incoming_attributes_analysis

May 6, 2019

This notebook investigates the student population given incoming attributes

1 Preparing the data

We need to load pre survey (incoming attitudes), post survey (demographics data), and worksheet (incoming knowledge).

In [1]:
```
from IPython.display import display
from IPython.display import HTML
import IPython.core.display as di

# This line will hide code by default when the notebook is converted to HTML
di.display_html('<script>jQuery(function() {if (jQuery("body.notebook_app").length == (
```

In [2]:
```
%load_ext autoreload
%autoreload 1
%aimport utils_read_parsing
%aimport utils_timeline_viz
from utils_timeline_viz import *
from utils_read_parsing import *
import matplotlib.pyplot as plt
from tabulate import tabulate
from scipy.stats import ranksums
import seaborn as sns
import statsmodels.api as sm
pd.set_option("display.width", 100)
import matplotlib.pyplot as plt
from statsmodels.formula.api import ols
from statsmodels.graphics.api import interaction_plot, abline_plot
from statsmodels.stats.anova import anova_lm
from statsmodels.discrete.discrete_model import Logit
from sklearn import decomposition
%matplotlib inline
matplotlib.style.use('ggplot')
matplotlib.rcParams['figure.figsize'] = 10, 6
pd.set_option('precision',3)
pd.set_option("display.width", 100)
```

```
pd.set_option('display.max_columns', 60)
np.set_printoptions(precision=3,suppress=True)
```

```
D:\Applications\Anaconda2\lib\site-packages\statsmodels\compat\pandas.py:56: FutureWarning: The
  from pandas.core import datetools
```

1.1 Loading all the data sources

```
In [3]: %reload_ext utils_read_parsing
        pre_survey_df = get_massaged_pre_survey()
        post_survey_df = get_massaged_post_survey()
        worksheet_df = get_massaged_worksheet_model_data()
```

1.2 Merging the data

```
In [4]: post_survey_df = post_survey_df[['age','english.0-writing','english.1-reading','gender
        post_survey_df.fillna(0,inplace=True)
        post_survey_df = post_survey_df[post_survey_df['sim_index']==2]
        data = pre_survey_df.merge(post_survey_df, how='outer', on=['sid']);

In [5]: worksheet_df = get_massaged_worksheet_highest_understanding_data()
        pre_know = worksheet_df[['sid','variable','pre_highest']].pivot(index='sid', columns='
        data = data.merge(pre_know['pre_highest'], how='outer', on=['sid']);

In [6]: data.rename(columns={"sim": "second sim"},inplace=True)
        # data.sort_values('sid').head(12)
```

1.3 Convert lickert scale and other values from strings to integers

```
In [7]: value_converter2 = {
            'Prefer not to answer':0,
            '20-22':21,
            '18-19':19,
            '17 and under':17,
            'Fluent':3,
            'Average':2,
            'Beginner':1,
            'Absorbance':2,
            'Capacitance':1,
            'Not at all':1,
            'Definitely':4,
            'Somewhat':2,
            'Mostly':3,
            'Almost always':4,
            'Sometimes':2,
            'Almost never':1,
            'Often':3,
        }
```

```
In [8]: for value,replacement in value_converter2.iteritems():
            data = data.replace(value,replacement)
        data.fillna(0,inplace=True)
        data.head()
```

```
Out[8]:    [prior_lab] What lab courses are you presently taking or have taken in the past? Che
        0                                                 1.0
        1                                                 1.0
        2                                                 0.0
        3                                                 1.0
        4                                                 0.0

           [prior_lab] What lab courses are you presently taking or have taken in the past? Che
        0                                                 1.0
        1                                                 1.0
        2                                                 1.0
        3                                                 1.0
        4                                                 1.0

           [prior_lab] What lab courses are you presently taking or have taken in the past? Che
        0                                                 1.0
        1                                                 0.0
        2                                                 1.0
        3                                                 1.0
        4                                                 0.0

           [prior_lab] What lab courses are you presently taking or have taken in the past? Che
        0                                                 0.0
        1                                                 0.0
        2                                                 0.0
        3                                                 0.0
        4                                                 0.0

           similar_L  similar_C  same_L  same_C  prior_number_virtual_labs  perceivedvalue.0-bo
        0          0          0       0       0                          2
        1          0          0       0       0                          0
        2          0          0       0       0                          1
        3          0          1       0       0                          3
        4          1          0       0       0                          2

           perceivedvalue.1-productive  perceivedvalue.2-useless  perceivedvalue.3-engaging  \
        0                          2.0                         1                        2.0
        1                          4.0                         1                        3.0
        2                          2.0                         1                        2.0
        3                          3.0                         1                        3.0
        4                          3.0                         1                        3.0

           taskinterpretation.0-investigate the basic mechanics of the topic at hand  \
```

```
0                                                 3
1                                                 4
2                                                 2
3                                                 3
4                                                 4

   taskinterpretation.1-design my own experiments that can help me understand the topi
0                                                 2
1                                                 4
2                                                 2
3                                                 4
4                                                 2

   taskinterpretation.2-memorize information about the topic at hand  \
0                                                 3
1                                                 3
2                                                 2
3                                                 3
4                                                 1

   taskinterpretation.3-complete a certain number of  questions  \
0                                                 2
1                                                 4
2                                                 3
3                                                 4
4                                                 3

   taskinterpretation.4-develop scientific reasoning skills  pocc.0-learning the basic
0                                               2.0
1                                               4.0
2                                               2.0
3                                               3.0
4                                               4.0

   pocc.1-testing my ideas and theories  pocc.2-answering given questions  \
0                                     3                                 3
1                                     2                                 2
2                                     3                                 3
3                                     4                                 3
4                                     3                                 3

   pocc.3-memorizing key information  pocc.4-exploring the topic       sid  age  \
0                                  2                           3  77047160   19
1                                  2                           3  23836160   19
2                                  3                           2  64006159   21
3                                  3                           3  24566161   19
4                                  2                           4  46792161   19
```

```
   english.0-writing  english.1-reading  gender-Gender non conforming/non-binary  gende
0                  3                  3                                       0.0
1                  2                  2                                       0.0
2                  2                  2                                       0.0
3                  3                  3                                       0.0
4                  3                  3                                       0.0

   gender-Prefer not to answer  gender-Woman                                    major  \
0                          0.0           1.0  *Non science or applied science major
1                          0.0           0.0                      Civil Engineering
2                          0.0           1.0                            *Undeclared
3                          0.0           0.0                    Engineering Physics
4                          0.0           0.0                 Mechanical Engineering

   year-1st year undergraduate  year-2nd year undergraduate  year-3rd year undergradua
0                          1.0                          0.0                          0
1                          1.0                          0.0                          0
2                          0.0                          1.0                          0
3                          1.0                          0.0                          0
4                          1.0                          0.0                          0

   year-4th year undergraduate second sim  sim_index  Area  Battery voltage  Concentra
0                          0.0          C          2   1.0              1.0
1                          0.0          L          2   0.0              3.0
2                          0.0          C          2   1.0              2.0
3                          0.0          L          2   2.0              3.0
4                          0.0          C          2   3.0              3.0

   Separation  Wavelength  Width
0         0.0         1.0    1.0
1         0.0         1.0    1.0
2         1.0         1.0    0.0
3         2.0         1.0    2.0
4         3.0         1.0    2.0
```

In [9]:
```
for c in data.columns:
    if data[c].dtype not in ['int64','float64']:
        print c, data[c].dtype
```

```
major object
second sim object
```

We could remove the “major” if we wanted to…

In [10]:
```
# data.drop('major',axis=1,inplace=True)
```

2 Analysis of all attributes

```
In [11]: N = len(set(data['sid']))
         print "The study includes {0} students.".format(N)
```

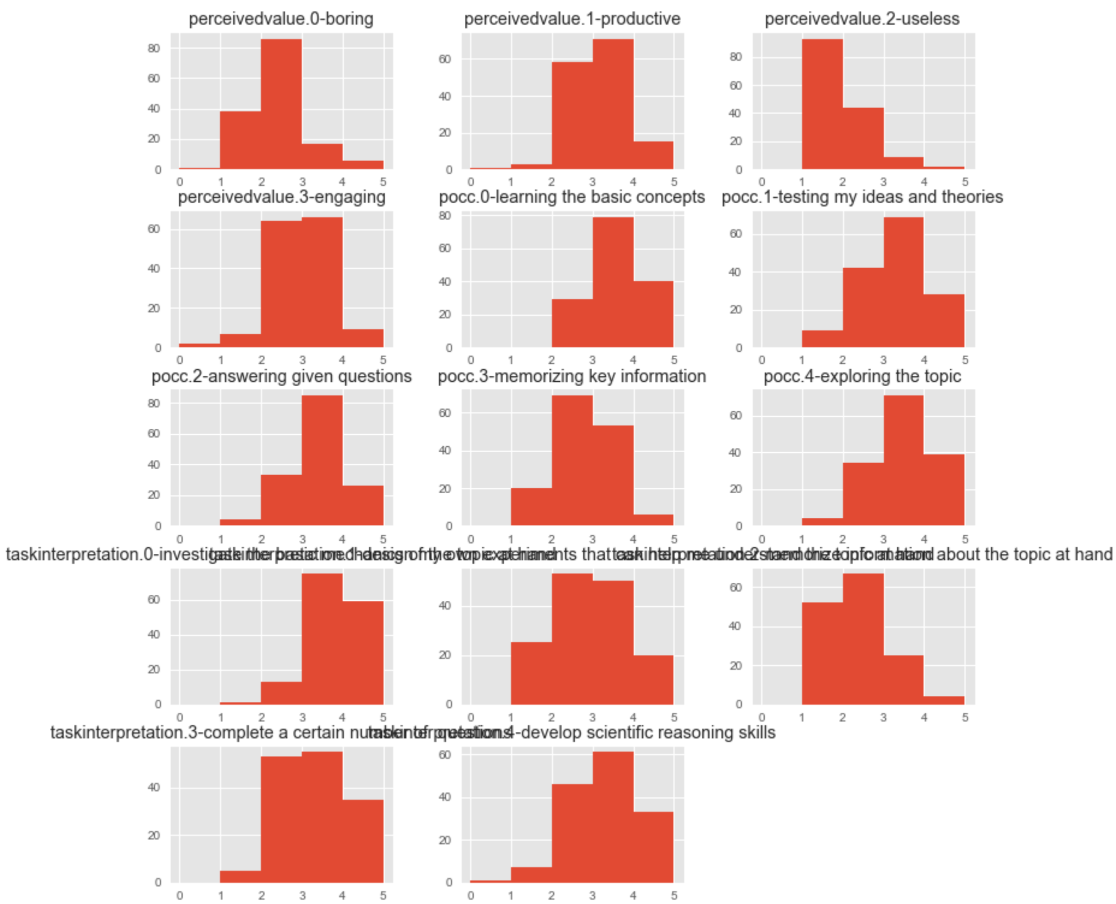
```
The study includes 148 students.
```

```
In [12]: demo_columns = ["age","english.0-writing","english.1-reading","gender-Gender non conf
         pv_columns = ["perceivedvalue.0-boring","perceivedvalue.1-productive","perceivedvalue
         ti_columns = ["taskinterpretation.0-investigate the basic mechanics of the topic at ha
         pocc_columns = ["pocc.0-learning the basic concepts","pocc.1-testing my ideas and theo
         know_columns = ['Concentration','Wavelength','Width','Area','Separation','Battery vol
         att_columns = pv_columns + ti_columns + pocc_columns
         all_columns = demo_columns + att_columns + know_columns
```

