DATA CLEANING

Entrepreneurial Competency in University Students

Alix ,Chahra and Mario

Content

- 1- Cleaning the Data
- 2- Exporting the Data into CSV File
- 3- Conducting a query on mysql
- 4- Some Insight
- 5- Challenges

Macro/ Micro Cleaning

Macro Level

1- Look at the raw file for preliminary insight

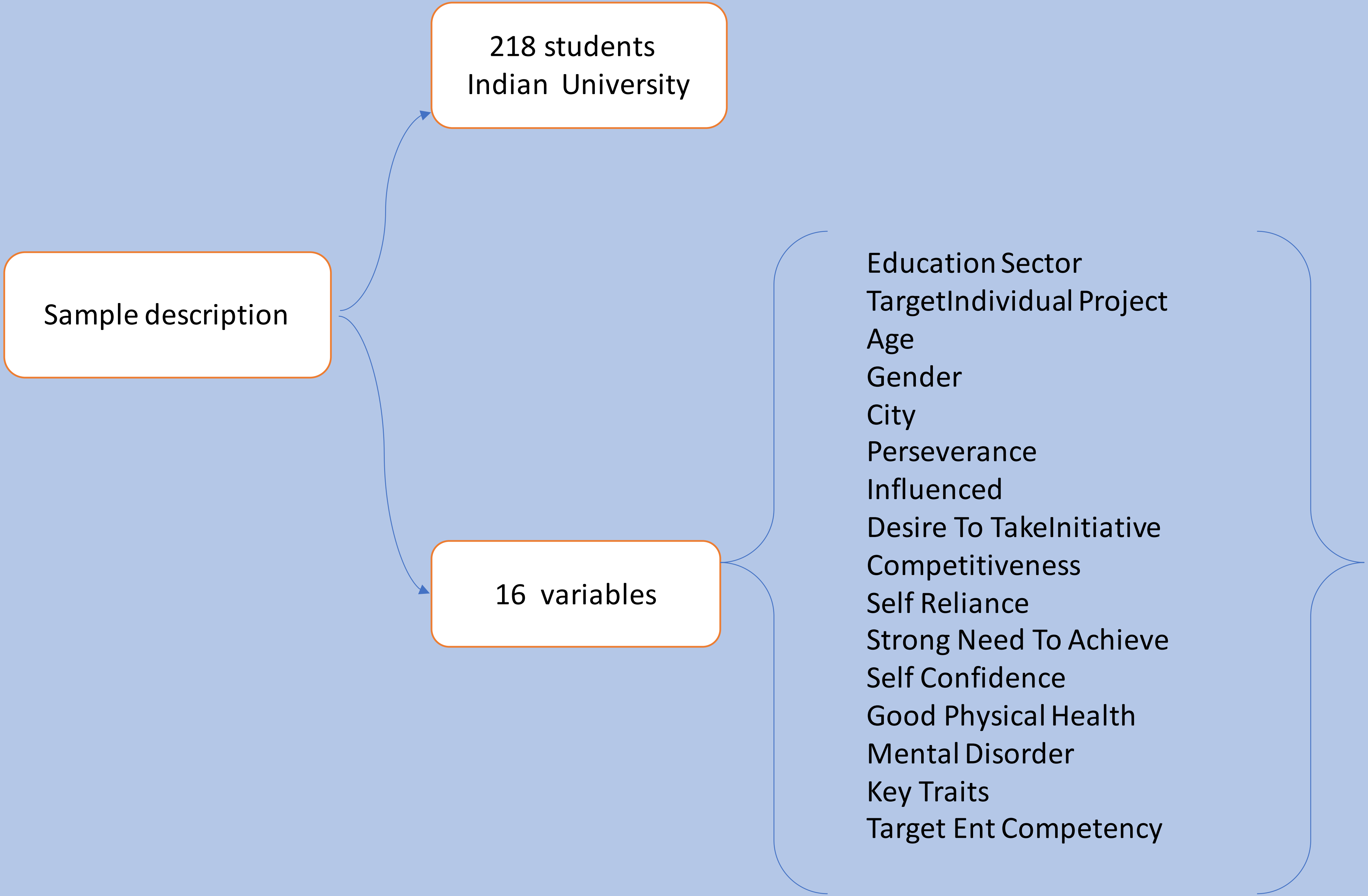
2- Fix structural errors:

a-

```
to_lower_letter=[‘TargetIndividualProject','Gender','City','Influenced','MentalDisorder','ReasonsForLack','ReasonsForLack']
for i in to_lower_letter:
    data[i]=data[i].str.lower()
```

b-

```
data.rename(columns={'Target-ent_competency':'TargetEntCompetency'}, inplace=True)
```



Sample description

218 students
Indian University

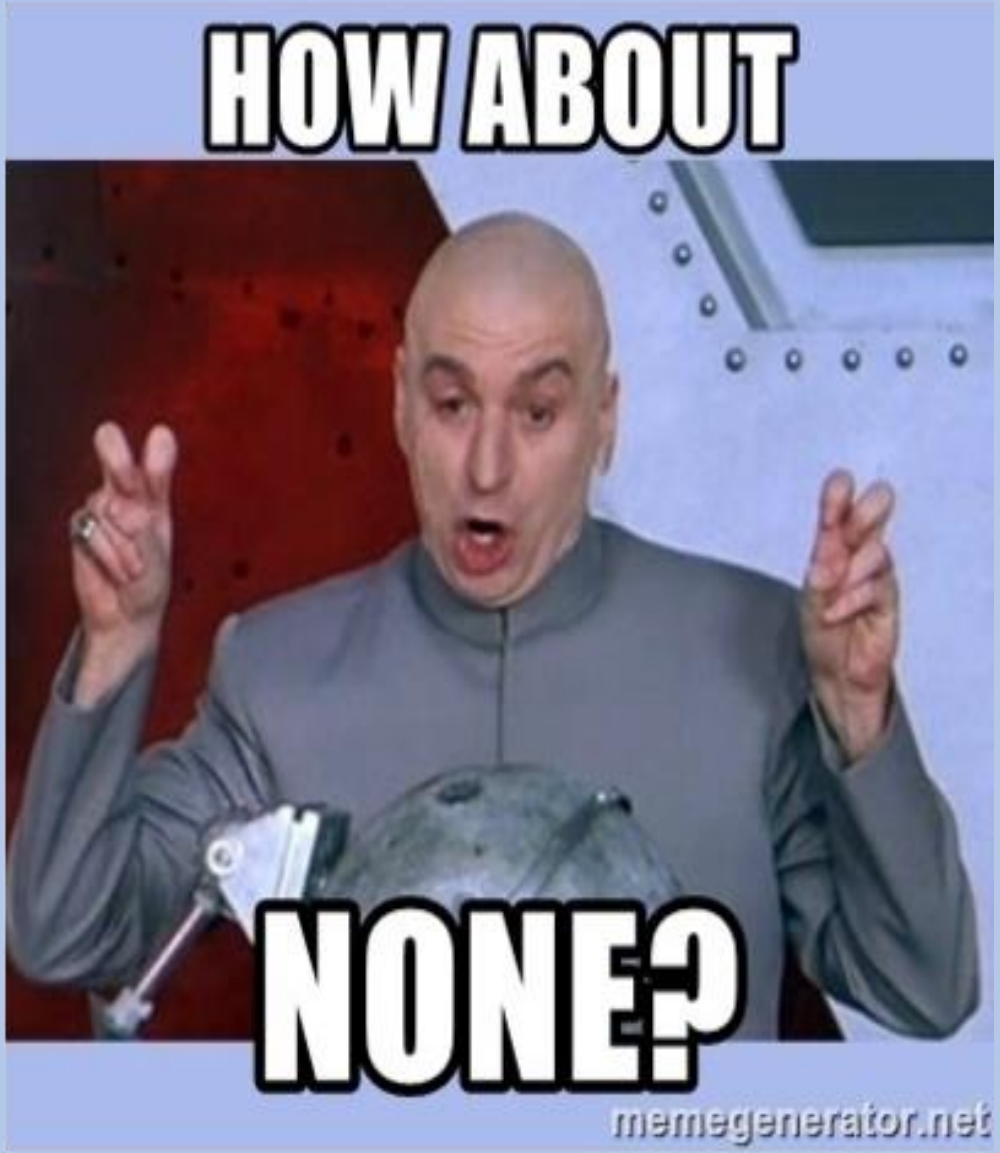
16 variables

- Education Sector
- TargetIndividual Project
- Age
- Gender
- City
- Perseverance
- Influenced
- Desire To TakeInitiative
- Competitiveness
- Self Reliance
- Strong Need To Achieve
- Self Confidence
- Good Physical Health
- Mental Disorder
- Key Traits
- Target Ent Competency

Micro Level Cleaning

Look for Outlier

- Cliquez pour ajouter un texte



Deal with missing Data

- data.loc[data['ReasonsForLack'].isna()==True,'ReasonsForLack']='no reason given’

