Persona Modeling

By:- Yumeng Guo
Vineet Balan
Bingjun Wang
Navneet Poddar

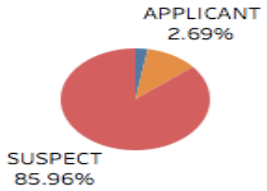
NYU | TANDON SCHOOL OF ENGINEERING

Table of Contents

1. Data Visualization
2. Challenges with Dataset
3. Data Cleaning
4. Persona Modeling
5. Budget Allocation

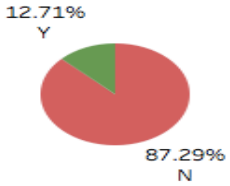
1. Data Visualization

Status 1

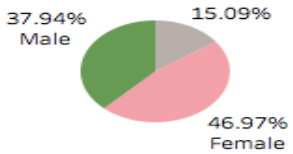


GPA

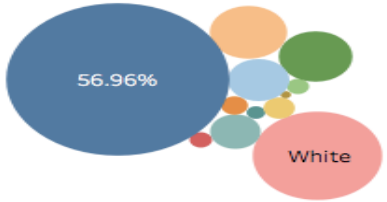
In State



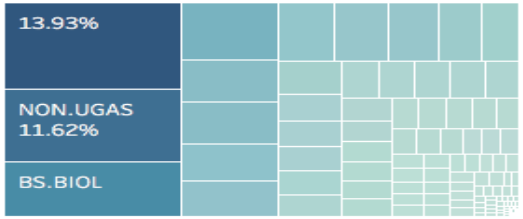
Gender



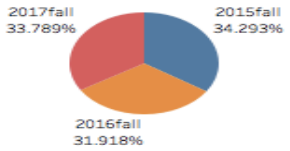
Race



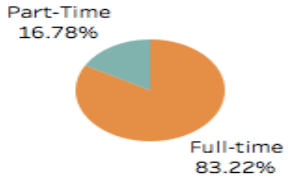
Program



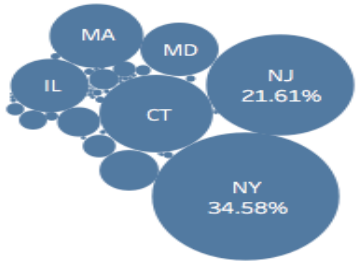
Term



Student Type



State



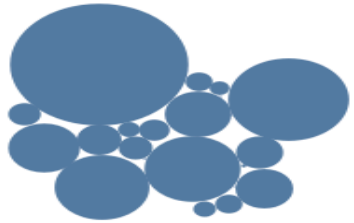
Religion

Religion	
Null	93.38%
Adventist	0.05%
African Methodist	0.04%
Assemblies of God	0.04%
Baptist	0.29%
Charismatic/Pentecostal	0.04%
Christian	0.78%
Episcopal	0.01%
Evangelical	0.02%
Friends/Quaker	0.01%
Household Religion - Cath..	3.43%

SAT Score

SAT Score	
Null	94.00%
990 - 1040	0.00%
930 - 980	0.03%
930 - 1070	3.12%
1410 - 1460	0.00%
1360 - 1530	0.34%
1350 - 1400	0.00%
1290 - 1340	0.00%
1230 - 1280	0.00%
1170 - 1220	0.00%
1110 - 1160	0.00%
1080 - 1350	2.51%
1050 - 1100	0.00%

Source



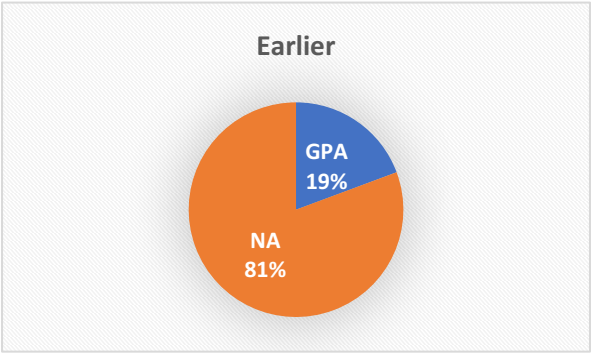
2. Challenges with Dataset

- 80% Null Values in GPA
- 94% Null Values in SAT Score
- 93% Null Values in Religion
- 57% Null Values in Race
- Formatting Problem in Zip code
- Many Outliers in Income Variable

