x

A close up of a logo

Description automatically generated

IST 707

**Project Report**

**Saving the Salmon: Efficiency of Collaborative Watersheds in Oregon**

Maaz Kamal

Anishka Tahiliani

**A close up of a logo

Description automatically generatedIST 707**

Table of Contents

**Problem3**

**Major Questions3**

**Dataset Information** **3**

**Exploring Data** **4**

Correlation Matrix4

Data Summary4

Data Visualization5

Other Observations6

**Preparing Data** **7**

**Exploring and Fine-tuning ML Models** **8**

Evaluation Metrics of Trained Models8

Decision Tree Plot9

PAM Clustering 11

**Conclusion** **12**

**Problem**

Decades of overfishing, pollution, lack of legislation and building of dams and barriers in northwest pacific resulted in endangerment and listing of numerous salmonid species in 1990’s.

To revitalize their common pool resources, the aquatic ecosystems and native salmonid species, Oregon Legislature and Governor John Kitzhaber pushed for establishment of communal and collaborative organization that were sensitive to local needs and could incorporate better natural resource management practices.

So far, no one has looked upon how the structural characteristics effects the actions of collaborative governance regimes. Our project will explore how collaborative structural characteristics effect the actions undertaken in watersheds of Oregon by different collaborative governance regimes.

**Major Questions**

* How does increase in number of unique collaborators effect cost of the projects? Do we see that increase in collaborative participant increase or decrease the cost per acre or mile?
* Do projects spearheaded by watershed councils fare better than those led by other actors?
* Many of the private landowners are still skeptical of these collaboratives and see them as another way through which government can interfere on their land. Does this imply that the collaborative might be performing better on projects that are done on public lands compared to private land?
* Does the length of project determine how efficiency is improved?
* How does a mix of state funding and in-kind contributions affect the outcome of the projects?

**Dataset Information**

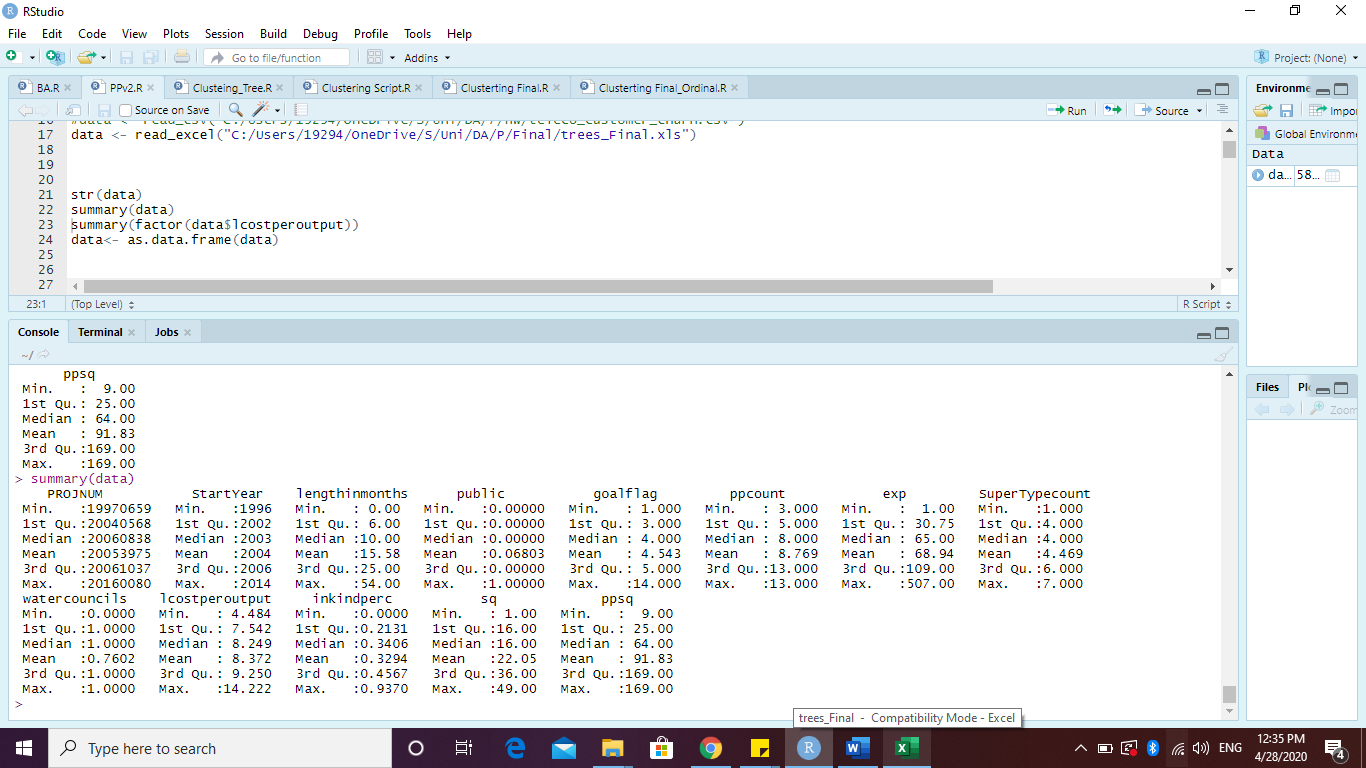
Watershed Enhancement Board’s dataset including structural characteristics of Riparian tree planting projects by different collaborative governance regimes, since late 1997, will be used for the purpose of this project.

