DS4ALL Final Project

Joseph Jung

Part 1: Introduce Data Set

Introduction of Data Set

- This dataset was found in Data.ca.gov under the title "Short-Term Occupational Employment Projections"
- Description: Short-term Occupational Projections for a 2-year time (2023-2025)produced for the state of California.
- Purpose: Data helps to make informed decisions on individual career and organizational program development

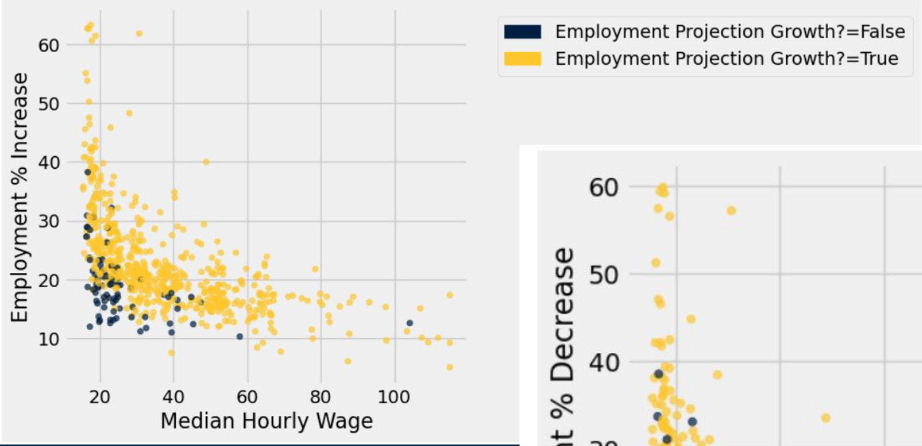
Variables

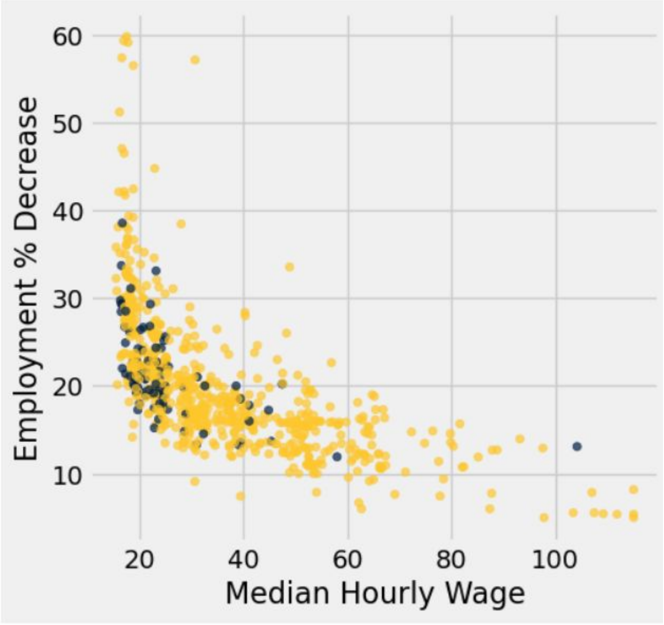
- "Occupational Title" - (Categorical)
- “"Base Quarter Employment Estimate" - measures the employment for an occupation in 2023 (Numerical)
- "Projected Quarter Employment Estimate" - measures the predicted employment for an occupation in 2025 (Numerical)
- "Numeric Change" - measures the predicted change in employment between the two years for occupations (Numerical)
- "Percentage Change" - measures the predicted percentage change in employment between the two years (Numerical)
- "Exits" - measures the amount of people that have left an occupation (Numerical)
- "Transfers" - measures the amount of people that left an occupation and transferred to a different one (Numerical)
- "Total Job Openings" - measures the openings (positions) for workers entering an occupation (Numerical)
- "Median Hourly Wage" - (Numerical)
- "Median Annual Wage" - (Numerical)
- "Entry Level Education" - (Categorical)
- "Work Experience" - (Categorical)
- "Job Training" - (Categorical)

