#### SER594: Exploratory Data Munging and Visualization

#### University Suggestion System

#### Vinay Chavhan

#### 10/23/2022

##Dataset 1

## Basic Questions

\*\*Dataset Title: \*\* Graduate school admission data

\*\*Dataset Author(s): \*\* Ravi (Owner)

\*\*Dataset Construction Date: \*\*October 23, 2017 (Updated 5 Years Ago)

\*\*Dataset Record Count: \*\* 400

\*\*Dataset Field Meanings: \*\*

* 1. Admit: This column represents only two values, 0 and 1. 0 means no admit and 1 means admitted.
  2. GRE: Score range from 200 - 800
  3. GPA: GPA range from 2.26 – 4
  4. Rank: Rank range from 1 – 4

\*\*Dataset File Hash(es): \*\* 22a4c8073be15429e4490da20a0f5418

## Interpretable Records

### Record 1

\*\*Raw Data: \*\* (0th row): This student has not been admitted (value is 0). They scored 380 in GRE and their GPA is 3.61 and rank is 3.

Interpretation: \*\* The values are reasonable and are not out of range, it also makes sense that student didn’t get admit because they have low GRE score. We can say that this row is genuine.

### Record 2

\*\*Raw Data: \*\* (1st row): This student has been admitted (Value is 1). They scored 660 in GRE and their GPA is 3.67 and rank is 3.

\*\*Interpretation: \*\* Here as well the values are reasonable and are not out of range. We can see all the values gives us valuable information. As contradictory to 0th row record this student got the admit, having string GRE score is one factor for that.

## Data Sources

<https://www.kaggle.com/datasets/malapatiravi/graduate-school-admission-data/code>

MD5 hash: 22a4c8073be15429e4490da20a0f5418

### Transformation N

\*\*Description: \*\* There were not missing values in this data and we have only four columns in this data set. We verify if there’s any missing values in data in code.

\*\*Soundness Justification: \*\* The dataset is itself is clean and we didn’t perform any munging on this dataset. We have total 3 datasets and we have performed the data munging in one of them cause other two datasets are clean and ready to use.

##Dataset 2

## Basic Questions

\*\*Dataset Title: \*\* Students' Academic Performance Dataset

\*\*Dataset Author(s): \*\* Ibrahim Aljarah (Owner)

\*\*Dataset Construction Date: \*\* 2016-11-8 (Updated 6 YEARS AGO)

\*\*Dataset Record Count: \*\* 480

\*\*Dataset Field Meanings: \*\*

1 Gender - student's gender (nominal: 'Male' or 'Female’)

2 Nationality- student's nationality (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’  
Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)

3 Place of birth- student's Place of birth (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’  
Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)

4 Educational Stages- educational level student belongs (nominal: ‘lowerlevel’,’MiddleSchool’,’HighSchool’)

5 Grade Levels- grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)

6 Section ID- classroom student belongs (nominal:’A’,’B’,’C’)

7 Topic- course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’)

8 Semester- school year semester (nominal:’ First’,’ Second’)

9 Parent responsible for student (nominal:’mom’,’father’)

10 Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)

11- Visited resources- how many times the student visits a course content(numeric:0-100)

12 Viewing announcements-how many times the student checks the new announcements(numeric:0-100)

13 Discussion groups- how many times the student participate on discussion groups (numeric:0-100)

14 Parent Answering Survey- parent answered the surveys which are provided from school or not  
(nominal:’Yes’,’No’)

15 Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:’Yes’,’No’)

16 Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

\*\*Dataset File Hash(es): \*\* f584cdb16ead3ecd8e1f6641c712e865

## Interpretable Records

### Record 1

\*\*Raw Data: \*\* (0th row): This Male student came from KW and has place of birth KuwaIT. The staged is lower level and Grade id is G-04, Section ID is A, Topic is IT, Semester is F, Relationship with student is Father, 15 times raised their hand, 16 times visited resources, 2 times viewed announcement, 20 times participated in discussion groups, Parent Answering Survey is taken, Parent school satisfaction is Good, Student absence days is Under-7, and class is M.

Interpretation: \*\* There are total 16 columns each stating unique aspect of the row. We need to filter the unrequired columns for out data analysis. This row state that this student is male and from IT and parent school satisfaction is Good.

### Record 2

\*\*Raw Data: \*\* (1st Row): 0th row: This Male student came from KW and has place of birth KuwaIT. The staged is lower level and Grade id is G-04, Section ID is A, Topic is IT, Semester is F, Relationship with student is Father, 20 times raised their hand, 20 times visited resources, 3 times viewed announcement, 25 times participated in discussion groups, Parent Answering Survey is taken, Parent school satisfaction is Good, Student absence days is Under-7, and class is M.

