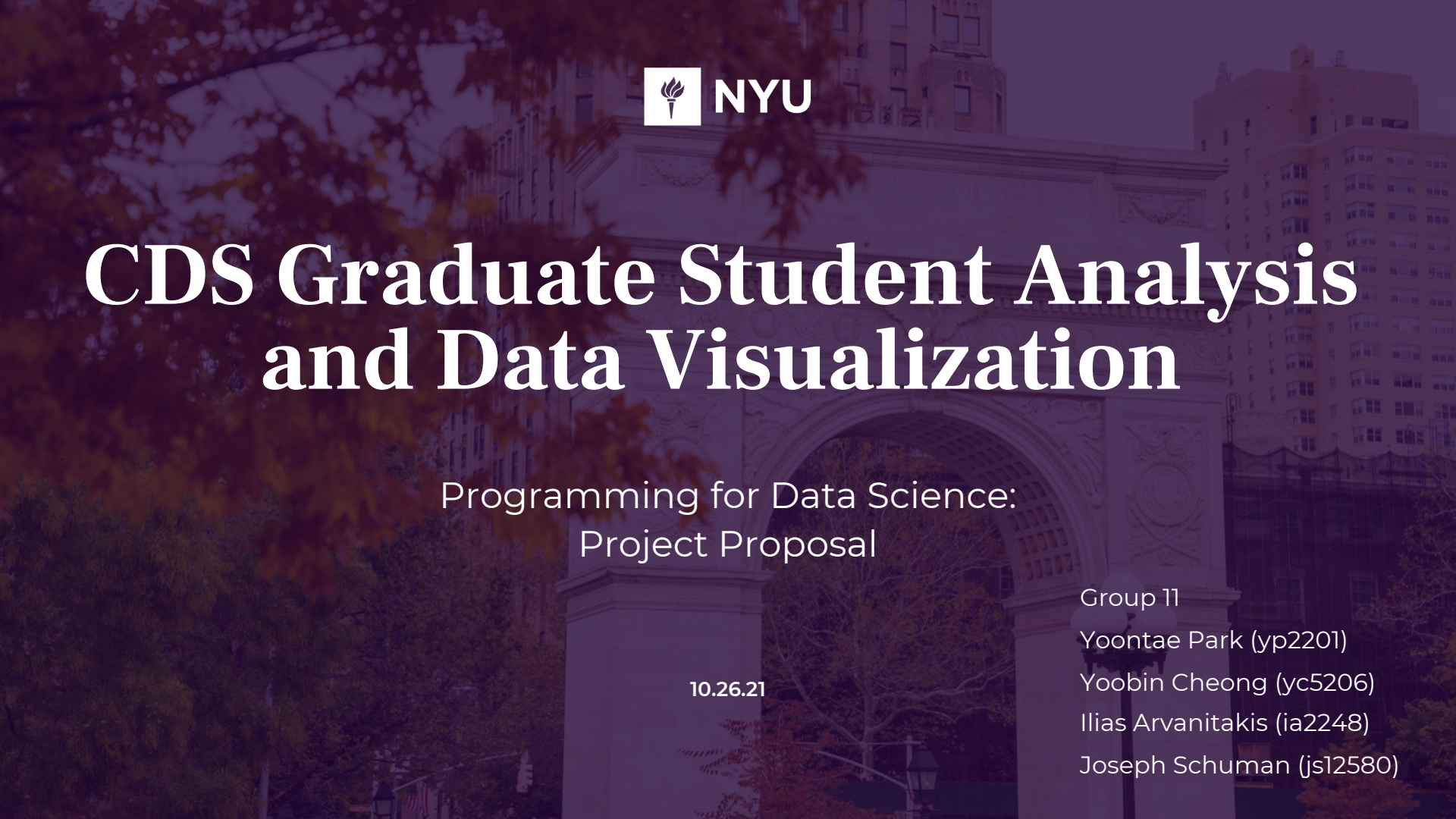


NYU

CDS Graduate Student Analysis and Data Visualization

Programming for Data Science:
Project Proposal

10.26.21

Group 11
Yoontae Park (yp2201)
Yoobin Cheong (yc5206)
Ilias Arvanitakis (ia2248)
Joseph Schuman (js12580)

Objective

Develop deeper insight for CDS administration, faculty, and students

- Develop an intuition of CDS students, both current and former
- Provide analysis and data visualization report to CDS administration and faculty

Student Backgrounds: CDS Webpage

Standardized Tests

We require that students submit standardized tests scores for the GRE. There are no exceptions: we do not accept "out of date" scores; nor do we accept scores of other, similar tests; nor do we allow waivers (regardless of previous educational attainment or circumstance).