* **(Ppcount)-** number of unique collaborators in a project
* **(SuperTypecount)-** number of Supertype collaborators in a project
* **(Public)-** Land public or private
* **(lengthinmonths)-** This is the length, in months, of project
* **(StartYear)-** When the project was started
* (**inkindperc)-** percentage of in-kind contribution to the cost of project
* **(watercouncils)-** if the project was led by a watershed council or not
* **(lcostperoutput)-** Cost per acre or cost per mile of the project

Where **lcostperoutput** is the target variable and all other variables are the potential explanatory variables.

**Exploring Data**

**Data Summary**



* Dataset includes information on projects starting between 1996 and 2014.
* Average duration of the projects is 16 months, with a maximum duration of up to 54 months.
* Number of unique collaborators in a project range between 3 and 13.
* Average in-kind contribution in a project is 30% and goes as high as 94%.
* Average cost per acre of a project is 8.3 and goes as high as 14.2 and as low as 4.5.

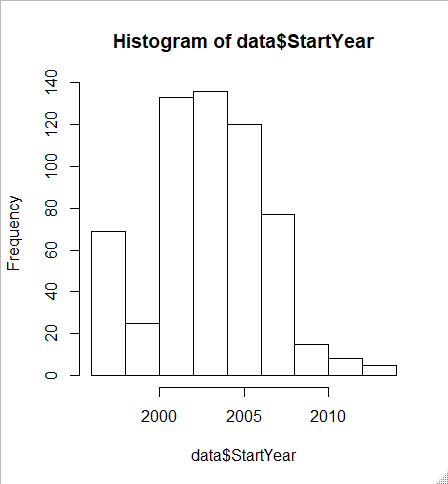
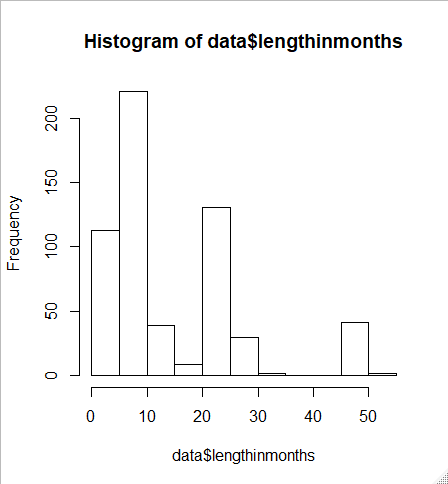
**Correlation Matrix**