2.1 Some descriptives

2.1.1 Incoming Attitudes

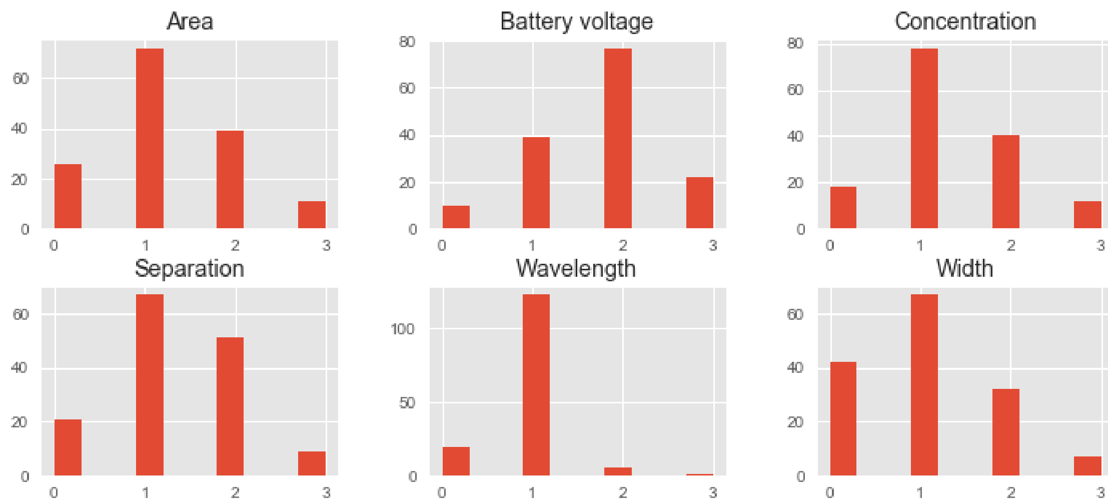
```
In [13]: data[att_columns].hist(figsize=(12,13),layout=(len(att_columns)/3+1,3),bins=[0,1,2,3,4
```



Overall, many students thought the activity would be productive and engaging. Some thought they would be able to explore the topic and test their theories. Many interpreted the task as "answer a certain number of questions".

2.1.2 Incoming knowledge

```
In [14]: data[know_columns].hist(figsize=(12,5),layout=(2,3));
```



The quantitative variables look pretty similar. Few know how to predict Wavelength however they know it's relevant (lots of 1s). Battery voltage seems to be a variable tha many describe qualitatively.

2.1.3 Demographics

See storyline_journal_paper notebook for details. Essentially, most students are first years (76%) with undeclared majors (52%), are fluent in English, have experience with virtual labs and experimental lab classes.

2.2 Overall correlations between attributes

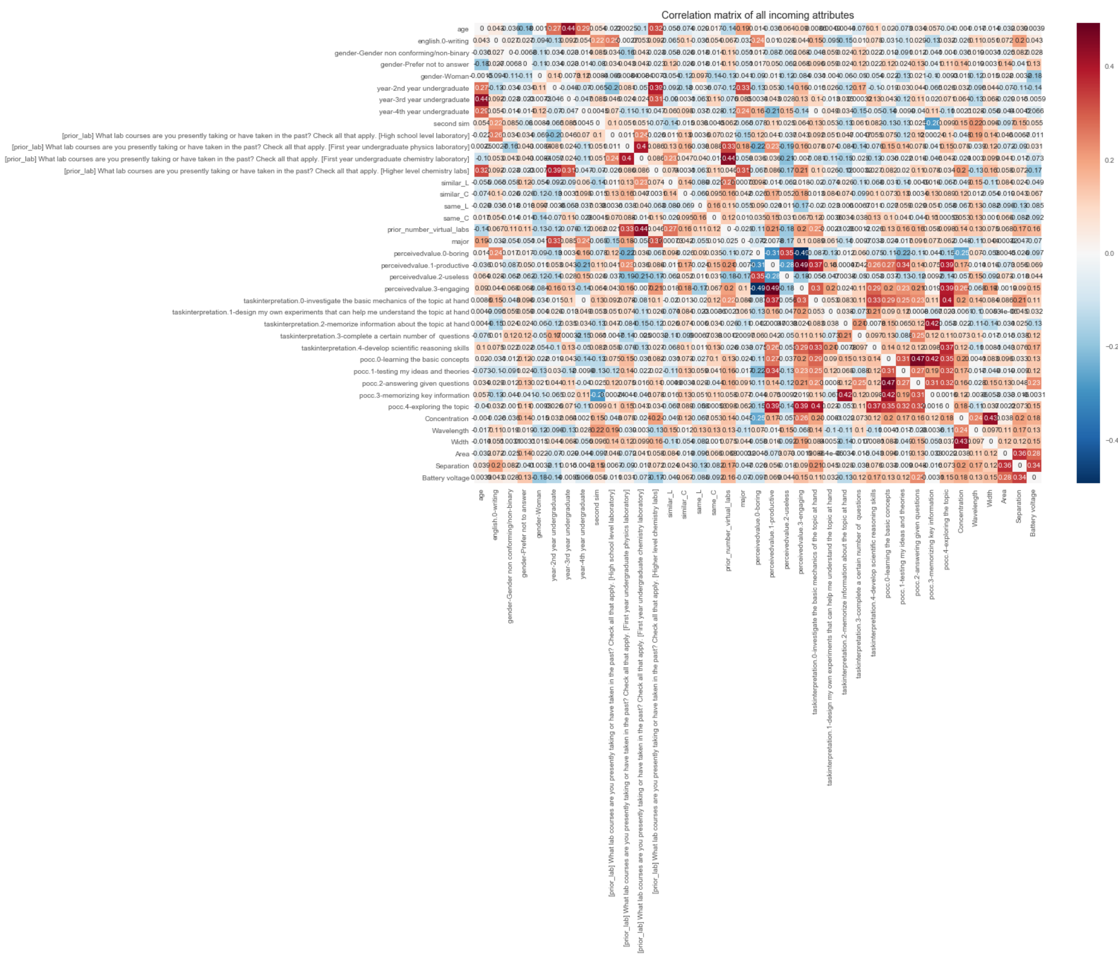
We remove columns that are obviously going to be highly correlated ("Man" wih other gender columns and "1st year undergraduate" with other undergraduate year columns, "english-reading" with "writing")

```
In [15]: all_columns_for_correlation = list(all_columns)
         all_columns_for_correlation.remove('gender-Man')
         all_columns_for_correlation.remove('year-1st year undergraduate')
         all_columns_for_correlation.remove('english.1-reading')
```

```
In [16]: from scipy.stats import spearmanr
         fig, ax = plt.subplots(figsize=(18,12))
         correlation_matrix = np.zeros((len(all_columns_for_correlation),len(all_columns_for_c
```

```
for i,att_i in enumerate(all_columns_for_correlation):
    for j,att_j in enumerate(all_columns_for_correlation):
        r,p = spearmanr(data[att_i],data[att_j])
        correlation_matrix[i,j] = r

np.fill_diagonal(correlation_matrix,0)
plt.title("Correlation matrix of all incoming attributes")
sns.heatmap(correlation_matrix,ax=ax,yticklabels=all_columns_for_correlation,xticklabe
```

```
D:\Applications\Anaconda2\lib\site-packages\scipy\stats\stats.py:253: RuntimeWarning: The inpu
  "values. nan values will be ignored.", RuntimeWarning)
```



Taking into consideration that we are doing 38*38 correlations and some are going to be from random chance, here are a few observations: * Age is correlated to Undergraduate year and 2nd year+ chemistry labs (duh) * Gender doesn't seem to be correlated with anything (yay!) * The fact that students have used similar or the same virtual labs before doesn't correlated with anything * The number of prior virtual labs done by students is correlated to having taken or taking first year chem and physics lab courses. * A lot of attitudinal measures are correlated. We will investigate this in more depth later * Some incoming knowledge seems to be correlated bery slightly with perceived value. For example, Concentration knowledge is correlated with the activity being engaging but not boring. (Keep in mind students got different screenshots of the virtual lab depending on activity order). * Some incoming knowledge on different variables are correlated. We will investigate this next.

