Multi-Institutional Study of Leadership

Done By:- Ankita Prasad, Purvee Agrawal, Rishika Multani

Introduction

In today's world higher education plays a significant role as it not only paves a way for our future but also lays foundation for a stronger nation. As the importance of education has been rising, many scientists have started to apply data analysis techniques in order to analyse the educational data. The concept of data analysis is basically combining all the information which helps to observe the trends and patterns that can be used to evaluate certain processes, create efficiencies, and also helps in the improvement of a particular process. Institutions have a great amount of data pertaining to the students grades, various skills and faculty information. With the help of this educational data many conclusions can be drawn that can not only improve the operational effectiveness of the educational institutions but also benefit students in their personal development.

By imbibing data analysis techniques, the following benefits could be seen in the education industry:-

a) Performance Evaluation:- At present the only method that is used by most of the institutions to evaluate a student's performance is by how the student has answered his/her exam or assignments. However, this method is not enough to analyse the student's performance as they might not excel academically but might do so in other extra-curricular activities. By incorporating big data techniques it is possible to monitor student actions, such as how long they take to answer a question, which sources they use for exam preparation, which questions do they skip, what all other extra-curricular activities are they involved in. Therefore this method of evaluation will help schools/universities to understand every student better and give every student an immediate feedback on their overall performance (academic & non-academic).

b) Designing Programs/Courses:- By carrying out a short survey related to student interest at the start of an academic year will help the institutions understand every student better. They will be able to analyse the weak and strong areas and by that data design/structure the courses where some of the courses could be hybrid in nature which is a combination of offline and online learning. This would give students the opportunity to take up classes that they are interested in, keep up with the class at their own pace, while still having access for guidance/problem solving with the professor remotely or in person. This will allow a student to excel in the area of their interest and work towards their career.

c) Skill Analysis:- Today many universities while applying for undergraduate/masters degree not only look at your academic scores but also the various skills that an individual has acquired over the years. Universities use this data to analyse the incoming students and work on creating programs that will improve various skills such as leadership, teamwork, cognitive

and technical etc. By working on these skills not only will students become more confident but will also impact the society.

Background/Motivation

The project that we decided to work on is related to a survey which is conducted by Rutgers' Student Affair (RSA). The program is known as Multi-Institutional Study of Leadership (MSL) which is an international research program that analyses the impact of higher education on leadership skills on college students. The survey of this study is also performed to examine influence of different college experiences on leadership-related outcomes such as complex cognitive skills, social perspective-taking, leadership efficacy. The survey incorporates questions to capture self-evaluation of the individuals on these skills prior to college, activities they engage in college, their demographics, their environment and evaluation on these skills after attending college. Some other skills that are studies through the survey are complex cognitive skills, leadership efficacy, social change behaviours, seeing alternative social perspectives, spiritual development, racial identity, resiliency, and agency which play an important role in leadership skill development. Therefore, this study is aimed to understand how the institute can make efforts (by arranging workshops, introducing courses etc.) in helping build such skills in the students.

The dataset that was provided to us had attributes=700, respondents=10,300 and students who completed the survey= approx 27%. To understand the needs of RSA better, we had three to four remote conversations with them. Post the conversation we decided to analyse the data and answer the following questions:-

a) To check whether an aspect like demographics (college year, gender, race) plays any role in the students' leadership or cognitive skills.
b) The various factors that will help in improving the students' leadership development be it by conducting various workshops.
c) Observe leadership skills pre-college and post-college and predict whether higher education impacts the students' leadership skills.

In order to answer all these questions, the first thing that we did was read all the MSL documents that were provided to us and then went on to understand the dataset. Post reading we saw that it was a survey that was conducted for all of the years of college students with a list of demographics, cognitive and psychometric question sets. The data that they had provided to us was in the form of :- single response, multiple response, numerical response, text response, memo response. We observed that the dataset had a large amount of null values which was there because many students had skipped certain questions in the survey or had left the survey half-way through. Therefore this project had a major chunk of data cleaning as we had to identify a technique as to how to handle the null, #null values as well as the missing values. Further, post cleaning of the data the algorithms that we

decided to apply on the data in order to get the answers to the desired questions are clustering, performing t-test and finally carrying out a comparison between supervised and unsupervised learning.

Literature review

In order to understand how to deal with survey data and to learn about the techniques that are being used on educational data and apply it to our project. We came across three papers that helped us understand educational data mining techniques and the importance/benefits of higher education to an individual as well as to the society. The papers are:-

A Survey and Future Vision of Data mining in Educational Field (Barahate Sachin R., Shelake Vijay M. 2012)

As the importance of higher education is rising many data scientists have been curious to find its benefits and outcomes and have therefore started to apply various data mining techniques to make conclusions and to see how higher education is impacting an individual as well as the society. The main objective behind applying these techniques in educational data is to discover and extract certain unique patterns in the huge data. The term that is used for handling educational data it is known as Educational data mining, abbreviated as EDM and at present it is a hot topic of research among data scientists to decipher the data that will not only help in predicting student performance but also the benefits/outcomes of education to society.
The main points that caught our attention which we decided to imbibe in our project are:-

1. Clustering:- It is basically the process where similar objects are grouped together into classes. Students with similar grades can be grouped together in order to determine the average or above average students. It can also be applied to find out the common aspects and the different aspects between schools.
2. Relationship Mining:- It basically refers to the relationship between variables in a data set. The main types that are discussed in this paper are association rule, correlation and sequential pattern mining. The main one being association rule mining in order to observe student learning disabilities and other learning patterns.

For our project we incorporated these two techniques. After reading the clustering algorithm in the paper we decided to use this technique to combine students who had similar traits related to grades, skills, class year (freshman, sophomore, junior, senior and graduate students). It also made it easier for us to analyse the number of male,females, transgenders and who claimed themselves as “unidentified”.In case of relationship mining it was helpful as it was able to answer one of our questions whether higher education impacts students’ leadership skills. Therefore these technique helped us visualize the relationship of pre-college and post-college students’ leadership skills.

Higher Education Plays Critical Role in Society: More Women Leaders Can Make a Difference (Leah Jackson Teague, 2015)

This paper states that higher education plays a prominent role in shaping the future of our society. It is important to be able to competing in this competitive world. In every field like social, economic or cultural, the leadership aspects are important in order to be on the top of everything and face the difficulties and challenges that come up. And to make women participate equally in everything and be able to face the challenges, there is a need of more women leaders. Women are as innovative, productive, as men but barriers still exist to their advancement. They too are well-qualified, the need is to make them prepared leaders.

From this paper it has come out that a disproportionately low percentage of students served by those colleges and universities are women. Educating citizens who will be more engaged in their communities through civic activities and public discourse and in developing leaders who will contribute to the advancement of business, organizations and society should be given more emphasis.

Form our project as well we are planning to see if gender of a student plays any role while determining leadership skills. This will be one of the parts in our analysis and we took this paper as reference.

Higher education as a change agent for sustainability in different cultures and contexts (Jennie C. Stephens, Maria E. Hernandez, Argentina Mikael, Amanda C. Graham, Roland W. Scholz,2008)

This paper emphasizes the need of higher education, in different cultures and contexts, to be change agents for sustainability. Since it is common that our society faces serious issues and challenges associated with environmental change, resource scarcity, increasing inequality and injustice, as well as rapid technological change and will make students ready to cope up with these.It also identifies the emerging critical issues like financing structure, institutional organization, the democratic processes, and communication and interaction with society. It puts more emphasis on how higher studies help an individual understand challenges associated with our interactions with the earth's natural systems. It is basically more about human-environment interaction and how it is affected by unequal, complex and interconnected societal structure and rapid technological change.

This paper presents a different perspective of higher education and it can certainly in some way promote leadership in an individual. From this paper we can also see and analyze if the student majors that are related to the environmental science affects their leadership skills. This is just a short example, we can take many more things into consideration like this and do our analysis.

Approach

The data received by us was a survey data with questions capturing demographic answers, score on various skills such as - leadership efficacy, cognitive skills, social perspective taking etc., environment related questions. Through analysing this survey we aim to provide insights to RSA stating the patterns in students who are more confident in leadership skills and whose cognitive skills have improved and compare it with students' who still lack in it, so that they can form initiatives such as - workshops, on-campus groups, classes to help them boost their skills.

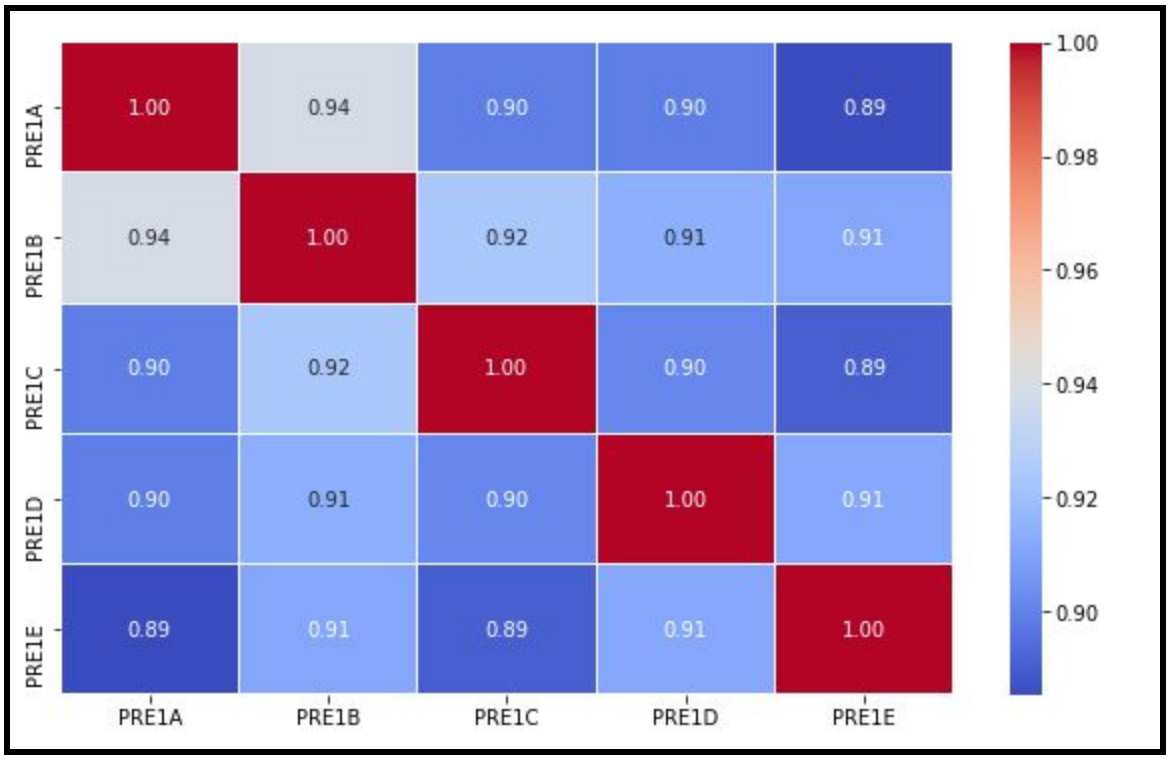
Cleansing of the data and Algorithms-

(i) Using Python and IDE Jupyter Notebook we imported the Rugers Only data.
(ii) For our analysis we have used demographic variables - DEM3, DEM7 and DEM10; Environmental Variable - ENV 4a,b,c,d,e,f,g, ENV 7a, 7f, g,b,c,q,p,e,n,d,d1,d2,d3,7h,7i,7j,7k,7l,7m,7o,7r,7u,7v,7w, ENV12,
Cognitive Skill Pretest - PRE1a,b,c,d,e - PRECOG(Mean)
Cognitive Skill Test Out - OUT 1a,b,c,d - OUTCOG(Mean)
Leadership Efficacy Pretest - PRE2a,b,c,d - PREEFF(Mean)
College - OUT2a,b,c,d-OUTEFF(Mean)
OMNIBUS -> Mean(SRLS) (y)
We have dropped all other columns other that these using python script and created a new data set.
(iii) This new created data set is our final data for analysis.
(iv) In case of all the numeric attributes we have replaced all the "#Null!" attributes with "-1" and in case of all the categorical attributes we have replaced "#Null!" with "NA".

```
for column in NumericAttributes:
  #Replacing Null Values
    NumericAttributes[column].replace(['#NULL!'], ['-1'], inplace=True)
    NumericAttributes[column]=pd.to_numeric(NumericAttributes[column])
```

```
for column in CategoricalAttributes:
  #Replacing Null with NA
    CategoricalAttributes[column].replace(['#NULL!'], ['NA'], inplace=True)
    CategoricalAttributes[column].replace(r'\s+','NA',regex=True)
```