- data.loc[data[‘MentalDisorder'].isna()==True,'MentalDisorder']='undisclosed

- data.loc[data['Age'].isna()==True,'Age']=round(data['Age'].mean())

- data['Age'].mean()

- #putting mean age to fill the empty age data

-
  ```
  data.loc[data[‘Age'].isna()==True,'Age']=round(data['Age'].mean())
  data[‘Age’].mean()
  ```

- #dropping unnecessary columns

  ```
  data.drop(columns="ReasonsForLack", inplace=True)
  ```

Uploading data Into the computer

```
data.to_csv(r'/Users/naitsaidifariza/Desktop/
data_cleaned.csv')
```

1 0 0
0 1 0
0 0 1

Get Dummies Encoder

- pd.get_dummies(data[“Gender”])

-pd.get_dummies(data[“EducationSector"])

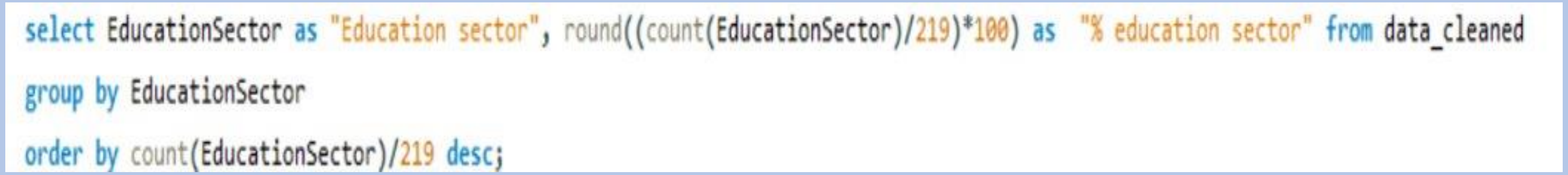
-

SKlearn

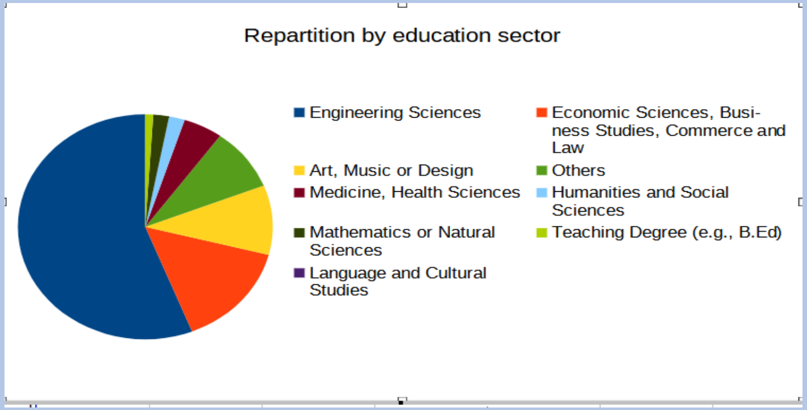
- **From sklearn.preprocessing import LabelEncoder**
- City/ Influenced/ Target Individual Project/ Mental Disorder