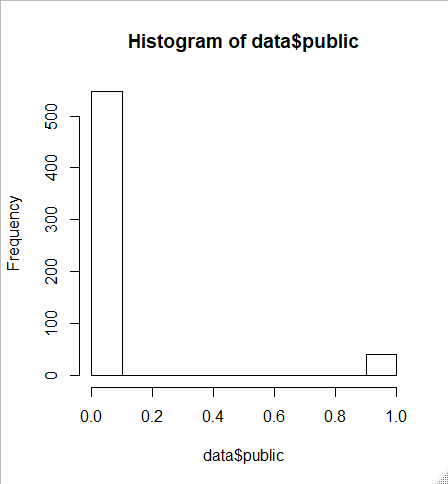
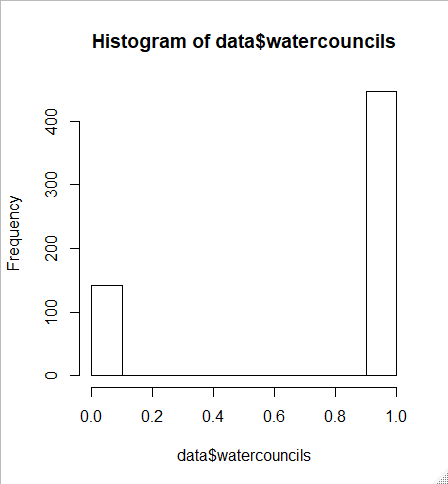
Following conclusions may be drawn from the correlation matrix of the variables.

* + StartYear is most correlated with duration of the project and cost per acre of the project. StartYear is negatively correlated with in-kind contribution of the project.
  + Duration of the project is most correlated with the number of unique and Supertype contributors in the project. Duration of the project is negatively correlated with cost per acre of the project.
  + Most significantly, the unique number of contributors and number of Supertype contributors are very highly correlated (over .94)- possibly due to duplication of information.

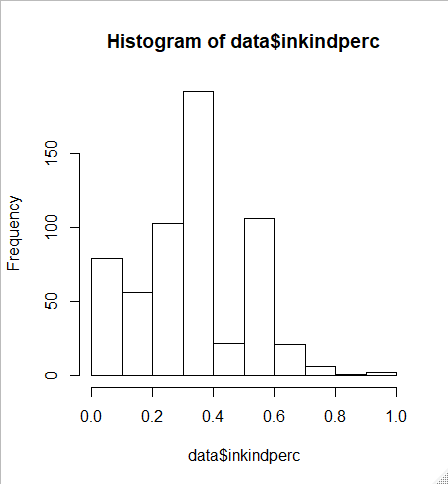
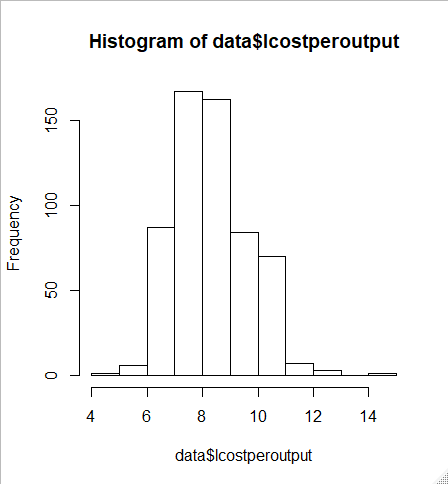
**Data Visualization**

* Most of the projects started between 2000 and 2008.
* The project duration variable is heavily skewewd.

* Almost 93% of the projects were implemented on private land – indicating a major imbalance in the dataset.
* 76% of the projects were lead by watershed councils.

* The target variable, cost per acre of the project, in the dataset approximately follows normal distribution.
* The in-kind distribution variable is right skewed. Transformation, using logarithmic function, might be neccesarry for better results.

**Other Obsersvations:**

Target variable is a continuous variable. It must be discretized into Low and High Cost per Acre so that we may test our hypothesis. And to make sure there is enough buffer between the two labels, another label for Midium Cost per Acre, will also be introduced. Hence the continuous target variable will be discretized under 3 labels.