\*\*Interpretation: \*\* This student have almost same info as the 0th row student but theya re slightly better cause they have raised 20 times which is 5 more than 0th row student. They also participated more in discussion groups and visited resources. As this student did better than previous student and previous student got parent school satisfaction as Good it’s very common that this student also must get parent school satisfaction as Good, and which is exactly the case. Hence, we can say that data is forming little patterns and we can predict some values.

## Data Sources

<https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data>

MD5 hash : f584cdb16ead3ecd8e1f6641c712e865

### Transformation N

\*\*Description:\*\* We have applied data filtering to drop unnecessary columns like 'NationalITy', 'PlaceofBirth'. As previous dataset this dataset also does not have any missing value and we have verified it in code

\*\*Soundness Justification:\*\* We have used filtering technique to drop 2 columns only which will not impact the data analysis. Those 2 columns are unnecessary and not required for our further analysis.

## Dataset 3

##Dataset 1

## Basic Questions

\*\*Dataset Title: \*\* “American University Data” IPEDS dataset

\*\*Dataset Author(s): \*\* Sumit Bhongale (Owner)

\*\*Dataset Construction Date: \*\*October 23, 2017 (Updated 5 YEARS AGO)

\*\*Dataset Record Count: \*\* 1534

\*\*Dataset Field Meanings: \*\*

Below are the total fields

ID number, Name, year, ZIP code, Highest degree offered, County name, Longitude location of institution, Latitude location of institution, Religious affiliation, Offers Less than one year certificate, Offers One but less than two years certificate, Offers Associate's degree, Offers Two but less than 4 years certificate, Offers Bachelor's degree, Offers Postbaccalaureate certificate, Offers Master's degree, Offers Post-master's certificate, Offers Doctor's degree - research/scholarship, Offers Doctor's degree - professional practice, Offers Doctor's degree - other, Offers Other degree, Applicants total, Admissions total, Enrolled total, Percent of freshmen submitting SAT scores, Percent of freshmen submitting ACT scores, SAT Critical Reading 25th percentile score, SAT Critical Reading 75th percentile score, SAT Math 25th percentile score, SAT Math 75th percentile score, SAT Writing 25th percentile score, SAT Writing 75th percentile score, ACT Composite 25th percentile score, ACT Composite 75th percentile score, Estimated enrollment, total, Estimated enrollment, full time, Estimated enrollment, part time, Estimated undergraduate enrollment, total, Estimated undergraduate enrollment, full time, Estimated undergraduate enrollment, part time, Estimated freshman undergraduate enrollment, total, Estimated freshman enrollment, full time, Estimated freshman enrollment, part time, Estimated graduate enrollment, total, Estimated graduate enrollment, full time, Estimated graduate enrollment, part time, Associate's degrees awarded, Bachelor's degrees awarded, Master's degrees awarded, Doctor's degrese - research/scholarship awarded, Doctor's degrees - professional practice awarded, Doctor's degrees - other awarded, Certificates of less than 1-year awarded, Certificates of 1 but less than 2-years awarded, Certificates of 2 but less than 4-years awarded, Postbaccalaureate certificates awarded, Post-master's certificates awarded, Number of students receiving an Associate's degree, Number of students receiving a Bachelor's degree, Number of students receiving a Master's degree, Number of students receiving a Doctor's degree, Number of students receiving a certificate of less than 1-year, Number of students receiving a certificate of 1 but less than 4-years, Number of students receiving a Postbaccalaureate or Post-master's certificate, Percent admitted - total, Admissions yield - total, Tuition and fees, 2010-11, Tuition and fees, 2011-12, Tuition and fees, 2012-13, Tuition and fees, 2013-14, Total price for in-state students living on campus 2013-14, Total price for out-of-state students living on campus 2013-14, State abbreviation, FIPS state code, Geographic region, Sector of institution, Level of institution, Control of institution, Historically Black College or University, Tribal college, Degree of urbanization (Urban-centric locale), Carnegie Classification 2010: Basic, Total enrollment, Full-time enrollment, Part-time enrollment, Undergraduate enrollment, Graduate enrollment, Full-time undergraduate enrollment, Part-time undergraduate enrollment, Percent of total enrollment that are American Indian or Alaska Native, Percent of total enrollment that are Asian, Percent of total enrollment that are Black or African American, Percent of total enrollment that are Hispanic/Latino, Percent of total enrollment that are Native Hawaiian or Other Pacific Islander, Percent of total enrollment that are White, Percent of total enrollment that are two or more races, Percent of total enrollment that are Race/ethnicity unknown, Percent of total enrollment that are Nonresident Alien, Percent of total enrollment that are Asian/Native Hawaiian/Pacific Islander, Percent of total enrollment that are women, Percent of undergraduate enrollment that are American Indian or Alaska Native, Percent of undergraduate enrollment that are Asian, Percent of undergraduate enrollment that are Black or African American, Percent of undergraduate enrollment that are Hispanic/Latino, Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander, Percent of undergraduate enrollment that are White, Percent of undergraduate enrollment that are two or more races, Percent of undergraduate enrollment that are Race/ethnicity unknown, Percent of undergraduate enrollment that are Nonresident Alien, Percent of undergraduate enrollment that are Asian/Native Hawaiian/Pacific Islander, Percent of undergraduate enrollment that are women, Percent of graduate enrollment that are American Indian or Alaska Native, Percent of graduate enrollment that are Asian, Percent of graduate enrollment that are Black or African American, Percent of graduate enrollment that are Hispanic/Latino, Percent of graduate enrollment that are Native Hawaiian or Other Pacific Islander, Percent of graduate enrollment that are White, Percent of graduate enrollment that are two or more races, Percent of graduate enrollment that are Race/ethnicity unknown, Percent of graduate enrollment that are Nonresident Alien, Percent of graduate enrollment that are Asian/Native Hawaiian/Pacific Islander, Percent of graduate enrollment that are women, Number of first-time undergraduates - in-state, Percent of first-time undergraduates - in-state, Number of first-time undergraduates - out-of-state, Percent of first-time undergraduates - out-of-state, Number of first-time undergraduates - foreign countries, Percent of first-time undergraduates - foreign countries, Number of first-time undergraduates - residence unknown, Percent of first-time undergraduates - residence unknown, Graduation rate - Bachelor degree within 4 years, total, Graduation rate - Bachelor degree within 5 years, total, Graduation rate - Bachelor degree within 6 years, total, Percent of freshmen receiving any financial aid, Percent of freshmen receiving federal, state, local or institutional grant aid, Percent of freshmen receiving federal grant aid, Percent of freshmen receiving Pell grants, Percent of freshmen receiving other federal grant aid, Percent of freshmen receiving state/local grant aid, Percent of freshmen receiving institutional grant aid, Percent of freshmen receiving student loan aid, Percent of freshmen receiving federal student loans, Percent of freshmen receiving other loan aid, Endowment assets (year end) per FTE enrollment (GASB), Endowment assets (year end) per FTE enrollment (FASB)