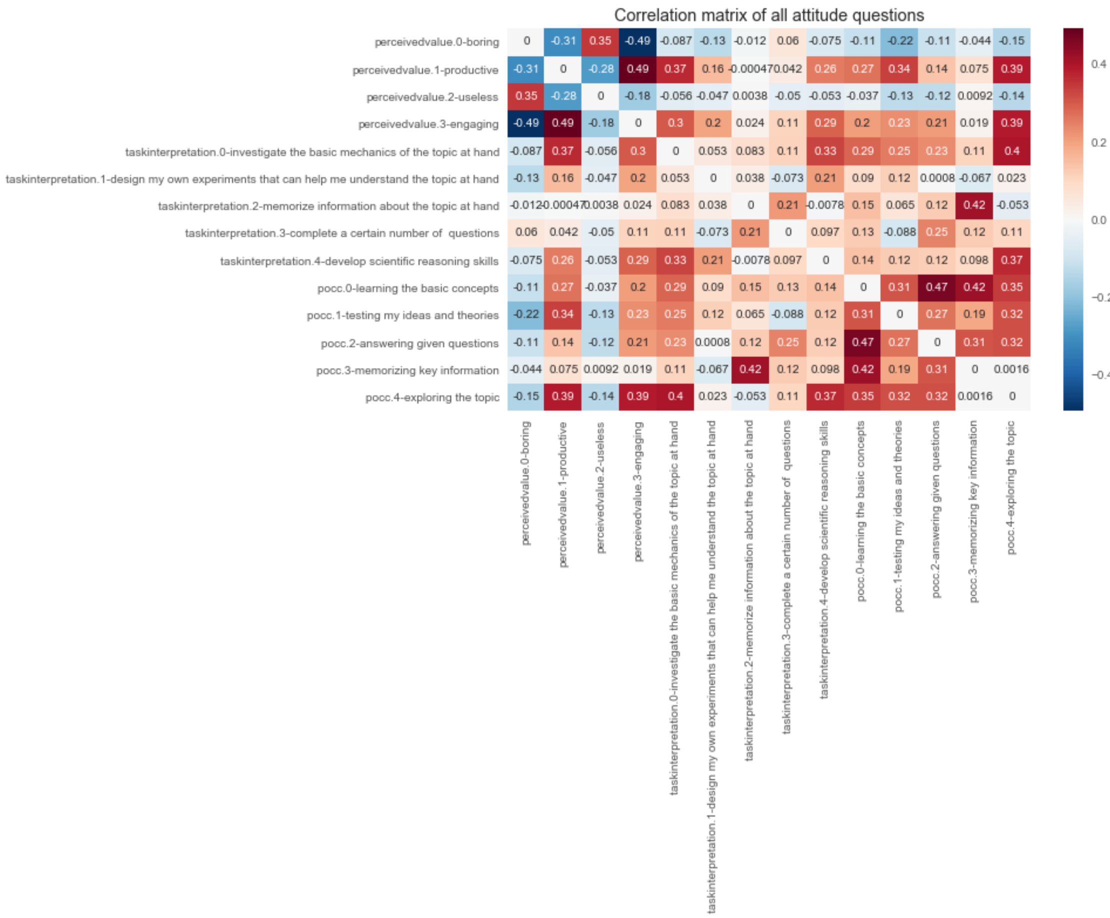
Overall no correlations seem to be out of the ordinary or troubling.

3 Analysis of incoming attitudes

Clearly, many attitude questions are related from our descriptives. Let's investigate how much:

3.0.1 Relationship between attitude measures

```
In [17]: from scipy.stats import spearmanr
         correlation_matrix = np.zeros((len(att_columns),len(att_columns)))
         for i,att_i in enumerate(att_columns):
             for j,att_j in enumerate(att_columns):
                 r,p = spearmanr(data[att_i],data[att_j])
                 correlation_matrix[i,j] = r

         np.fill_diagonal(correlation_matrix,0)
         plt.title("Correlation matrix of all attitude questions")
         sns.heatmap(correlation_matrix,yticklabels=att_columns,xticklabels=att_columns,annot=
```



All the POCC type questions are highly correlated between each other (0.3,0.4). All the perveiced value questions are also highly correlated or anti-correlated (0.3,0.4) which is great since they were deisgned that way ("Do you think the activity will be productive/useless, boring/engaging"). What is interesting is that if they find it productive they also find it engaging.

3.0.2 Reliability scores of attitude questions

To get good reliability measures, we need to reverse some questions values with negative correlations to everything else

```
In [18]: data['perceivedvalue.0-boring_reversed'] = 4-data['perceivedvalue.0-boring']
         data['perceivedvalue.2-useless_reversed'] = 4-data['perceivedvalue.2-useless']
         pv_fixed_columns = ["perceivedvalue.0-boring_reversed","perceivedvalue.1-productive",
```

```
In [19]: def cronbach_alpha(scores):
             if scores.isnull().values.any():
                 print "Cannot compute Cronbach alpha: your dataframe has NaN values."
                 return
             K = scores.shape[1]
             sum_of_var = float(scores.apply(lambda x: np.var(x,ddof=1), axis=0).sum())
```

```
        var = float(np.var(scores.sum(axis=1),ddof=1))
        alpha = (K/float(K-1.0))*(1.0-sum_of_var/var)
        return round(alpha,2)
```

```
In [20]: t = [['questions combined','reliability score']]
         t.append(['all (14)',cronbach_alpha(data[att_columns]) ])
         t.append(['POCC (5)',cronbach_alpha(data[att_columns]) ])
         t.append(['Perceived value (4)',cronbach_alpha(data[pv_fixed_columns]) ])
         t.append(['Task interpretation (5)',cronbach_alpha(data[ti_columns]) ])
         t.append(['Task interpretation and perceived value (9)',cronbach_alpha(data[ti_column

         print tabulate(t)
```

```
-------------------------------------------  -----------------
questions combined                           reliability score
all (14)                                     0.61
POCC (5)                                     0.61
Perceived value (4)                          0.66
Task interpretation (5)                      0.33
Task interpretation and perceived value (9)  0.57
-------------------------------------------  -----------------
```

3.0.3 What are the principle factors of all attitude questions? (PCA)

```
In [21]: columns_for_pc = att_columns
         pca = decomposition.PCA(n_components=6)
         pca.fit(data[columns_for_pc])
         pca.explained_variance_ratio_.cumsum()
```

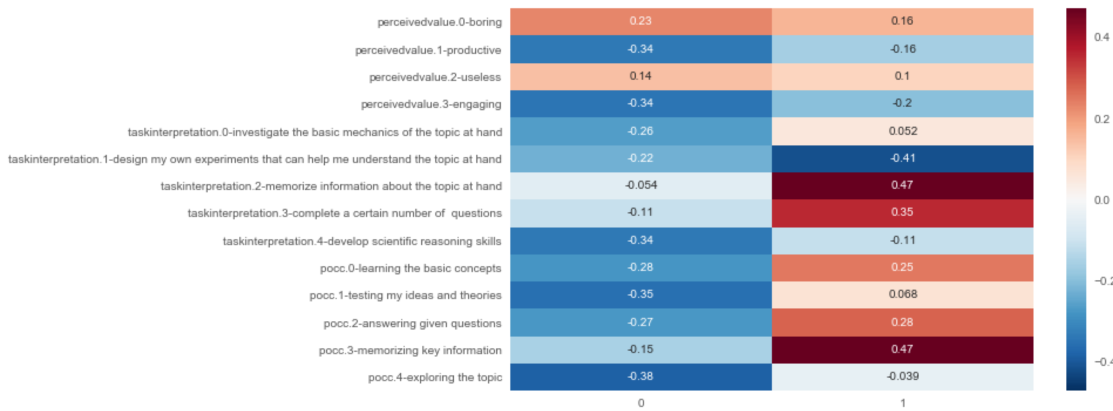
```
Out[21]: array([ 0.23 ,  0.366,  0.463,  0.555,  0.639,  0.702])
```

The first three components of the PCA exaplined almost 50% of the data. Let's stick to 2 to simplify our analysis

```
In [22]: NC = 2
         pca = decomposition.PCA(n_components=NC)
         pca.fit(data[columns_for_pc])
         X = pca.transform(data[columns_for_pc])
         data['PC1'] = zip(*X)[0]
         data['PC2'] = zip(*X)[1]
```

```
In [23]: sns.heatmap(pca.components_.T,yticklabels=columns_for_pc,annot=True)
```

```
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0xfdcb860>
```



Looking at items that are part of the PC (by a factor of 0.3 percent). We can interpret the PC this way:

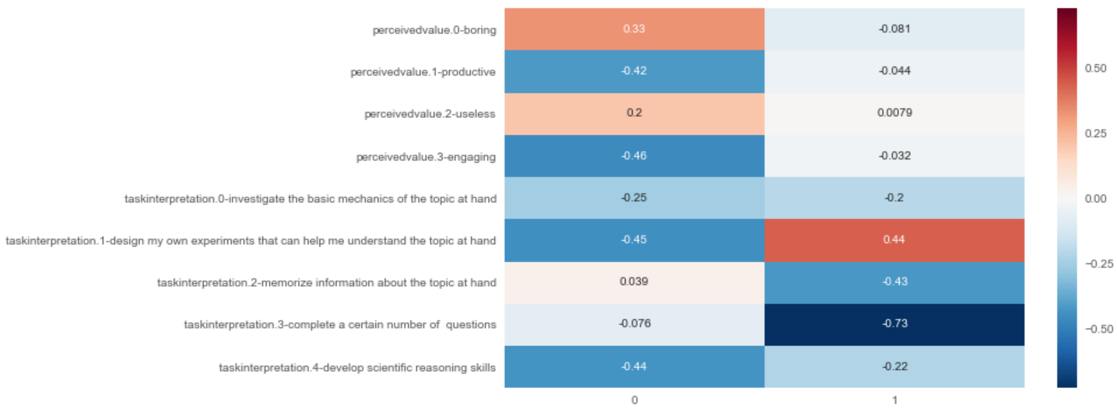
PC 1 - Having a low PC1 means a student thinks: * the activity will be engaging and productive * the activity is design to develop scientific reasoning skills * they can do a good job of exploring the topic and testing their ideas. "the engaged explorers" have low PC1 "bored and not engaged" have high PC1

PC 2 - Having a high PC2 means a student thinks: * the activity is NOT designed to design their own experiments * the activity is design to memorize information and complete a certain number of questions * they can do a good job of memorizing key information "the expecting to be assessed" have high PC1 "ready to design experiments and understand" have low PC1

