Students' Academic Success Forecast

This paper offers a thorough analysis of 643 students from Forsberg College with 9 variables: gpa, reading, math, composite, gender, level, highschool, ACTcomposite, and STAT2103 (pass, fail, or succeed a class). By calculating sample statistics, creating graphics and models, the results of the analysis determine how well a given student - Bob - will do in STAT2103.

Sample Statistics

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from scipy.cluster.vq import whiten

import sklearn as skl
from sklearn import svm
from numpy import arange

import statistics as ss

### Load data

url="https://raw.githubusercontent.com/nanathecutecow/final/main/forsbergCollege.csv"
dt = pd.read_csv(url)

### Sample Statistics

## Numerical Variables
print("GPA")
gpa = dt[dt.columns[0]]
plt.hist(dt[dt.columns[0]])
plt.xlabel('GPA')
print("Min:", min(gpa))
print("Mean:", round(ss.mean(gpa),2))
print("Quantiles:", ss.quantiles(gpa))
print("Max:", max(gpa))
print()
```

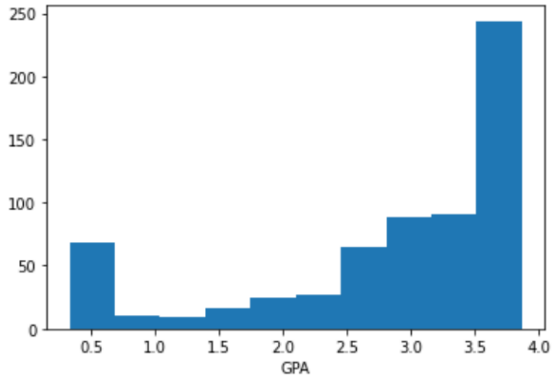
```
print("Reading")
reading = dt[dt.columns[1]]
plt.hist(dt[dt.columns[1]])
plt.xlabel('Reading')
print("Min:", min(reading))
print("Mean:", round(ss.mean(reading),2))
print("Quantiles:", ss.quantiles(reading))
print("Max:", max(reading))
print()

print("Math")
math = dt[dt.columns[2]]
plt.hist(dt[dt.columns[2]])
plt.xlabel('Math')
print("Min:", min(math))
print("Mean:", round(ss.mean(math),2))
print("Quantiles:", ss.quantiles(math))
print("Max:", max(math))
print()

print("SATcomposite")
sat = dt[dt.columns[3]]
plt.hist(dt[dt.columns[3]])
plt.xlabel('SATcomposite')
print("Min:", min(sat))
print("Mean:", round(ss.mean(sat),2))
print("Quantiles:", ss.quantiles(sat))
print("Max:", max(sat))
print()

print("ACTcomposite")
ACT = dt[dt.columns[7]]
plt.hist(dt[dt.columns[7]])
plt.xlabel('ACTcomposite')
print("Min:", min(ACT))
print("Mean:", round(ss.mean(ACT),2))
print("Quantiles:", ss.quantiles(ACT))
print("Max:", max(ACT))