(v)Since we have 5 attributes resulting in leadership efficacy skills and cognitive skills we thought of using heat maps to analyse the correlation between the variables, in case the variables are highly correlated we decided to carry our analysis using the mean values of the attributes.



(vi) To perform preliminary analysis and plotting we have used manipulation techniques such as dummy variables for our categorical variables such that the values are in form of 0 or 1.

```
dummies1=pd.get_dummies(d_RSA_MSL['DEM3'])
dummies1=dummies1.add_prefix("{}#".format('DEM3'))
```

(vii) Used Pivot table to create the mean of numerical different attributes corresponding to different categorical variables.

```
#Comparing Leadership Efficacy before and after college by Class Level

pd.pivot_table(d_RSA_MSL,index=["DEM3"],values=["PREEFF","OUTEFF"],aggfunc=np.mean)
```

Out[16]:

DEM3	OUTCOG	PRECOG
1	1.869594	2.550590
2	1.958120	2.603761
3	2.159450	2.636067
4	2.256846	2.659084
5	1.500000	2.446154
6	1.643939	2.424242
NA	-0.995323	-0.995323

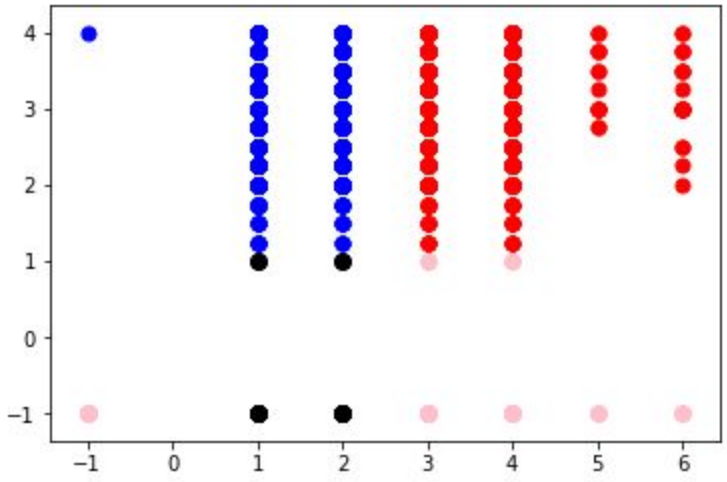
(viii) We then implemented linear regression using single variables such as DEM3 and OMNIBUS to see how can they predict OMNIBUS, but we expected the R squared value not to be very high.

```
#Linear regression with only DEM3 which is class level

from statsmodels.formula.api import ols


model_Dem3 = ols("OMNIBUS ~ data_RSA_MSL['DEM3#1']+data_RSA_MSL['DEM3#2']+data_RSA_MSL['DEM3#3']+data_RSA_MSL['DEM3#4']+ data_RSA

print(model_Dem3.summary())
```

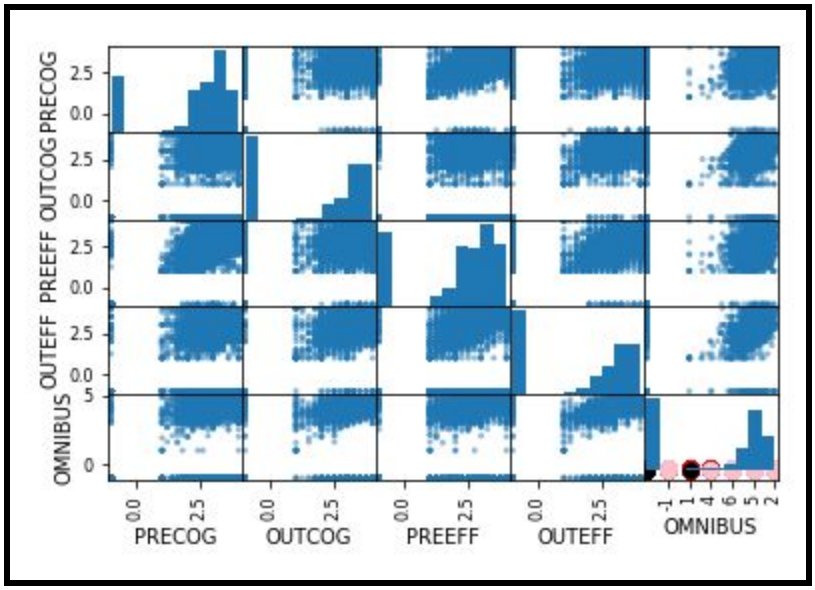
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                OMNIBUS   R-squared:                       0.243
Model:                            OLS   Adj. R-squared:                  0.242
Method:                 Least Squares   F-statistic:                     441.7
Date:                Thu, 13 Dec 2018   Prob (F-statistic):               0.00
Time:                        05:19:27   Log-Likelihood:                -18173.
No. Observations:                8271   AIC:                         3.636e+04
Df Residuals:                    8264   BIC:                         3.641e+04
Df Model:                           6
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    1.7227      0.073     23.583      0.000       1.580       1.866
data_RSA_MSL['DEM3#1']       0.8318      0.088      9.498      0.000       0.660       1.003
data_RSA_MSL['DEM3#2']       0.8204      0.086      9.560      0.000       0.652       0.989
data_RSA_MSL['DEM3#3']       0.9969      0.085     11.791      0.000       0.831       1.163
data_RSA_MSL['DEM3#4']       1.0999      0.085     12.953      0.000       0.933       1.266
data_RSA_MSL['DEM3#5']       0.1927      0.377      0.511      0.609      -0.547       0.932
data_RSA_MSL['DEM3#6']       0.4985      0.336      1.482      0.138      -0.161       1.158
data_RSA_MSL['DEM3#NA']     -2.7174      0.093    -29.191      0.000      -2.900      -2.535
==============================================================================
Omnibus:                     1661.319   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1408.377
Skew:                          -0.922   Prob(JB):                    1.50e-306
Kurtosis:                       2.173   Cond. No.                     1.77e+15
==============================================================================
```

(ix) So, we performed linear regression by grouping few variables together and then checked how the value of R square changes.
(x) We have then performed k means clustering using various categorical variables and the leadership efficacy skills and cognitive skills.