Since the PCs are orthogonal, we have 4 types of students: 1. The engaged and expecting to be assessed (-+) 2. The engaged and looking for understanding (–) 3. The not engaged and looking for understanding (+-) 4. The not engaged and expecting to be assessed (++)

3.0.4 What are the principle factors of task interpretation and perceived value? (PCA)

```
In [24]: columns_for_pc = pv_columns + ti_columns
         pca = decomposition.PCA(n_components=6)
         pca.fit(data[columns_for_pc])
         pca.explained_variance_ratio_.cumsum()

Out[24]: array([ 0.254,  0.413,  0.548,  0.67 ,  0.762,  0.841])

In [25]: NC = 2
         pca = decomposition.PCA(n_components=NC)
         pca.fit(data[columns_for_pc])
         X = pca.transform(data[columns_for_pc])
         data['PC1_wo_pocc'] = zip(*X)[0]
         data['PC2_wo_pocc'] = zip(*X)[1]

In [26]: sns.heatmap(pca.components_.T,yticklabels=columns_for_pc,annot=True)

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x12c1a898>
```



Looking at items that are part of the PC (by a factor of 0.3 percent). We can interpret the PC this way:

(low) **PC 1** - Students think: * the activity will be engaging, productive * the activity is design to develop scientific reasoning skills

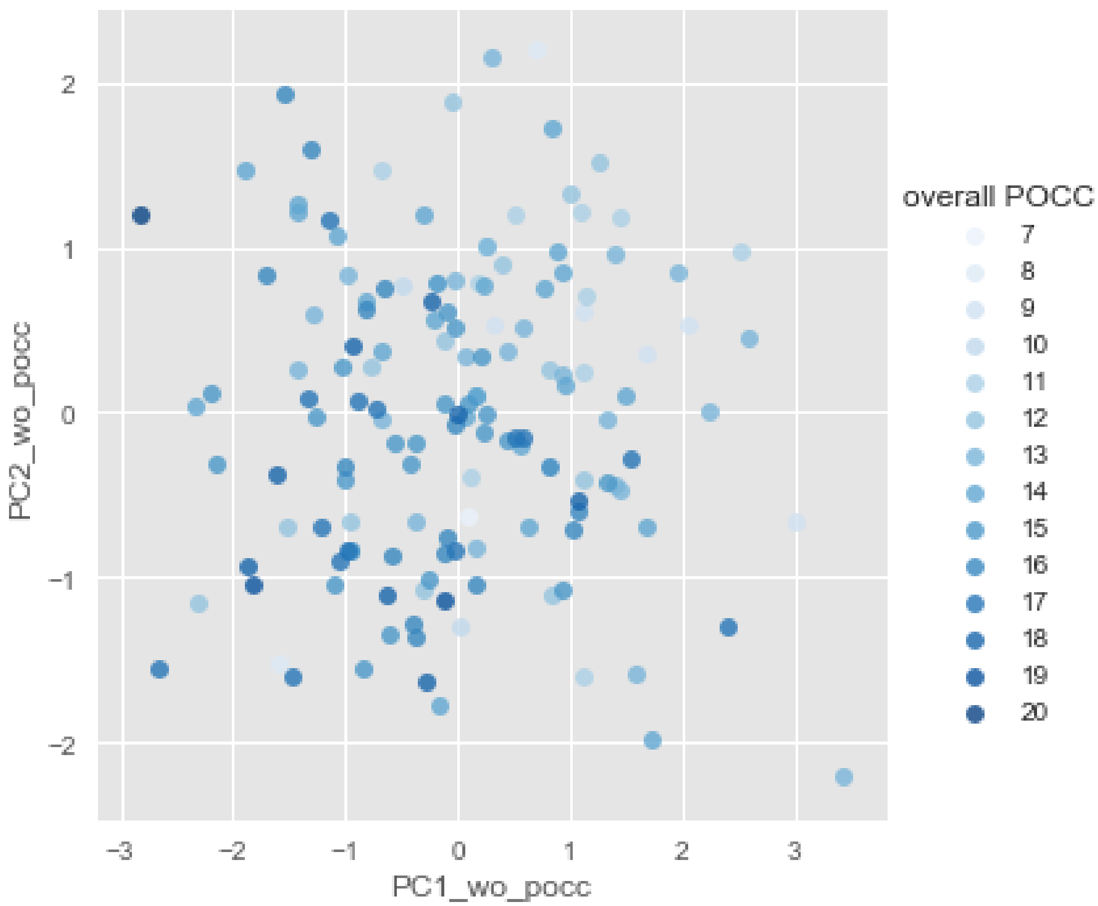
(low) **PC 2** - Students think: * the activity is not designed to design their own experiments * the activity is designed to memorize information and complete a certain number of questions

3.0.5 Investigating the two attitudinal groups in terms of overall POCC

In [27]:
```
# from sklearn.cluster import KMeans
# clusters = KMeans(n_clusters=2, n_init=10, verbose=0).fit(data[columns_for_pc])
# data['cluster_2_label'] = clusters.labels_
```

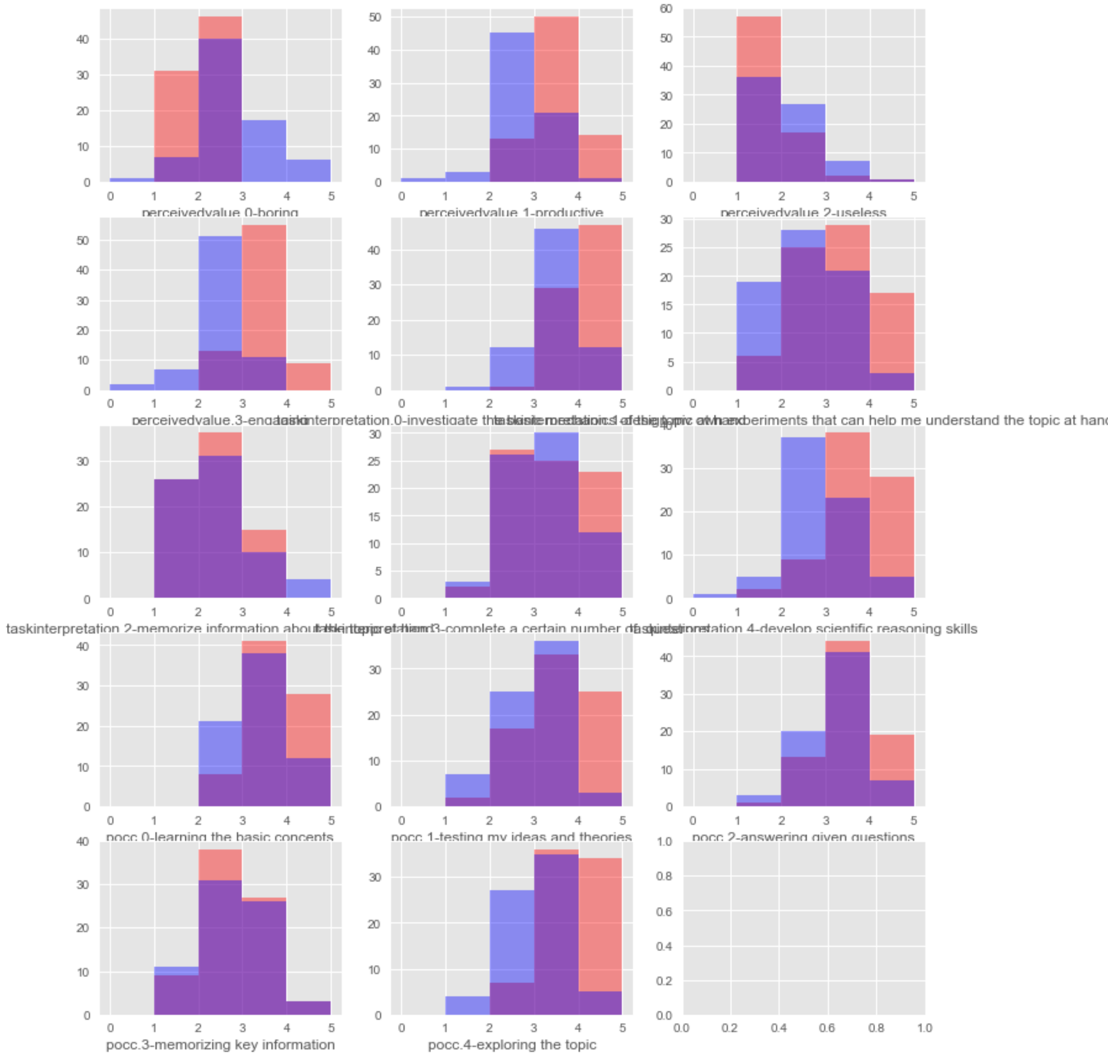
In [28]:
```
data['overall POCC'] = data[pocc_columns].sum(axis=1)
sns.lmplot(x='PC1_wo_pocc', y='PC2_wo_pocc',data=data,fit_reg=False,hue='overall POCC
```

Out[28]:
```
<seaborn.axisgrid.FacetGrid at 0x12b28a58>
```




```
In [29]: fig,axes = plt.subplots(nrows=len(att_columns)/3+1,ncols=3,figsize=(12,15))
         for ax,col in zip(axes.reshape(-1),att_columns):
         #     sns.distplot(data[data['cluster_2_label']==0][col],ax=ax,bins=[0,1,2,3,4,5],kde=
         #     sns.distplot(data[data['cluster_2_label']==1][col],ax=ax,bins=[0,1,2,3,4,5],kde=
             sns.distplot(data[data['PC1_wo_pocc']<0][col],ax=ax,bins=[0,1,2,3,4,5],kde=False,
             sns.distplot(data[data['PC1_wo_pocc']>=0][col],ax=ax,bins=[0,1,2,3,4,5],kde=False
         plt.legend()
```

```
D:\Applications\Anaconda2\lib\site-packages\matplotlib\axes\_axes.py:545: UserWarning: No label
  warnings.warn("No labelled objects found. "
```