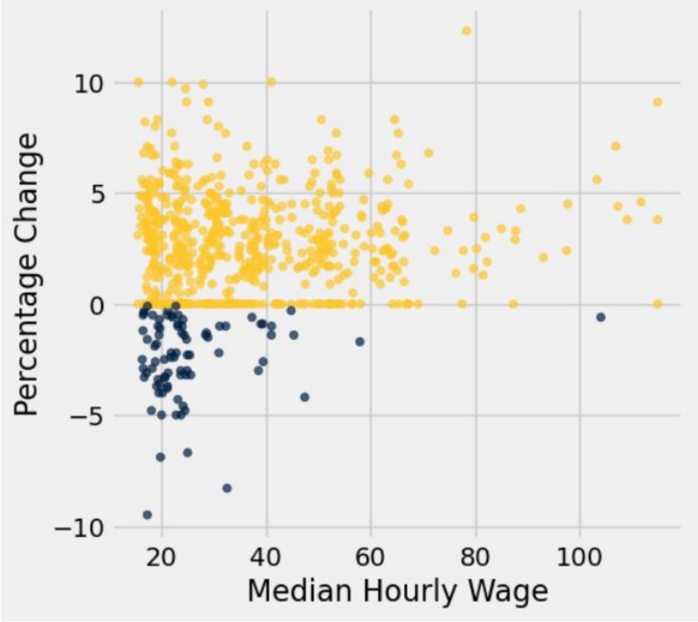
Initial Raw Data cleaned:

Occupational Title	Base Quarter Employment Estimate	Projected Quarter Employment Estimate	Numeric Change	Percentage Change	Exits	Transfers	Total Job Openings	Median Hourly Wage	Median Annual Wage	Entry Level Education	Work Experience	Job Training
Total, All Occupations	19920000	20412400	492400	2.5	1981550	2469420	4943370	24.73	51450	nan	nan	nan
Management Occupations	1669600	1705600	36000	2.2	102730	153320	292050	63.9	132918	nan	nan	nan
Top Executives	335600	342000	6400	1.9	17520	34380	58300	0	0	nan	nan	nan
Chief Executives	51800	51500	-300	-0.6	3190	3590	6480	104.13	0	Bachelor's degree	5 years or more	nan
General and Operations Managers	281500	288100	6600	2.3	14180	30600	51380	57.11	118793	Bachelor's degree	5 years or more	nan
Legislators	2300	2300	0	0	140	190	330	0	61446	Bachelor's degree	Less than 5 years	nan
Advertising, Marketing, Promotions, Public Relations, an ...	187100	189500	2400	1.3	9010	18930	30340	0	0	nan	nan	nan
Advertising and Promotions Managers	5200	5200	0	0	210	690	900	65.4	136038	Bachelor's degree	Less than 5 years	nan
Marketing Managers	60300	61100	800	1.3	2880	6580	10260	81.59	0	Bachelor's degree	5 years or more	nan
Sales Managers	108600	109900	1300	1.2	5320	10500	17120	63.81	132734	Bachelor's degree	Less than 5 years	nan