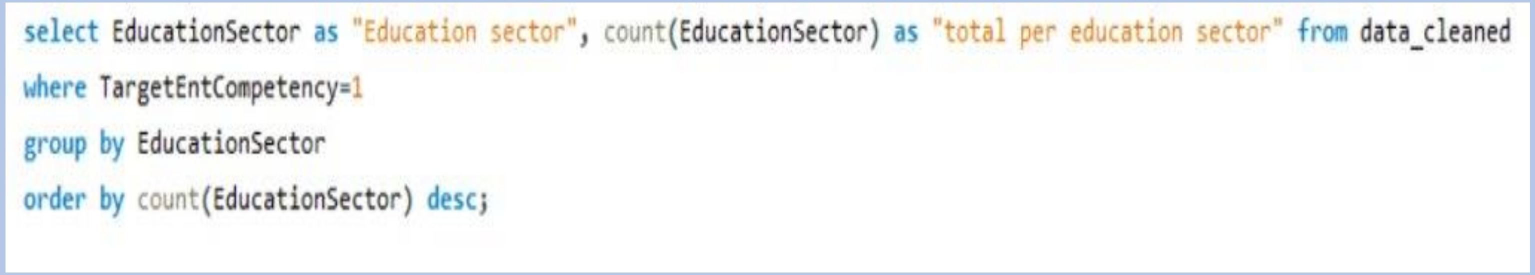
Our Query

what are the factors that influence the entrepreneurial drive?

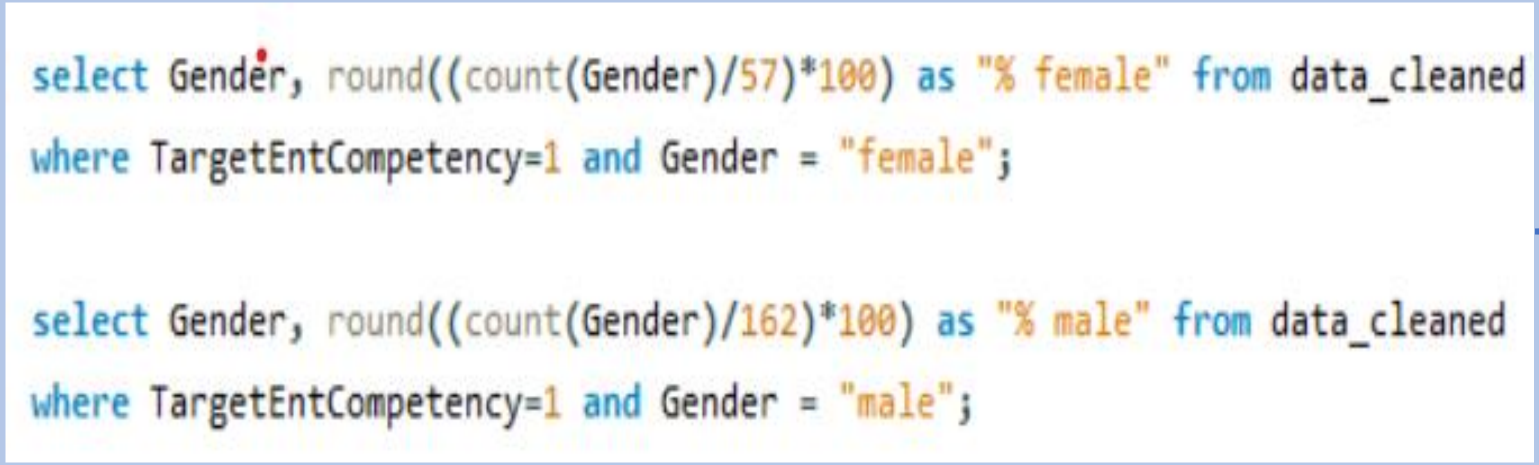


```
select EducationSector as "Education sector", round((count(EducationSector)/219)*100) as  "% education sector" from data_cleaned
group by EducationSector
order by count(EducationSector)/219 desc;
```





```
select EducationSector as "Education sector", count(EducationSector) as "total per education sector" from data_cleaned
where TargetEntCompetency=1
group by EducationSector
order by count(EducationSector) desc;
```