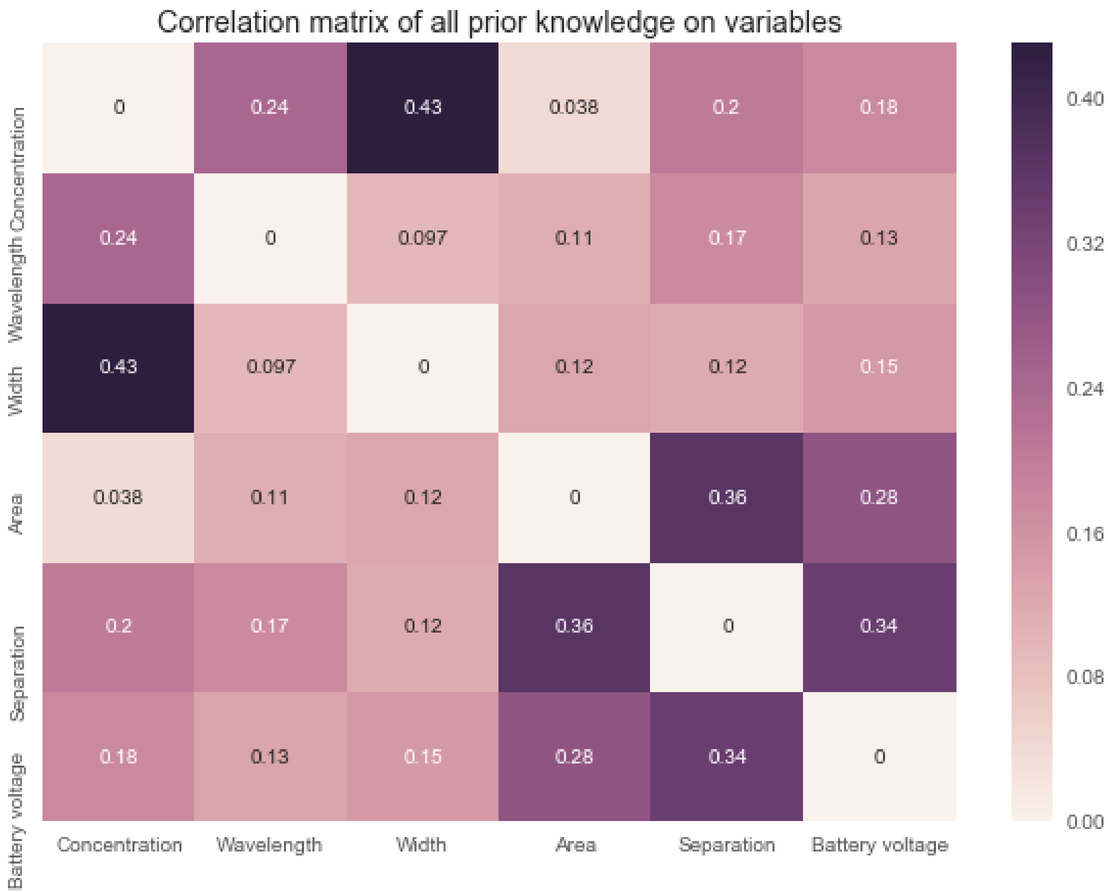
Clearly, the “red” students (low PC1) think: * the activity will be productive and engaging * the activity was design to develop scientific reasoning skills and test their ideas * they can do a good job at the activity

Clearly, the “blue” students (high PC1) think: * the activity will be boring, useless * they tend to have a low perception of their control and competence

4 Analysis of incoming knowledge

```
In [30]: from scipy.stats import spearmanr
         fig, ax = plt.subplots(figsize=(10,7))
         correlation_matrix = np.zeros((len(know_columns),len(know_columns)))
         for i,att_i in enumerate(know_columns):
             for j,att_j in enumerate(know_columns):
                 r,p = spearmanr(data[att_i],data[att_j])
                 correlation_matrix[i,j] = r
```

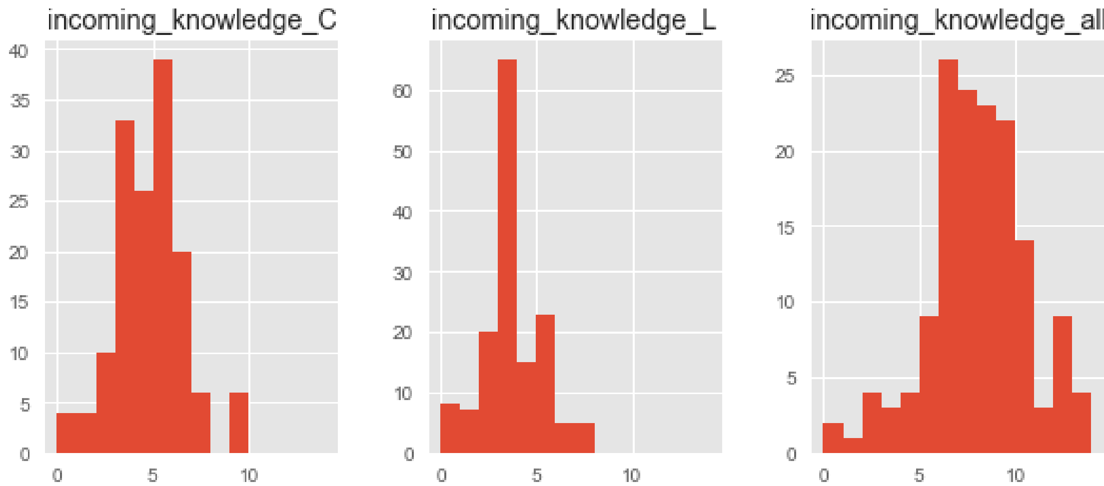
```
np.fill_diagonal(correlation_matrix,0)
plt.title("Correlation matrix of all prior knowledge on variables")
sns.heatmap(correlation_matrix,ax=ax,yticklabels=know_columns,xticklabels=know_columns
```



Clearly, knowledge of variables of the same sim are correlated, particularly Concentration and Width.

```
In [31]: data['incoming_knowledge_L'] = (data['Concentration']+data['Width']+data['Wavelength']
         data['incoming_knowledge_C'] = (data['Area']+data['Separation']+data['Battery voltage
         data['incoming_knowledge_all'] = (data['Area']+data['Separation']+data['Battery voltag
         data['incoming_knowledge_quant'] = (data['Area']+data['Separation']+data['Concentratio
```

```
In [32]: data[['incoming_knowledge_L','incoming_knowledge_C','incoming_knowledge_all']].hist(f
```



We have a somewhat normal distribution.

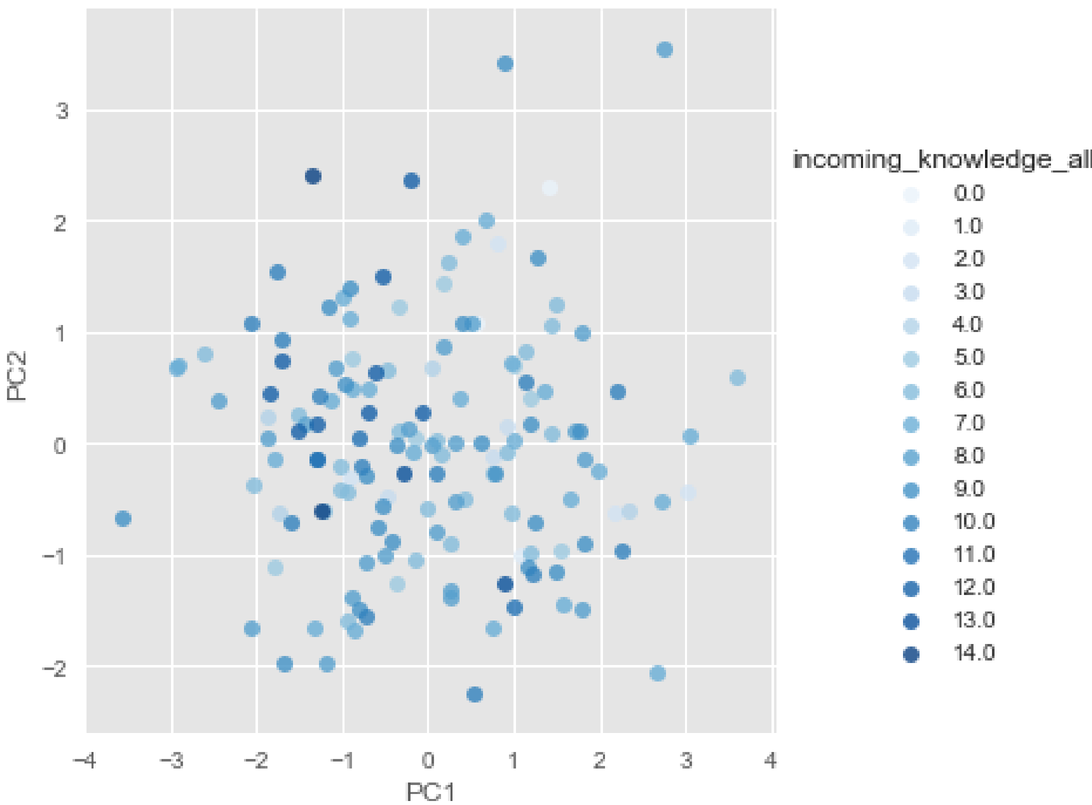
5 Analyzing all attributes

5.1 Relating attitudes to incoming knowledge

Given ou PCA above for all attitude variables (TI, POCC, PV), let's see how incoming knowledge is related.

```
In [33]: sns.lmplot(x='PC1', y='PC2',data=data,fit_reg=False,hue='incoming_knowledge_all', pal
```

```
Out[33]: <seaborn.axisgrid.FacetGrid at 0xe6f1710>
```



From our PCA of attitudes, there doesn't seem to be any natural forming groups in the data. When overlaying the total of their incoming knowledge score, there's no much trend. We notice that the high scoring students (dark blue) generally don't have a high PC1 and high PC2 (are not the boring and expecting to be assessed).

From this, we don't suspect to have well defined clusters in our dataset, but let's investigate with K-means and silhouette analysis.

