DS 3000 Project

•••

How does undergraduate major and alma mater affect salary throughout different stages of one's career?

Aanay Anandpara, Kaiya Clarke, Lon Pierson, Joshua Yu

Our Dataset

- From the Wall Street Journal
- 3 CSV Files
  - Undergraduate major vs. salary (starting career median, mid-career median, mid 10th 25th 75th 90th percentiles)
    - 50 x 8, 350 data points
  - School type (Liberal Arts, Ivy League, Engineering, etc.) vs. salary
    - 269 x 8, 1883 data points
  - School region vs. salary
    - 320 x 8, 2240 data points
- For NaN values, imputation with means was performed using .transform function
  - School type and school region datasets had missing values

ML Algorithms

KNN Classifier

Used to classify either school type or region based on salaries

Modified k neighbors and used best accuracy

See if school type or region and salary are related

Linear Regression

Regressed mid-career median salary from starting career median salary for each major

See if early career salary is indicative of expected future salaries

k-Means Clustering

Used to cluster different majors based on their salaries

See if there is a pattern in clustering of majors such that salaries of certain “types” (engineering, business, humanities) of majors are similar

Insight: School type can dictate pay, region does not (KNN)

School Type: KNN Model had 80% accuracy at 17 neighbors for school type. School type, then, can be said to be a determinant of salary level throughout one's career

Accuracy Report:

	precision	recall	f1-score	support
Engineering	1.00	0.50	0.67	6
Ivy League	1.00	1.00	1.00	2
Liberal Arts	0.80	0.57	0.67	14
Party	0.00	0.00	0.00	6
State	0.79	0.98	0.87	53
accuracy			0.80	81
macro avg	0.72	0.61	0.64	81
weighted avg	0.75	0.80	0.76	81

School Region: KNN Model had just 39% accuracy at 9 neighbors. Region of school is not as important in determining salary level.

Accuracy Report:

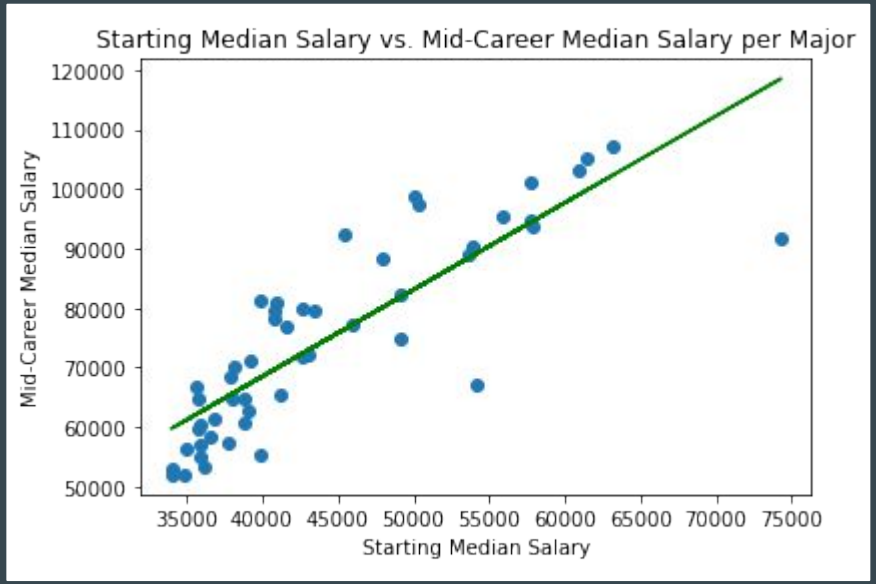
	precision	recall	f1-score	support
California	0.27	0.38	0.32	8
Midwestern	0.29	0.43	0.35	21
Northeastern	0.53	0.53	0.53	30
Southern	0.30	0.25	0.27	24
Western	0.75	0.23	0.35	13
accuracy			0.39	96
macro avg	0.43	0.36	0.36	96
weighted avg	0.43	0.39	0.38	96

Insight: Starting career salary for each major is largely indicative of mid-career salary (Linear Regression)