```
#Gender and Post college Cognitive skills
x=CategoricalAttributes['DEM7']
y=NumericAttributes['OUTCOG']
finalDf = pd.concat([x,y], axis = 1)
from sklearn.cluster import KMeans
cluster=KMeans(n_clusters=5)
finalDf['cluster']=cluster.fit_predict(finalDf)
plt.scatter(finalDf[finalDf.cluster==0]["DEM7"], finalDf[finalDf.cluster==0]["OUTCOG"], s=50, c='red')
plt.scatter(finalDf[finalDf.cluster==1]["DEM7"], finalDf[finalDf.cluster==1]["OUTCOG"], s=50, c='black')
plt.scatter(finalDf[finalDf.cluster==2]["DEM7"], finalDf[finalDf.cluster==2]["OUTCOG"], s=50, c='blue')
plt.scatter(finalDf[finalDf.cluster==3]["DEM7"], finalDf[finalDf.cluster==3]["OUTCOG"], s=50, c='pink')
plt.scatter(finalDf[finalDf.cluster==4]["DEM7"], finalDf[finalDf.cluster==4]["OUTCOG"], s=50, c='pink')
```

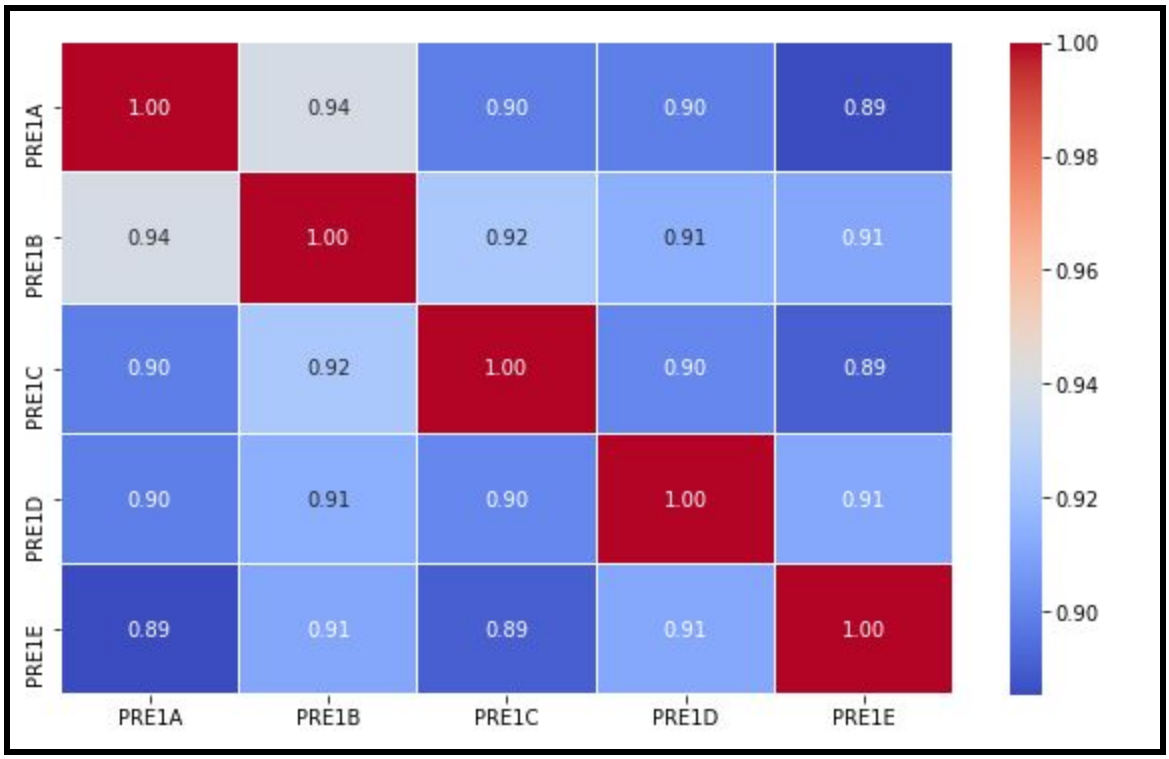


(xi) Next, we have performed hierarchical clustering and compared the results of the two clustering.
(xii) Using Tableau, we have created graphs for visual comparison between various groups, which will provide insights to the RSA about the groups they need to focus on.
(xiii) We have performed pair t-tests to verify the correlation between the two variables.
(xiv) We have created scatterplots to check correlation between attributes.



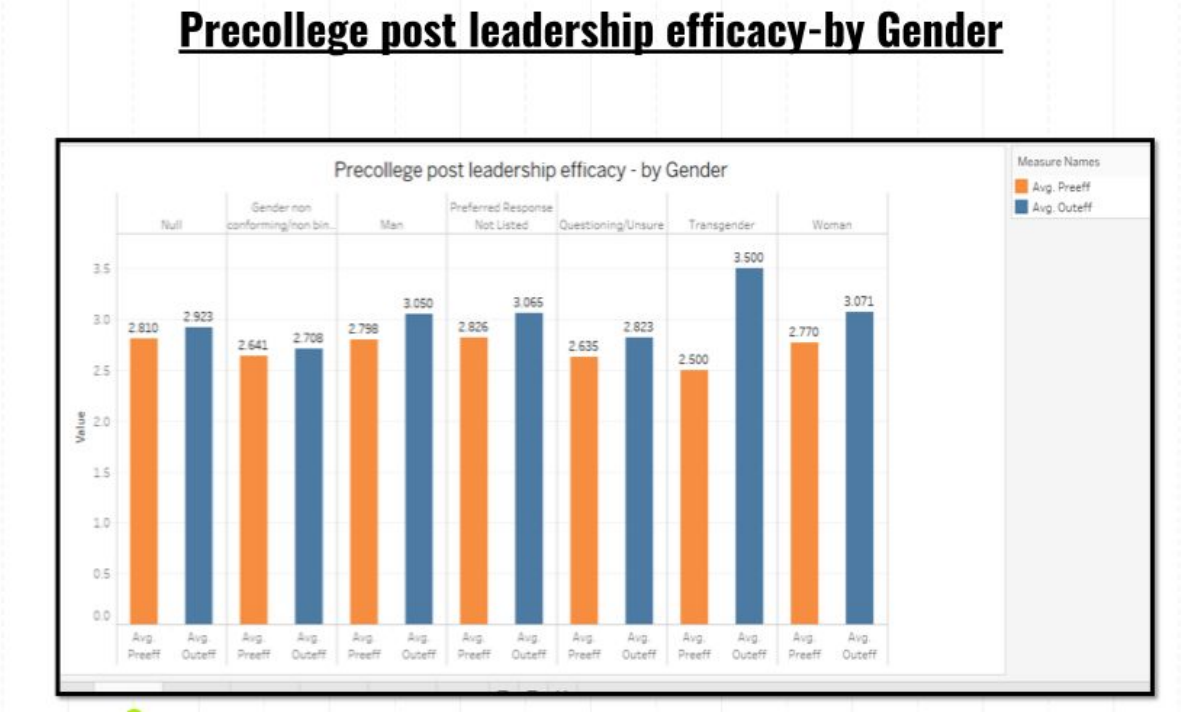
Results

The heatmap plot between various Cognitive Skills and Leadership Efficacy showed us strong correlations between the the variables. Hence, we decided to drop the variables and use the mean of the score.



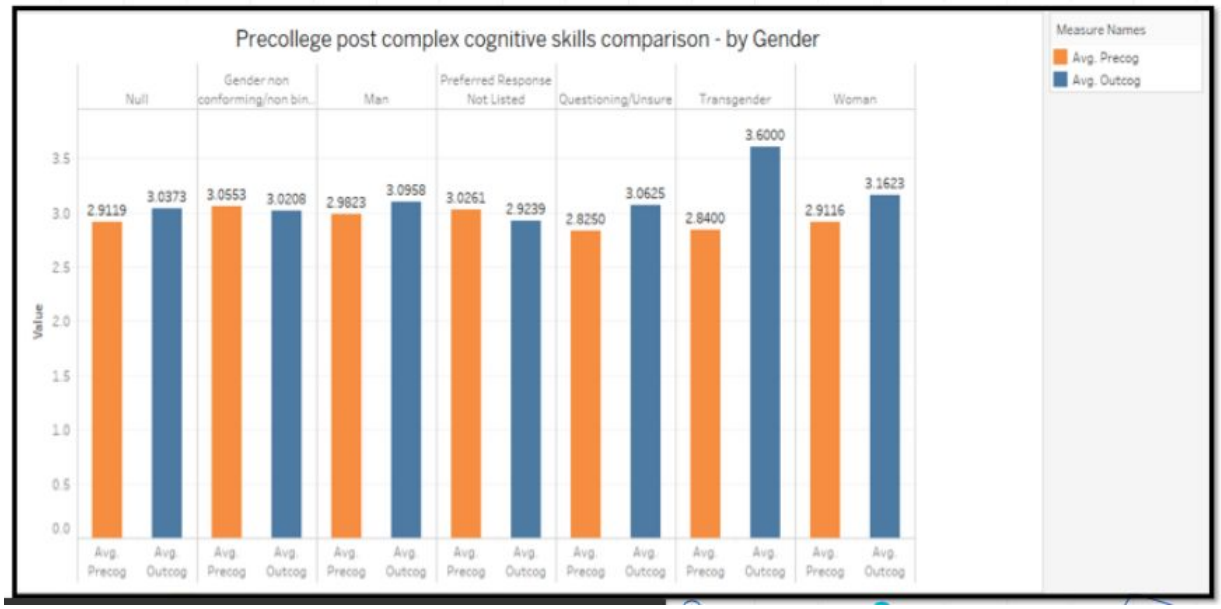
For example, precollege cognitive skills variable PRE1A,1B,1C,1D,1E are all strongly correlated as we can see so we use PRECOG mean of the values instead.
We created graphs using Tableau for ease of comparing leadership efficacy and cognitive skills between various demographic groups, tabular output of which we received using our python code. And we got some amazing insights,

When a graph between precollege and post college leadership efficacy was created against Gender, we found that (i)the highest increase in leadership efficacy was in people who identified themselves as “Transgender”.
(ii) The increase in “men” is more as compared to “women”.



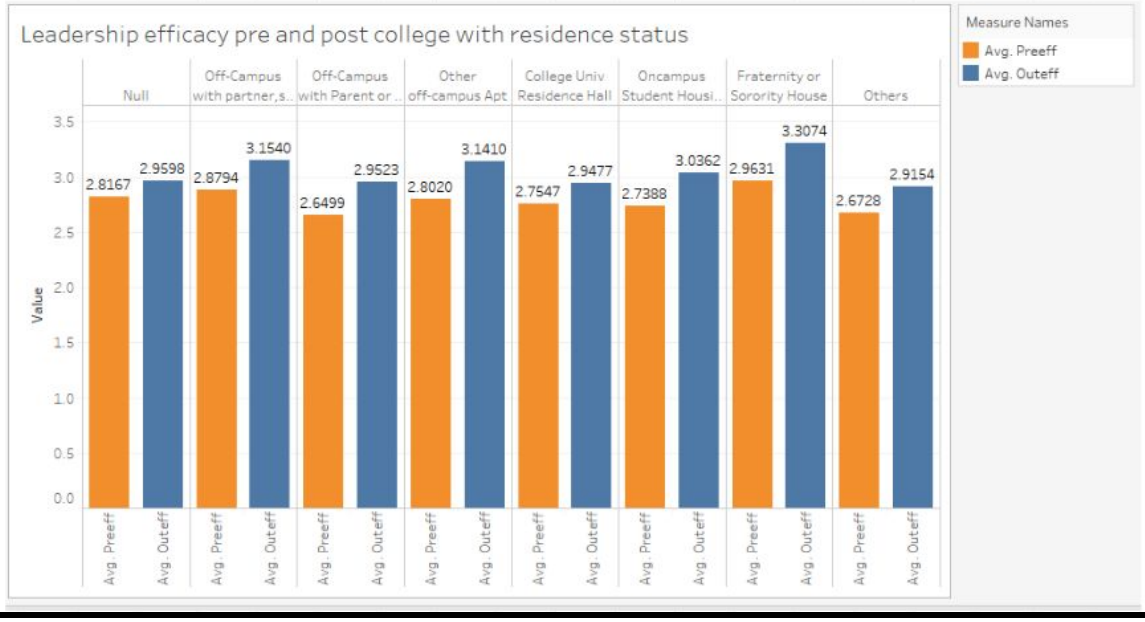
When we plotted the graph between precollege and post college cognitive skills and gender we got the following insights - (i) there was a decrease in cognitive skills score of people who identified themselves as - “Gender non conforming” and “preferred response not listed” (ii) Yet again, the increase of the score in the people who identified themselves as “transgender” was the highest. (iii) Increase of score is more in “Women” as compared to “Men”

Precollege post complex cognitive skills comparison-by Gender

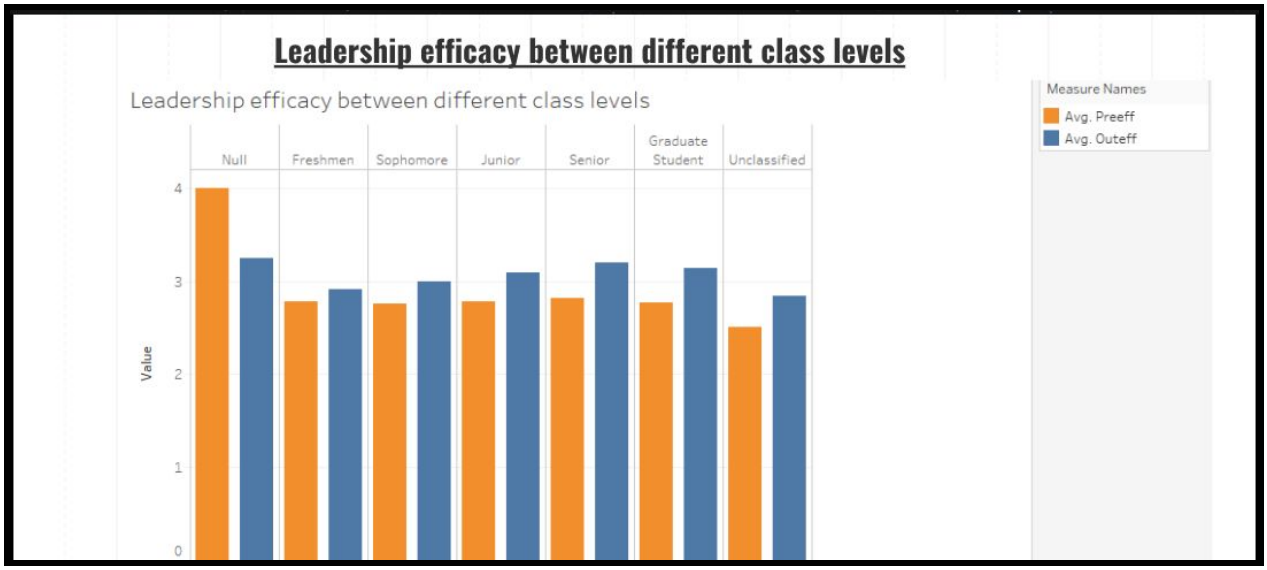


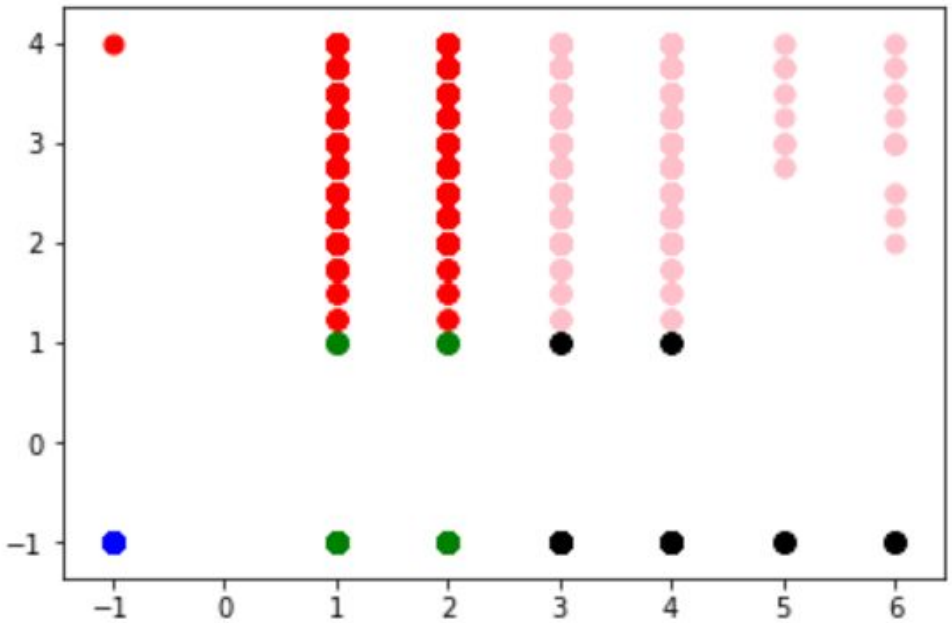
When a graph was plotted between leadership efficacy score and residence status we saw (i) increase in all groups (ii) But it was very interesting to note that the highest increase was in people staying at the Fraternity or Sorority house (iii) Lowest increase was in people who stayed at College University residence hall. This could be due to the fact that students residing at a Fraternity or Sorority house generally manage all things at their end and this could help in building their leadership skills.

Leadership efficacy between different residence status



The next graphical representation is between the class level and leadership efficacy. The insights which we got from this - (i) There has been constant increase in the leadership efficacy with increase in class year (ii) However, it is very interesting to note that the there is least increase in graduate students and unclassified(may be part-time students). This could be due to the fact that not enough programs are available on campus for them to get involved in.

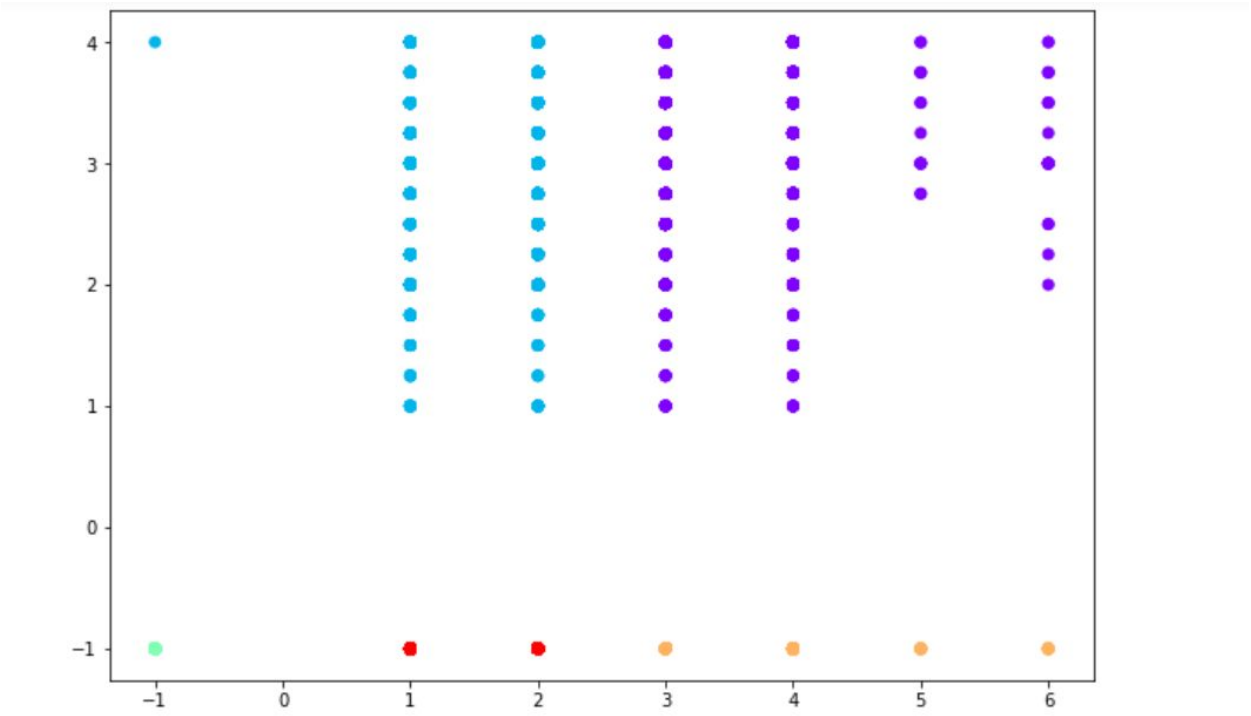




The above graph shows the clustering between DEM3 (in X-Axis)(the current class level of the students) and OUTCOG(in Y-Axis)(Complex cognitive skills. DEM3 value ranges from 1-6, with 1 aas freshman, 2 as sophomore and so on and OUTCOG value ranges from 1-4 where 1 is “not grown at all”, 2 is “grown somewhat” and so on. For both of these attributes, the #NULL! Values have been converted to -1.