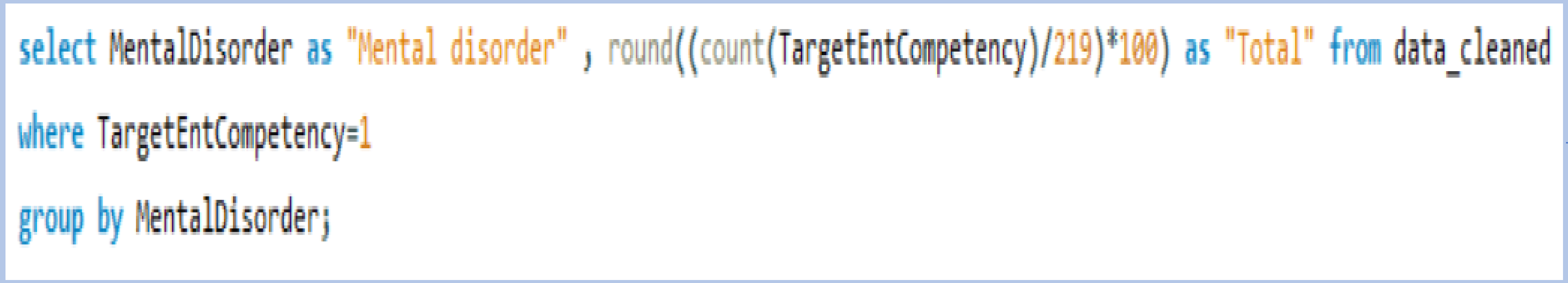
```
select Gender, round((count(Gender)/57)*100) as "% female" from data_cleaned
where TargetEntCompetency=1 and Gender = "female";

select Gender, round((count(Gender)/162)*100) as "% male" from data_cleaned
where TargetEntCompetency=1 and Gender = "male";
```