Lastly, the dataset has mixed datatpes (that is both, categorical and numeric ). Hence appropriate tranformations (categorical to numeirc and numeric to categorical) will be required.

**Preparing Data**

* The data was screened for possible outliers in numerical columns. Some of the columns turned out to be heavily skewed or unbalanced.
* Skewed numerical variables (“lengthinmonths” and “inkindperc”) were transformed using logarithmic function, where needed.
* StartYear, Public and Watercouncil variables were factorized.

“sq”, “ppsq”, “exp” and “PROJNUM” columns were removed from the data as they do not provide any generalize-able information.

* The target variable (lcostperoutput) was discretized into 3 labels: “Low”, “Mid” and “High” using its mean, minimum and maximum values.
  + The lower bound for the Low-label was the minimum value and upper bound was the mean of minimum and average.
  + The lower bound for the Mid-label was the average and upper bound was the mean of average and maximum value.
  + The lower bound for the High-label was the mean of average and maximum value and upper bound was the maximum value.
* Another version of dataset was created where numerical variables were discretized. So that dataset could be optimally utilized for machine learning algorithms like Naive Bayes Classifier.
  + “Lengthinmonths” and “inkindperc” were also discretized into Low, Medium and High labels using the values of minimum, maximum and mean- as explained above.
* Another version of dataset was created where categorical variables were converted using dummy variables, so that dataset could be used for machine learning algorithms like Support Vector Machine and k-Nearest Neighbors.
  + “StartYear”, “ppcount” and “SuperTypecount” were converted to dummy variables.
  + The data was randomly divided into 2 sets: training data (80%) and test data (20%). However, using “set.seed()” it was made sure that both sets reflect the same target classifications ratio as it was in the original dataset.

**Exploring and Fine-tuning** **ML Models**

A total of 4 models were trained using Decision Tree, Naive Bayes Classifier, Support Vector Machine and k-Nearest Neighbors algorithms, resulting in the following evaluation metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | F-Score of Low  Cost Per Acre | Recall of Low  Cost Per Acre |
| Decision Tree | 81.36% | 77.27% | 85.00% |
| Bayes Model | 99.15% | 97.87% | 100.00% |
| SVM Model | 76.27% | 69.77% | 78.95% |
| KNN Model | 79.66% | 73.68% | 100.00% |

Upon inspection of our target classifications, it was found that “Low” Cost per Acre project was only present in 21% of the all records. Hence, instead of accuracy (which would be mainly contributed by large number of True Negatives), using F-Score might be a good idea.