So from above graph, we have 5 clusters in 5 different colors. So we can see that students who has DEM3 values 1 and 2 and OUTCOG value 1 fall into same cluster(green coloured). That means, that Sophomore and Junior class level students fall into same cluster and they have the OUTCOG value 1, i.e they have rated their complex cognitive skills as 1(not grown at all). And in this way we can analyze for all the clusters as to what class level of students with what outcog values, fall into the same cluster and we can use this to see if can be done something for the students falling into the cluster with low outcog values.

Cluster	ENV12	DEM3	OUTCOG
0	4.153370	1.744589	2.986549
1	-0.998003	-0.422104	-0.891312
2	3.502630	2.669058	3.049737
3	-0.996678	3.030565	-0.753654
4	2.778278	3.503213	3.284222

The above is the result of clustering done between the attributes ENV12(where are the students currently living while attending the college), DEM3 and OUTCOG. We have again 5 clusters, 0-4. And the mean value of each of ENV12, DEM3 and OUTCOG is calculated. So from the above table we can interpret that students with mean ENV12 value, DEM3 value and OUTCOG value equal to 4.15,1.74 and 2.9 respectively fall into same cluster(cluster 0). And it can be seen that students with highest OUTCOG values are the one whose mean values of DEM3 and ENV12 fall in the range of almost 3-4. That means that students who are living in Other off-campus home, apartment, or room or in College/university residence hall and are Junior or senior have rated their OUTCOG skills as highest.



And again is the graph of clustering between DEM3 and OUTCOG again but this is done using Hierarchical Clustering. We tried hierarchical clustering to see if we are getting any different clustering patterns as compared to K-Means, but realised that the result was no different from K-Means and in fact, Hierarchical clustering consumed comparatively much time as compared to K-Means. We tried Hierarchical clustering for 2 instances and got almost same results as K-Means, so we decided to go ahead with K-Means only as it is more efficient.

Statistical results Discussion and conclusion

Key Results

(i) From the analysis of the survey using the leadership efficacy and cognitive skills we can see that there is a general increase in the score from before college to after college.
(ii) It is very interesting to note that demographic variables such as - Gender, class level and ethnicity do play a role in the score and increase in the score is not even in all the sections.

(iii) When we tried linear regression using one demographic variable and leader efficacy, the R squared value as we expected was less.
(iv) This value increases, when we include more variables - DEM3, DEM7, ENV7, ENV12. The R square rises to 77% from 34%.
(v) It is however, interesting to note that when we used two dependent variables to calculate independent variable it also gave similar value of R square.
(vi) The environment such as where the student stays during education, which all group they were involved in during their college plays a vital role in building the skills.
(vii) RSA can introduce some workshops for students residing in the College residence halls to help them get the same level of exposure as students staying at fraternity house or sorority house.
(viii)From clustering we got that students who are freshman have low complex cognitive skills as compared to students of other class levels and in this way for other demographic and environmental values, the cognitive skills score of the students was found out.
ix) Also from clustering based on both Environmental and Demographic values against the Cognitive skill scores, it was found out that the students who are seniors and live in College/university residence hall have the highest cognitive skills score.

Future work

For future work we would like to -

(i) Since this analysis has been done using Rutgers data only, we would like to take the data from Big10 and then compare results of Rutgers with other Big10 institutes.
(ii) We have performed our analysis using using three key demographic variables - we would like to incorporate more variables in future.
(iii) We have performed our analysis using using three key environmental variables - we would like to incorporate more variables in future.
(iv) We have take only two key skills - Leadership Efficacy and Cognitive skills. In future we would like to use more skills as independent variables.
