Education Attainment and Income

Rachel Leach, Elizabeth Wu, Adair Hancock, Kendle Schooler, & Anisha Sapkota

Data Selection

- **American Community Survey** - US Census Bureau
- Accessed through Social Explorer
- 5-year estimate (2018-2022)
- Median earnings by state & education level

- **College Scorecard -** US Dept. of Education
- 2022-2023
- Institution data (graduate earnings, student demographics, admissions, financial aid)

➢ Wide range of variables
➢ Credible sources
➢ Personal relevance
➢ Large samples

Motivation: Are the income benefits of college education worth the investment?

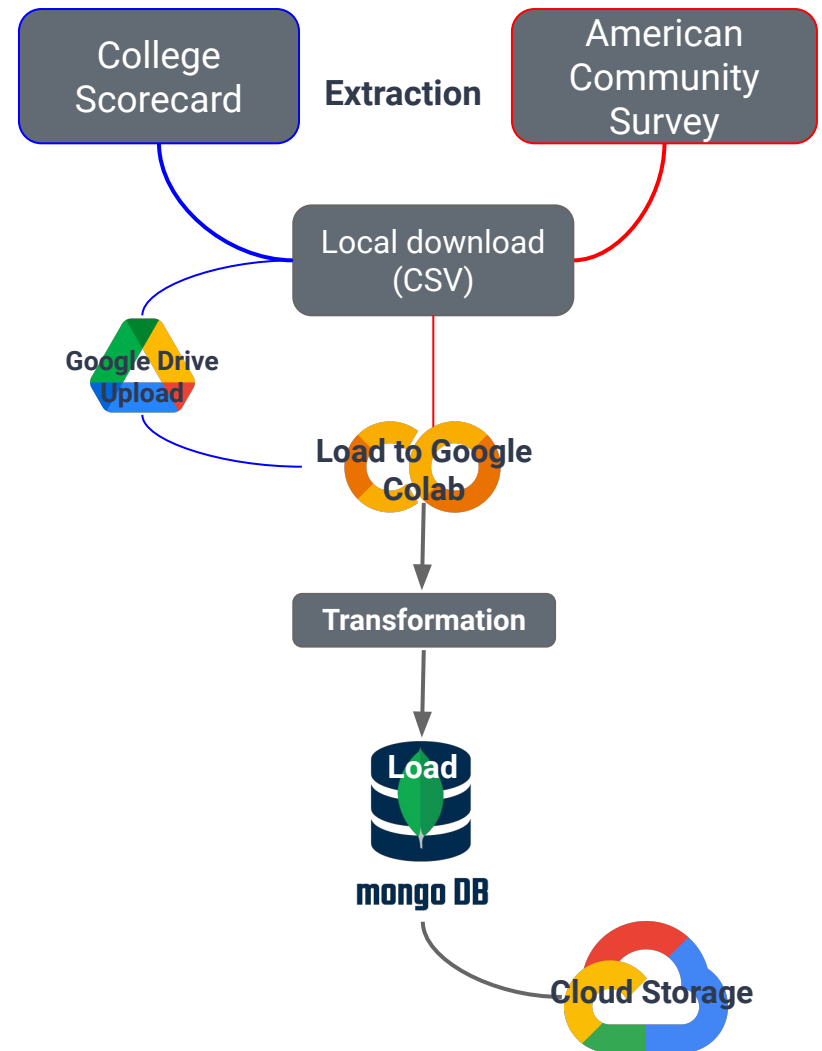How is this affected by institution and location?

ETL Process

Cleaning/Transformation:

➢ Selected relevant variables
➢ Renamed columns
➢ Removed NA values
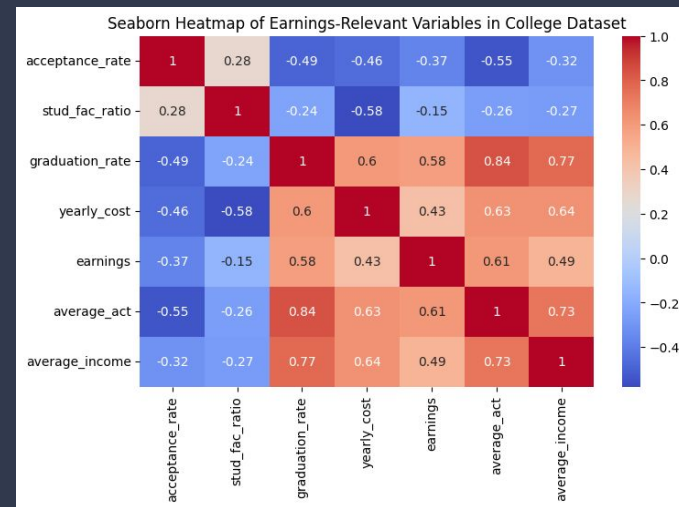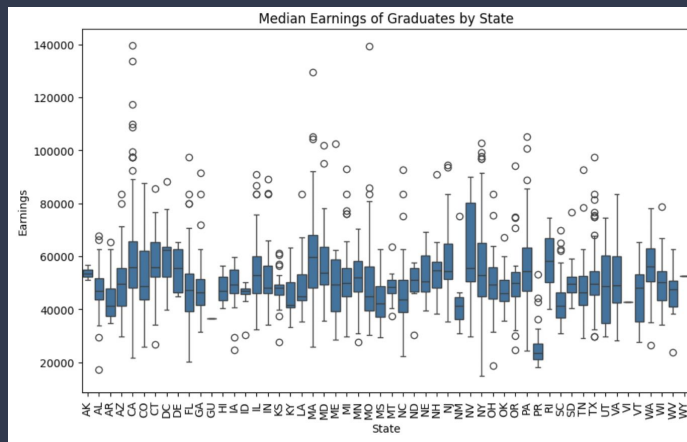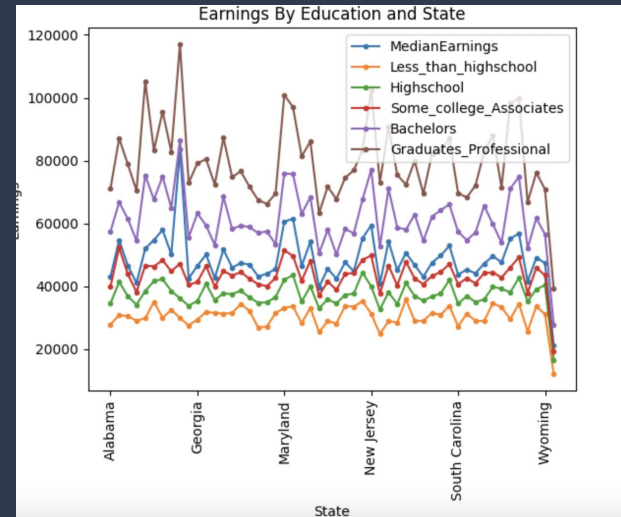➢ Filtered by predominant degree type (Bachelors)

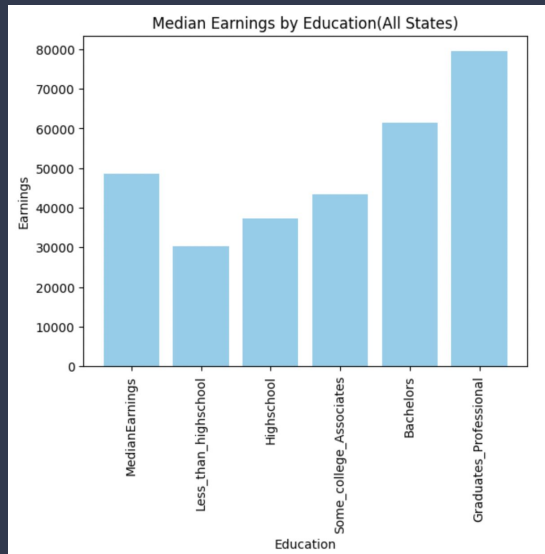**Google Cloud Storage:**

➢ Uploaded transformed CSV's to bucket
➢ Settings:
  - Allow public reading access
  - Uniform access
  - Location: multi-region US
➢ High availability and low latency

Extraction

College Scorecard → Local download (CSV) ← American Community Survey

Google Drive Upload

Load to Google Colab

Transformation

Load — mongo DB

Cloud Storage

Analysis: Trends in Earnings

Bachelor's degree or higher is best for the most income benefit.

Earnings Comparison

	earnings_diff
earnings_diff	1.000000
earnings	0.834622
average_act	0.425048
graduation_rate	0.408408
average_income	0.321494
yearly_cost	0.245225
student_pop	0.204584
acceptance_rate	-0.325741

Return on Investment

	ROI_diff
ROI_diff	1.000000
stud_fac_ratio	0.504205
part_time	0.454715
student_pop	0.411810
earnings	0.270257
acceptance_rate	0.219887
graduation_rate	-0.216312
average_act	-0.229038
average_income	-0.320571
yearly_cost	-0.753877

- Correlation coefficients for institutional variables and difference between graduate earnings and state Bachelor's median earnings
- **Earnings: strong positive correlation**
- **Average ACT score and graduate rate: weak positive correlation**

- Correlation coefficients for institutional variables and ROI [difference between graduate earnings and cost of attendance]
- **Yearly cost:** strong negative correlation
- **Student-to-faculty ratio**: moderate positive correlation

Main Challenges & Takeaways

- **Dataset selection**
    - Finding large datasets that were also relevant to our topic
- **ETL Process**
    - Handling large files
    - Mounting to Colab
    - Importance of renaming columns and isolating variables
- **Analysis**
    - Deciding which visualizations would be most impactful
    - Accurately interpreting visualizations
- **Google Cloud Storage**
    - Large datasets
    - Choosing appropriate settings
- **Team collaboration**
    - Difficult to find time for meetings
    - Needed regular communication & clear delegation of tasks