P4 10 Peer Tutor Trend

Group Member: Wenting Yue, Xiaofei Xie, Yuxi Shen

Data Exploration

- **Data types:** Review the data and answer the following:
    1. What type of data is in each column (categorical, ordinal, or quantitative)?

There are three separate sheets in our partner's dataset.

For the "PTP Tutors" sheet, one column is categorical data of the tutor's name, and other columns including the sum of payroll, number of sessions, and number of hours are quantitative data. The number of hours is the continuous data, and others are the discrete data.

For the "PTP by Subject" sheet, one column of categorical data about the name of the college, course number, and the number of professors. The number of hours is continuous quantitative data, and other columns associated with the number of tutor appointments are discrete quantitative data.

For the "summary" sheet, one categorical data for the name of the college and other columns are the quantitative data for the number of tutoring sessions (one-to-one, small group, in-person, online).

2. Write a few sentences (3–5) summarizing the data you are working with and how the data was collected/generated (e.g. survey, statistical, internal revenue, etc.).

The data is collected by Peer Tutoring Program at Northeastern University from last semester from the internal management system. Each information of happening tutor session would be recorded and the supervisor collected all the tutor's sessions at end of the semester to record how many tutor sessions, the payroll, and hours have been taken for each course taught by each professor under each college. The data is cleaned and ready to use with no missing data or misleading information. The data from each sheet should be connected from tutor sessions and courses to each student.

- **Potential issues:** While reviewing the data, look for missing data, variables you are confused about, missing metadata, etc.

There are some confusing points that we try to understand before the interview, and most of them have been solved by our partner's explanation.

1. confusion of the term of columns:
    a. The PTP organization used letters to replace the names of instructors of each class to protect the privacy of instructor identities.
    b. Difference between appointment sessions and hours. The number of sessions refers to the number of meetings that the tutors held, while the number of hours refers to the time length tutors spend with students.

2. confusion about the missing data
    a. The first sheet is about the sum of tutor information in each college. However, the grand total of some columns is not the sum of all of the elements. There may need an extra column to store the “other meetings” tutors held.

Overall, the dataset provided by our partner PTP is complete, clear, and clean.

- **Insights:**
    1. What trends and patterns do you see? Did anything surprise you during the exploration?
    2. Did you identify any further errors or messy/confusing data past what you noted in Potential issues? If so, sort it out ASAP if possible!

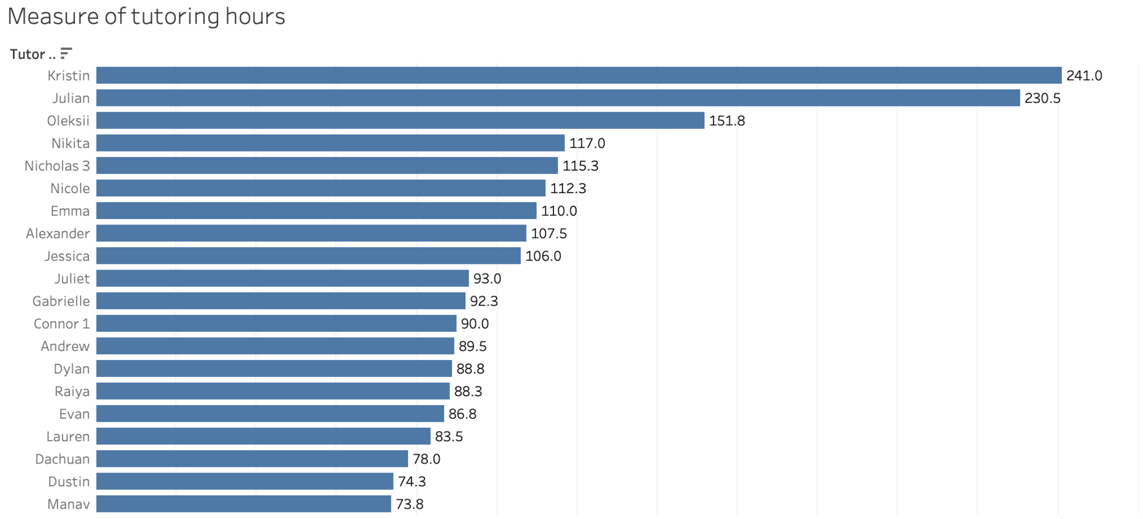
During our exploration, we witness some useful trends inside the data. For example, we see a positive correlation between the sum of payroll and the number of working hours. However what surprised me most during the exploration is that although the sum of payroll and number of sessions have a positive correlation, they are not linked so tight. For example, when sorting the sum of payroll in descending order, we found that the number of hours is in descending order, but the number of sessions is not perfectly in the descending order. So we want to explore more about how does the number of sessions affects the sum of payroll.