Linear Regression shows that major's starting median salary can be used to predict mid-career median salary. Equation: y = 1.46x+ 10171.97

Moderate-to-strong r-squared of 0.72

However, variability could serve as a point of further exploration to explore which majors have disproportionately large jumps between starting and mid-career salary



Insight: Certain “kinds” of majors pay differently (KMeans)

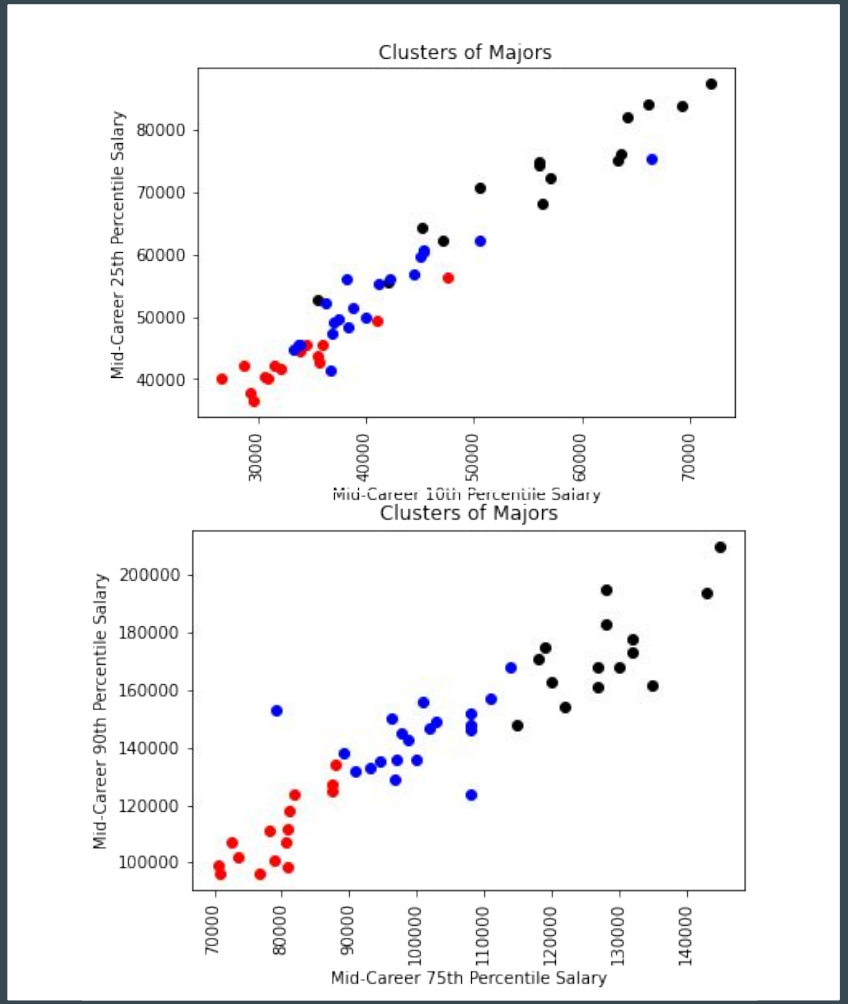
Clusters were well-defined and were clearer as percentile of salary increased (75th and 90th)

- **Increase in cluster definition as percentile increases shows there are some majors which have lower pay on the low end but pay very well at the upper end of the scale**

Cluster 1 (Red)
- Lowest paying, typically humanities and social sciences

Cluster 2 (Blue)
- Mid-tier, typically lower-end of business majors (MIS, business management) and STEM (chemistry, anthropology) or higher-end of humanities (Poli Sci, IR, Comms)

Cluster 3 (Black)
- Highest paying, almost entirely STEM (engineering and computer science) with higher-end of business majors (finance, marketing)



Conclusion

kNN Classifier	• School type plays a role in salaries one can expect • Region of school is not important in predicting salaries
Linear Regression	• Major starting career salary is correlated with mid-career salary • Some variability in correlation leaves room for exploration of large increases in salary from early to mid-career for some majors
K-Means	• Majors in related fields pay similarly • Undergrads should explore variety of majors within a field to find best fitting majors aligned with goals and interests