## Categorical Variables
print(dt['gender'].value_counts())
print()
```

```
print(dt['level'].value_counts())
print()
print(dt['highschool'].value_counts())
print()
print(dt['STAT2103'].value_counts())
```

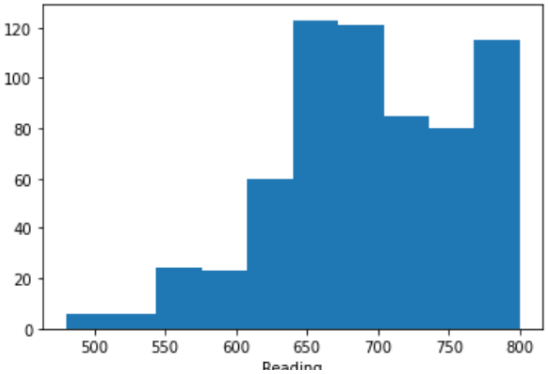
The data includes 5 numerical variables, the summary statistics of which are shown below.

	Min	Q1	Median	Mean	Q2	Max
GPA	0.33	2.66	3.33	2.88	3.66	3.87
Math	300	580	630	626.27	680	800
Reading	480	650	700	695.01	750	800
SAT	780	1250	1330	1321.28	1410	1600
ACT	6	18	21	21	24	36

GPA covers a wide range (0.33 to 3.87), but it is highly asymmetrical. With a difference of 0.45 between the median and the mean, the variable is skewed to the left. However, since GPA is a significant indicator for the overall valuation of a student's academics, we normalize the data to use it as the first predictor variable for our analysis.

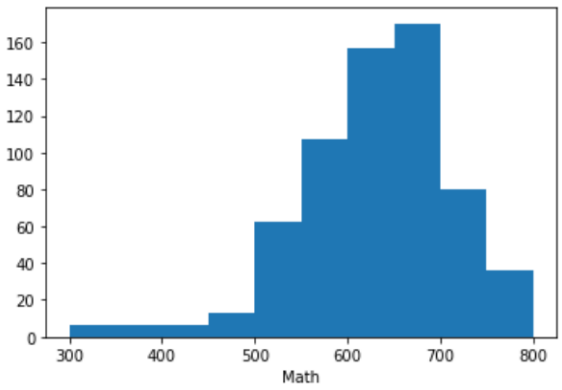


The SAT Reading score also has a wide range from 480 to 800 out of 800. Although the data is not normally distributed, a median of 700 and a mean of 695.1 shows a more symmetrical distribution in the data with a left

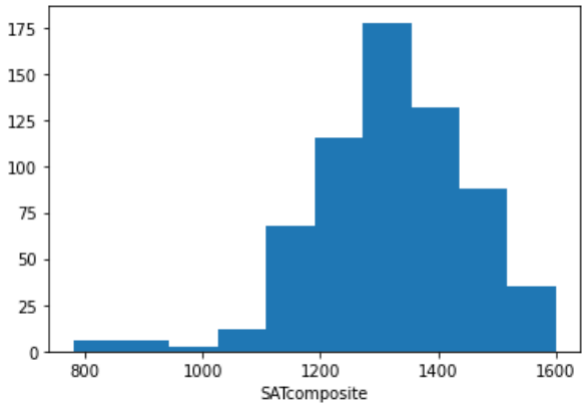


skewness.

The histogram of the SAT Math score shows a bell-shaped distribution of the variable. The data ranges from 300 to 800 and has the mean nearly equal to the median (626.27 and 600), which clearly shows the variable's normality.

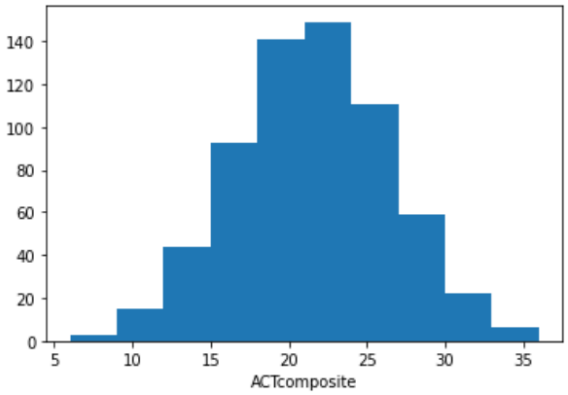


The SAT Composite variable also has a normal distribution with a mean of 1321.28 and a median of 1330, the small difference between which indicates a light left skewness in the data.



Regarding the three correlated variables SAT Reading, SAT Math, and Sat Composite, SAT Math has a comparatively normal distribution with the smallest difference between the mean and the medium. Moreover, mathematical skills lay the foundation for statistics, so we choose SAT Math as the second predictor variable for our later analysis.

ACT composite ranges from 6 to 36. It has a perfectly normal distribution with the mean and medium both represented at the peak of the bell-shaped curve (21). For this reason, ACT composite serves as our third indicator variable.



The statistics of the 4 categorical variables in the data are included in the tables below.