From this point, we haven't identified any further errors or messy/confusing data. So probably we have already got a clean and useful dataset. But once we notice there exists any confusing data, we will sort it out and ask for explanations as soon as possible.

- **Screenshots**:
    1. What data, or a subset of data, you were exploring.
    2. What visual encoding(s) do you use and why.
    3. What trend or pattern (or lack of trend/pattern!) does the visualization show.

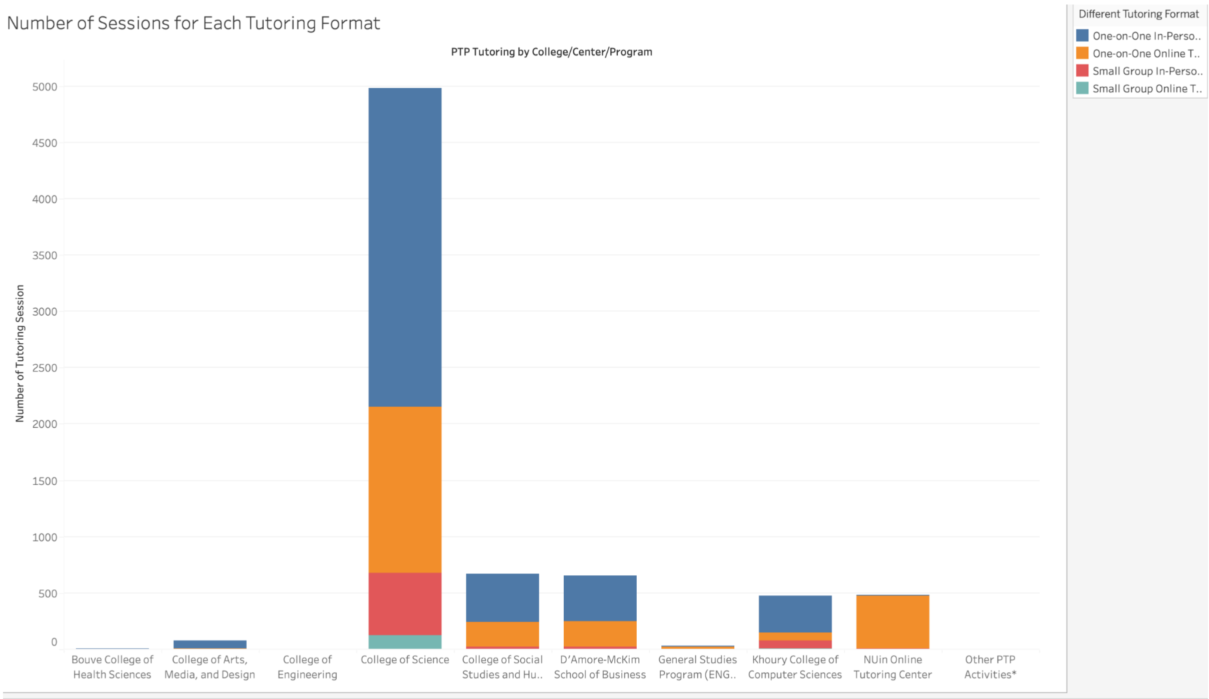
>> Screenshot_1

The measure of tutoring hours per tutor



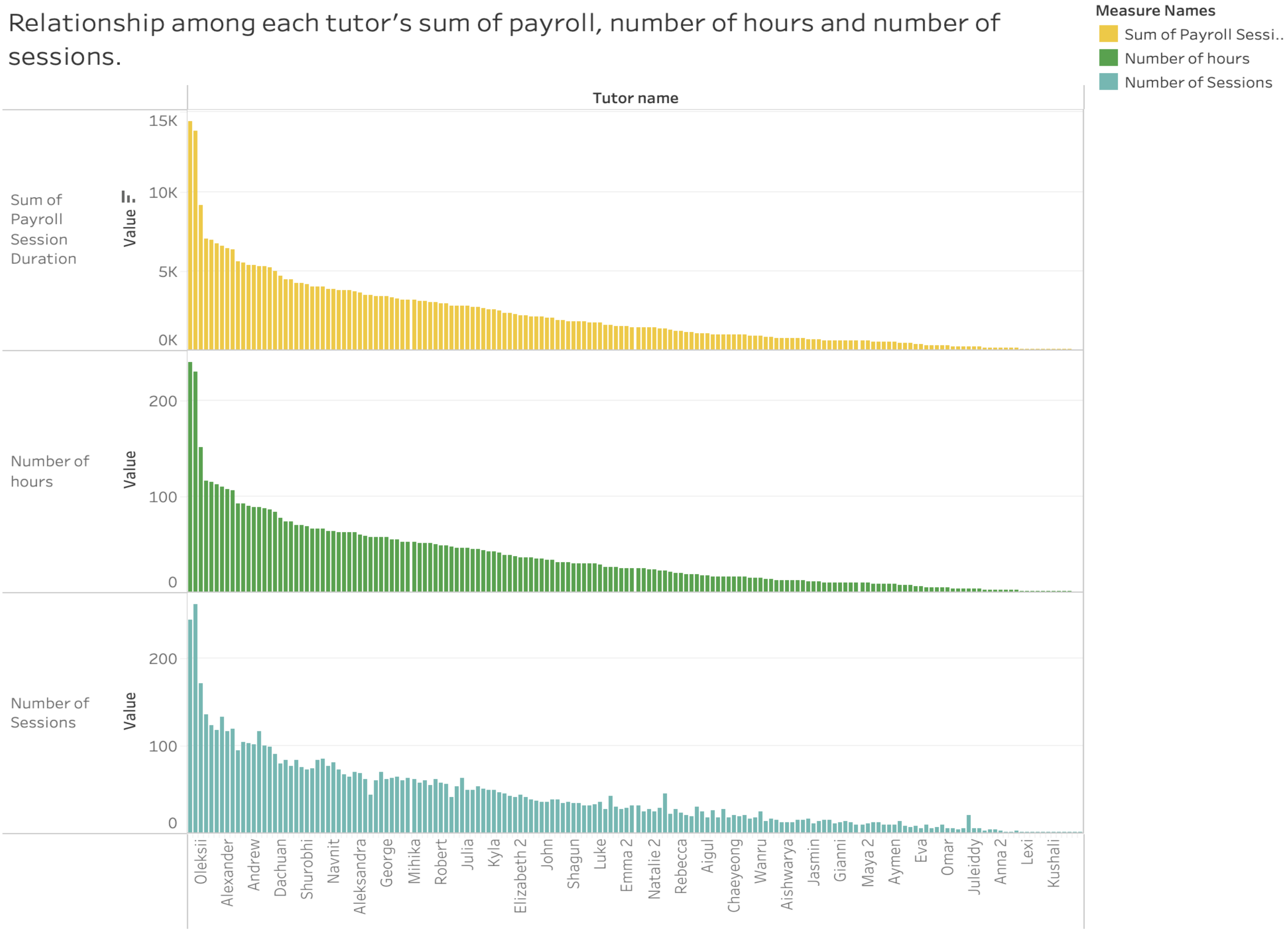
We explored the third sheet which summarizes the total hours of each tutor working during the 2021 fall semester. We used the bar chart and the length of each bar to display the time length each tutor worked during their service. The categorical channel is the name of each tutor, and the magnitude channel is the length of the line of each bar. This chart surprised me that the top two tutors worked around 240 hours in the 2021 fall semester. By calculation, this couple of students worked 18 hours each week if we assume that there were 13 weeks in one semester. If we divide the hours into week manner, the tutors worked 3-4 hours each day. We appreciate what the tutors did for all the students to succeed in their academic performance and we also admire the time-organization skill that those peer tutors develop during this semester. On the other side, we are also curious about which class Kristin and Julian were tutoring and why they have to work so many hours on these classes. If the PTP could find out the classes that Kristin and Julian are taking charge of, PTP could assign other tutors to assist Kristin and Julian to relieve their stress.