\*\*Dataset File Hash(es): \*\* ffae4e40f444551d2210a65eef11c687

## Interpretable Records

### Record 1

\*\*Raw Data: \*\* (0th row): This row is about Alabama A & M University which is founded in 2013. They offer Doctorate as highest degree. The university is situated in Madison County. Total 6142 students applied to this university and 5521 students got into it, and from that 1104 enrolled.

Interpretation: \*\* There are multiple columns, there are unnecessary columns which are not required. Not all columns are filled there are some missing fields. There are some outliers as well in that data which needs cleaning.

### Record 2

\*\*Raw Data: \*\* (1st row): This row is about University of Alabama at Birmingham which is founded in 2013. They offer Doctorate as highest degree and also scholership. The university is situated in Jefferson County. Total 5689 students applied to this university and 4934 students got into it, and from that 1773 enrolled.

\*\*Interpretation: \*\* We observed same things as we observed for 0th row. There are some missing values and outliers in the data. We will need to remove the unnecessary columns from the dataset.

## Data Sources

<https://www.kaggle.com/datasets/sumithbhongale/american-university-data-ipeds-dataset>

MD5 hash: ffae4e40f444551d2210a65eef11c687

### Transformation N

\*\*Description: \*\* We used smoothing, filtering and imputation.

\*\*Soundness Justification: \*\*We have dropped several columns from the dataset which are irrelevant for our analysis. We have also filled the missing values for all missing values columns. We have not modified the other data which states that the operations have not changed the semantics of the data.

## Visualization

### Visual N

\*\*Analysis:\*\*

We have six visualizations graphs.

1. GRE and GPA Plot: We can see that the relation is not much dense, we see the student has low GPA also have high GRE score and vice versa. We see the average students who have good GPA and GOOD GRE score are denser than another category.
2. GRE and Rank Plot: There are only fours ranks and GRE score is ranging between 200-800. We see one outlier where GPA is 3 and score is very low. Other dots looks symmetrical for all the ranks.
3. GPA and Rank Plot: The data is very dense in rank 2. There are seven outliers totally present in all ranks.
4. Hist

(duplicate above as many times as needed; remove this line when done)