5.2 Kmeans clustering

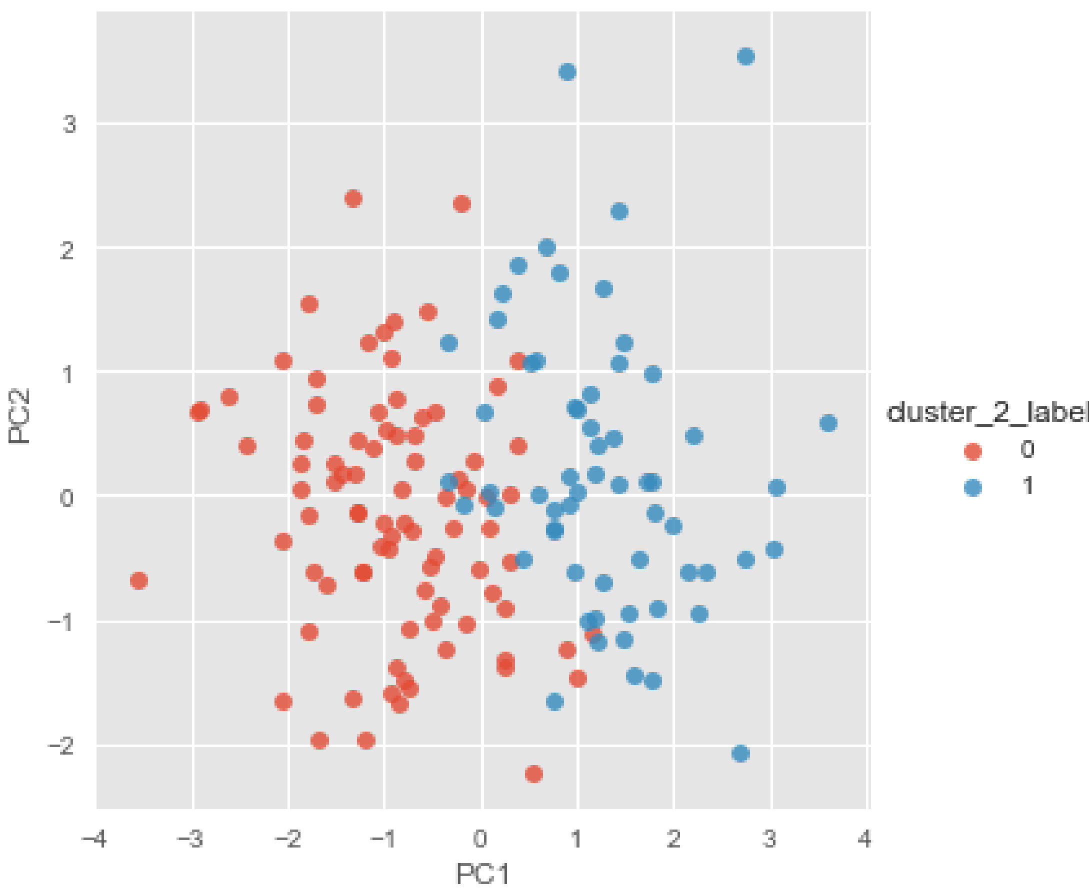
We run clustering on all incoming factors (attitudes, knowledge, year undergradte, experimental labs exerience, virtual lab experience, english writing and reading fluency.

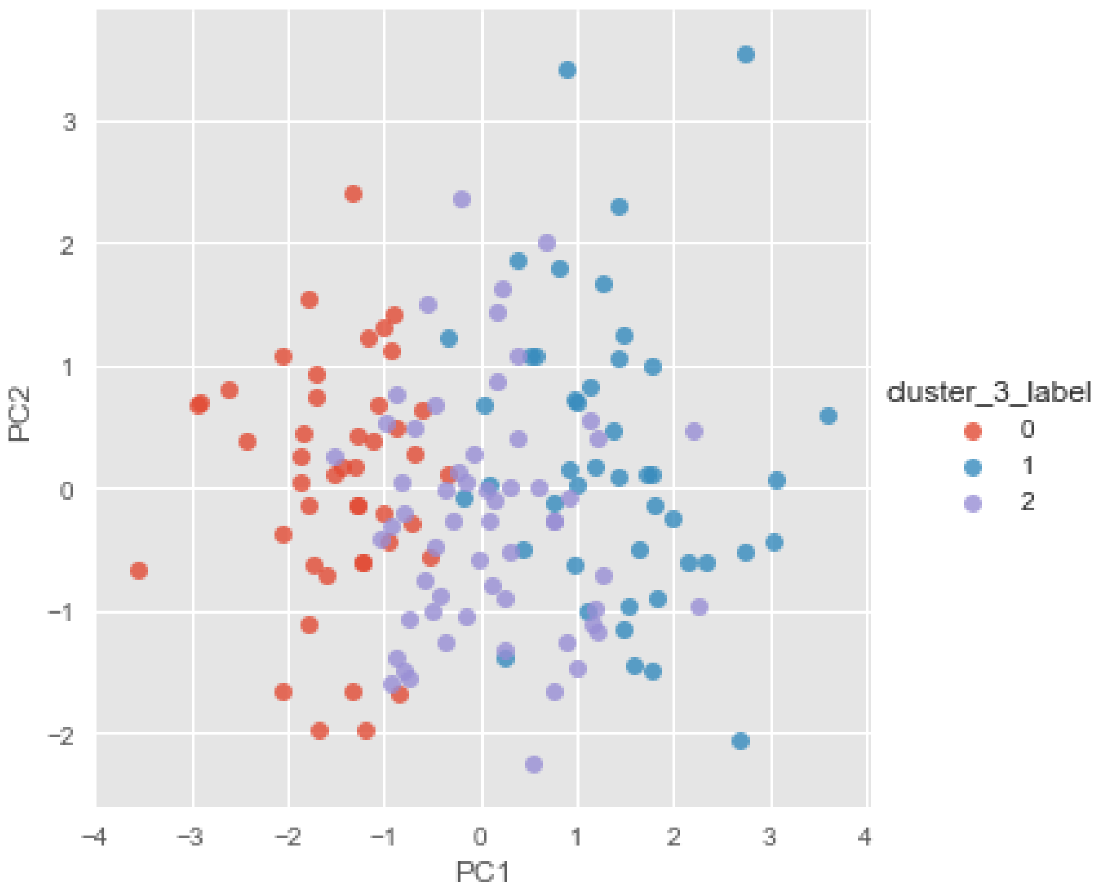
We remove age and Gender because one student answer "0" and "Prefer not to answer" for these questions. We also remove major since it's a qualitative variable

```
In [34]: import numpy as np
         import matplotlib.pyplot as plt
         from sklearn.cluster import KMeans
         from sklearn.metrics import silhouette_score

         cluster_columns = [x for x in all_columns if x not in ['second sim','major','sid','age

         for i in [2,3,4,5]:
```

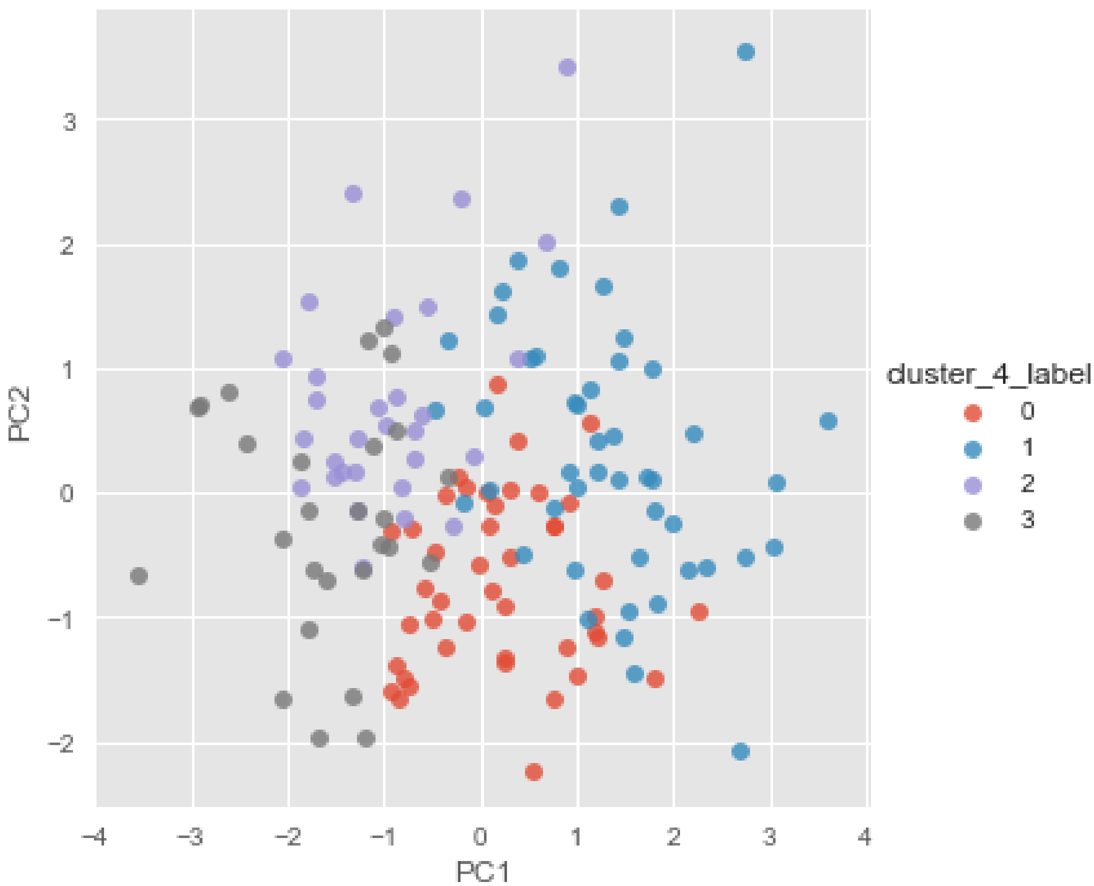
```
        y_pred = KMeans(n_clusters=i).fit_predict(data[cluster_columns])
        data['cluster_'+str(i)+'_label'] = y_pred
        silhouette_avg = silhouette_score(data[cluster_columns], y_pred)
        print("For n_clusters =",i,
              "The average silhouette_score is :", silhouette_avg)
        sns.lmplot(x='PC1', y='PC2',data=data,fit_reg=False,hue='cluster_'+str(i)+'_label
```

```
('For n_clusters =', 2, 'The average silhouette_score is :', 0.098663705806131596)
('For n_clusters =', 3, 'The average silhouette_score is :', 0.058400307335511999)
('For n_clusters =', 4, 'The average silhouette_score is :', 0.053903909496143466)
('For n_clusters =', 5, 'The average silhouette_score is :', 0.059340822782365014)
```