GENDER	Male	380
	Female	263

STAT2103	Pass	371
	Succeed	161
	Fail	111

LEVEL	Freshman	22
	Sophomore	136
	Junior	291
	Senior	194

HIGHSCHOOL	Home School	523
	Private High School	72
	Public High School	28
	Transfer	20

These categorical variables are unevenly distributed. However, regarding the already given data of our subject that Bob is a male senior at a private highschool, we will only take STAT2103 into consideration among the 4 categorical variables.

Therefore, we conduct our data analysis with the following 4 variables: GPA, SAT Math, ACT Composite, and STAT2103.

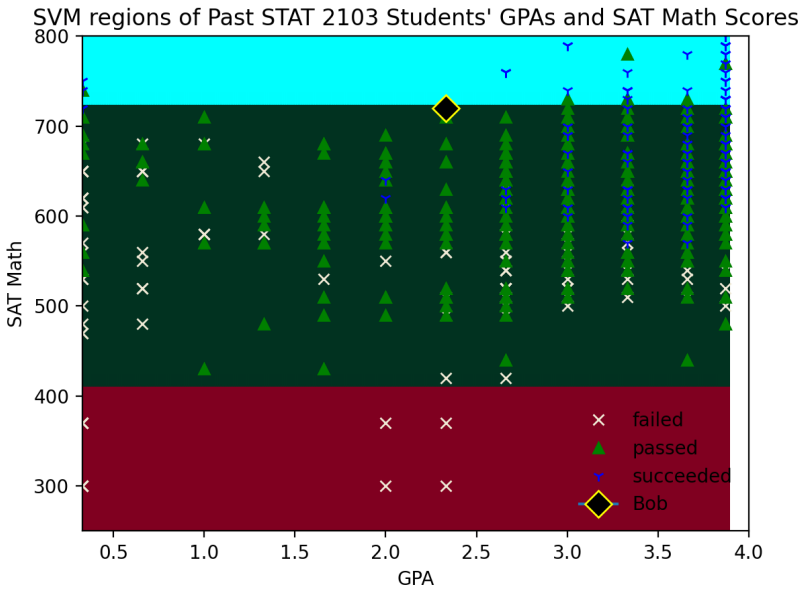
Data Analysis and Prediction

We applied SVM both to generate our prediction and to create the SVM regions for the cluster graph of students' GPAs and SAT Math results:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from scipy.cluster.vq import whiten

import sklearn as skl
from sklearn import svm
from numpy import arange
### Load data

url  = "https://courses.kvasaheim.com/stat195/project/forsbergCollege.csv"
dt = pd.read_csv(url)

n = dt.shape[0]

dx = dt.loc[:, ['gpa', 'math']]
tt = dt[dt.columns[8]]
dv = dx.values

xx = dv

plt.xlabel("GPA")
plt.ylabel("SAT Math")

plt.xlim([0.33,4.0])
plt.ylim([250,800])

sp0 = (tt=="Fail")
sp1 = (tt=="Pass")
sp2 = (tt=="Succeed")
svmModel = svm.SVC()
svmModel.fit(dv, tt)
for xval in arange(0.33,3.87,0.01):
   for yval in arange(250,800,1):
       esp = svmModel.predict([[xval, yval]])
```

```
        if(esp=='Fail'):     gCol="#800020"
        if(esp=='Pass'): gCol="#013220"
        if(esp=='Succeed'):  gCol="#00FFFF"

        plt.plot( xval, yval, color=gCol,  marker="s", linestyle="None" )

plt.plot( xx[sp0,0], xx[sp0,1], color="#f0ead6", marker="x",
linestyle="None", label="failed" )
plt.plot( xx[sp1,0], xx[sp1,1], color="green", marker="^",
linestyle="None", label="passed" )
plt.plot( xx[sp2,0], xx[sp2,1], color="blue", marker="1",
linestyle="None", label="succeeded" )
plt.plot(2.33, 720, marker="D", markersize=10, markeredgecolor="yellow",
markerfacecolor="black", label="Bob")
print( svmModel.predict([[2.33, 720]]) )



plt.legend(frameon=False)
plt.title("SVM regions of Past STAT 2103 Students' GPAs and SAT Math
Scores")

plt.show()

plt.close()
```