3. Data Cleaning

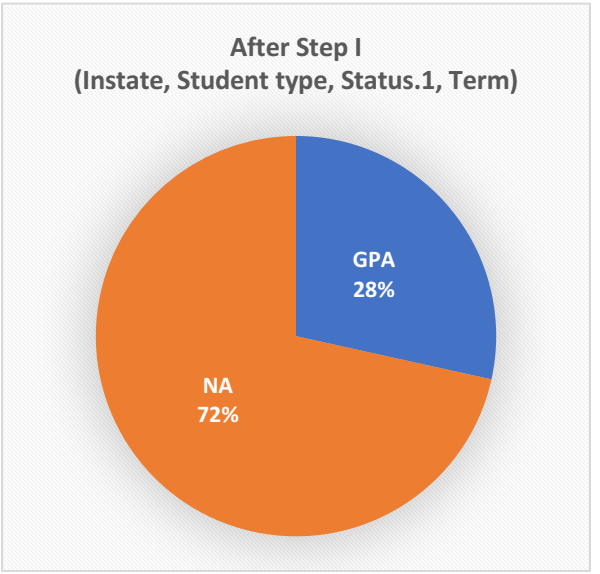
❑Null Value Replacement in GPA

```
> summary(uni$GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    2.0     3.0     3.0     3.1     3.3     4.0  538896
```



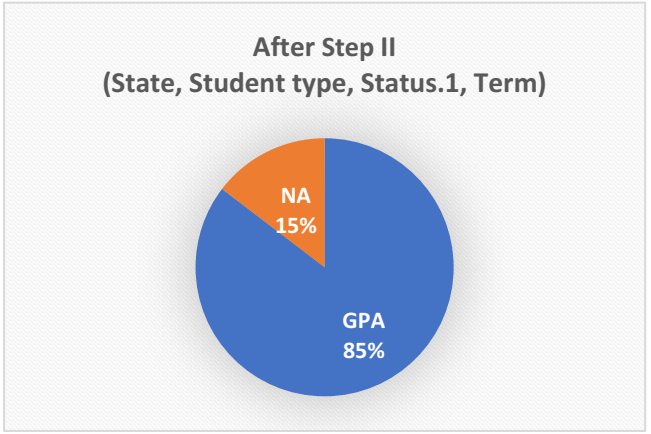
Step I : Instate, Student type, Status.1, Term

```
> #1.InState Y, Full time, 2015, Applicant
> iNgpa1 <- which(uni$InState == "Y" & uni$StudentType == "Full-time" & uni$Term == "2015fall" &
uni$Status.1 == "APPLICANT" & is.na(uni$GPA))
> iYgpa1 <- which(uni$InState == "Y" & uni$StudentType == "Full-time" & uni$Term == "2015fall" &
uni$Status.1 == "APPLICANT" & !is.na(uni$GPA))
> uni$GPA[iNgpa1]=mean(uni$GPA[iYgpa1])
> summary(uni$GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    2.0     3.0     3.0     3.1     3.3     4.0  538896
> #18.InState Y, Part time, 2017, Suspect
> iNgpa18 <- which(uni$InState == "Y" & uni$StudentType == "Part-Time" & uni$Term == "2017fall" &
uni$Status.1 == "SUSPECT" & is.na(uni$GPA))
> iYgpa18 <- which(uni$InState == "Y" & uni$StudentType == "Part-Time" & uni$Term == "2017fall" &
uni$Status.1 == "SUSPECT" & !is.na(uni$GPA))
> uni$GPA[iNgpa18] = mean(uni$GPA[iYgpa18])
> summary(uni$GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    2.0     3.0     3.0     3.1     3.2     4.0  478773
```