(v) We would like to try other clustering techniques to compare our results.

References

“Data used in this article were collected as part of the Multi-Institutional Study of Leadership 2018. For further information regarding that study, please visit www.leadershipstudy.net."

Dayal, P. (n.d.). 5 Dramatic Impacts of Big Data on Education. Retrieved from https://www.newgenapps.com/blog/5-dramatic-impacts-of-big-data-on-education

Importance of Using Data and Analysis in Higher Education. (2018, February 26). Retrieved from https://insidebigdata.com/2018/02/25/importance-using-data-analysis-higher-education/

Joshinav. (2017, August 28). 4 ways big data is transforming the education sector. Retrieved from https://www.allerin.com/blog/4-ways-big-data-is-transforming-the-education-sector

(n.d.). Retrieved from https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

Ramshaw, A. (2018, August 31). How to Analyze Survey Data in Excel. Retrieved from https://www.genroe.com/blog/analyze-survey-data-in-excel/11483

Sachin, R. B., & Vijay, M. S. (2012, January). A survey and future vision of data mining in educational field. In Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on (pp. 96-100). IEEE.

Stephens, J. C., Hernandez, M. E., Román, M., Graham, A. C., & Scholz, R. W. (2008). Higher education as a change agent for sustainability in different cultures and contexts. International Journal of Sustainability in Higher Education, 9(3), 317-338.

Teague, L. J. (2015). Higher Education Plays Critical Role in Society: More Women Leaders Can Make a Difference. In Forum on Public Policy Online (Vol. 2015, No. 2). Oxford Round Table. 406 West Florida Avenue, Urbana, IL 61801.

http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schulte/theses/phd/algorithm.pdf
Trevino,A.(n.d.).Introduction to K-means Clustering. Retrieved from https://www.datascience.com/blog/k-means-clustering

Zapier. (n.d.). How to Design and Analyze a Survey. Retrieved from https://zapier.com/learn/forms-surveys/design-analyze-survey/#analyze

Multi-Institutional Study of Leadership

Group Members:- ANKITA PRASAD, PURVEE AGRAWAL, RISHIKA MULTANI

Outline

1. Introduction
2. Approach
3. Evaluation Methodology
4. Results
5. Conclusion & Future Work

1. Introduction

- The project is based on a research known as Multi-Institutional Study of Leadership (MSL) which is an international research program to examine the influence of higher education on the leadership skills of the college going students.

- Dataset:-**700 Attributes** and **10300 respondents** and approx **27% of students completed survey.**

- And out of these we extracted around **80 attributes** categorical and numerical attributes for our analysis.

Aim for Data Analysis

- Survey incorporated questions to assess many **skills** of the students **prior-college** and **post-college.**
  a) Analyse and **compare** the **precollege cognitive skills and leadership efficacy skills** with **post college cognitive skills and leadership efficacy skills.**
  b) If **demographics** such as - gender, class year of the student, ethnicity has any effect on the skills.
- If environment attributes such as - being a part of sorority,club participation, extracurricular, residency status during college have any impact.
- If using the attributes can we calculate OMNIBUS(overall measure of leadership skills).
- Performing K mean clustering on individual attributes and then together to see the clusters formed and then deciphering common patterns in the clusters.

2. Approach

- Data is a set of survey questions, due to which the cleansing of the data was a major part this also included deciding which attributes to use for analysis.

- In the **numerical columns**, using python we have replaced **“#NULL!”** values with **-1** and in **categorical columns** we have replaced **null values** with **“NA”**.

- Since we had multiple attributes to measure cognitive skills and leadership efficacy, at preliminary stage, **we wanted to perform correlation test between them to** see if they are strongly correlated.

- Using **heat maps we checked correlation** between various attributes of “Leadership Efficacy skills” and “cognitive skills” in order to find correlated attributes and remove the correlated attributes.

2.Approach

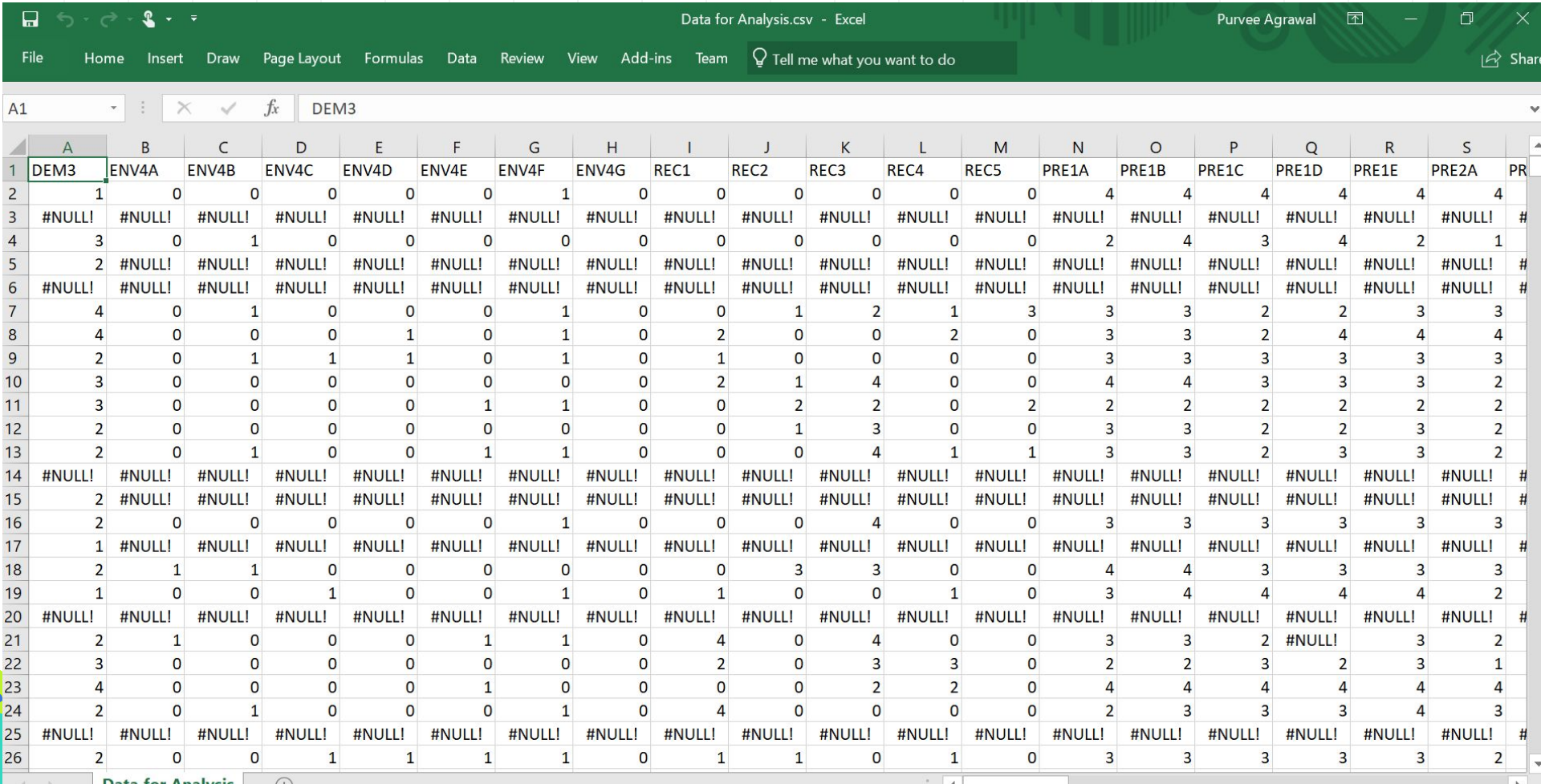
