**Short-Term Occupational Employment Projections Report**

**Pham Duy Anh Vu**

**A Data Analytics Report**

**I. Introduction:**

The link: [Short-Term Occupational Employment Projections | Data.gov](https://catalog.data.gov/dataset/short-term-occupational-employment-projections/resource/9628b1ce-9fed-40b2-bf76-67aae817e793)

The CSV: [Short.Term\_Occupational\_Employment\_Projections.csv](https://data.edd.ca.gov/api/views/guh4-bakw/rows.csv?accessType=DOWNLOAD)

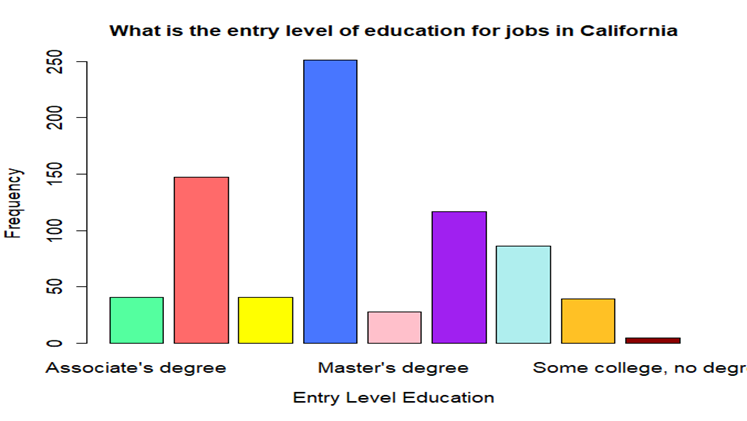
Occupational employment projections data is an integral part of the government’s database. This methodology is developed in the Labor Market Information (LMI) sections of each State Employment Security Agency (SESA) to gather information about the staffing patterns used in projections, employment change, occupational openings, education, training, and wages for each detailed National Employment Matrix occupation.

Developed by the California Employment Development Department, this dataset is about the short-term projections about occupational employment for a 2-year time horizon produced by the state of California for individuals and organizations with an occupational outlook to make informed decisions on careers and development. It is comprised of 18 different variables and a total of 755 objects, ranging from different job qualifications (entry level education, work experience, total job openings) to the specifics of the business cycles of each job (base and projected quarter employment estimate, exits, transfers, wages)

This report will go more in depth, classifying, showing the purpose and relationship between each factor.

**II. Visualizations:**

**2.1. Entry level education for jobs in California:**



*Figure 1: Entry level education of jobs in California (Full image:* [*[Image]*](https://drive.google.com/file/d/1LyFmWegpAo7rk5WGktN-UCm0C6NnWzZm/view?usp=sharing)*)*

Based on figure 1, most jobs require some type of degree from their employees with most requiring a college-level degree or higher. Despite this, the most common requirement Californian jobs require is just a high school diploma or a degree of equivalent value, totaling approximately 250 different jobs. The second most common requirement is a bachelor's degree by around 150 jobs, and surprisingly followed by just under 100 jobs not requiring any formal educational credentials. Despite this, and ignoring N/A values, it can be extrapolated that the higher the difficulty of the credentials, the fewer it is required.

**2.2. Median Wage – Hourly vs Annual:**

Chart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

*Figure 2: Median Hourly Wage Histogram Figure 3: Median Annual Wage Histogram*

Figure 2 and figure 3 are histograms of the median hourly wage and median annual wage

of different jobs in California, respectively. While not entirely the same, both histograms do exhibit similar patterns. It can also be noted that the most jobs have a median hourly wage of just under $25, coinciding with a median annual wage of just below $50000.

**III. Clustering model:**

**3.1. Hierarchical Clustering:**

A picture containing text

Description automatically generated

*Figure 4: Dendrogram of the business cycle – California*

With the variety of jobs offered in California to choose from, it is important to understand how stable the economy of each job is. The purpose of the hierarchical clustering model is to divide the 755 different jobs offered into groups based on their economic stability. In order to get desired results, as shown by the dendrogram in figure 4, 5 variables are chosen: Base.Quarter.Employment.Estimate, Projected.Quarter.Employment.Estimate, Exits, Transfers, and Total.Job.Openings. From the dendrogram, the business employment cycles of Californian jobs can be distributed into six distinct groups based on the level of stability. From there, the groups are used in hierarchical clustering.