Pairs of Numerical Variable I Looked At

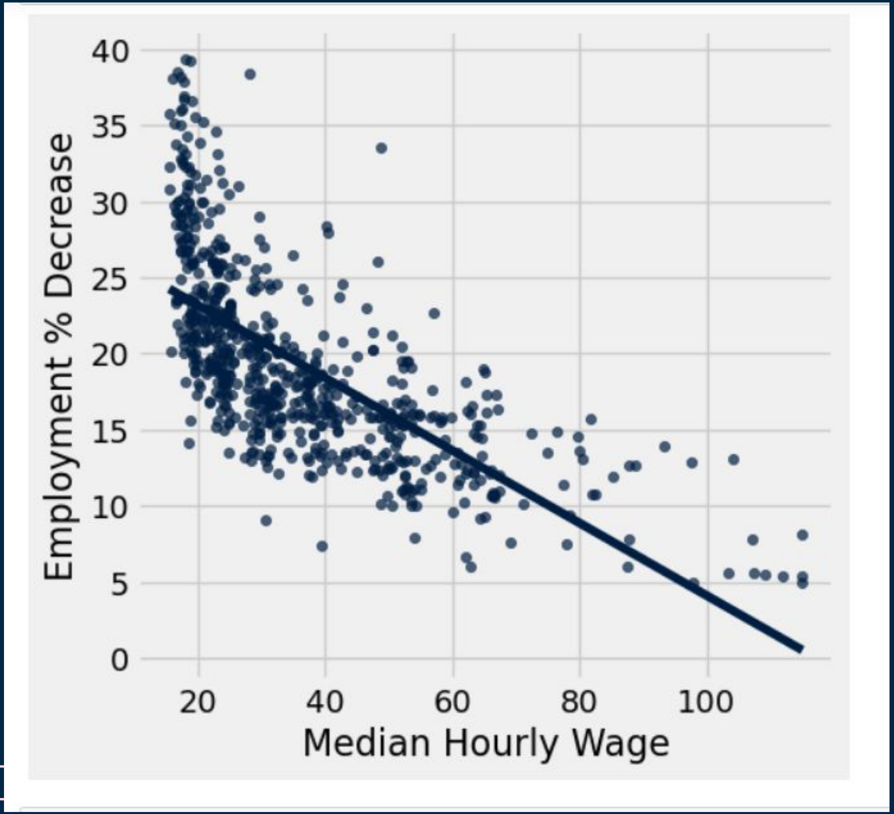






Part 2: Regression / Correlation Results

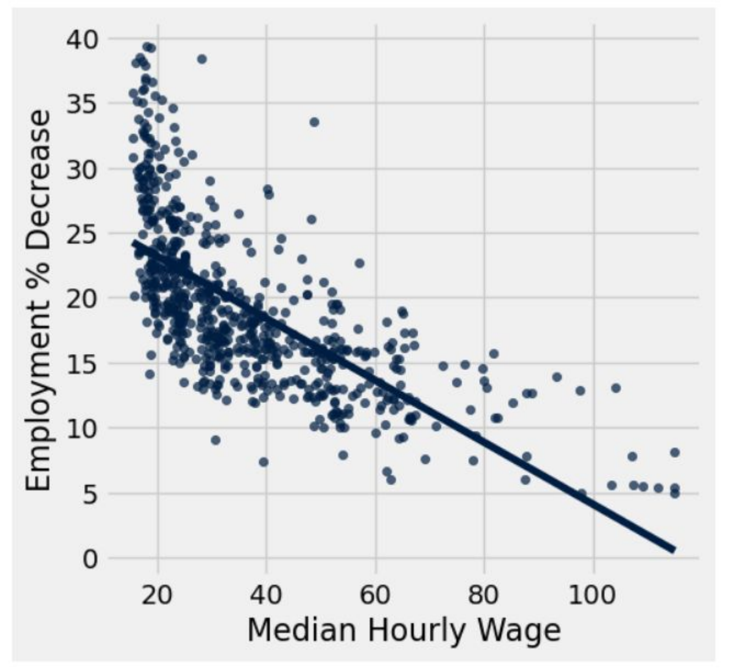
Linear Regression



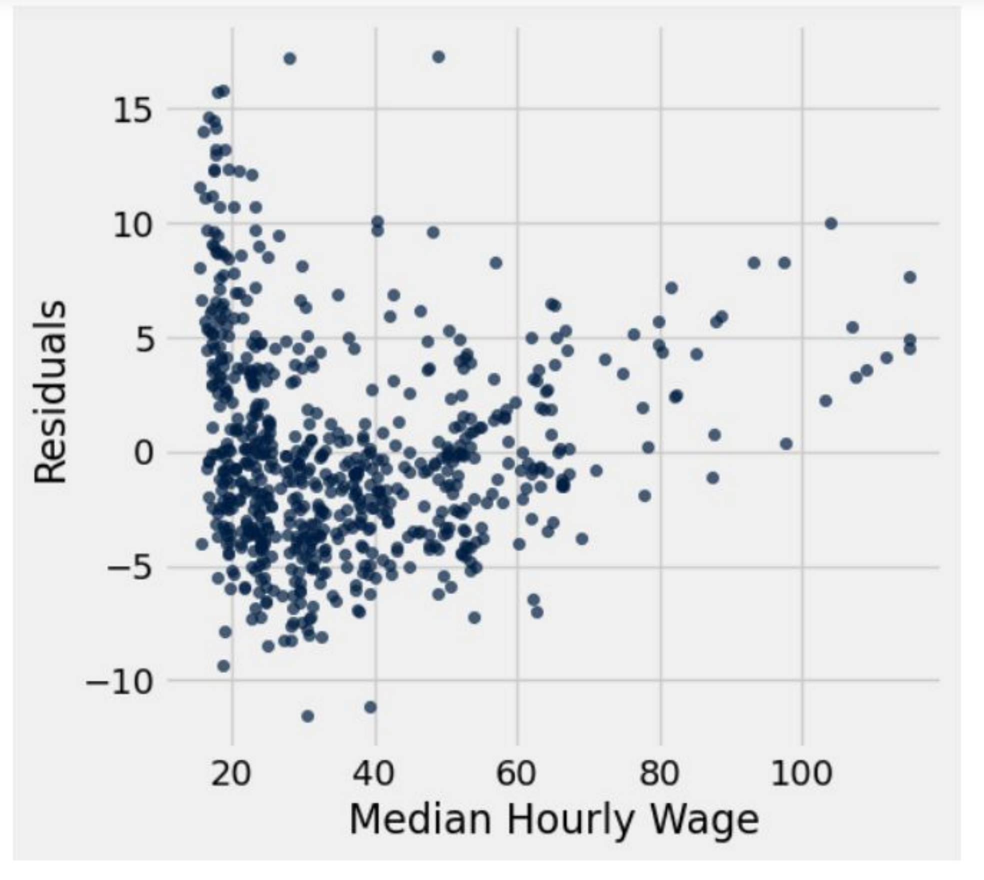
- Null Hypothesis: There is no correlation/association between the variables, and the data values are what they are solely due to chance.
- Alternative Hypothesis: There is a correlation/association between the variables. Jobs with higher hourly wage will likely have employment % decrease that are changing towards one direction (higher or lower) and jobs with lower hourly wage will likely have employment % decrease that are changing towards the other direction.
- Slope: -0.23864
- Intercept: 27.96

