How to Choose A College Major: The Story Behind the FiveThirtyEight College Majors Dataset

Alison Qiu, Maggie Bao

Abstract:

Our project analyzed the FiveThirtyEight College Majors dataset to better understand the relationship between college majors and future career prospects. Exploratory data analysis revealed trends in earning potential, employment levels, and job types across majors. K-means clustering grouped majors with similar expected outcomes and provided probability of each major category found in different clusters. The methods link different job types and salary levels with common major categories and thus provide insights to guide students in selecting majors aligned with their career goals.

Introduction:

As economic policies and access to higher education continue to improve, more and more people are entering the job market with bachelor's degrees. While recruiters benefit from this influx of qualified candidates, it also increases competition among recent graduates. As a result, the skills acquired through academic pursuits have become an important part of the job-search process, making it even more important for students to choose a major when they enter college. In 2021, the US Bureau of Labor Statistics officially published an article discussing the linkage between college majors and career prospects including survey results that cover the employment growth and wage changes in more than 10 occupations (Torpey et al., 2021). Such a wide-spread social phenomenon calls attention to the current college students as well as prospective students to reflect on the economic values of their college majors and start to think about one question: is getting my major really worth it?

Students' decisions about their college major are heavily influenced by data about the job market. Researchers have found that college students choose majors first based on their own interest and belief on course enjoyment and grade, followed by concerns of the labor market outcomes, but the weight of the job concern is growing rapidly (Baker et al., 2018). Moreover, the same study shows that students tend to pay special attention to the average pay of jobs associated with their direct major, increasing their possibilities to choose a major with higher future salary by 10-20%. To better predict the major choice of college students given the current statistics like salaries and employment rate in the job market, a comprehensive analysis of data related to college majors and post-graduate employment outcomes promises to provide valuable insights into the future of the job market.

The focus of this project is the exploration of questions surrounding which majors are likely to yield the highest employment rates, salaries, and long-term career benefits in current circumstances. Using such metrics, we can predict the probability of certain major graduates entering job positions with different levels of pay rate. Not only will this study help students make informed

decisions about their majors, but it will also help them strategically plan their educational and career paths.

Related Work:

Previous research on employment of college graduates has focused on science, technology, engineering, and mathematics (STEM) majors, with the focus on engineering majors (Casselman, 2014). In this project, we're expanding our analysis to include non-STEM majors as well. We'll be looking at things like male-to-female graduate rates, graduate age distributions, and the percentage of graduates working in their field of study versus those working outside of it.

While previous studies have focused on calculating simple statistics (average annual earnings, median, percentiles) on annual earnings and unemployment rates (Torpey et al., 2021) (Hussar, 2021), our approach is to look at a college major's economic potential. This means combining multiple factors, such as annual pay and the employment rate, to get a comprehensive view of the prospects for each major.

In the upcoming sections, we will introduce the methodologies, results, and discussions surrounding our study, shedding light on the intricacies of predicting employment rates, unemployment rates, and salaries for different college majors based on factors like gender and age group.

Methods:

The dataset contains real-life data with both categorical and numerical components. The data features and variable types are summarized in *Supplementary Fig. 1*. Appropriate features that are highly correlated with major categories like median salary, unemployment rate, and number of low wage jobs are selected to load on the K-Means model. Feature engineering converted count features like employed, low wage jobs, and full time jobs to rate features to remove the effect of different numbers of observations within each major, as shown in *Supplementary Table 1*. Other numerical values, like median earnings, were also standardized and scaled using *sklearn StandardScaler()* for comparison. The full database was split into training and testing data using *sklearn train_test_split()*.

Our main model clusters the majors based on selected features using the K-means algorithm and outputs a probability matrix containing the probabilities of each major category in each cluster. The elbow method was applied to determine the optimal number of clusters (K), as shown in *Supplementary Fig. 2*. Input features for the K-means algorithm include median salary, unemployment rate, employed rate, rate of jobs requiring a college degree, low wage job rate, full and part time job rate.

Training of the K-means model was achieved by the *sklearn Kmeans()* function with n_clusters = 3. The probability of each major appearing in each cluster was calculated and stored in a probability matrix. Major categories absent in a certain cluster were assigned with the probability equal to an adjusting constant $\varepsilon = 0.01$, whose value was subtracted equally from the major categories with a real probability to ensure the total probability for all categories still equals 1.

After our model has grouped majors into clusters, we analyzed the Clusters by extracted cluster centers and inversely transformed them to interpret the original

feature scales. We organize the cluster center data into a DataFrame named cluster_data to help with the understanding of the characteristics associated with each cluster. This DataFrame has columns corresponding to the features used in the clustering process. Each row in this DataFrame represents a cluster, and the values within each column are the average of the respective features for that cluster.