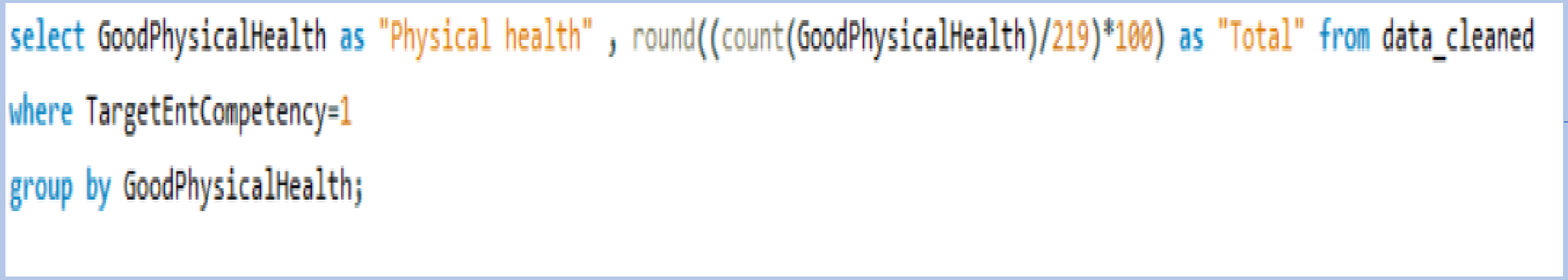
35 % females

45 % males

Physical and mental health



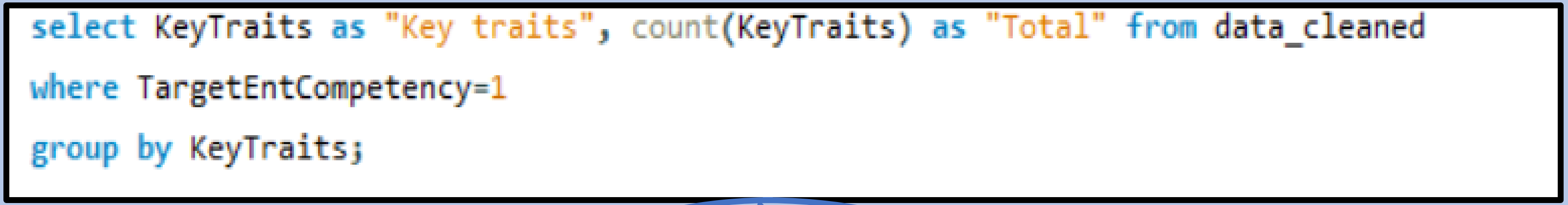
```sql
select MentalDisorder as "Mental disorder" , round((count(TargetEntCompetency)/219)*100) as "Total" from data_cleaned
where TargetEntCompetency=1
group by MentalDisorder;
```



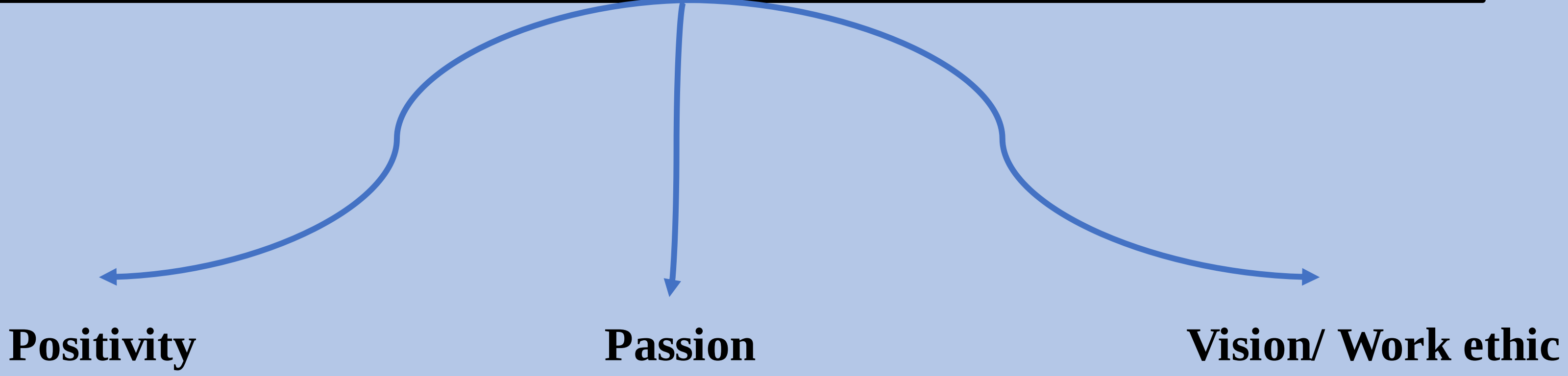
```sql
select GoodPhysicalHealth as "Physical health" , round((count(GoodPhysicalHealth)/219)*100) as "Total" from data_cleaned
where TargetEntCompetency=1
group by GoodPhysicalHealth;
```

