIN3   -0.4354698
DOMAIN4   -0.3960522
DOMAIN5   -0.1517512
DOMAIN6   -0.4477828
USERWIKI1  0.5451421
```

```
          Use3test
lda.class   0   1
        0 163  26
        1  47 218
```

The LDA model was able to correctly predict the faculty's feelings towards Use3 381 out of 454 times or 83.92% this is slightly better than the logistic regression model.

Sensitivity	Specificity	Accuracy
89.34%	**77.62%**	**83.92%**

Quadratic Discriminant Analysis

```
Call:
lda(Use3 ~ PU2 + BI2 + Exp1 + Exp2 + Vis3 + DOMAIN + USERWIKI,
    data = wiki, subset = trainwiki)

Prior probabilities of groups:
        0         1
0.4575163 0.5424837

Group means:
       PU21      BI21      Exp11     Exp21      Vis31     DOMAIN2    DOMAIN3   DOMAIN4
0 0.5047619 0.3571429 0.3952381 0.6523810 0.07142857 0.03333333 0.09523810 0.1142857
1 0.9076305 0.9196787 0.8835341 0.9518072 0.49799197 0.08433735 0.06827309 0.1847390
     DOMAIN5    DOMAIN6   USERWIKI1
0 0.13809524 0.4476190 0.05238095
1 0.08835341 0.3493976 0.22489960

Coefficients of linear discriminants:
                 LD1
PU21       0.7699849
BI21       1.2591342
Exp11      0.8251365
Exp21      0.4059278
Vis31      0.7763169
DOMAIN2    0.1970657
DOMAIN3   -0.4354698
DOMAIN4   -0.3960522
DOMAIN5   -0.1517512
DOMAIN6   -0.4477828
USERWIKI1  0.5451421
```

```
         Use3test
qda.class   0   1
        0 163  26
        1  47 218
```

The QDA model was able to correctly predict the Use3 class 381 out of 454 times or 83.92%. This is the same percentage that is seen in LDA.

Sensitivity	**Specificity**	**Accuracy**
89.34%	**77.62%**	**83.92%**

Conclusion

From the different analysis, Faculty use of Wikipedia was able to be predicted accurately 84% of the time. The use of Wikipedia improves student learning (PU2), In the future, I will use Wikipedia in my teaching (BI2), I consult Wikipedia for issues related to my field of experience (Exp1), I consult Wikipedia for other academic related issues (Exp2), I cite Wikipedia in my academic papers (Vis3), DOMAIN and USERWIKI were the most significant predictors of whether or not a faculty member will use Wikipedia.

The LDA and QDA models have the highest rate of correct predictions with 83.92%. They also have the highest rates of sensitivity (89.34%) and specificity (77.62%).They both contain a smaller subset of variables which makes for a simpler model that is helpful for the prediction process. LDA can be similar to logistic regression with the normality assumption is met, but the survey variables are rated on 5 point scale, making it difficult to examine the normality assumption; LDA also performs better when the classes are well defined. Logistics regression has less restrictive assumptions, which can at times make it a better classification model, but when the assumption are met LDA performs better.

LDA and QDA both use a covariance matrix for their estimates. LDA uses a common covariance matrix whereas QDA creates a different covariance matrix for each class. In this analysis, there is not a higher percentage of predictability because of the separate covariance matrix. In cases where there are a small number of training observations or a non-linear relationship in the data, QDA models will perform significantly better since it only makes an assumption about the form of the decision boundary, but in this case, there are a large number of training observations.

Overall the LDA seems to be the best model for the data. The common covariance matrix provides a higher percentage of correct predictions with enough variables to provide correct predictions nearly 84% of the time without being a large complicated model. The LDA model is able to correctly predict the use of Wikipedia in faculty members 89.34% of the time, and it able to correctly predict the non use of Wikipedia at a rate of 77.61%.

Code

```
wiki = read.csv("wiki4HE.csv", header=T, sep=";", na.strings="?")

sum(is.na(wiki))

summary(wiki)

for(i in 1: ncol(wiki)){

 wiki[is.na(wiki[,i]),i]=mean(wiki[,i],na.rm=TRUE)