In addition to sending your official scores to the Graduate School of Arts and Science please upload a PDF of the unofficial scores, which are made available upon completion of your test, to the "Additional Information Section" of your application.

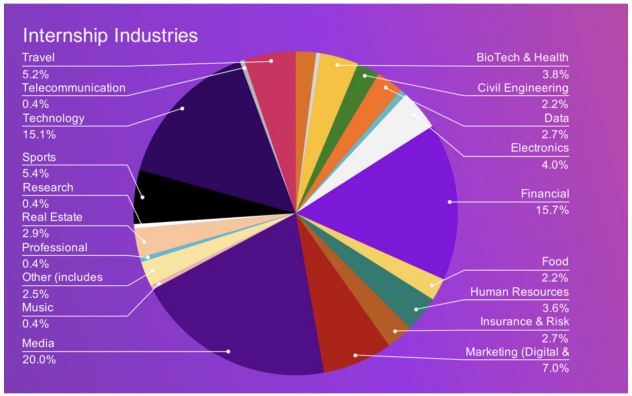
We wish to emphasize that *we have no set minimums for the GREs, and we consider the totality of an application when making a decision about admission.* Nonetheless, to the extent that it is helpful to give applicants a sense of things, what follows are the averages for the current cohort of MSDS students:

Average GRE Verbal: 159.3 (80th percentile)
Average GRE Quantitative: 167.4 (90th percentile)
Average GRE Analytical: 4.14 (61st percentile)
Average TOEFL (where required): 111

Career Outcomes: CDS Webpage

Our Students

INTERNSHIP INDUSTRY STATISTICS

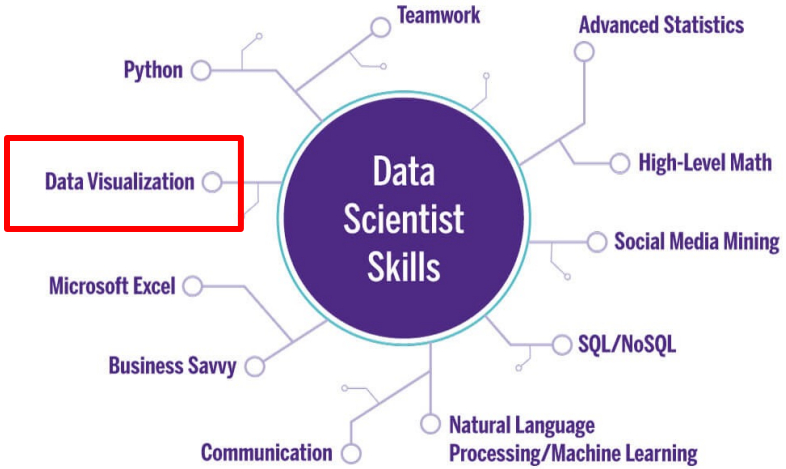


Since 2016, students have completed internships in various industries.

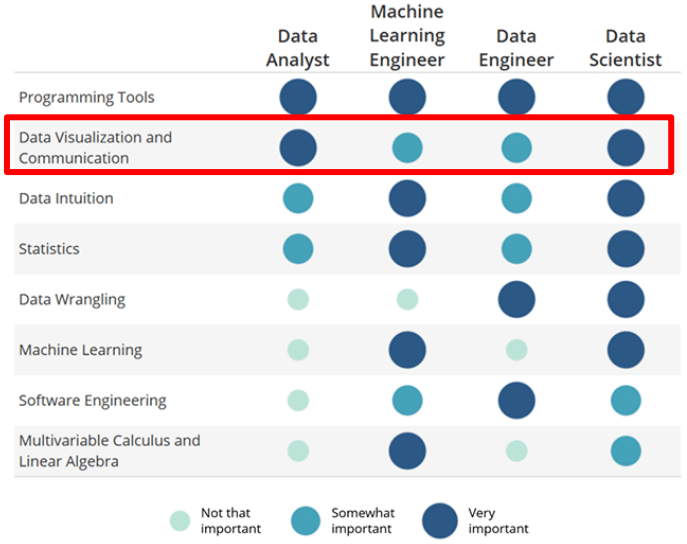
Why data visualization

- Visualization is one of the most important skills for Data Scientists
- Easier to comprehend the data; improves insights and decision-making process

