Initial EDA

Before running any unsupervised algorithm, we should look at basic descriptive stats that answer questions like:

* what are the most commonly taken elective courses across the entire population?

|  |  |
| --- | --- |
| Capital Markets & Investm | FINCB8306 |
| 6166 | 6062 |

* How does that change when aggregated by [insert any other field e.g. gender, race, future industry etc.].
* Who are the most popular instructors?
* Does this change over time? Etc.
* Maybe we want to do a heatmap with all elective courses to find which courses are most highly correlated initially?

Cleaning:

28 July 2025

* The new course dataset has some Unis as dates. Im going to manually change them (It should be the first 3 letters of the month (12 = dec, 3 = mar) followed by the year)
* But this one student has null : C004457768 : thn2114 (missing in SIS—only in admissions data)
* Re-ran EDA file, data wrangling file, correlations file.