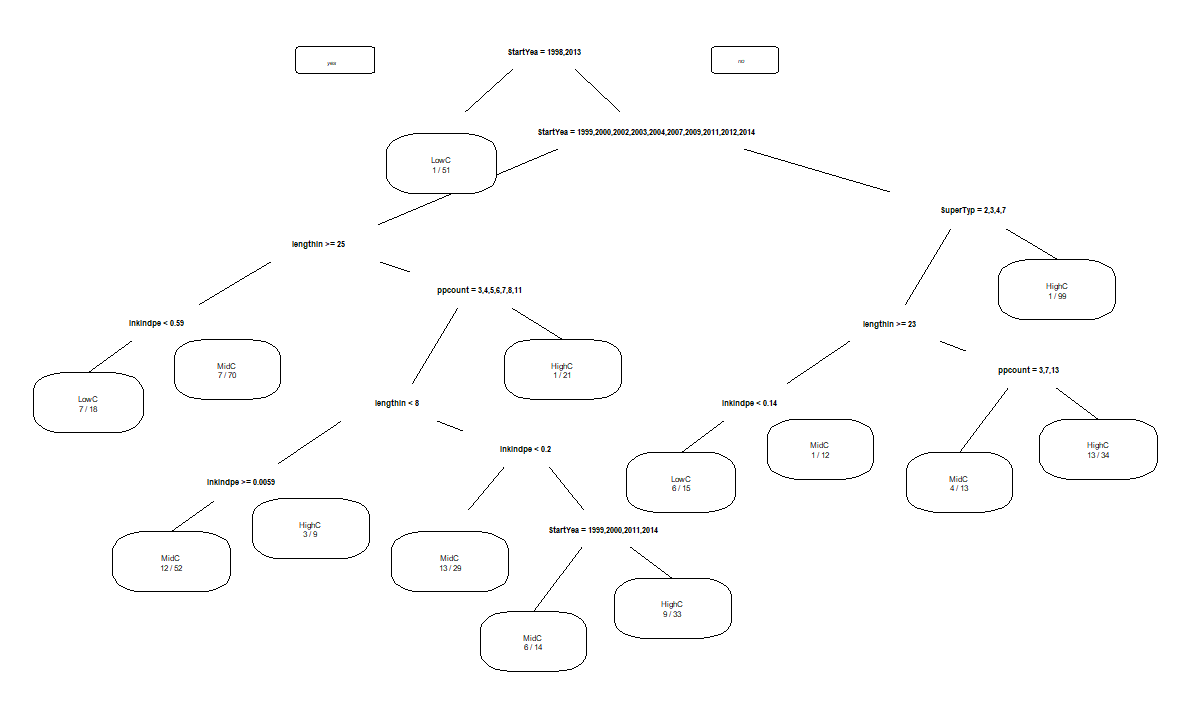
Furthermore, since we are most concerned with Low Cost per Acre projects, we can’t afford to misclassify these projects (False Negative), hence “Recall” for “Low” Cost per Acre projects should also be considered.

As far as the prediction is concerned, clearly Bayes Classifier is most suitable with its high Recall and F-Score. However, it does not do a great deal as far as explaining the model or underlying patterns are concerned.

For the purpose of explaining underlying data patterns, Decision Tree (provided below) is much more effective. It has approximately second highest Recall and F-score values among the 4 tested Models.

Since the main purpose of this project was to identify patterns for Low Cost per Acre projects, we will go ahead with Decision Tree Model and try to tweak variables so that we may deduce/predict most effectively.