More than 50 % Good mental and physical health

Key traits



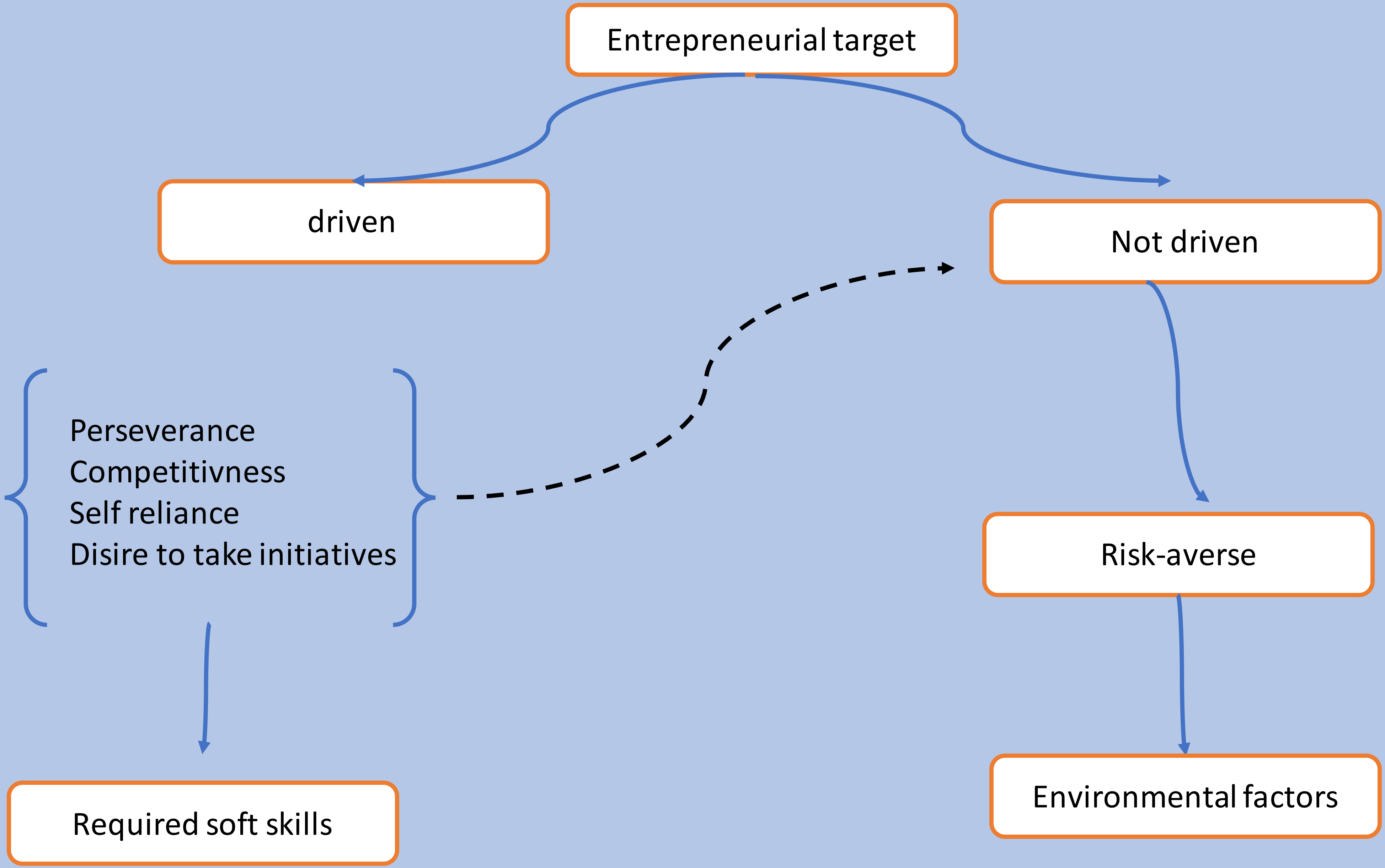
```
select KeyTraits as "Key traits", count(KeyTraits) as "Total" from data_cleaned
where TargetEntCompetency=1
group by KeyTraits;
```



Positivity

Passion

Vision/ Work ethic



Entrepreneurial target

driven

Not driven

Perseverance
Competitivness
Self reliance
Disire to take initiatives

Risk-averse

Required soft skills

Environmental factors