- At the preliminary stage we created **pivot tables** to see the **aggregate values** of some of the **attributes grouped by their types - demographic, environment.**
- We created **dummy variables** for **demographic variables**.
- Implemented **linear regression** between **single variables** and **OMNIBUS.**
- Performed **multiple regression** between combination of **demographic variables, environment variables and OMNIBUS.**
- We captured how the values of **R square changed** from first scenario to second.

2. Approach

- We performed **unsupervised learning** on the data, to analyse **main clusters** in our data based on demographic and environment variables and how good they are in “leadership skills” and “cognitive skills”.

- We implemented **K-Means** and then tried **hierarchical clustering** but we found out that it’s didn’t give any different clustering results for our dataset. Therefore choose K-Means only.

A SNAPSHOT OF THE DATASET THAT WE HAVE TAKEN:



Data for Analysis.csv - Excel

Purvee Agrawal

File Home Insert Draw Page Layout Formulas Data Review View Add-ins Team Tell me what you want to do

A1 DEM3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	DEM3	ENV4A	ENV4B	ENV4C	ENV4D	ENV4E	ENV4F	ENV4G	REC1	REC2	REC3	REC4	REC5	PRE1A	PRE1B	PRE1C	PRE1D	PRE1E	PRE2A
2	1	0	0	0	0	0	1	0	0	0	0	0	0	4	4	4	4	4	4
3	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
4	3	0	1	0	0	0	0	0	0	0	0	0	0	2	4	3	4	2	1
5	2	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
6	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
7	4	0	1	0	0	0	1	0	0	1	2	1	3	3	3	2	2	3	3
8	4	0	0	0	1	0	1	0	2	0	0	2	0	3	3	2	4	4	4
9	2	0	1	1	1	0	1	0	1	0	0	0	0	3	3	3	3	3	3
10	3	0	0	0	0	0	0	0	2	1	4	0	0	4	4	3	3	3	2
11	3	0	0	0	0	1	1	0	0	2	2	0	2	2	2	2	2	2	2
12	2	0	0	0	0	0	0	0	0	1	3	0	0	3	3	2	2	3	2
13	2	0	1	0	0	1	1	0	0	0	4	1	1	3	3	2	3	3	2
14	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
15	2	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
16	2	0	0	0	0	0	1	0	0	0	4	0	0	3	3	3	3	3	3
17	1	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
18	2	1	1	0	0	0	0	0	0	3	3	0	0	4	4	3	3	3	3
19	1	0	0	1	0	0	1	0	1	0	0	1	0	3	4	4	4	4	2
20	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
21	2	1	0	0	0	1	1	0	4	0	4	0	0	3	3	2	#NULL!	3	2
22	3	0	0	0	0	0	0	0	2	0	3	3	0	2	2	3	2	3	1
23	4	0	0	0	0	1	0	0	0	0	2	2	0	4	4	4	4	4	4
24	2	0	1	0	0	0	1	0	4	0	0	0	0	2	3	3	3	4	3
25	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
26	2	0	0	1	1	1	1	0	1	1	0	1	0	3	3	3	3	3	2

Data for Analysis

3. Evaluation Methodology

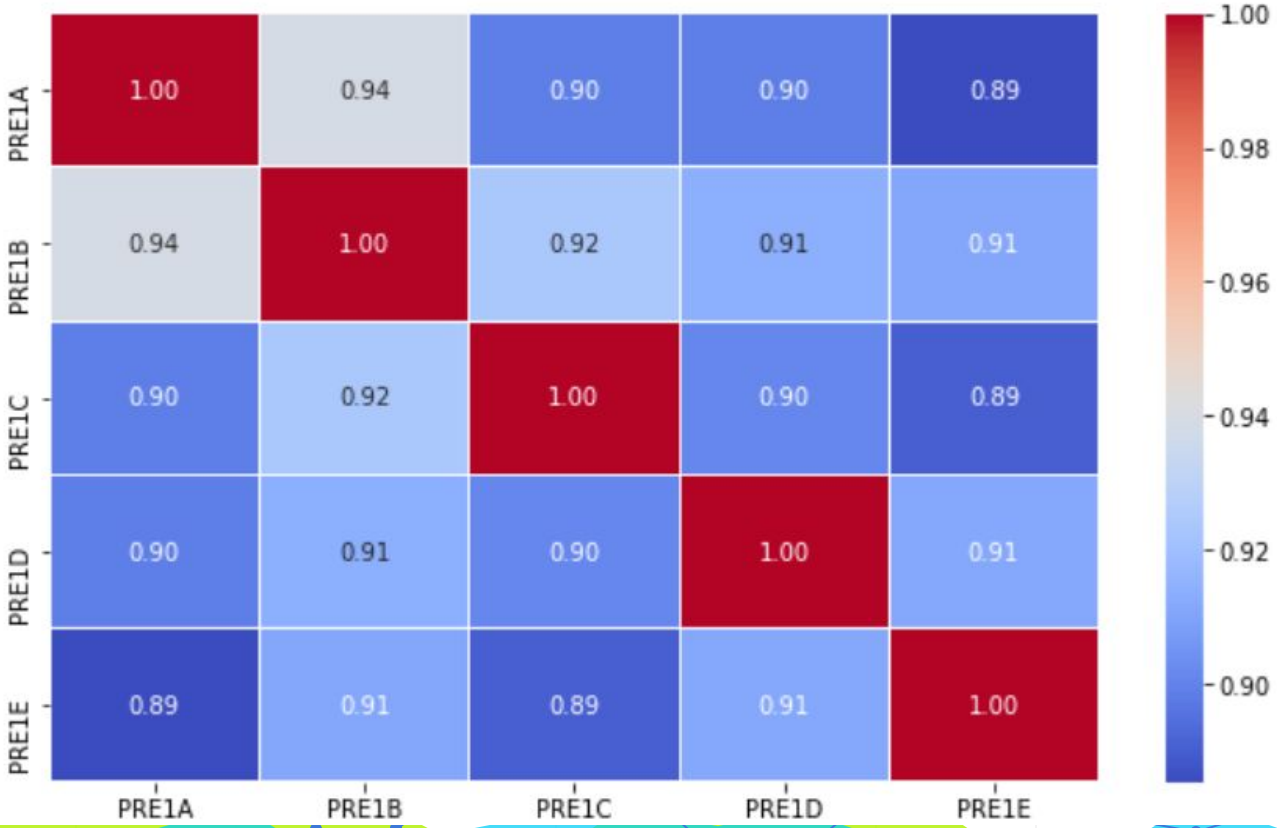
- We implemented **t-tests** on dependent variable OMNIBUS and independent variables - DEM3, DEM7, DEM10A, ENV13 and kept 90% confidence interval as the attributes are sociological .

- **Scatter plots** to see the **correlation between attributes**.

- **Linear regression** between **omnibus** and **env/dem** attributes to check the values like **R square.**

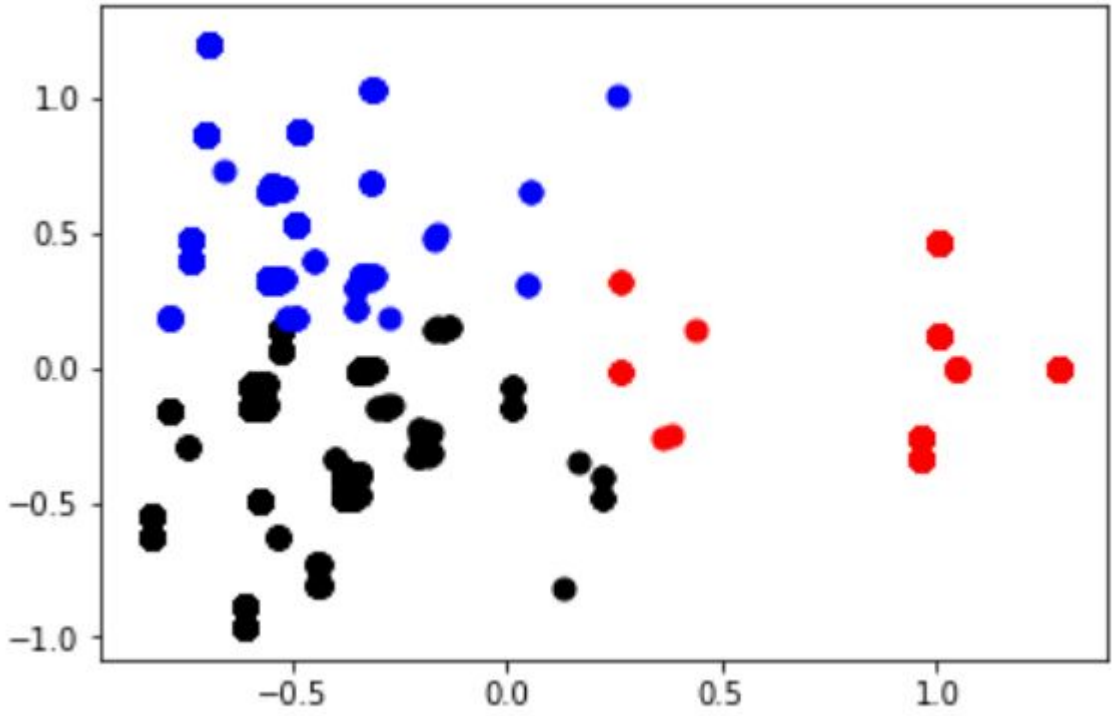
- **K-Means** to see which pair of attribute values fall into same clusters to **find common patterns** between them.

4. Results

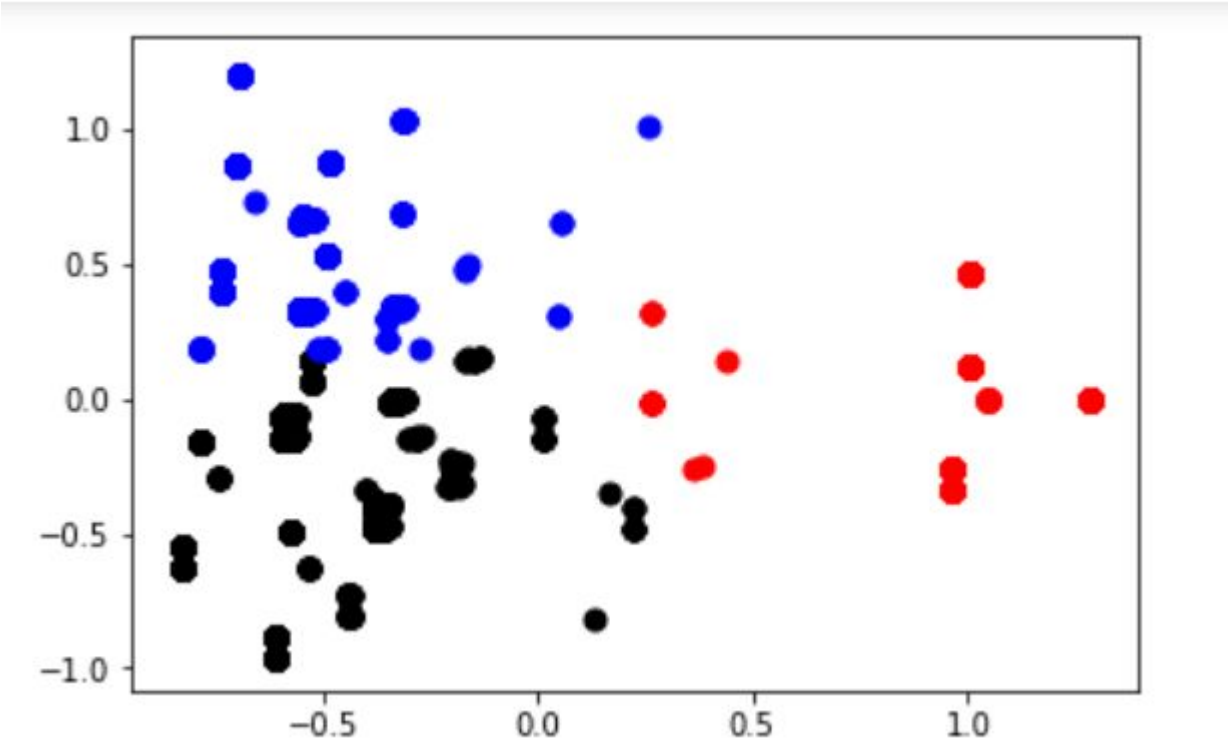


- We created a heatmap for all the Precognitive determining attributes and as we can see they are highly correlated with each other. Hence, we use the mean of these attributes which is column 'PRECOG' in our survey data.

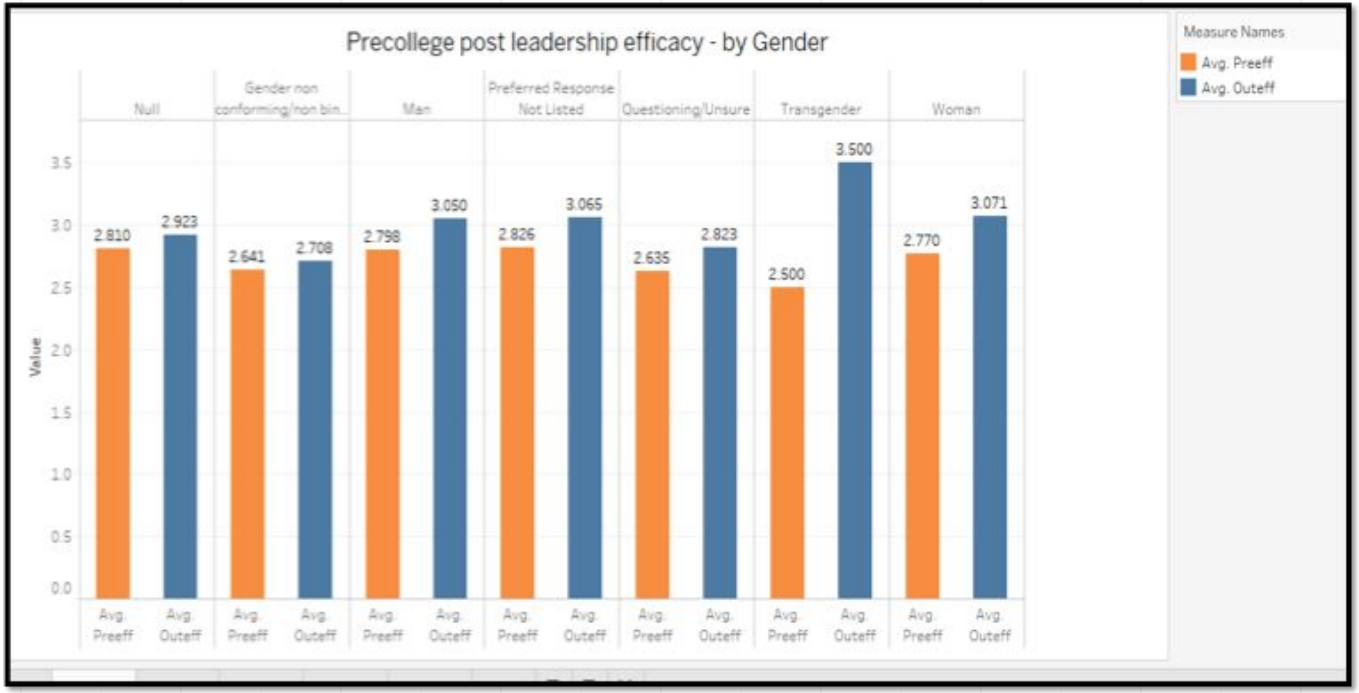
Result of clustering between #Demographic and Pre College Cognitive Skills



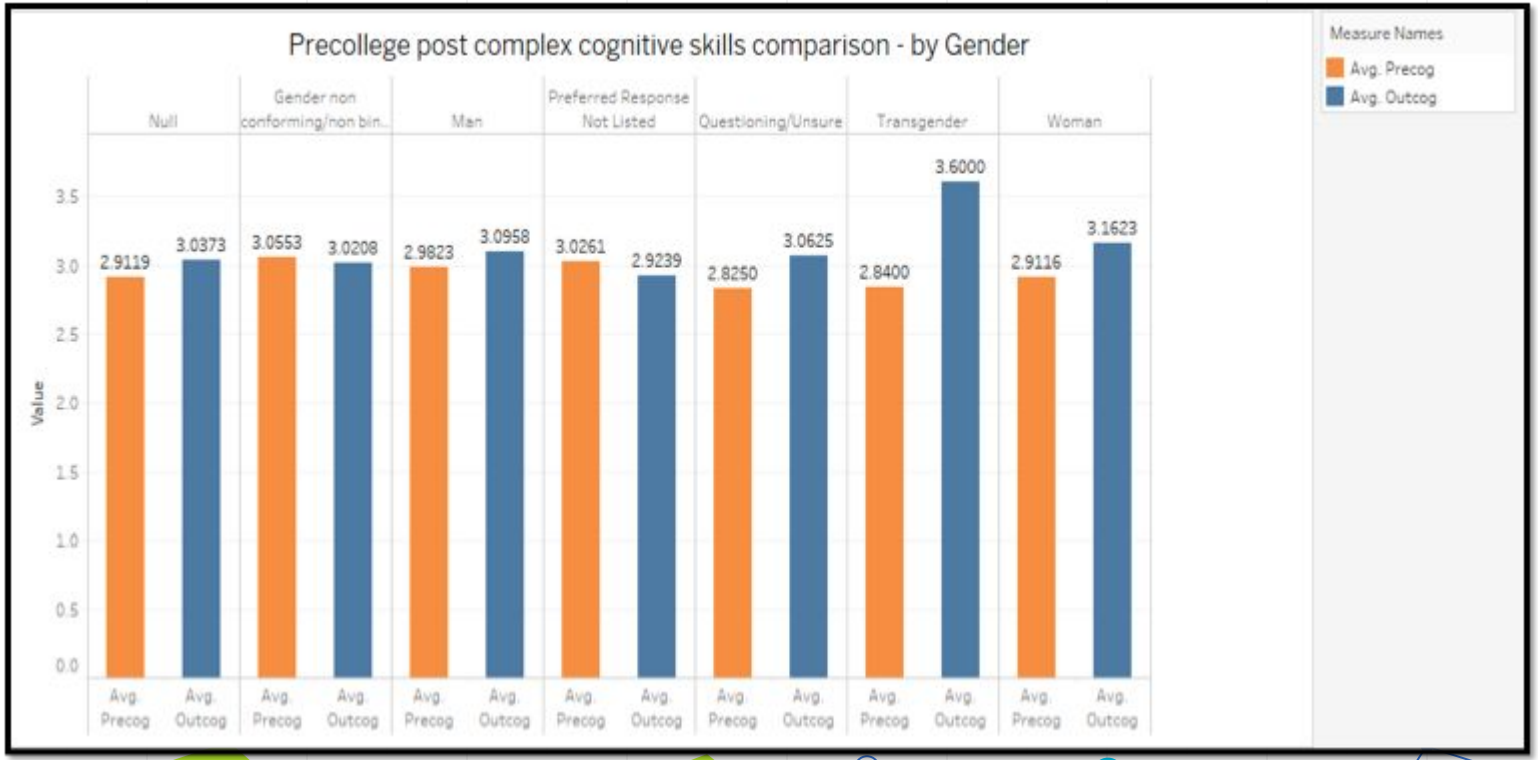
Clustering between #Demographic and Post College Cognitive Skills



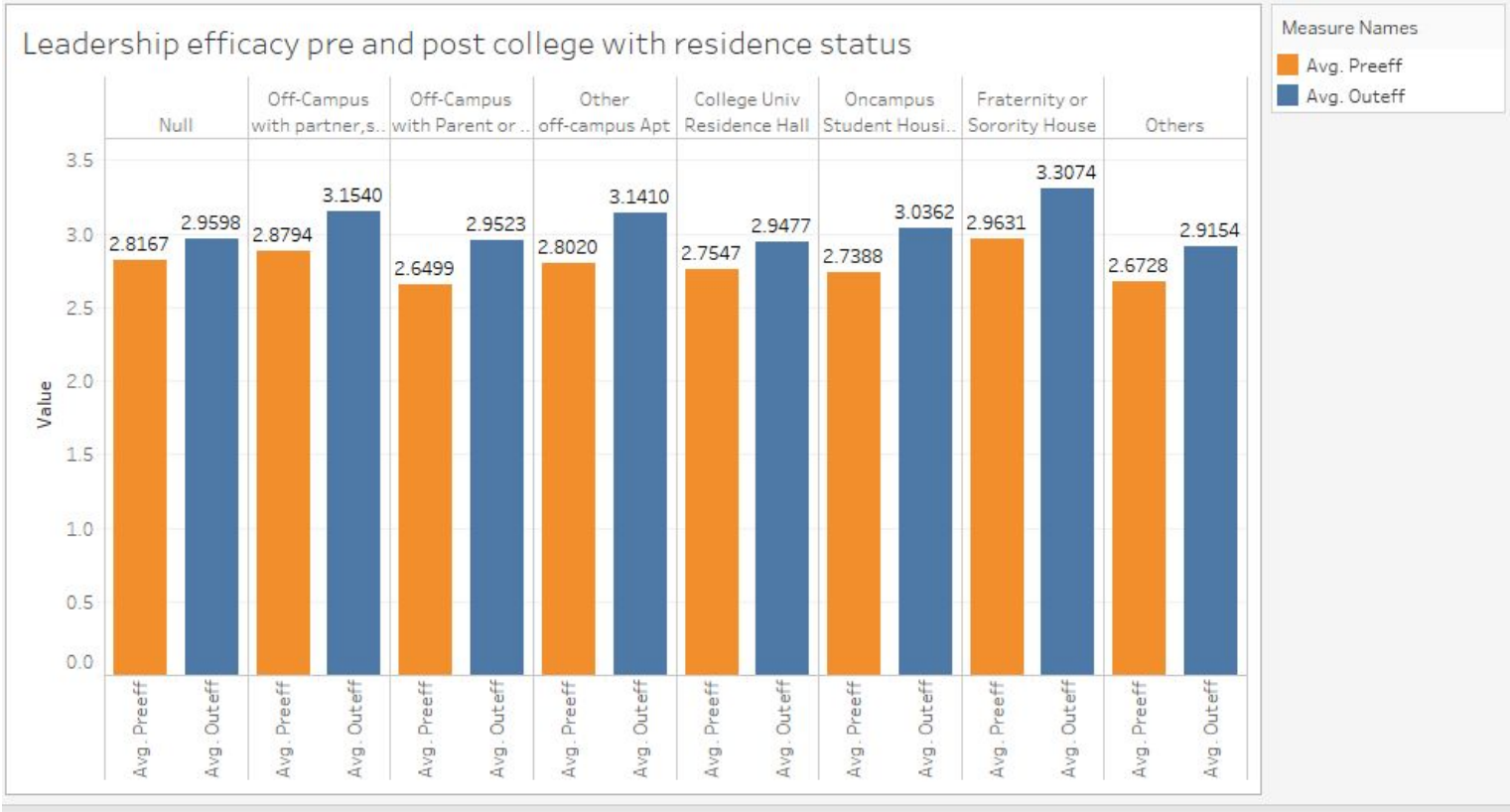
Precollege post leadership efficacy-by Gender



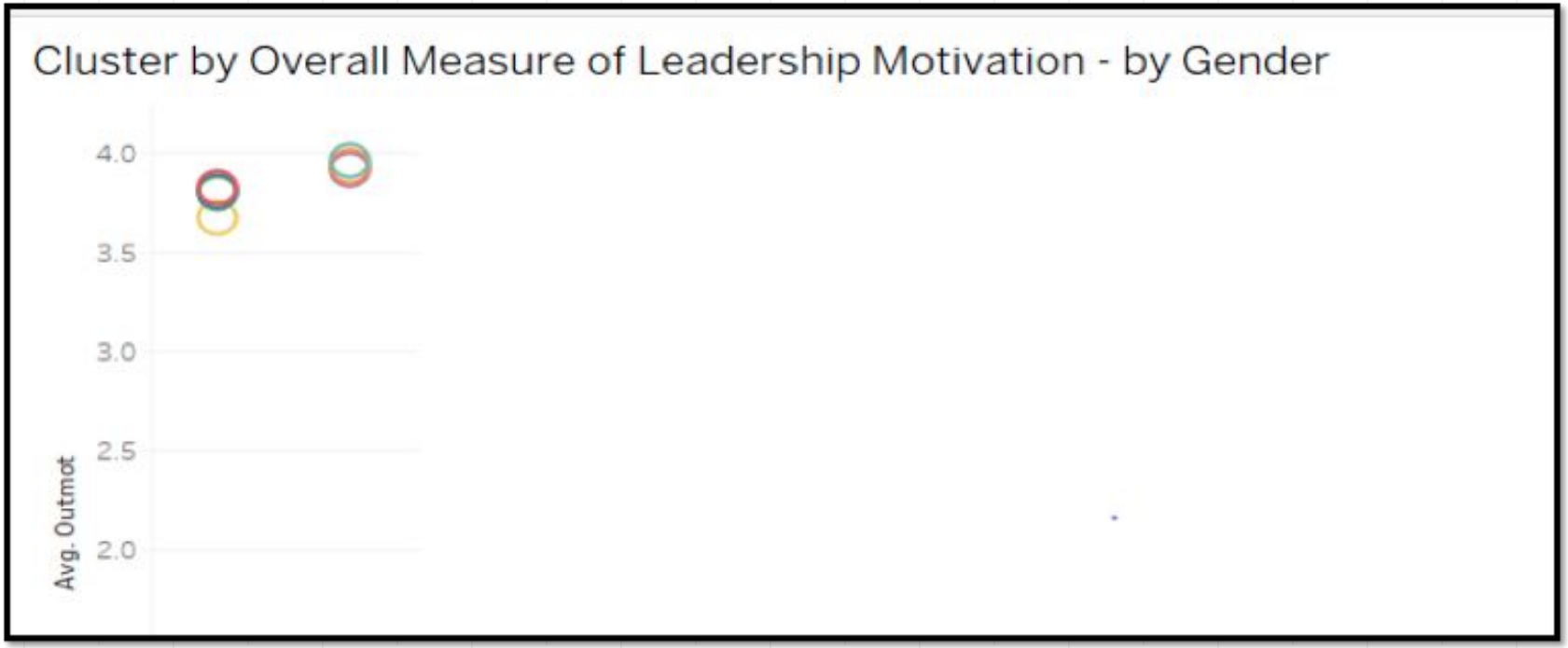
Precollege post complex cognitive skills comparison-by Gender



Leadership efficacy between different residence status



Leadership efficacy between different class levels


Cluster by Overall Measure of Leadership Motivation - by Gender
4.0
3.5
3.0
2.5
2.0
Avg. Outmot

5. Conclusion & Future Work

- Demographic variables such as gender, ethnicity play a role in leadership and cognitive skills in college.
- There has been in general increase in the leadership efficacy and cognitive skills of the students after higher education.
- Environment variable such as where students stay, which groups they are part of also plays an important role.