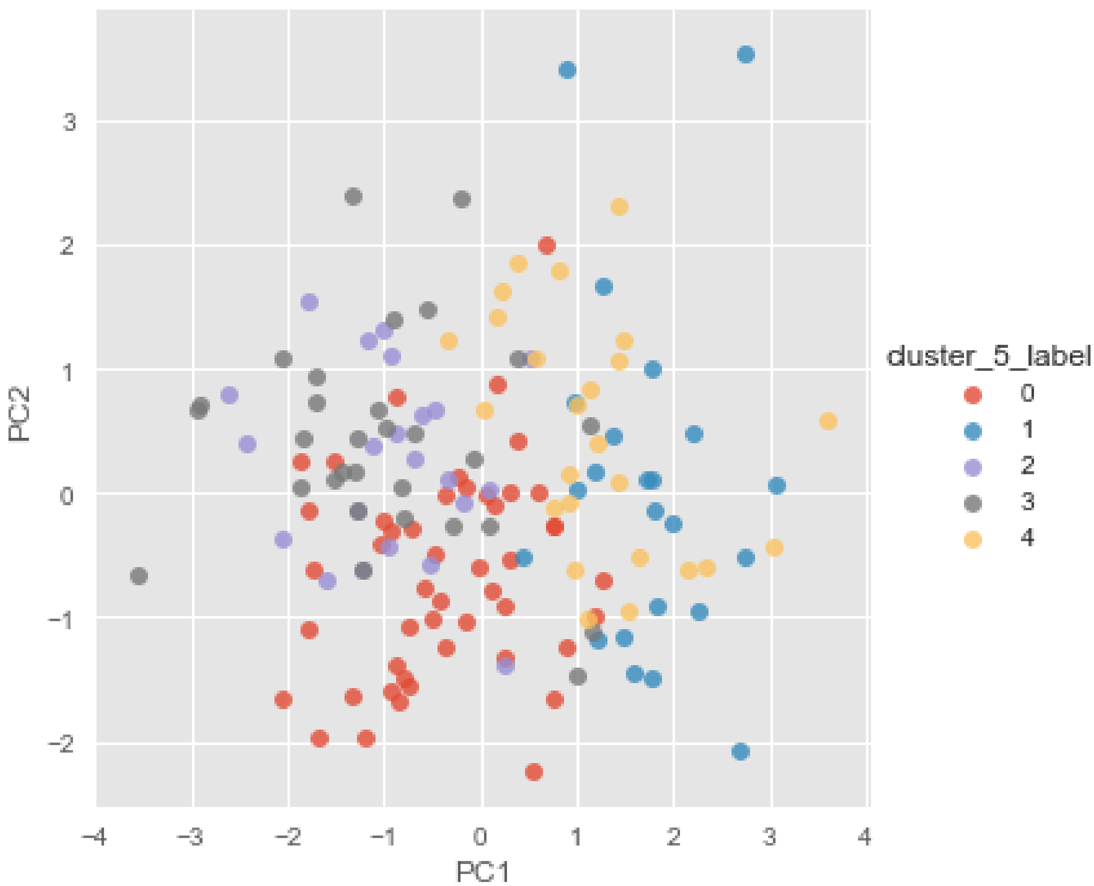




PC2
PC1
duster_3_label
0
1
2
3
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4



PC2
PC1
cluster_4_label
0
1
2
3



Two clusters seems to follow PC 1 divisions. Three clusters as well. Though we still need to investigate what differentiates the clusters in terms of knowledge and education background. Four and five clusters gets messier...

HOWEVER, the silhouette scores are really low (max is 1, near 0 means most points are as close to a neighboring cluster than their own, negative means most points are closer to neighboring cluster). For more info: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

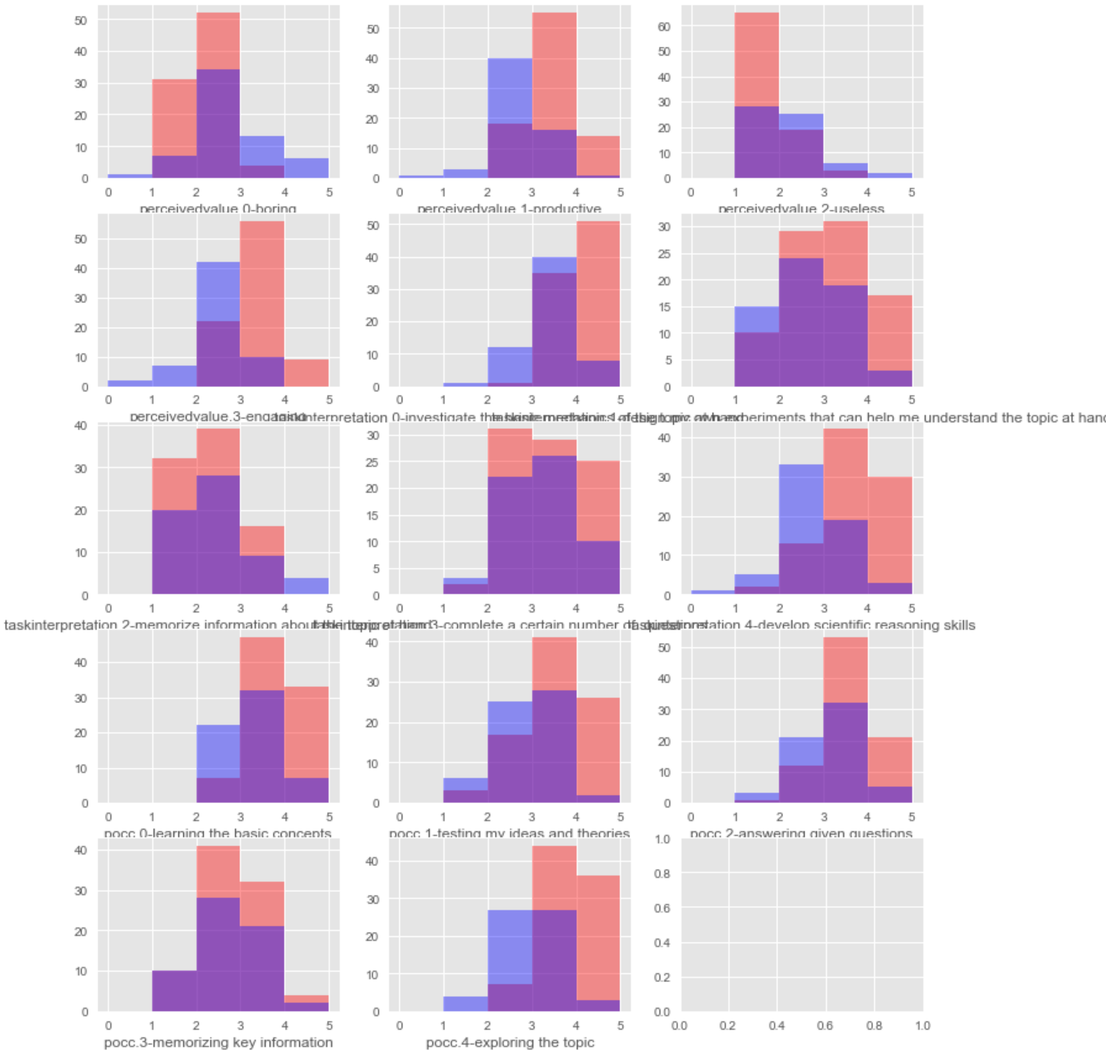
All to say, our data doesn't have clusters... Other analysis (ex. hierarchical analysis) would probably lead to similar results.

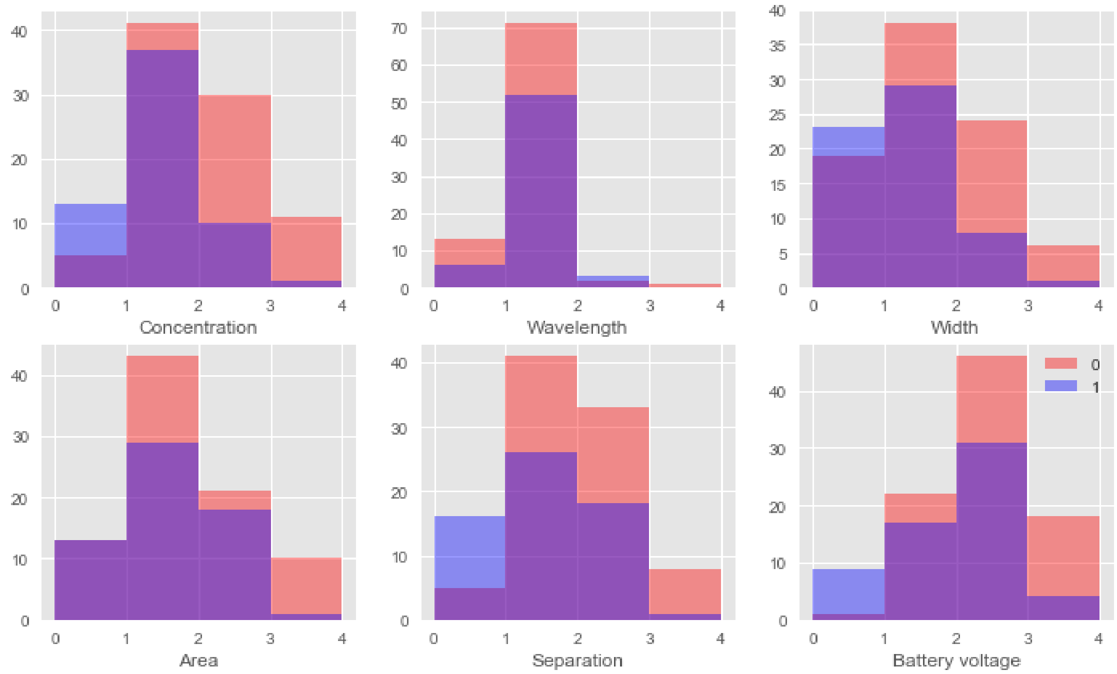
5.3 Picking 2 clusters

```
In [35]: fig,axes = plt.subplots(nrows=len(att_columns)/3+1,ncols=3,figsize=(12,15))
         for ax,col in zip(axes.reshape(-1),att_columns):
             sns.distplot(data[data['cluster_2_label']==0][col],ax=ax,bins=[0,1,2,3,4,5],kde=Fa
             sns.distplot(data[data['cluster_2_label']==1][col],ax=ax,bins=[0,1,2,3,4,5],kde=Fa
         plt.legend()
         fig,axes = plt.subplots(nrows=2,ncols=3,figsize=(12,7))
         for ax,col in zip(axes.reshape(-1),know_columns):
             sns.distplot(data[data['cluster_2_label']==0][col],ax=ax,bins=[0,1,2,3,4],kde=Fals
```

```
        sns.distplot(data[data['cluster_2_label']==1][col],ax=ax,bins=[0,1,2,3,4],kde=Fals
    plt.legend()
```