>> Screenshot_2
Number of Sessions for Each Tutoring Format



We explore the summary sheet from the dataset which summaries how many sessions took place for each college and each format. We choose the stacked bar having different colleges as categorical data as the x-axis and the quantitative data, the number of sessions, as the x-axis. We use different tutoring formats as color dimensions so that each bar is separated by color to show the percentage of each tutoring format happening in each college. The reason why we choose this encoding is that we want to compare the total number of tutoring sessions for each college while having a sense of students' preference of tutoring format in each college. The trend we found is that most of the tutoring is in-person and students from the college of sciences schedule the most tutor sessions. Also, students from NU particularly prefer One-on-one Online Tutoring sessions and rarely have in-person tutoring.
>> Screenshot_3

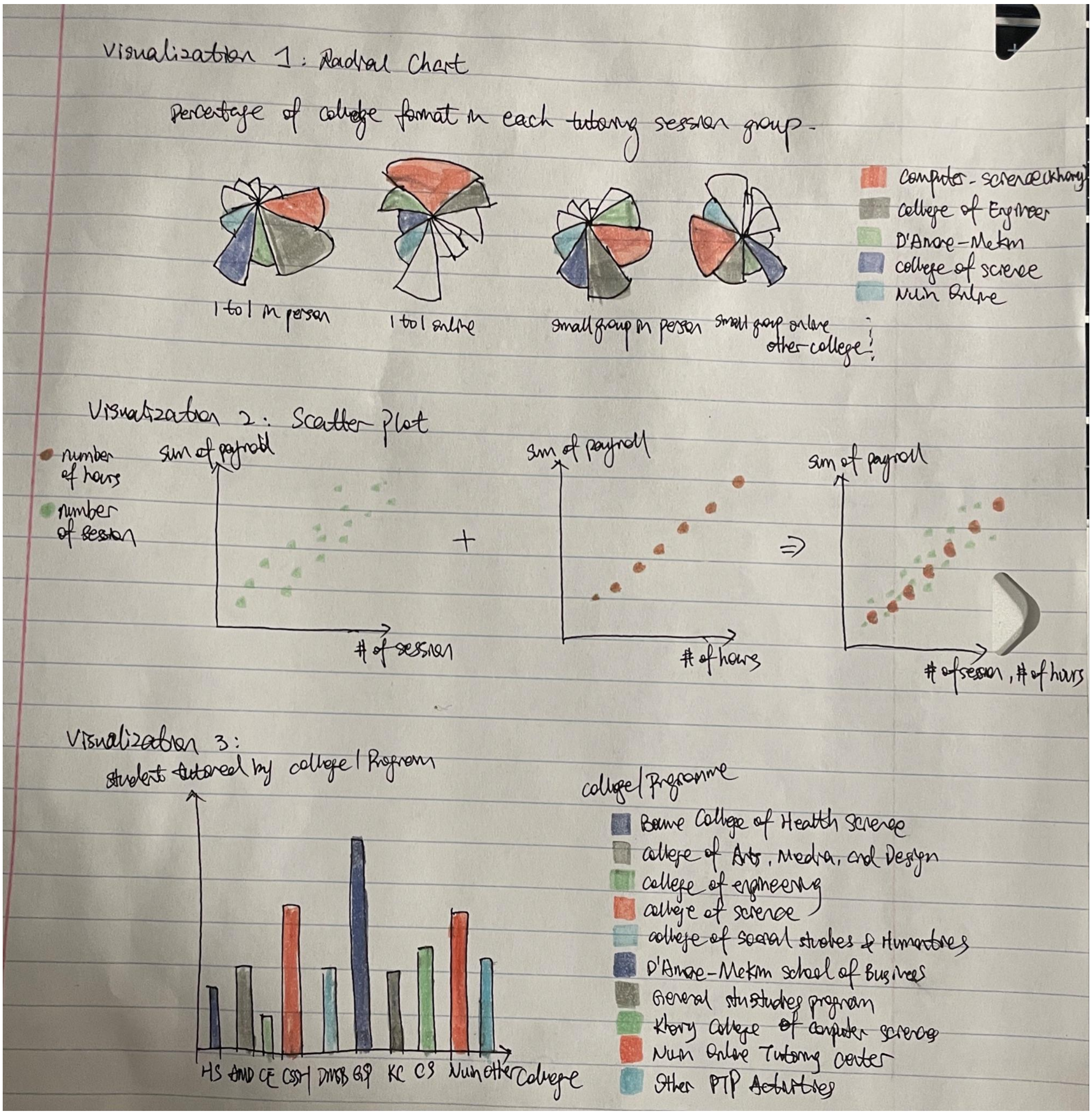
Relationship among each tutor’s sum of payroll, number of hours, and number of sessions.