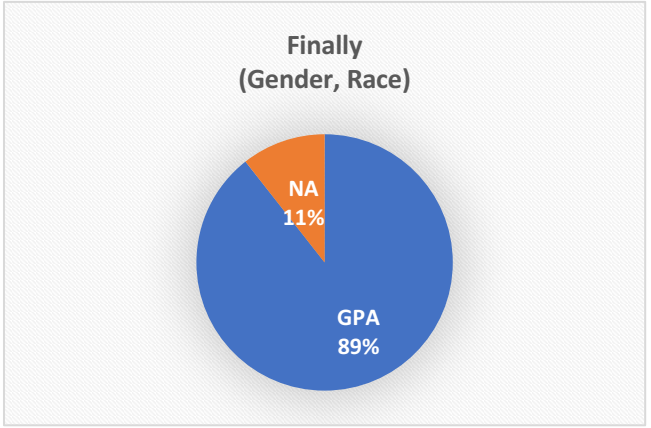
Step II: State (NY,NJ,MA MD, IL), Student type, Status.1, Term

```
> #NEW YORK
> #1.State NY, Full time, 2015, Applicant
> iNgpa1 <- which(uni$State == "NY" & uni$StudentType == "Full-time" & uni$Term == "2015fall" & uni$Status.1 == "APPl
ICANT" & is.na(uni$GPA))
> iYgpa1 <- which(uni$State == "NY" & uni$StudentType == "Full-time" & uni$Term == "2015fall" & uni$Status.1 == "APPl
ICANT" & !is.na(uni$GPA))
> uni$GPA[iNgpa1]=mean(uni$GPA[iYgpa1])
> summary(uni$GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    2.0     3.0     3.0     3.1     3.2     4.0  475772
> #18.State IL, Part time, 2017, Suspect
> iNgpa18 <- which(uni$State == "IL" & uni$StudentType == "Part-Time" & uni$Term == "2017fall" &
uni$Status.1 == "SUSPECT" & is.na(uni$GPA))
> iYgpa18 <- which(uni$State == "IL" & uni$StudentType == "Part-Time" & uni$Term == "2017fall" &
uni$Status.1 == "SUSPECT" & !is.na(uni$GPA))
> uni$GPA[iNgpa18] = mean(uni$GPA[iYgpa18])
> summary(uni$GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2.00    3.02    3.08    3.09    3.12    4.00   97814
```



Final Step: Gender, Race

```
> #Male, American Indian/Alaskan Native
> MNgpa4 <- which(uni$Gender == "Male" & uni$Race == "American Indian/Alaskan Native" & is.na(uni$GPA))
> MYgpa4 <- which(uni$Gender == "Male" & uni$Race == "American Indian/Alaskan Native" & !is.na(uni$GPA))
> uni$GPA[MNgpa4]=mean(uni$GPA[MYgpa4])
> summary(uni$GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2.00    3.02    3.08    3.09    3.12    4.00   70878
```



❑Data Augmentation

External Dataset = Average Household Income, State wise (*Source: https://factfinder.census.gov*)

```
> avg.HH<-read.csv(file ="E:\\BA assingment\\Project 1\\Household average income.csv", header = T, stringsAsFactors = F)
> uni<- merge(x=uni1, y=avg.HH,by ="State",all=T)
> low<-which(uni$HouseholdIncome<0.8*uni$Household.average.income)
> uni["class"] <- NA
> uni$class[low]="LOW"
> High<-which(uni$HouseholdIncome>1.2*uni$Household.average.income)
> uni$class[High]="HIGH"
> Mid<-which(uni$HouseholdIncome<=1.2*uni$Household.average.income & uni$HouseholdIncome>=0.8*uni$Household.average.income)
> uni$class[Mid]="MID"
```

Final Outcome :-

InState	Zipcode	Race	Religion	Term	StudentType	DistancetoCampus_miles	HouseholdIncome	Household.average.income	class
N	'01810	White		2017fall	Full-time	147.35864	165602	97295	HIGH
N	'01105			2016fall	Full-time	71.90244	24153	97295	LOW
N	'01960			2015fall	Full-time	148.22974	80713	97295	MID
N	'01915			2017fall	Full-time	153.40307	117447	97295	HIGH