Confidence Interval for Correlation

- Correlation: -0.674 (Around)
- 95% CI: [-0.7180702538883847, -0.6569951027043751] ,
- Reject the null: True



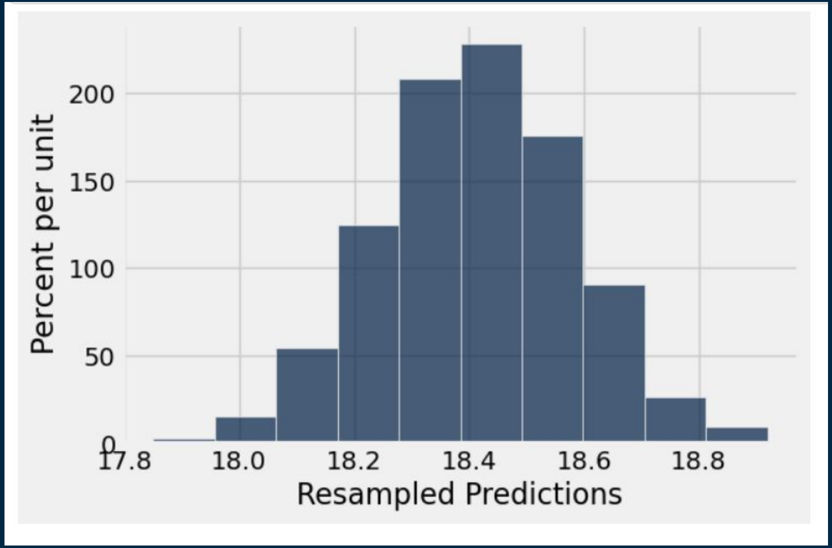
Residuals Plot



Value of X for Prediction

My x value: 40 ($) (Median Hourly Wage)

- In California, an Hourly Wage of $40 is categorized under Top Earners according to ZipRecruiter.
- Y value Result: 18.345



99% CI for Y Prediction:

- [17.987474720542956, 18.845318591845107]

Potential Biases

- occupations with greater than or equal to 40% employment decrease are not accounted for.
- This causes my regression model to not fully resemble every occupations from my raw data set
- rows in my data are not conventional individual beings, but rather different occupations.
    - occupations have various different factors. One might have a lower end wage and have a higher employment % decrease while another occupation with similar wage may have a significantly lower employment % decrease.

Part 3: First Pass Classifier

KNN Classifier Algorithm

Attributes:

- Base Quarter Employment Estimate', 'Projected Quarter Employment Estimate', 'Numeric Change', 'Percentage Change', 'Exits', 'Transfers', 'Total Job Openings', 'Median Hourly Wage', 'Median Annual Wage', 'Exit %', 'Transfer %', 'Employment % Increase', 'Employment % Decrease'

Training/Testing Set:

- Training set: 506 examples
- Test set: 169 examples

Accuracy/Evaluation

- K: 3
- Attributes: previous slide
- Accuracy: 0.8816568047337278

Employment Projection Growth?	Exit %	Transfer %	Employment % Increase	Employment % Decrease	Prediction	Was correct
False	13.5484	11.6129	21.9355	25.1613	True	False
False	8.04487	6.76282	13.5256	14.8077	True	False
True	9.15749	11.8526	23.3078	21.0101	False	False
False	8.40517	12.5259	19.1207	20.931	True	False
False	6.40758	12.7773	19.09	19.1848	True	False
False	8.11189	12.7506	20.3963	20.8625	True	False