We explored the third sheet of the excel chart which lists the information about each tutor’s sum of payroll, number of hours, and number of sessions during the 2021 fall semester. We used the horizontal bar chart to compare the trend of each three attributes. The categorical channel is the name of each tutor, and the magnitude channel is numerical values that contain the sum of payroll, number of hours, and number of sessions. I sort the value in descending order based on the sum of payroll for each tutor. We see the trend for some of the payroll match the trend for the number of hours quite well, but what surprised me most is that the trend for the number of sessions does not shows a perfectly matching pattern compared to the above two attributes. This might indicate that the sum of payroll and number of hours have a strong relationship with each other, but the number of sessions is relatively independent of the other two attributes.

Individual Sketches

Artist: Xiaofei Xie



1. Radial Chart
mark: area
channel:

magnitude channel: Each part of the area of this circle shows the percentage of college format in each tutoring session group.

encoding:

Area and radius encoding of the area of each part of the pie.

I choose this pie chart to show the percentage of each college in each tutoring session group.

From this pie chart, users could understand which tutoring session group can better fit for which college. So we can find the most efficient way of tutoring for each college.

2. Scatter plot
mark: point
channel:

categorical channel: the position of points on a common scale

encoding:

position encoding to show the location of each TA in the sum of Payroll session duration with the number of sessions ranges and the location of each TA in the sum of Payroll session duration with the number of hours.

I want to use this scatter plot to explore the relationship between the sum of payroll session duration with the number of sessions, and the relationship between the sum of payroll sessions with duration and the number of hours. By overlapping the number of hours and number of sessions together, we can see which attributes have the most direct correlations with the sum of Payroll.

3. Bar chart
mark: lines
channel:

categorical channel: color hues represent different kinds of college

magnitude channel:

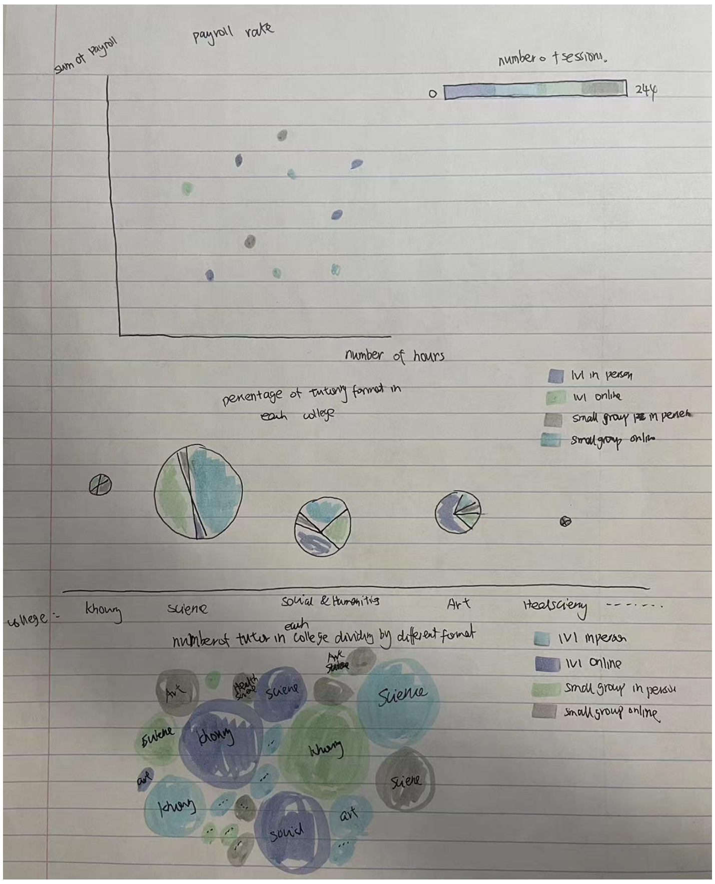
length of the bars represent the number of students tutored by college or program

encoding:

length encoding of length/height of different bars.

I chose this bar chart to visualize the distributions of how many students are tutored by each college or program. From this bar chart, we and the users can easily find out which college is more popular and needs more tutors.