❑Zip code Formatting

Earlier

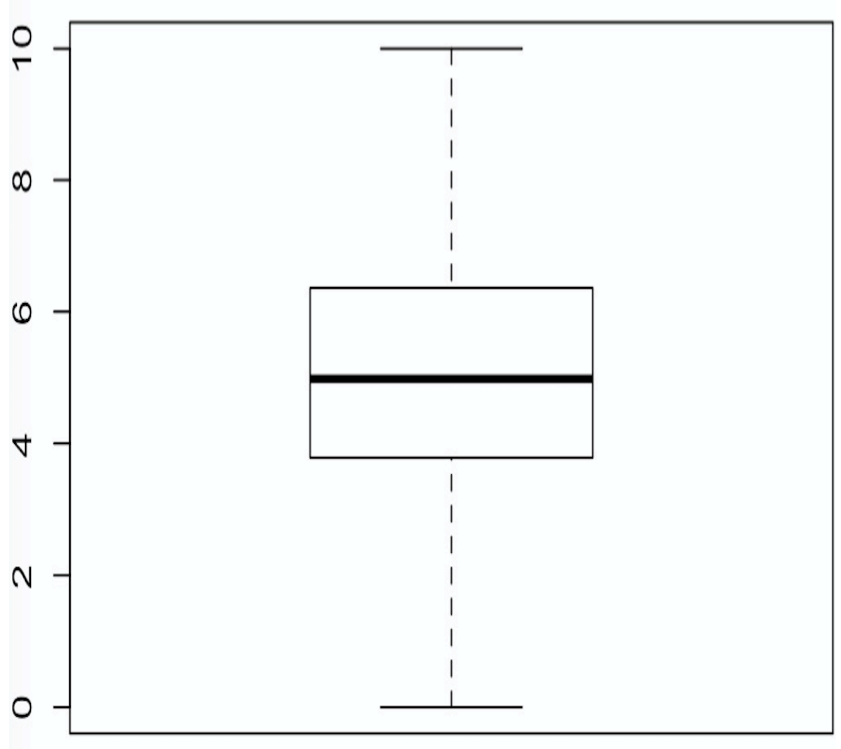
```
> unique(uni$Zipcode)
   [1] "No Data" "'06468"  "'10454"  "'06606"  "'11210"  "'11233"  "'00641"  "'14224"  "'06095"  "'07105"  "'00382"
  [17] "'01015"  "'02194"  "'00693"  "'00730"  "'00976"  "'02209"  "'43200"  "'10000"  "'63303"  "'09034"  "'04405"
  [33] "'01023"  "'16725"  "'47800"  "'00716"  "'39114"  "'11371"  "'65000"  "'00971"  "'00677"  "'10220"  "'10250"
  [49] "'81100"  "'99999"  "'10230"  "'47500"  "'04313"  "'15811"  "'11110"  "'10530"  "'10500"  "'21110"  "'10130"
  [65] "'91204"  "'02750"  "'03187"  "'02191"  "'10240"  "'03104"  "'10600"  "'01219"  "'10140"  "'00953"  "'20019"
  [81] "'00791"  "'00717"  "'12000"  "'00778"  "'00792"  "'70000"  "'04024"  "'10300"  "'28003"  "'00234"  "'04033"
```

After

```
> NewZip<-gsub(pattern = "'", replacement = "", uni$Zipcode)
> unique(NewZip)
   [1] "No Data" "06468"   "10454"   "06606"   "11210"   "11233"   "00641"   "14224"   "06095"   "07105"   "00382"
  [12] "50023"   "01807"   "11333"   "60121"   "00725"   "01015"   "02194"   "00693"   "00730"   "00976"   "02209"
  [23] "43200"   "10000"   "63303"   "09034"   "04405"   "01141"   "02040"   "10400"   "05320"   "33013"   "01023"
  [34] "16725"   "47800"   "00716"   "39114"   "11371"   "65000"   "00971"   "00677"   "10220"   "10250"   "20150"
```