**Decision Tree with StartYear**

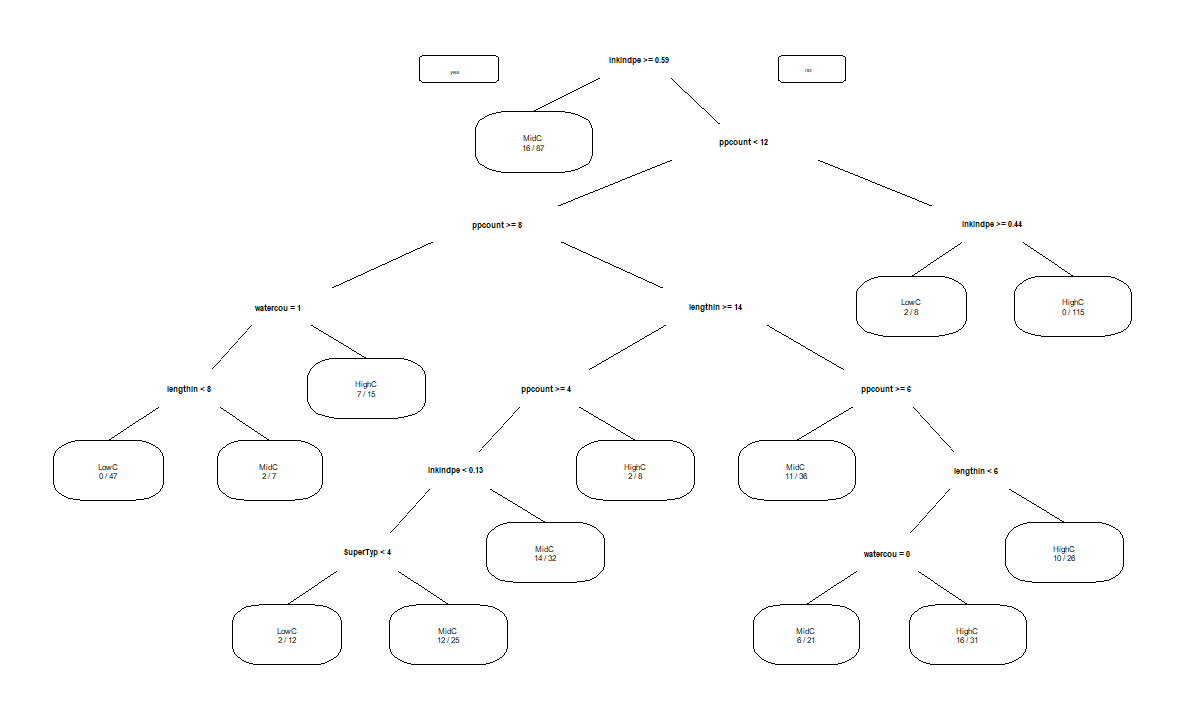


While Decision Tree including “StartYear” variable results in higher accuracy (81.36%), it might not be the most usful in terms of deduction. While clearly “StartYear” of the project has significance, it does not add any generalizable information as far as understanding the patterns or predicting new projects are concerend. Nevertheless, following are the most significant patterns:

* Projects started in 1998 or 2013 were expected to have low cost per acre.