Artist: Wenting Yue



1. Scatter plot:

I choose the two quantitative attributes number of tutoring hours and the sum of payroll. The point marks the vertical and horizontal spatial position. I add color channels to represent the number of sessions. The reason I choose this encoding is that there are three quantitative attributes that make one attribute to the color dimension let the plot see the relation between the other two attributes while having the sense of another metric. This plot could help us find the trending payroll rate. For example, if there is the trend that the rate of increasing payroll is lower than the rate of increasing tutoring hours. (Task 1)

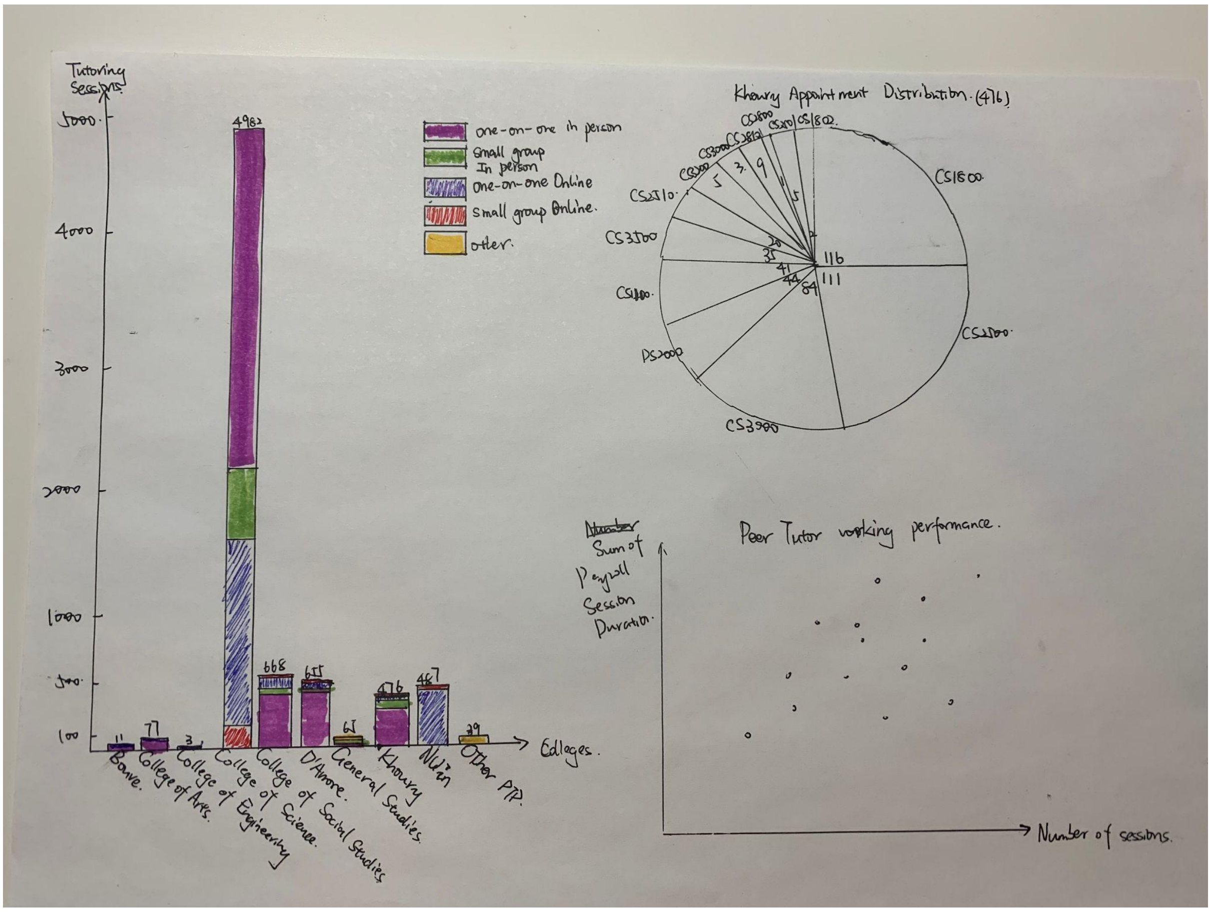
2. Pie Chart:

The x-axis is the categorical data for each college, and each college has a pie chart representing the percentage of each tutoring format. The size of the pie represents the total tutoring hours in each college. The chart has the area mark and color and area dimension. This encoding perfectly compares the total tutoring hours for each college, and the pie chart helps understand which type of tutoring takes the biggest portion of total tutoring in each college. The chart addresses the task that asks to find out students' preferences in each college in terms of the type of tutoring. (Task 2)