❑Distance Normalization

```
> uni_withoutinf<-subset(uni, is.finite(uni$DistancetoCampus_miles))
> Distance_withoutinf<-subset(uni_withoutinf$DistancetoCampus_miles,uni_withoutinf$DistancetoCampus_miles<85&uni_withoutinf$DistancetoCampus_miles>8)
> normalization<-function(x){
+   return((x-min(x))/(max(x)-min(x)))}
> Nordata<-normalization(Distance_withoutinf)*10
> boxplot(Nordata)
> near<- which(uni_withoutinf$DistancetoCampus_miles<8)
> uni$Normalization[near]="NEAR"
> unique(uni$Normalization)
[1] NA     "NEAR"
> far<-which(uni_withoutinf$DistancetoCampus_miles>85)
> uni$Normalization[far]="FAR"
> unique(uni$Normalization)
[1] NA     "NEAR" "FAR"
> near.1<-which(Nordata<3)
> uni$Normalization[near.1]="NEAR"
> medium<-which(Nordata>=3&Nordata<=7)
> uni$Normalization[medium]="MEDIUM"
> unique(uni$Normalization)
[1] "NEAR"   "MEDIUM" "FAR"    NA
> far.1<- which(Nordata>7)
> uni$Normalization[far.1]="FAR"
> summary(uni$Normalization)
   Length     Class      Mode
   669304 character character
> nrow(uni)
[1] 669304
> unique(uni$Normalization)
[1] "NEAR"   "MEDIUM" "FAR"    NA
```



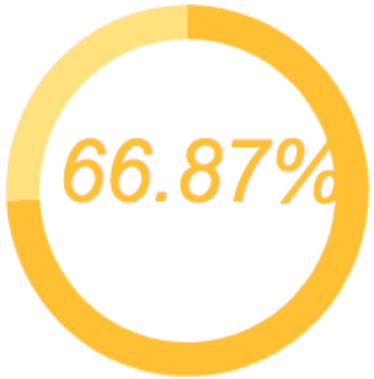
4. Persona Modeling

Will Rooney

Undergraduate

Bio

Will is a young exuberant candidate who in on the hunt for a university for pursuing her undergraduate degree.



66.87% Will

23.23% Other

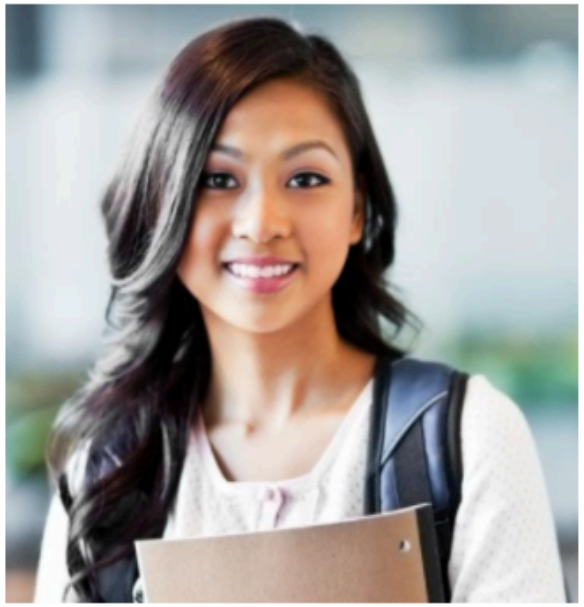
Challenges

- Flexibility in choosing courses.
- Getting admitted into desired course in the desired term..
- Distance of University from home.
- Food.

Motivation

- Flexibility in Choosing courses after the First year.
- Events organized specially for undergraduate students
- Health care resources.
- Jobs and internship opportunities after graduation.
- Campus Life.

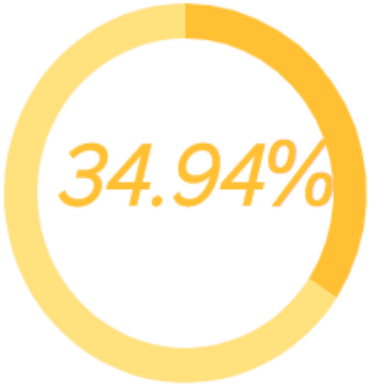
Stacy Anderson



Outstate | Female | Full-Time

Bio

Stacy is a young exuberant female who is on the hunt for a university for pursuing her higher education. She intends to be a full-time student but does not belong to Connecticut.



34.94% Stacy

65.06% Other

Challenges

- Getting admitted into her desired course and desired term.
- Finding the university of her choice to apply.
- Safe Neighborhood around university.
- Accommodation near the university

Motivation

- Women In Engineering clubs.
- Womentorship Programs.
- Women in STEM benefits.
- Self Defense Programs.