We evaluated our clustering performance using Inertia and Silhouette Score.

Inertia measures the sum of squared distances within clusters, with lower values indicating better-defined clusters. Silhouette Score, on the other hand, assesses how similar an object is to its own cluster compared to other clusters, with higher scores being more favorable.

Our Inertia value is 358.55. It indicates that the points within each cluster are closer to the centroid, suggesting compact and well-defined clusters. Our Silhouette Score is 0.31. It is not extremely high but still indicates well-defined clusters.

The trained model was then applied to assign clusters for the testing data and output the same probability matrix. Model performance was evaluated by the same loss function as the logistic regression, where y^i is the probability of a major appearing in a certain cluster and $h(x^i)$ is the predicted probability of applying the model on testing data, summarized across all major categories m.

$$-\frac{1}{m}\sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))]$$

The additional reference/baseline model was also constructed to compare with the performance of the K-means model. The single feature, median salary, was used in this model. Training data was split 3-way equally based on low, middle, and high median wage and the probability of each major category in each wage level was computed. The same wage cutoff found in the training data was applied on the testing data to split it into 3 wage groups and the same probability matrix was calculated. The same loss function comparison was conducted on this reference model.

Results:

By grouping majors based on shared characteristics, we can visualize the clusters formed by the K-means algorithm and then identify the clusters with the best career prospects.

Therefore, we created a scatter plot visualizing the 'Average Earnings' and 'Average Unemployment Rate' for each major (*Figure 1*). Points are colored based on their assigned cluster and major names are annotated on the plot, but to avoid clutter, only every 20th point is annotated.

Majors categorized into the best clusters are likely to have high average earnings and low unemployment rates, indicating strong career prospects. Cluster 3 (green) was found to be the best cluster, where major categories of “Engineering”, “Education”, and “Computers & Mathematics” are found in highest probabilities. These three major categories together occupy over 80% majors in cluster 3, exceeding all other major categories by a large margin.

Major categories of “Biology & Life Science” and “Humanities & Liberal Arts” are most likely to appear in cluster 1 (red), the cluster with the lowest average salary and highest

average unemployment rate. However, these two categories together only occupy less than 40% of all majors in cluster 1 and the rest of major categories have a relatively even probability distribution.

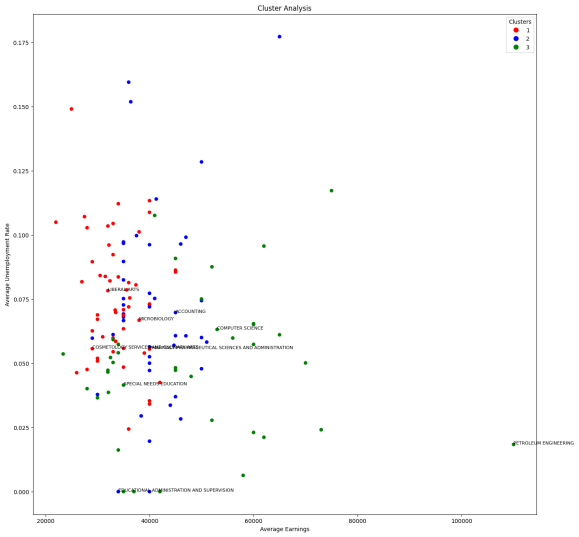


Figure 1. Visualization of assigned clusters by average earning and average unemployment rate of the majors. Clusters 1, 2, 3 are colored red, blue, and green.

Model performance was evaluated using the loss function described above. Comparison of loss function values are summarized across clusters for the K-means and baseline model in *Table 1*.

Table 1. Summary of loss function calculated for each cluster in both models.

Loss Function	K-Means Model	Baseline/Reference Model
Cluster 1	0.32262	0.18555
Cluster 2	0.30884	0.22296
Cluster 3	0.23084	0.16847

Discussion:

We chose K-means as our main model because it looks at multiple metrics simultaneously (earning potentials, unemployment rates, etc.) without having to manually assign feature importances. This allows us to find majors with similar overall expected career outcomes. Also, K-means is an unsupervised learning algorithm, so it looks for inherent patterns in the data without being limited by any predefined labels.

While exploratory analysis of this dataset revealed general trends of how different features are correlated and which features are associated more tightly with the mean salary and unemployment rate, the k-means clustering provided nuanced groupings of majors that go beyond just separating by major categories. It helps identify less obvious majors that perform similarly to top majors as well as majors that seem high-potential but actually underperform. For example, majors like “Soil Science”, “United States History”, and “Construction Services” were grouped with engineering and education majors in the best-performing cluster 3. Still, these nuances don’t harm the big picture formed by this model, which provided top major categories that are most frequently seen in top-performing clusters. Undeniably, engineering and educational majors still outperform most other major categories and are highly associated with high median salary and low unemployment rate.