**3.2. Scatter plots of the hierarchical clustering:**

Chart

Description automatically generatedChart

Description automatically generated

*Figure 5: Scatter plot of employment estimations Figure 6: Scatter plot of exits / transfers*

Going more in depth with the data provided from hierarchical clustering, the scatter plot from figure 5 reflect the estimate between the base quarter employment and the projected quarter employment, showing that on average, the base quarter employment is less than the projected quarter employment except for one where the opposite occurred. A similar pattern can be seen in Figure 6, which reflects the relationship between the number of employees quitting and the number of employees transferring into the job. On average, there are more transfers compared to exits except for one where the opposite also occurred.

**IV. Regression Model:**

**4.1. Pairs clustering:**

To start with the regression model, a pairs clustering model is needed to figure out which specific terms (interaction terms, polynomial, etc.) to use based on the line of best fit of the relationship between each variable. To get the appropriate dataset to us to create the pairs clustering model, categorical sample units and N/A values needed to be removed. This ensures that the pairs clustering is based on an adjusted dataset with more sensible variables. Below is figure 7, the pairs model of an adjusted dataset.

Diagram

Description automatically generated

*Figure 7: Pairs clustering of an adjusted dataset*

**4.2. Linear regression model:**

Hierarchical clustering can only be used to compare two variables with one another to see how the numbers might relate. Nothing more can be inferred from the results of clustering, like whether they might affect each other and how significant their relationship is. The linear regression model shown in Figure 8 was created to solve this problem.

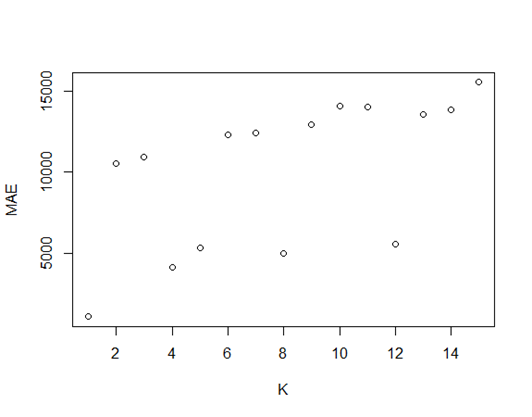
Figure 8 reflects the linear regression model using the Median.Annual.Wage variable as the response variable and comparing it with the predictor variables from a revised short-term occupational employment projections dataset. Overall, this model has an adjusted R-squared value of 0.815 (81.5%), meaning that the estimations of the model accounts for 81.5% variability in the relationship between the response variable and the predictors. In the context of the Median.Annual.Wage variable, it can be inferred that the base quarter employment estimate, median hourly wage, entry level education of either Bachelor’s degree, Doctoral, Highschool diploma, and jobs with internship/residency trainings significantly impacts the response variable. Based on the coefficient of each variable of significance, the base quarter employment estimate, median hourly wage, entry level education of either Bachelor’s degree or Doctoral all have a positive relationship with the response variable, whereas jobs with internship/residency trainings or only need high school diplomas or no formal education credentials would decrease the mean annual wage as they have a negative relationship.

Text

Description automatically generated

*Figure 8: Linear regression model*

**V. Classification Model:**

 **5.1: Model set-up:**

To set up the classification models, all the categorical variables would need to be removed. Afterwards, the dataset is split into three different datasets, a test set, a training set, and a validation set, each with the purpose of testing various aspects of the dataset for classification and making sure that the results line up with each other.

*Figure 9: KNN regression model*

Figure 9 is a k nearest neighbor regression model which uses the training set and the validation set to figure out the k value (in this case k = 1) which best illustrates the closest distance between the Total.Job.Openings of different jobs as it is the k value with the least discrepancies.

**5.2. The Classification Model:**

Chart, scatter chart

Description automatically generatedFor the next step of the classification model, the variable Job.Training had to be adjusted to be a binary variable for classification. In this case, a value of 1 in Job.Training would mean that the job offered training whereas the value of 0 would mean that there was not any training offered by the company. Based on the KNN – classification model in figure 10, the threshold of k = 1 would yield the highest accuracy of just under 100%. *Figure 10: KNN classification model*

Comparing it with the validation set and

training set (with roughly 75% and 85% accuracy respectively), it can be concluded that this method is the most efficient method. The results from the KNN – classification declares that out of the total 604 jobs (with 151 jobs having data that aren’t sensible for classification), 282 companies had a value of 0, meaning that they didn’t offer job trainings, and 322 jobs had a value of 1, meaning that they offered some kind of training for their employees.