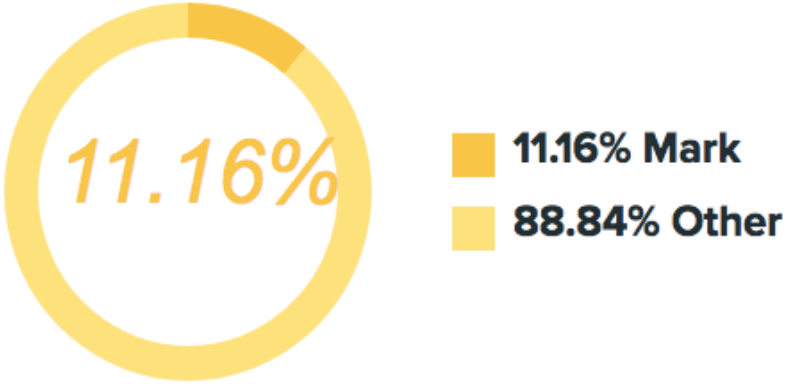
Mark Rhee

Part-Time

Distance from University < 43 Miles

Bio

Mark is a working professional who is looking to enroll into the university as a Part-time student to complete his degree. He stays within 43 Miles of the university.



Challenges

- Flexibility in Class timings
- Low interaction with professors due to schedule.
- Distance of University from his house and place of work

Motivation

- Flexibility in choosing courses.
- Lectures to be held online or on weekends.
- Experienced Faculty for thesis.
- Special Professor availability for Part-Time students.

Kim Betty

Applicant | Full-Time | GPA > 3

Average Household income - Low

Bio

Kim is a blossoming student who seeks to take a step towards her passion for higher studies. She intends to be a Full-Time student and has applied to the university with a GPA of greater than three. Her average household income is lower than the average income of her state.

1.13% Kim

98.87% Other

Challenges

- Financing her education due to financial condition of her household.
- Getting admitted into desired course in the desired term.
- Accommodation near the university.
- Jobs and internships during education.

Motivation

- Scholarship which covers a part of tuition fee.
- Course-work according to industry demand.
- Career fair.
- Special Events On-campus.

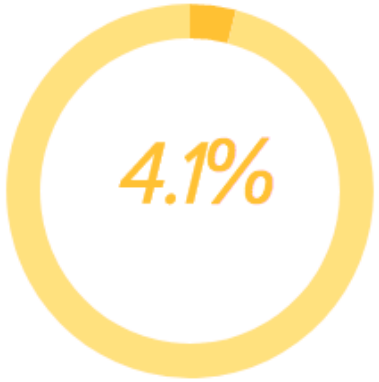
Steve Smith

Prospect | Full-Time | GPA > 3

Average Household income - Low

Bio

Steve is a passionate young student who is looking forward to taking his next step towards his Higher Education. He intends to be a Full-time student but is a prospective candidate for the university with a GPA greater than 3 and an average household income lower than the average household income of his state.



4.1% Steve

95.9% Other

Challenges

- Finding the university of his choice to apply.
- Financing his education due to financial condition of her household.
- Getting admitted into desired course in the desired term.
- Accommodation near the university.
- Jobs and internships during education.

Motivation

- Scholarship which covers a part of tuition fee.
- Course-work according to industry demand.
- Career fair.
- Special Events On-campus.

Alex D'cruz

Suspect | Full-Time | GPA > 3

Average Household income - Mid

Bio

Alex is a jovial high school pass-out who is in search of a university for pursuing his higher education He intends to be a full-time student but does not know about our university and is a suspect candidate He has a GPA of greater than 3 and an average income lower than the average household income of his state.

20.02% Alex

79.98% Other

Challenges

- Finding the university of his choice to apply..
- Getting admitted into desired course in the desired term.
- Accommodation near the university.
- Jobs and internships during education.

Motivation

- Course-work according to industry demand.
- Career fair.
- Life on Campus
- Special Events On-campus.
- University Location

5. Budget Allocation

Persona	Allocation of Budget
Total Budget	**800000$**
Will Rooney	360000$
Stacy Anderson	187500$
Mark Rhee	60525$
Kim Betty	6052$
Steve Smith	21750$
Alex Dcruz	108600$
Reserve Fund	50000$

Thank You!