The output was the following graph:



We can see that Bob sits almost exactly between the “Success” and “Passing” SVM regions, but leaning slightly more towards the “Passing” region, which is what the SVM predicts when trained on GPA and SAT Math data.

Further Data Analysis and Prediction

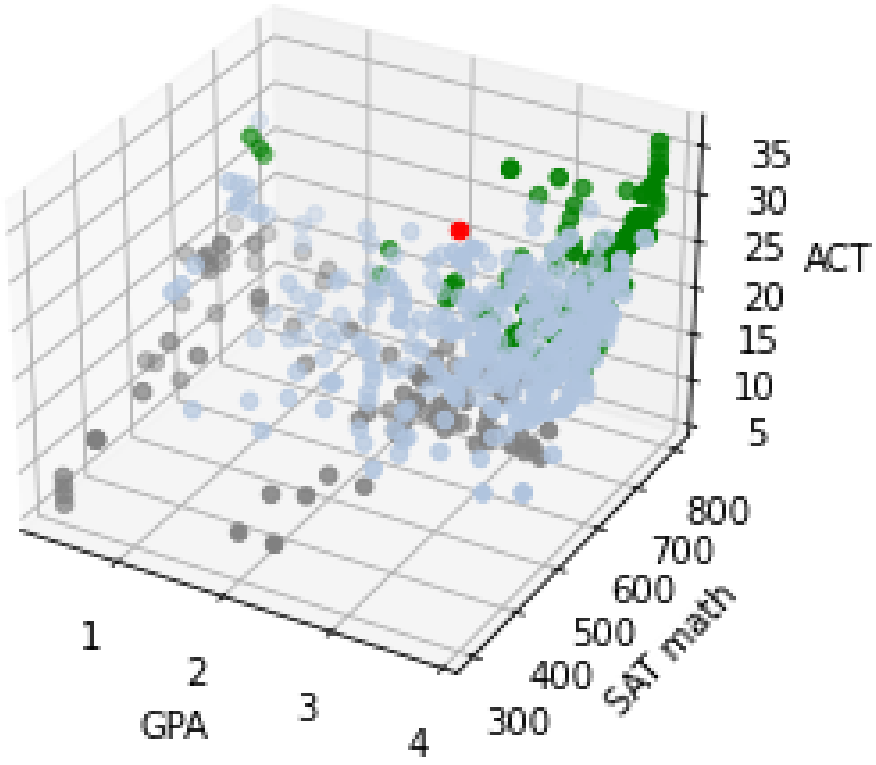
We conduct the following steps to acquire the analysis of the data and prediction for how Bob will be doing in STAT2103 regarding all 3 indicator variables GPA, SAT Math, and ACT Composite:

- Import the library
- Get the data
- Cluster into 3 categories
- Draw the graph
- Predict by using SVM

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.vq import kmeans,vq,whiten
import sklearn as skl
from sklearn import svm
```

```
dt=pd.read_csv('https://raw.githubusercontent.com/nanathecutecow/final/main/forsbergCollege.csv')
n = dt.shape[0]
dv = dt.loc[:, ['gpa','math', 'ACTcomposite']]
tt = dt[dt.columns[8]]

# get the data each category: Pass, Fail, Succeed
p = dt[dt['STAT2103'] == 'Pass']
f = dt[dt['STAT2103'] == 'Fail']
s = dt[dt['STAT2103'] == 'Succeed']
# get value each category: Pass, Fail, Succeed
p1 = p.loc[:, ['gpa','math', 'ACTcomposite']]
f1 = f.loc[:, ['gpa','math', 'ACTcomposite']]
s1 = s.loc[:, ['gpa','math', 'ACTcomposite']]
#####
p2 = p1.values
f2 = f1.values
s2 = s1.values

# graph 3D
ax = plt.axes(projection='3d')
ax.set_xlabel('GPA')
ax.set_ylabel('SAT math')
ax.set_zlabel('ACT')

ax.scatter3D(p2[:,0], p2[:,1], p2[:,2], c ="lightsteelblue", marker ='o');
ax.scatter3D(f2[:,0], f2[:,1], f2[:,2], c ="gray", marker ='o');
ax.scatter3D(s2[:,0], s2[:,1], s2[:,2], c ="green", marker ='o');
#graph BOB
ax.scatter3D(2.33, 720, 25, c ="red", marker ='o');