Skills of data scientists



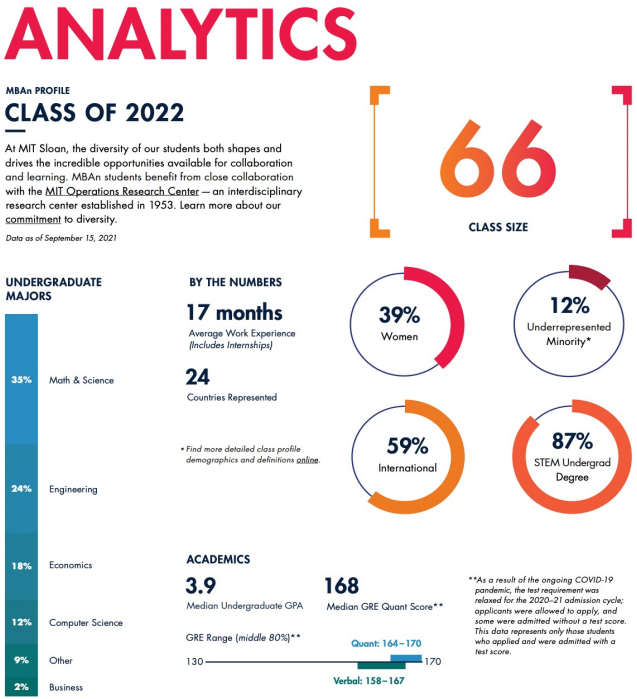
Visualization among most important skills for DS



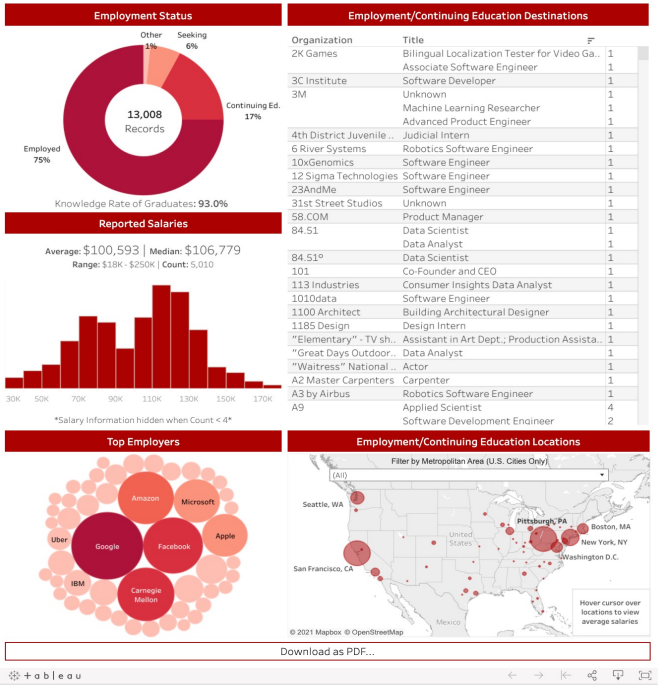
Final report draft example

Class analytics and data visualization

Enrolled Class Analysis



Student Career Outcomes



Overview of our dataset

- Student background information: Legal sex, US citizenship, GPA, GRE
- Student professional outcomes: Summer internship outcomes, Career outcomes
- **NOTE:** Any data that can be used to identify an individual is NOT given. (Name, DOB, etc)