* Projects not started in 1998 or 2013, were expected to have low cost per acre only if they had long duration (“lengthinmonths”> 25) and in-kind funding was less than 56%.

Next, “StartYear” was removed from the explanatory variables and Decision Tree was remodeled. Accuracy reduced to 75.42%, however inference became much more relevant.

**Decision Tree without StartYear**



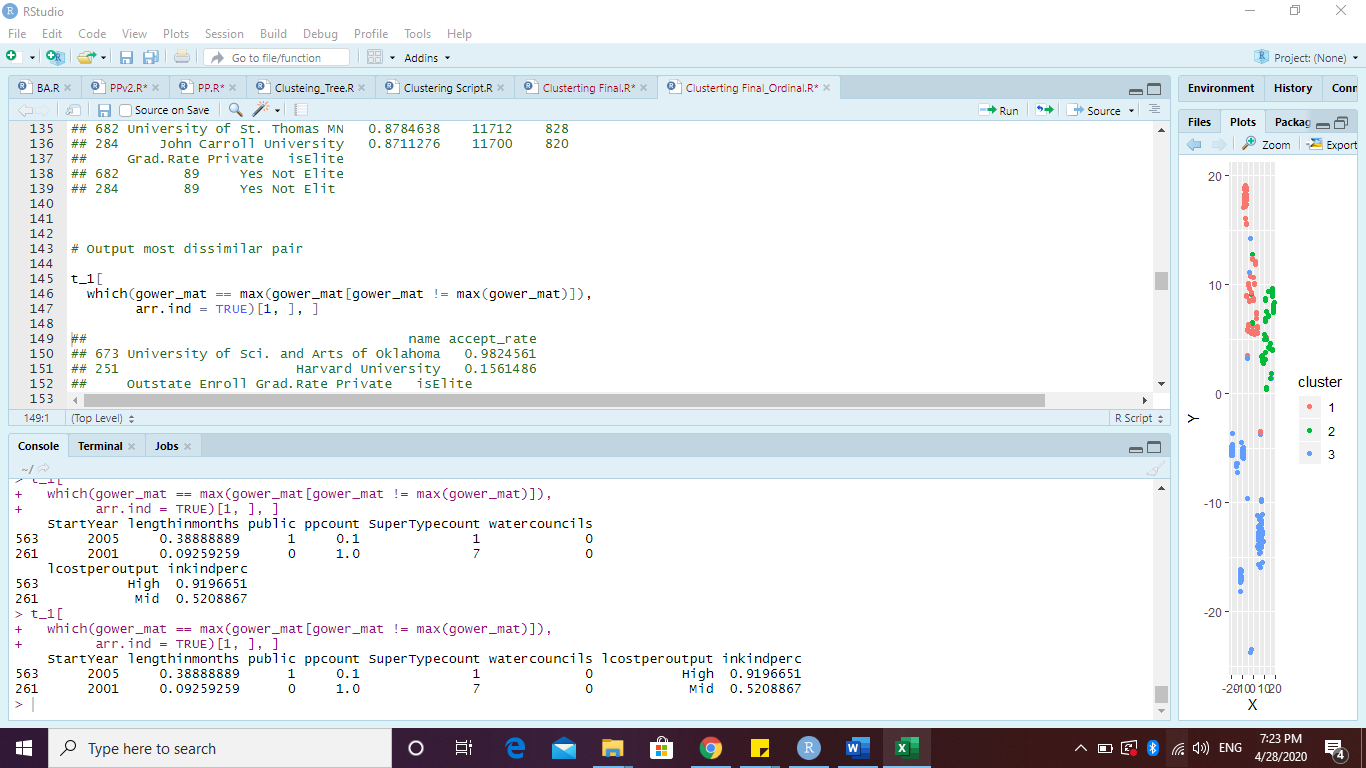
Following are the most significant patterns:

* Projects with “ppcount” less than 12 and in-kind funds between 44% and 60%, are expected to have Low cost per acre.
* Projects that had less than 60% of in-kind funding, had “ppcount” between 8 and 12, included “watercouncils” and were short duration (lengthinmonths<8) are expected to have Low cost per acre.

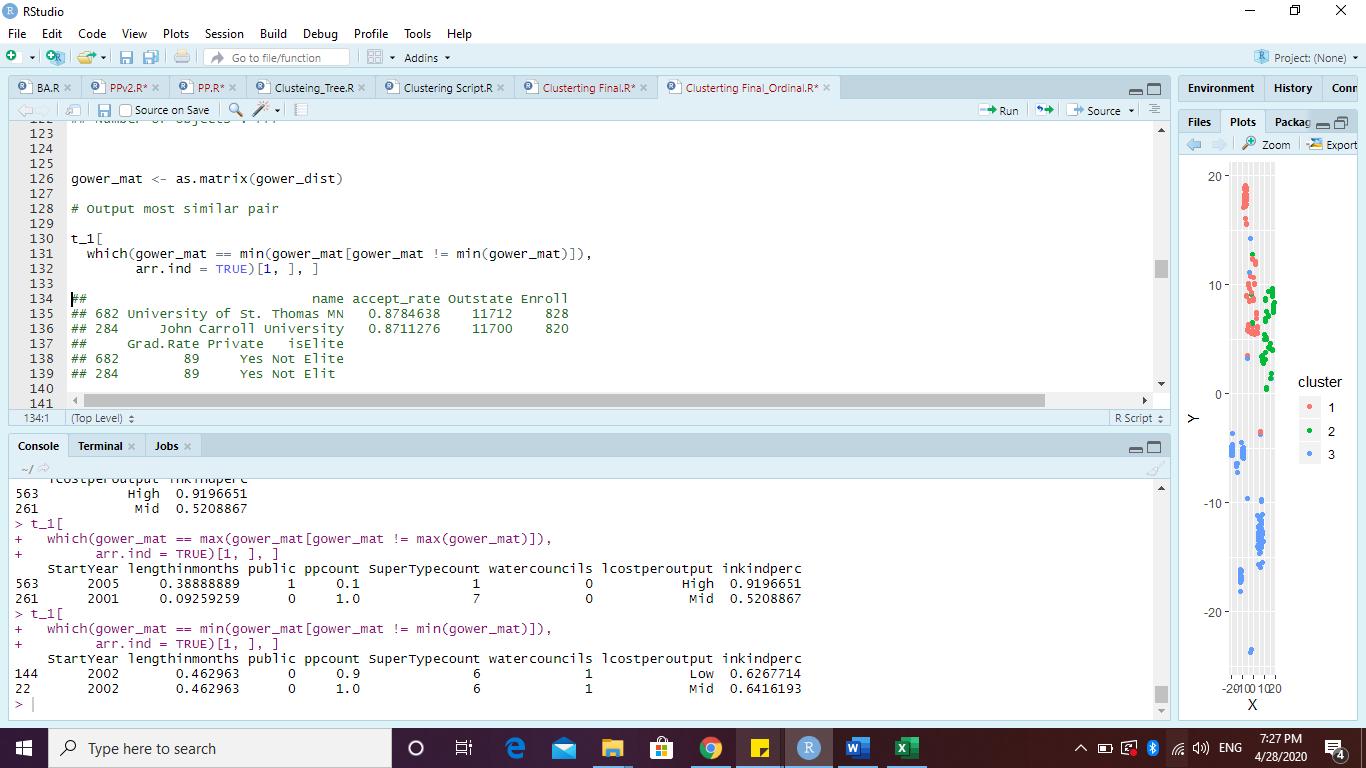
**PAM Clustering**

A separate attempt was made to apply clustering algorithm. However, the challenge was that our data consisted of mixed data types- both categorical and numeric. After some research, PAM (or k -medoids) clustering algorithm was finalized.

Gower distance was computed using daisy () function provided in the “cluster” package. Based on the Grower distance, following turned out to be the most dissimilar projects:

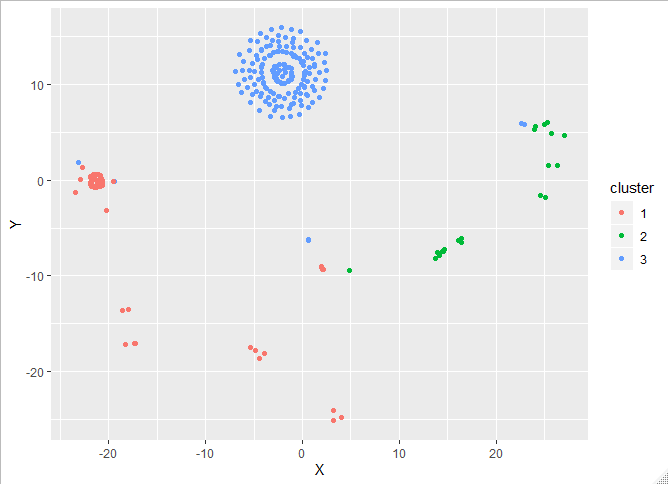


And following turned out to be the most similar projects:



Based on the labels of most similar and dissimilar projects, it may be seen that while it is easier to differentiate between High cost per acre projects, it is not easy to differentiate between Medium cost per acre and Low cost per acre projects.

Using Gower distance of the explanatory variables and K=3 (since we have 3 labels) clusters were computed. Following is a 2-dimensional plot of the clusters, which was created with the help of t-distributed stochastic neighborhood embedding.



As expected, High cost per acre projects (Cluster 3, marked in blue) can easily be differentiated from the rest. Comparing these clusters with the actual labels of the projects resulted in a low accuracy of 54.59%. However, it had a decent Low cost per acre project Recall value, of 76.38%, and F-Score of 61.59%.

**Conclusion**

The given dataset was small as well as highly skewed which made training algorithms much difficult.

Based on our models following conclusions may be drawn.

* For Low cost per acre projects, ideal number of unique collaborators is between 8-12 and must be spearheaded by watershed councils.
* Long Term projects with more than 56% direct funding (less than 56% in-kind) are expected to have low cost per acre.
* There is no statistically significant pattern as to how does the land status (public/private) affects cost of the project. This may be because “public” variable was unbalanced, with 93% of the projects implemented on private land.
* While Naïve Bayes Classifier does not provide any insight on the underlying patterns, it is ideal for predicting cost effectiveness of a new project.

**Acknowledgement**

For the code of PAM (k-medoid) clustering, help was taken from the following blog:

<https://www.r-bloggers.com/clustering-mixed-data-types-in-r/>