- We can give suggestion to RSA that special initiatives can be introduced in for of workshops or extra curricular to help students who identified themselves as “queer gender” as the increase has been lowest in that gender of all. Similarly for students residing in college residence halls programs can be designed as the increase in the leadership efficacy is the lowest.

Task Log- Ankita

Date	Timespan	Hours	Task Description
11/03/2018	1:00 PM – 4:00 PM	3 hrs.	Getting the access to Sakai; reading the background material provided.
11/04/2018	2:00 PM – 6:00 PM	3 hrs.	Finding research papers and reading them
11/05/2018	10:00 AM – 12:00 Noon	2 hrs.	Brainstorming on how to proceed with the data
	12:00 Noon – 12:30 PM	0.5 hrs.	Call with Prof. Christie regarding our approach
	12:30 PM – 3:00 PM	2 hrs.	Collecting information and reading various sources on Survey Analysis, evaluation metrics, cleansing of data.
11/06/2018	1:00 PM – 2:30 PM	1.5 hrs.	Assembling the information collected and summarizing it for the proposal
	3:00 – 5:00 PM	2 hrs.	Writing Proposal
11/07/2018	10:30 AM – 11:00 AM	0.5 hrs.	Creating sample dataset to include in the presentation and proposal
	4:00 PM – 4:30 PM	0.5 hrs.	Suggesting and making changes to proposal and presentations
11/11/2018	2:00 PM - 4:00 PM	2 hrs	Research about survey analysis
11/17/2018	11:00 AM - 1:00 PM	2 hrs	Understanding the data and devising cleansing techniques in python using Jupyter Notebook
11/30/2018	1:00 PM- 2:30 PM	1.5 hrs	Preliminary data analysis reports. Poster designing
11/30-12/13		14hrs	Applying Algorithms on the data set. Creating Jupyter notebook

Task Log- Purvee

Date	Timespan	Hours	Task Description
11/03/2018	2:30 PM – 5:00 PM	2.5 hrs.	Getting the access to Sakai; reading the background material provided on the portal
11/04/2018	1:00 PM – 5:00 PM	4 hrs.	Going through MSL Resources
11/04/2018	7PM-10:30PM	3.5 hrs.	Literature Research, reading papers
11/05/2018	11:30AM-2:30PM	3 hrs.	Writing proposal, contributing ideas
11/06/2018	2:30 PM – 5:00 PM	2.5 hrs.	Perusing dataset, read code book
11/27/2018	3:00-4:30pm	1.5 hours	Call with Dayna Weintraub
11/30/2018	12:00pm- 4:00pm	4 hours	Reading the MSL files
12/01/2018	4:00pm-7:30pm	3 hours	Elevator Pitch & Poster
12/08/2018	5:00pm- 11:00pm	6 hours	Final testing and presentation
12/10/2018	10AM-4PM	6 hrs	Researching on various approaches, writing and executing code in python for our project.
12/12/2018	11AM-10PM	8hrs	Preparing presentations, executing and checking the final codes and interpreting the results.

Task Log- Rishika

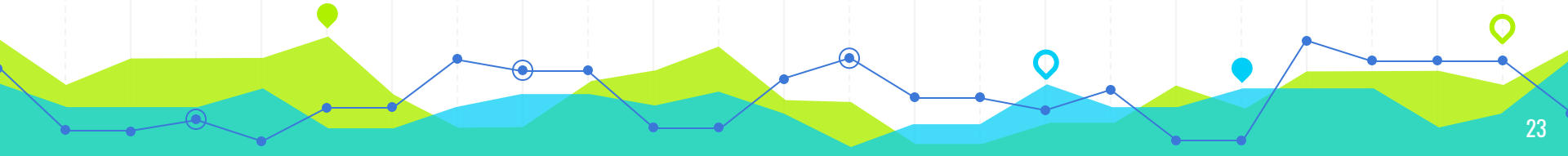
Date	Timespan	Hours	Task Description
11/04/2018	4:00pm-8:00pm	4 hours	Going through the material
11/05/2018	12:00-12:30pm	0.5 hours	Call with Dr.Christie
	2:00pm-7:00pm	5 hours	Finding Research papers
11/06/2018	5:00pm-10:00pm	5 hours	Reading Papers found
11/07/2018	7:00pm-9:00pm	2 hours	Writing Proposal and making presentation
11/27/2018	3:00-4:30pm	1.5 hours	Call with Dayna Weintraub
11/30/2018	12:00pm- 4:00pm	4 hours	Reading the MSL files
12/01/2018	4:00pm-7:30pm	3 hours	Elevator Pitch & Poster
12/06/2018	5:00pm- 11:00pm	6 hours	Researching different approaches
12/08/2018	1:00pm-5:30pm	4.5 hours	Final testing and presentation
12/15/2018	8pm-11:30pm	3.5 hours	Final Report
12/16/2018	11:30am- 3:30pm	4 hours	Final Report

References

- https://files.eric.ed.gov/fulltext/EJ1091521.pdf
- https://www.datascience.com/blog/k-means-clustering
- https://zapier.com/learn/forms-surveys/design-analyze-survey/#analyze
- https://www.ilri.org/biometrics/TrainingResources/Documents/University%20of%20Reading/Guides/Guides%20on%20Analysis/ApprochAnalysis.pdf
- https://www.genroe.com/blog/analyze-survey-data-in-excel/11483
- http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schulte/theses/phd/algorithm.pdf
- https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

THANK YOU!

Any questions?



Elevator Pitch

Done by:- Ankita Prasad, Purvee Agrawal, Rishika Multani

1) **Short description of the project:-** The project is based on a research that is carried out by the Rutgers Student Affairs**.** It is known as Multi-Institutional Study of Leadership (MSL) which is an international research program to examine the influence of higher education on college going students leadership development. The study is also done for the examination of experiences during college and their influences on skills like complex cognitive skills, social perspective-taking, leadership efficacy.