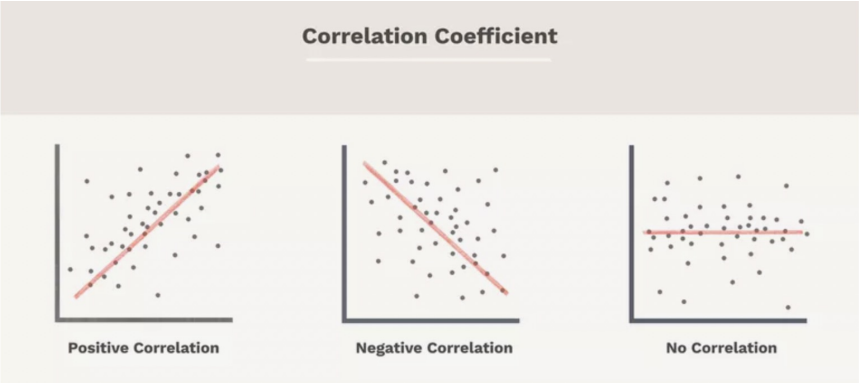
Anonymous Student Data

	A	B	C	D	E	F
1	2021-2022 MS Enrolled					
2	Student	Legal Sex	Undergraduate GPA	Quant	Verbal	Writing
3	Anonymous 1	F	3.79	168	169	4
4	Anonymous 2	F	3.71	168	169	3
5	Anonymous 3	M	3.79	163	159	3
6	Anonymous 4	M	3.71	163	159	4.5
7	Anonymous 5	M	3.89	163	159	3
8	Anonymous 6	F	3.85	166	163	4
9	Anonymous 7	F	3.77	168	154	3.5
10	Anonymous 8	F	3.75	163	157	3.5
11	Anonymous 9	M	3.64	168	157	2.5
12	Anonymous 10	M	3.71	168	157	3.5
13	Anonymous 11	F	3.28	163	150	4
14	Anonymous 12	M	3.86	163	150	4
15	Anonymous 13	M	3.45	165	150	3
16	Anonymous 14	M	3.91	166	150	4
17	Anonymous 15	F	3.72	168	150	3.5
18	Anonymous 16	F	3.71	152	151	4

What will we do with this dataset

- Data analysis: Identify possible correlations between student features
- Visualization: Distributions of students by visualizing mean/min/max values of data

Correlations between features



Visualization of CDS Class statistics

	GPA	GRE	...
count	///	///	///
mean	///	///	///
std	///	///	///
min	///	///	///
25%	///	///	///
50%	///	///	///
75%	///	///	///
max	///	///	///

Obtaining our results

- Getting basic statistics by using functions that we've learned in class (i.e. numpy, pandas)
- Trying some regression methods and figure out correlations between columns
- Using visualization tools such as Tableau, MS Power BI, etc.

Numpy Functions

1.7 Statistical Functions

1. **np.mean**(data,axis=0)
2. **np.var**(data,axis=0)
3. **np.sum**(data,axis=0)
4. **np.max**(data,axis=0)
5. **np.min**(data,axis=0)
6. **np.percentile**(data, percentage,axis=0)

In [46]:

```
mat = np.arange(9).reshape((3,3))
print ('Original matrix is \n%s\n'%mat)
print ('Overall mean of matrix is \n%s\n'%np.mean(mat))
print ('Row mean of matrix is \n%s\n'%np.mean(mat, axis=0))
print ('Column mean of matrix is \n%s\n'%np.mean(mat, axis=1))
print ('~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n')
print ('Overall varience of matrix is %s\n'%np.var(mat))
print ('Overall sum of matrix is %s\n'%np.sum(mat))
print ('Overall min of matrix is %s\n'%np.min(mat))
print ('Overall max of matrix is %s\n'%np.max(mat))

print ('~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~\n')
marks = np.array([30,31,32,40,90,95,97,98,99,100])
print ('Marks = %s\n'%marks)
print ('Overall 30 percent quantile of marks is %s\n'%(str(np.percentile(marks,30))))
```

```
Original matrix is
[[0 1 2]
 [3 4 5]
 [6 7 8]]

Overall mean of matrix is
```

Pandas / Tableau

Pandas

Power BI

QlikView

+ableau

Sources:

- CDS faculty anonymous student dataset (data that cannot distinguish students)
- LinkedIn profile of NYU MSDS graduate students
- NYU websites (i.e. CDS webpage, Career website)
- Reference websites (MIT sloan student analytics report, CMU career website)