 }



library(caret)

set.seed(2)

trainwiki=createDataPartition(paste(wiki$UNIVERSITY, wiki$DOMAIN),p=.5, list=FALSE, times=1)

wikitrain= wiki[trainwiki,]

 wikitest= wiki[-trainwiki,]

table(wikitrain$DOMAIN)

table(wiki$DOMAIN)

table(wikitrain$UNIVERSITY)

 table(wiki$UNIVERSITY)

y=paste(wiki$DOMAIN, wiki$UNIVERSITY)

 table(y)


myvars=names(wiki)%in% c("AGE", "GENDER", "DOMAIN", "PhD", "YEARSEXP", "UNIVERSITY",
"UOC_POSITION","OTHER_POSITION","OTHERSTATUS", "USERWIKI","Use1","Use2", "Use3", "Use4",
"Use5")
```

```
newtrainingwiki=wikitrain[!myvars]

newtestingwiki=wikitest[!myvars]

## PCA on training data

library(FactoMineR)

trainingpca=PCA(newtrainingwiki, scale.unit=TRUE, graph=TRUE)

trainingpca$eig

trainingpca$var

loadings=sweep(trainingpca$var$coord,2,sqrt(trainingpca$eig[1:5,1]),FUN="/")

loadings

library(factoextra)

fviz_eig(trainingpca, addlabel=TRUE)

## Logistic Regression

library(car)

 wiki$PU1= recode(wiki$PU1, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$PU2= recode(wiki$PU2, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$PU3= recode(wiki$PU3, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$PEU1= recode(wiki$PEU1, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$PEU2= recode(wiki$PEU2, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$PEU3= recode(wiki$PEU3, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$ENJ1= recode(wiki$ENJ1, '1=0; 2=0;3=1; 4=1;5=1')

 wiki$ENJ2= recode(wiki$ENJ2, '1=0; 2=0;3=1; 4=1;5=1')

wiki$Qu2= recode(wiki$Qu2, '1=0; 2=0;3=1; 4=1;5=1')
```

```
wiki$Qu1= recode(wiki$Qu1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Qu3= recode(wiki$Qu3, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Qu4= recode(wiki$Qu4, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Qu5= recode(wiki$Qu5, '1=0; 2=0;3=1; 4=1;5=1')

wiki$Vis1= recode(wiki$Vis1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Vis2= recode(wiki$Vis2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Vis3= recode(wiki$Vis3, '1=0; 2=0;3=1; 4=1;5=1')

wiki$Im1= recode(wiki$Im1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Im2= recode(wiki$Im2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Im3= recode(wiki$Im3, '1=0; 2=0;3=1; 4=1;5=1')

wiki$SA1= recode(wiki$SA1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$SA2= recode(wiki$SA2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$SA3= recode(wiki$SA3, '1=0; 2=0;3=1; 4=1;5=1')

wiki$Use1= recode(wiki$Use1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Use2= recode(wiki$Use2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Use3= recode(wiki$Use3, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Use4= recode(wiki$Use4, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Use5= recode(wiki$Use5, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Pf1= recode(wiki$Pf1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Pf2= recode(wiki$Pf2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Pf3= recode(wiki$Pf3, '1=0; 2=0;3=1; 4=1;5=1')

wiki$JR1= recode(wiki$JR1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$JR2= recode(wiki$JR2, '1=0; 2=0;3=1; 4=1;5=1')
```

```
wiki$BI1= recode(wiki$BI1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$BI2= recode(wiki$BI2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Inc1= recode(wiki$Inc1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Inc2= recode(wiki$Inc2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Inc3= recode(wiki$Inc3, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Inc4= recode(wiki$Inc4, '1=0; 2=0;3=1; 4=1;5=1')

wiki$Exp1= recode(wiki$Exp1, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Exp2= recode(wiki$Exp2, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Exp3= recode(wiki$Exp3, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Exp4= recode(wiki$Exp4, '1=0; 2=0;3=1; 4=1;5=1')
wiki$Exp5= recode(wiki$Exp5, '1=0; 2=0;3=1; 4=1;5=1')