3. Packed Bubbles

For this visualization, the quantitative data is the number of tutoring sessions, and the categorical data is the type of tutoring. The mark is the area that the size of the bubble represents the number of sessions. The channel is the color that each type of tutoring has a different color. By using this encoding, the type of tutor in which college has the biggest size gets highlighted. This visualization also helps users address the task of how to compare the number of tutoring happening in each college. (Task 3)

Artist: Yuxi Shen



1. Bar chart (**Favorite**)
mark: lines
channel:
> categorical channel: color hues represent different kinds of tutoring
> magnitude channel: length of the bars represent the number of tutoring sessions
encoding:
> length encoding of length/height of different bars.
I choose this stacked bar chart to visualize the distributions of different types of tutoring services among all colleges in Northeastern University. From this bar chart, we and the users could easily find out which tutoring session most students registered for from different colleges. Only this stacked bar chart could achieve this goal(task2).

2. Pie chart(**Favorite**)
mark: area
channel:
> magnitude channel: Each part of the area of this circle shows the percentage of tutoring sessions of a particular class.
encoding: Area encoding of the area of each part of the pie.

I choose this pie chart to show the percentage of sessions of each course in Khoury College. From this pie chart, users could understand which class may need more tutors to help students succeed in this particular class. This pie chart could also be enhanced to show precisely which instructor is the most popular one in each class. (Task 4 and partially solved Task No.5)

3. Scatter plot (**Favorite**)
mark: point
channel:
> categorical channel: the position of points on a common scale
encoding:
> position encoding to show the location of each TA in the sum of Payroll session duration and number of sessions ranges
I want to use this scatter plot to display each tutor's contribution to this tutoring program in 2021 fall by presenting their position on both the " sum of Payroll Session Duration" scale and the "Number of Sessions" scale. Therefore, this chart could tell us the tutor's information and who are the hard-working, busy ones(task 1).

Favorite Sketches

One of our favorite charts is Yuxi’s bar chart. This bar chart **visualizes the distributions of different types of tutoring services among all colleges in Northeastern University,** which help to solve **task No.2** that we mentioned in P3. This chart summarizes the current situation of each college in NU. We can display the distribution of tutors in this table by looking at the heights of the bars. Additionally, the chart displays the percentage of tutors in each college for each session type. Users can therefore better understand the needs of students in each college for each session. In a nutshell, this bar chart can provide a general view and details of the dataset to our stakeholder PTP at the same time, which indicates that we can put this table at the forefront of all future presentations to summarize the whole database. This chart also utilizes color hues as a categorical channel to distinguish the sessions, which can also attract viewers' attention.

Another chart we like is Yuxi’s scatter plot that shows the **count of peer tutoring sessions for each tutor**. Each point on the scatter plot represents a tutor. The x-axis is the number of tutoring session the tutor held over the semester, and the y-axis is the total salary the tutor get paid. The scatter plot helps us address **task No.1** that how each tutor contributes to the program. By analyzing the trend of the point, we can see the average count of sessions among the program and how many tutors have a highlighting contribution by holding much more sessions. Our partner can use this visualization to evaluate and estimate for the further new tutor how many sessions they would be expected to contribute. This scatter plot fits the data since then we have two quantitative attributes, and we want to see the correlation between the two attributes and find any cluster of points in the plot. This visualization uses the horizontal and vertical location of the point, which shows the distribution of all tutors and at which level the specific tutor lies on the program.

The other favorite chart is Yuxi’s pie chart. This bar chart **visualizes Khoury College appointment distribution,** which helps to solve **task No.4** that we mentioned in P3. This chart summarizes each course's total appointments at Khoury College. We can display the distribution of appointments in this table by looking at each segment of the chart. This chart delivers the classes that are in need of more tutors in the Khoury College, and it might be a useful insight for Peer Tutor Program for future recruitment. Also, if we visualize every course appointment information in every college, we can see which course is the most popular class, and solve **task No.5**. Overall speaking, this pie chart can provide a general view and details of the PTP by subject dataset to our stakeholder Nicole and her teams. And for future implementation, we can try to use the radial chart to exemplify the proportion of each course, and also we can use color hues to distinguish each course.