True	10	14	24	24	False	False
True	7.27273	14.7727	22.0455	22.0455	False	False
False	13.5	16.25	27.25	29.75	True	False
False	8.36667	14.3833	22.25	22.75	True	False
True	14.3137	14.3137	34.5098	28.6275	False	False
False	10	9.5	14.5	19.5	True	False
False	10.7143	17.7381	27.2619	28.4524	True	False
False	13.6885	6.52095	16.02	20.2095	True	False
False	7.27273	12.7273	16.9697	20	True	False

Evaluation of misclassifications

Majority of the occupations in the table are projected to have an employment shrinkage (being marked with 'False' for the column 'Employment Projection Growth?')(11 out of the first 15 rows). With the classifier failing to predict correctly for such occupations (despite there being significantly less occupations that are projected to shrink) we can assume that the classifier is specifically biased against occupations projected to have employment shrinkage.

Part 4: Improvements Tried

Accuracy/Evaluation

- K: 15
- Attributes: "Median Hourly Wage", "Employment % Increase", "Employment % Decrease"
- Accuracy: 0.9408284023668639

Employment Projection Growth?	Exit %	Transfer %	Employment % Increase	Employment % Decrease	New Prediction	Was correct
False	13.5484	11.6129	21.9355	25.1613	True	False
False	8.04487	6.76282	13.5256	14.8077	True	False
False	6.40758	12.7773	19.09	19.1848	True	False
False	8.11189	12.7506	20.3963	20.8625	True	False
False	13.5	16.25	27.25	29.75	True	False
False	8.36667	14.3833	22.25	22.75	True	False

False	10.7143	17.7381	27.2619	28.4524	True	False
True	7.27273	10.9091	18.1818	18.1818	False	False
False	13.6885	6.52095	16.02	20.2095	True	False
False	7.07071	13.9394	20	21.0101	True	False
True	5.74572	10.2934	17.5061	16.0391	True	True
True	11.1744	16.3345	28.3986	27.5089	True	True
True	5.73477	9.13978	16.3082	14.8746	True	True
True	11.7391	18.2609	30	30	True	True
True	11.3529	15.5882	29.2941	26.9412	True	True

Evaluation of misclassifications + Comparison

My new classifier has about a 6% (0.06) increase in accuracy from my previous classifier. Similar to the old classifier, majority of the incorrect predictions are made for jobs projected to face employment shrinkage, but the new classifier was able to predict correctly more of such jobs (9 out of the first 15 rows had "false" for "Employment Projection Growth?" column)(compared to 11 out of the first 15 for the first classifier).

Conclusion

Changes I Made to Improve the Classifier

I tried increasing the K from the previous value of 3, and learned that at ranges 11 to 17 (odd only), we get the highest accuracy. For the features, I removed all attributes from the first classifier except for "Median Hourly Wage", "Employment % Increase", and "Employment % Decrease." Due to the fact that the result in "Employment Projection Growth?" is dependent on how many employees left and joined an occupation during a certain year, attributes like "Employment % Increase"/"Employment % Decrease", and "Exits" + "Transfers"/"Total Job Openings" (that measure the joining/leaving of employees) can be used to improve the classifier. K values 11 - 17 had the most consistency in highest accuracy, and thus, using a value from that range can also improve the classifier.

What I learned about my Dataset + Initial Question Answered

By conducting regression inference on my dataset, I learned that variables like 'income level' and 'occupation employment projection' are not as closely correlated as I had expected. I initially chose my research topic being curious about whether technological advancements have really shrunk the employment numbers of lower end jobs compared to higher end jobs. However, the correlation was only moderately strong, and its value was achieved by removing occupations that had more than 40 in 'Employment % Decrease' as they would otherwise reduce the linearity. Through the classification section, I learned that one can make descently accurate predictions about occupation employment projection by using transfers %, exits %, and job openings % to derive variables like "Employment % Decrease" and "Employment % Increase" and using them as attributes (+ Hourly Wage) to a classification algorithm. I also learned that using hourly wage as attributes is better than using annual wage and employment % increase as attributes is better than using employment % decrease at classifying accurate employment projections.