data.frame(wiki)
wiki$USERWIKI=as.factor(wiki$USERWIKI)
wiki$PhD=as.factor(wiki$PhD)
wiki$GENDER=as.factor(wiki$GENDER)
wiki$UNIVERSITY=as.factor(wiki$UNIVERSITY)
wiki$DOMAIN=as.factor(wiki$DOMAIN)
wiki[,11:53]=lapply(wiki[,11:53],factor)

glm.full=glm(Use3~PU1+PU2+PU3+Qu1+BI1+BI2+Exp1+Exp2+AGE+GENDER+DOMAIN+PhD+YEARSEXP+
UNIVERSITY+USERWIKI, family=binomial, data=wiki, subset=trainwiki)
summary(glm.full)
Use3test=wikitest$Use3

glm.prob=predict(glm.full, wikitest, type="response")
summary(glm.prob)
glm.pred=rep("No",454)
```

```
glm.pred[glm.prob>.5]="Yes"

table(glm.pred,Use3test)



##Logistic Regression 2

glm.full2=glm(Use3~PU2+BI2+Exp1+Exp2+DOMAIN+USERWIKI, family=binomial, data=wiki,
subset=trainwiki)


 glm.prob2=predict(glm.full2, wikitest, type="response")

 summary(glm.prob2)


glm.pred2=rep("No",454)

glm.pred2[glm.prob2>.5]="Yes"


table(glm.pred2,Use3test)



##Logistic Regression 3

glm.full3=glm(Use3~PU2+BI2+Exp1+Exp2+Vis3+DOMAIN+USERWIKI, family=binomial, data=wiki,
subset=trainwiki)


 glm.prob3=predict(glm.full3, wikitest, type="response")

 summary(glm.prob3)


glm.pred3=rep("No",454)

glm.pred3[glm.prob3>.5]="Yes"


table(glm.pred3,Use3test)

glm.pred3=rep("No",454)
```

```
glm.pred3[glm.prob3>.5]="Yes"

table(glm.pred3,Use3test)



## LDA

library(MASS)

 lda.fit=lda(Use3~ PU2+BI2+Exp1+Exp2+Vis3+DOMAIN+ USERWIKI, data=wiki, subset=trainwiki)


lda.fit

lda.pred=predict(lda.fit, wikitest)

lda.class=lda.pred$class

table(lda.class, Use3test)


##QDA


qda.fit=lda(Use3~ PU2+BI2+Exp1+Exp2+Vis3+DOMAIN+ USERWIKI, data=wiki, subset=trainwiki)


qda.fit

qda.pred=predict(qda.fit, wikitest)

qda.class=qda.pred$class

table(qda.class, Use3test)
```

Additional Output

Logistic Regression 2

```
Call:
glm(formula = Use3 ~ PU2 + BI2 + Exp1 + Exp2 + DOMAIN + USERWIKI,
    family = binomial, data = wiki, subset = trainwiki)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4964  -0.4384   0.3012   0.6503   2.5970

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9587     0.5743  -6.893 5.45e-12 ***
PU21          1.4645     0.3368   4.348 1.38e-05 ***
BI21          2.1551     0.3136   6.873 6.29e-12 ***
Exp11         1.4375     0.3246   4.429 9.48e-06 ***
Exp21         1.0433     0.4462   2.338  0.01936 *
DOMAIN2       0.8292     0.6999   1.185  0.23610
DOMAIN3      -0.8825     0.5629  -1.568  0.11698
DOMAIN4      -0.6955     0.4732  -1.470  0.14168
DOMAIN5      -0.1938     0.5129  -0.378  0.70553
DOMAIN6      -0.8430     0.3891  -2.166  0.03028 *
USERWIKI1     1.6244     0.4959   3.276  0.00105 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 632.99  on 458  degrees of freedom
Residual deviance: 362.30  on 448  degrees of freedom
AIC: 384.3

Number of Fisher Scoring iterations: 5

         Use3test
glm.pred2   0   1
      No  163  36
      Yes  47 208
```

371/454= 81.72%