Out[35]: <matplotlib.legend.Legend at 0xe6ba550>





The blue group (contrary to red): * has higher POCC and think it will be productive but boring * has slightly higher incoming knowledge for Witdh and Concentration, but not much else.

6 Exporting the data

```
In [36]: data.head()
```

```
Out[36]:    [prior_lab] What lab courses are you presently taking or have taken in the past? Cl
         0                                                1.0
         1                                                1.0
         2                                                0.0
         3                                                1.0
         4                                                0.0

            [prior_lab] What lab courses are you presently taking or have taken in the past? Cl
         0                                                1.0
         1                                                1.0
         2                                                1.0
         3                                                1.0
         4                                                1.0

            [prior_lab] What lab courses are you presently taking or have taken in the past? Cl
         0                                                1.0
         1                                                0.0
         2                                                1.0
         3                                                1.0
```

```
4                                               0.0

   [prior_lab] What lab courses are you presently taking or have taken in the past? C
0                                               0.0
1                                               0.0
2                                               0.0
3                                               0.0
4                                               0.0

   similar_L  similar_C  same_L  same_C  prior_number_virtual_labs  perceivedvalue.0-
0          0          0       0       0                          2
1          0          0       0       0                          0
2          0          0       0       0                          1
3          0          1       0       0                          3
4          1          0       0       0                          2

   perceivedvalue.1-productive  perceivedvalue.2-useless  perceivedvalue.3-engaging
0                          2.0                         1                        2.0
1                          4.0                         1                        3.0
2                          2.0                         1                        2.0
3                          3.0                         1                        3.0
4                          3.0                         1                        3.0

   taskinterpretation.0-investigate the basic mechanics of the topic at hand  \
0                                                  3
1                                                  4
2                                                  2
3                                                  3
4                                                  4

   taskinterpretation.1-design my own experiments that can help me understand the top
0                                                  2
1                                                  4
2                                                  2
3                                                  4
4                                                  2

   taskinterpretation.2-memorize information about the topic at hand  \
0                                                  3
1                                                  3
2                                                  2
3                                                  3
4                                                  1

   taskinterpretation.3-complete a certain number of  questions  \
0                                                  2
1                                                  4
2                                                  3
```

```
3                                                4
4                                                3

   taskinterpretation.4-develop scientific reasoning skills  pocc.0-learning the basic
0                                                2.0
1                                                4.0
2                                                2.0
3                                                3.0
4                                                4.0

   pocc.1-testing my ideas and theories  pocc.2-answering given questions  \
0                                     3                                 3
1                                     2                                 2
2                                     3                                 3
3                                     4                                 3
4                                     3                                 3

   pocc.3-memorizing key information  pocc.4-exploring the topic       sid  age  \
0                                  2                           3  77047160   19
1                                  2                           3  23836160   19
2                                  3                           2  64006159   21
3                                  3                           3  24566161   19
4                                  2                           4  46792161   19

   english.0-writing  english.1-reading  gender-Gender non conforming/non-binary  genc
0                  3                  3                                      0.0
1                  2                  2                                      0.0
2                  2                  2                                      0.0
3                  3                  3                                      0.0
4                  3                  3                                      0.0

   gender-Prefer not to answer  gender-Woman                                  major  \
0                          0.0           1.0  *Non science or applied science major
1                          0.0           0.0                      Civil Engineering
2                          0.0           1.0                            *Undeclared
3                          0.0           0.0                    Engineering Physics
4                          0.0           0.0                 Mechanical Engineering

   year-1st year undergraduate  year-2nd year undergraduate  year-3rd year undergradua
0                          1.0                          0.0                          0
1                          1.0                          0.0                          0
2                          0.0                          1.0                          0
3                          1.0                          0.0                          0
4                          1.0                          0.0                          0

   year-4th year undergraduate second sim  sim_index  Area  Battery voltage  Concentra
0                          0.0          C          2   1.0              1.0
1                          0.0          L          2   0.0              3.0
```

```
2                            0.0          C           2   1.0              2.0
3                            0.0          L           2   2.0              3.0
4                            0.0          C           2   3.0              3.0

   Separation  Wavelength  Width  perceivedvalue.0-boring_reversed  \
0         0.0         1.0    1.0                               1.0
1         0.0         1.0    1.0                               2.0
2         1.0         1.0    0.0                               3.0
3         2.0         1.0    2.0                               2.0
4         3.0         1.0    2.0                               2.0

   perceivedvalue.2-useless_reversed    PC1    PC2  PC1_wo_pocc  PC2_wo_pocc  overall
0                                  3  0.997  0.704        1.486        0.107
1                                  3 -1.221 -0.611       -2.315       -1.158
2                                  3  0.975  0.726        0.965        0.167
3                                  3 -1.703  0.744       -1.205       -0.694
4                                  3 -1.232 -0.610       -1.007       -0.404

   incoming_knowledge_L  incoming_knowledge_C  incoming_knowledge_all  incoming_knowle
0                   3.0                   2.0                     5.0
1                   3.0                   3.0                     6.0
2                   3.0                   4.0                     7.0
3                   5.0                   7.0                    12.0
4                   5.0                   9.0                    14.0

   cluster_2_label  cluster_3_label  cluster_4_label  cluster_5_label
0                1                1                1                4
1                0                0                3                2
2                1                1                1                1
3                0                0                2                3
4                0                0                2                3
```

In [37]:
```
export_data = data[['sid','overall POCC','PC1','PC2']].copy()
export_data.to_csv(os.path.join(BIG_FOLDER,'all_massaged_data\\incoming_attitudes.txt
```

7 Picking out students

We want 8 students, 2 for each type: * high attitude (cluster=0, PC1>2), high knowledge (incoming_knowledge_L ==8) * high attitude (cluster=0, PC1>2), low knowledge (incoming_knowledge_L ==3) * low attitude (cluster=1, PC1<-2), low knowledge (incoming_knowledge_L ==3) * low attitude (cluster=1, PC1<-2), high knowledge (incoming_knowledge_L ==8)

On comment the following to save viz for 4 different types of students:

In [38]:
```
# print 'a',data[(data['PC1']>2)&(data['incoming_knowledge_L']>=6)]['sid']
# print 'b',data[(data['PC1']>2)&(data['incoming_knowledge_L']==2)]['sid']
# print 'c',data[(data['PC1']<-2)&(data['incoming_knowledge_L']==3)]['sid']
# print 'd',data[(data['PC1']<-2)&(data['incoming_knowledge_L']>=6)]['sid']
```

```
In [39]: # columns = ['sid',"Concentration","Wavelength","Width","perceivedvalue.0-boring","pe
         # exploration = data[data['sid'].isin([19989152,10537160,13654167,11929166])][columns
         # exploration['Fakename'] = ['Saturn','Tatouine','Ursula','Venus']
         # exploration['knowledge'] = ['low','high','low','high']
         # exploration['incoming_attitude'] = ['low','high','high','low']
         # exploration.sort_values('Concentration',inplace=True)
         # exploration

In [40]: # %aimport utils_timeline_viz
         # matplotlib.style.use('ggplot')
         # matplotlib.rcParams['figure.figsize'] = 25, 15
         # from matplotlib.backends.backend_pdf import PdfPages

         # to_plot_beers = ['Pause','','Log axis','Inverse axis','Linear axis','Other axes','A

         # def save_multipage_viz(students_to_explore):
         #     sim_name = {"beers":"Light absorbance",'capacitor':'Charge'}
         #     with PdfPages('multipage_timeline_viz_{0}.pdf'.format('_'.join([str(n) for n in
         #         for sim,to_plot in [('beers',to_plot_beers)]:#,('capacitor',to_plot_caps)]:
         #             for i,row in students_to_explore.iterrows():
         #                 studentid = row['sid']
         #                 name = row['Fakename']
         #                 att = row['incoming_attitude']+' attitude'
         #                 know = row['incoming_knowledge']+' knowledge'
         #                 filename = find_student_log_file(sim,studentid)
         #                 date = date = re.search(r'\d{7,8}_([\d\-\.\_]+)\.txt', filename).gr
         #                 df = prep_parsing_data(filename)
         #                 plt.figure(figsize=(20,12))
         #                 plt.title("{1} \t {0} \t {2} \t {3}".format(name,sim_name[sim],att,
         #                 plot(df,to_plot,family_name_to_code,function_to_use,colors)
         # #                 plt.show()
         #                 plt.tight_layout()
         #                 pdf.savefig()
         #                 plt.close()

         # save_multipage_viz(exploration)
```