The loss functions computed for each cluster vary slightly and suggest that cluster 3 predicts the major category probabilities the best, probably due to the dominant occupancy of major categories “Engineering” and “Education” in this cluster. However, the K-means model does not perform better than the baseline model, which in this way can be called the

reference model. The reference model intakes only one feature, median salary, as the determining factor that splits major categories into 3 groups. Its outperformance of the K-means model suggests that sometimes simplicity is the key, and that median salary is probably the most important feature to cluster majors in terms of their economic potentials. Future expansion on this study includes the incorporation of more recent dataset and potential inclusion of additional features that help determine a major's economic potential like promotion opportunities, job stability, and local versus international jobs.

Conclusion:

In conclusion, our K-means and reference models provide insight into major-major relationships in terms of economic payback after graduation and help predict one's major given the median salary, employment success, and job type. It confirmed the dominance of engineering as well as educational majors in the job market in recent years. In the end, this project helps us gain more insight into the economic value of different major categories and have great potential for continued study when future data are available.

References:

[1] Torpey, E. (2021, January). *Linking college majors to careers : Career outlook*. U.S. Bureau of Labor Statistics. https://www.bls.gov/careeroutlook/2021/article/field-of-degree-and-careers.htm

[2] Baker, R., Bettinger, E., Jacob, B., & Marinescu, I. (2018, June 5). *The effect of labor market information on Community College Students' major choice*. Economics of Education Review. https://www.sciencedirect.com/science/article/abs/pii/S0272775718300566

[3] Casselman, B. (2014, September 12). *The Economic Guide to picking a college major*. FiveThirtyEight. https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/

[4] Hussar, W. J., & Nachazel, T. (2021). *The condition of Education, 2020*. National Center for Education Statistics, U.S. Department of Education, Institute of Education

Contributions:

Alison Qiu ran the K-means clustering and determined the appropriate parameters using the elbow method and the silhouette method. Maggie constructed the baseline model and calculated the output probability matrix and loss function for both models. Both members contributed to the writing of this report as well as the oral presentation.

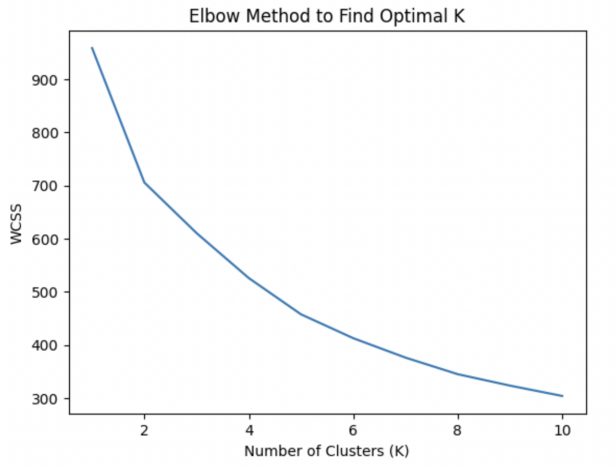
Supplementary Materials:

Supplementary Figure 1. Snip shut from the running code, presenting all data types and variables contained in the FiveThirtyEight College Majors dataset.

```
1 print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 173 entries, 0 to 172
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Rank                  173 non-null    int64
 1   Major_code            173 non-null    int64
 2   Major                 173 non-null    object
 3   Total                 172 non-null    float64
 4   Men                   172 non-null    float64
 5   Women                 172 non-null    float64
 6   Major_category        173 non-null    object
 7   ShareWomen            172 non-null    float64
 8   Sample_size           173 non-null    int64
 9   Employed              173 non-null    int64
 10  Full_time             173 non-null    int64
 11  Part_time             173 non-null    int64
 12  Full_time_year_round  173 non-null    int64
 13  Unemployed            173 non-null    int64
 14  Unemployment_rate     173 non-null    float64
 15  Median                173 non-null    int64
 16  P25th                 173 non-null    int64
 17  P75th                 173 non-null    int64
 18  College_jobs          173 non-null    int64
 19  Non_college_jobs      173 non-null    int64
 20  Low_wage_jobs         173 non-null    int64
dtypes: float64(5), int64(14), object(2)
memory usage: 28.5+ KB
None
```

Supplementary Figure 2. The output from the elbow method, with within-cluster sum of squares (WCSS) plotted against the number of clusters (K) used to run the K-Means model. Helps determine the optimal parameter to use (K = 3).



Supplementary Table 1. Conversion of count features to rate features to account for different sample sizes within each major.

Original Features	Engineered Features
Employed	Employed_rate
College_jobs	College_jobs_rate
Low_wage_jobs	Low_wage_jobs_rate
Full_time	Full_time_rate
Part_time	Part_time_rate