2) **Application Area:-** To decipher patterns in the data highlighting the students' leadership skills which will help the Students Affair organization to design strategies in helping incoming students to improve their leadership skills.

3) **Short discussion on the approach**:-Our data is survey based therefore a lot of cleansing and munging is required and is major portion of the project. The following approach has been carried out:-
    - In the numerical columns, using python we have replaced "#NULL!" values with -1 and in categorical columns we have replaced null values with "NA".
    - Using heat maps, we have tried to find correlation between various attributes of "Leadership Efficacy skills", "cognitive skills", "social perspective-taking" attributes in order to find correlated attributes and remove the extra attributes.
    - Carried out hypothesis testing to analyse relation between two subsets - leadership efficacy and cognitive skills.
    - For the basic analysis of the data we have used cross validation approach to compare various attributes by different demographic and environment variables.
    - We aim to perform unsupervised learning on the data, to analyse main clusters in our data based on demographic and environment variables and how good they are in "leadership skills" and "cognitive skills".
    - After this, based on professor Bill's advice we will use instance based learning approach and then compare the results with unsupervised results.

4) **Interesting result:-** While performing preliminary analysis we have found the following:-
    - For cognitive skills, for all genders categories other than students who have identified themselves as "genderqueer/non conforming " and " preferred response not listed", has increased. Highest increase in cognitive skills was in people who identified themselves as "transgender".
    - Improvement in women is more as compared to that in men.
    - For leadership efficacy, there has been an increase in all the gender categories, with the highest increase in people who have identified themselves as "transgender".

5) **Hook:-** Based on the preliminary analysis, we found our data has a great amount of highly correlated questions, we have tried to identify some selected environment, demographics attributes, and then focussed our analysis using the pruned attributes. It was interesting to compare the skills between various demographics. It gave us an insight on who benefitted the most and for whom the special programs must be designed to help them in gaining confidence about their leadership skill.

RU STUDENT AFFAIRS:- MSL

Ankita Prasad, Purvee Agrawal, Rishika Multani

RUTGERS

MULTI-INSTITUTIONAL STUDY OF LEADERSHIP

Objectives

1. Examine effect of higher education on students' leadership skills.
2. Analysis on students to check whether attending college and demographics have any relation with leadership skills.
3. To help the institution in designing the strategies in order to improve leadership skills in the students.

Data

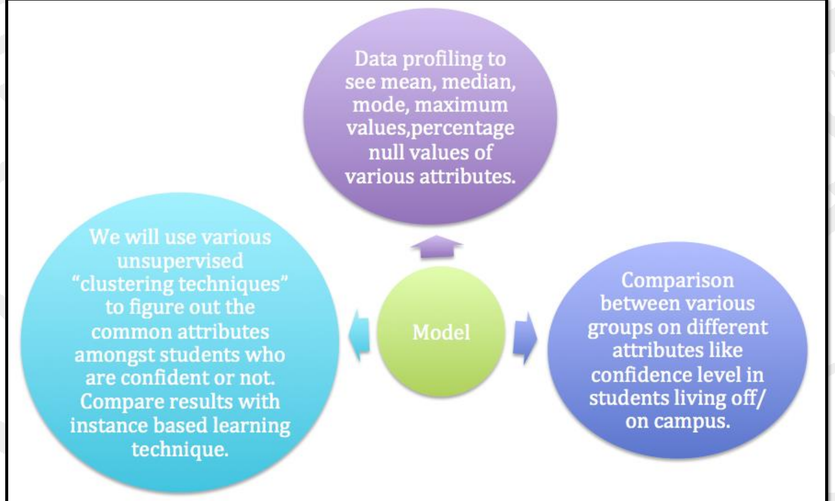
1) Attributes=700,
2) Respondents=10300
3) Completed survey=27%

Final Dataset

It is a survey conducted for all the years of college students with a list of demographics, cognitive,psychometric question sets.

Data in the form of:-
a) single response
b) multiple response
c) numerical response
d) text response,
e) memo response.

Models



Data Profiling & Supervised Learning

Response Composition - by Gender	
Null	27.80%
Gender non conforming/non bi..	0.66%
Man	27.75%
Preferred Response Not Listed	0.32%
Questioning/Unsure	0.33%
Transgender	0.07%
Woman	43.08%

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.596673
R Square	0.356018
Adjusted F	0.355381
Standard I	1.140824
Observatio	2024

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	2	1454.125	727.0627	558.6436	7.4E-194
Residual	2021	2630.288	1.301478		
Total	2023	4084.413			

	Coefficients	*ndard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	1.012271	0.072684	13.92697	3.62E-42	0.869727	1.154815	0.869727	1.154815
3	0.550407	0.040061	13.73928	3.92E-41	0.471842	0.628972	0.471842	0.628972
3	0.324966	0.038378	8.467545	4.75E-17	0.249702	0.400231	0.249702	0.400231

Comparison between pre college and post college skills based on demographic variables and Clustering