#prediction
svmModel = svm.SVC()
svmModel.fit(dv, tt)
print( svmModel.predict([[2.33, 720, 25]]) )
```

The output was the following graph:



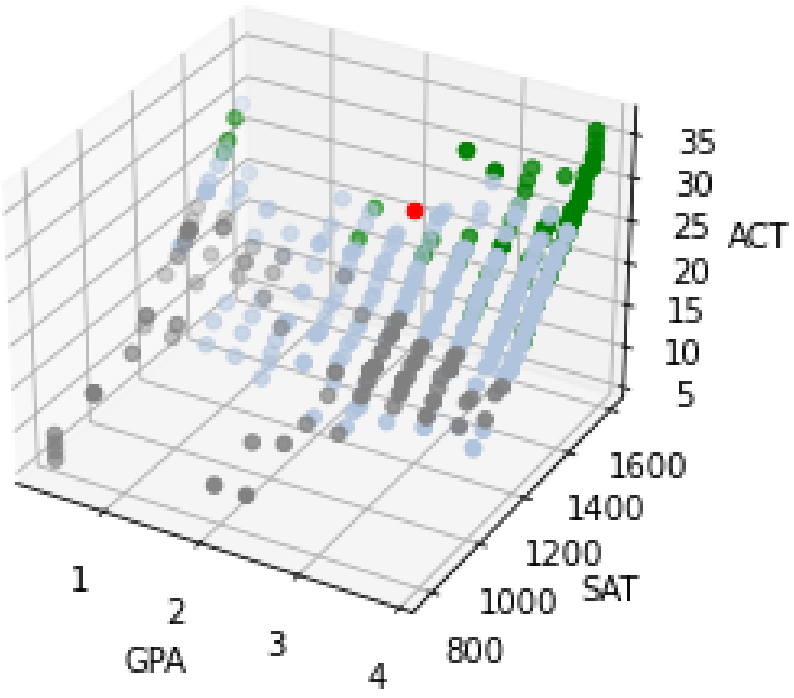
From the graph, we can see that Bob is in either the [pass] cluster or the [succeed] cluster.

After running the SVM, we can conclude that the prediction is correct: Bob is likely to PASS the course.

HOWEVER: If we used the **SAT,ACT,GPA** instead of the **SATmath,ACT,GPA**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.vq import kmeans,vq,whiten
import sklearn as skl
from sklearn import svm
dt=pd.read_csv('https://raw.githubusercontent.com/nanathecutecow/fina
l/main/forsbergCollege.csv')
n = dt.shape[0]
dv = dt.loc[:, ['gpa','composite', 'ACTcomposite']]
tt = dt[dt.columns[8]]

# get the data each category: Pass, Fail, Succeed
p = dt[dt['STAT2103'] == 'Pass']
f = dt[dt['STAT2103'] == 'Fail']
s = dt[dt['STAT2103'] == 'Succeed']
# get value each category: Pass, Fail, Succeed
p1 = p.loc[:, ['gpa','composite', 'ACTcomposite']]
f1 = f.loc[:, ['gpa','composite', 'ACTcomposite']]
s1 = s.loc[:, ['gpa','composite', 'ACTcomposite']]
#####
p2 = p1.values
f2 = f1.values
s2 = s1.values

# graph 3D
ax = plt.axes(projection='3d')
ax.set_xlabel('GPA')
ax.set_ylabel('SAT')
ax.set_zlabel('ACT')

ax.scatter3D(p2[:,0], p2[:,1], p2[:,2], c ="lightsteelblue", marker
='o');
ax.scatter3D(f2[:,0], f2[:,1], f2[:,2], c ="gray", marker ='o');
ax.scatter3D(s2[:,0], s2[:,1], s2[:,2], c ="green", marker ='o');
#graph BOB
ax.scatter3D(2.33, 1470, 25, c ="red", marker ='o');
```

```
#prediction
svmModel = svm.SVC()
svmModel.fit(dv, tt)
print( svmModel.predict([[2.33, 1470, 25]]) )
```



The graph is more stable.

From the graph, we can see that Bob is still in either the [pass] cluster or the [succeed] cluster.

SVM: Bob is surprisingly predicted to be [Succeed] in the course.