Comp309 Assignment 2

Part 1 : Core

Business understanding:

-Background: “Overfishing” refers to the fact that human fishing activities result in a fish population that is not found in the ocean to breed and replenish populations. The marine life captured by modern fisheries has exceeded the amount that the ecosystem can balance, and as a result, the entire marine system is ecologically degraded. The vast ocean provides the most space for biological growth. The fishing and hunting life that humans have been since ancient times. It has been replaced by large-scale industrial fishery production to this day. Humans are increasingly demanding the oceans as they march into the ocean. When the human's claim exceeds the limit of the ocean's ability to load, the marine fishery resources begin to shrink and eventually lead to extinction. In recent years, with the rapid growth of the world's population, the world's fishery production has developed rapidly, and many fishing areas have produced excessive fishing.

-Business objective: My aim is to prove whether overfishing happened in New Zealand.

-Data mining goal : I want to use data to prove whether people's fishing activities have affected fish in the New Zealand. And has the government taken appropriate measures to face this issue?

Data understanding:

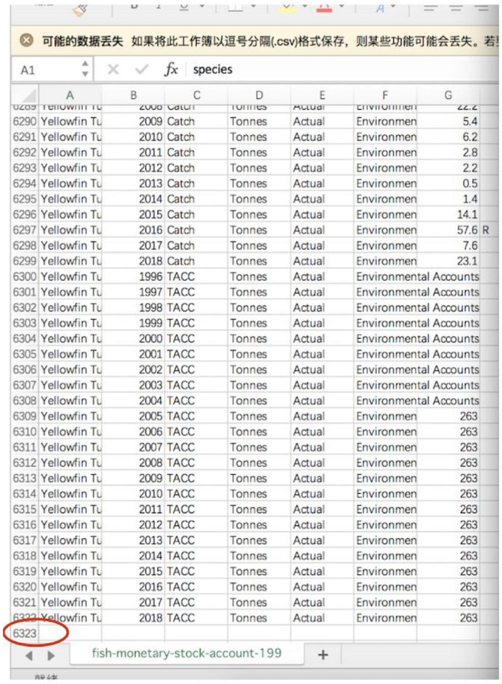
Source of the dataset:

My dataset is Environmental-economic accounts.I get the dataset from this link in data.govt.nz.(https://catalogue.data.govt.nz/dataset/environmental-economic-accounts-2019-tables/resource/9c0fdf15-1d92-4163-a32c-0e2f98e75b76)
I chose Fish monetary stock account, 1996–2018–CSV in a series of datasets.

-Description of the dataset :

This data set shows the catch and fishing trends of some fish from 1996 to 2018. And the annual government limits on the fishing of various fish species, as well as the benefits from these fish.Environmental-economic accounts show how our environment contributes to our economy, the impacts of economic activity on our environment, and how we respond to environmental issues.

-When I open the dataset in the first time I find a lot of missing values and the dataset is too huge(6322 rows) for my analysis purposes. and complicated (too much species)which may mask the important features I want to figure out.So the first thing after loading the dataset is to filter my data.

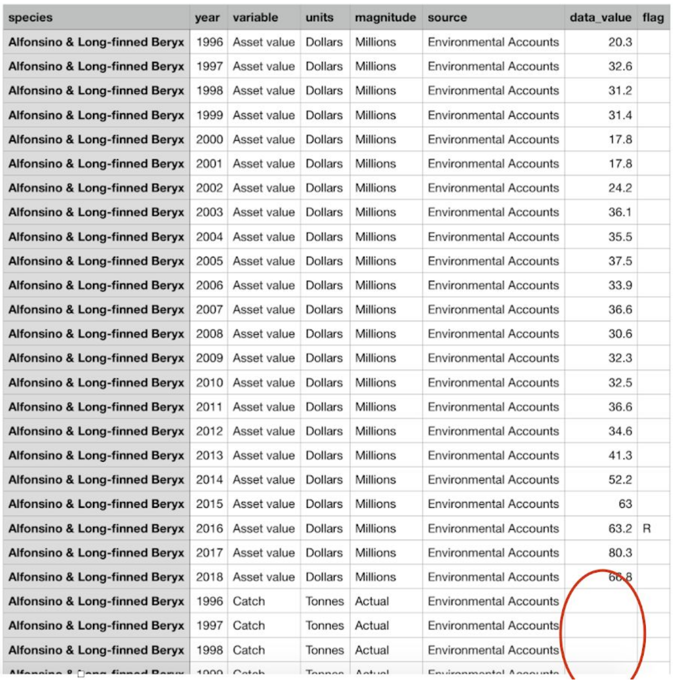


可能的数据丢失 如果将此工作簿以逗号分隔(.csv)格式保存，则某些功能可能会丢失。若

A1 species

	A	B	C	D	E	F	G	
6290	Yellowfin Tu	2009	Catch	Tonnes	Actual	Environmen	5.4	
6291	Yellowfin Tu	2010	Catch	Tonnes	Actual	Environmen	6.2	
6292	Yellowfin Tu	2011	Catch	Tonnes	Actual	Environmen	2.8	
6293	Yellowfin Tu	2012	Catch	Tonnes	Actual	Environmen	2.2	
6294	Yellowfin Tu	2013	Catch	Tonnes	Actual	Environmen	0.5	
6295	Yellowfin Tu	2014	Catch	Tonnes	Actual	Environmen	1.4	
6296	Yellowfin Tu	2015	Catch	Tonnes	Actual	Environmen	14.1	
6297	Yellowfin Tu	2016	Catch	Tonnes	Actual	Environmen	57.6	R
6298	Yellowfin Tu	2017	Catch	Tonnes	Actual	Environmen	7.6	
6299	Yellowfin Tu	2018	Catch	Tonnes	Actual	Environmen	23.1	
6300	Yellowfin Tu	1996	TACC	Tonnes	Actual	Environmental Accounts		
6301	Yellowfin Tu	1997	TACC	Tonnes	Actual	Environmental Accounts		
6302	Yellowfin Tu	1998	TACC	Tonnes	Actual	Environmental Accounts		
6303	Yellowfin Tu	1999	TACC	Tonnes	Actual	Environmental Accounts		
6304	Yellowfin Tu	2000	TACC	Tonnes	Actual	Environmental Accounts		
6305	Yellowfin Tu	2001	TACC	Tonnes	Actual	Environmental Accounts		
6306	Yellowfin Tu	2002	TACC	Tonnes	Actual	Environmental Accounts		
6307	Yellowfin Tu	2003	TACC	Tonnes	Actual	Environmental Accounts		
6308	Yellowfin Tu	2004	TACC	Tonnes	Actual	Environmental Accounts		
6309	Yellowfin Tu	2005	TACC	Tonnes	Actual	Environmen	263	
6310	Yellowfin Tu	2006	TACC	Tonnes	Actual	Environmen	263	
6311	Yellowfin Tu	2007	TACC	Tonnes	Actual	Environmen	263	
6312	Yellowfin Tu	2008	TACC	Tonnes	Actual	Environmen	263	
6313	Yellowfin Tu	2009	TACC	Tonnes	Actual	Environmen	263	
6314	Yellowfin Tu	2010	TACC	Tonnes	Actual	Environmen	263	
6315	Yellowfin Tu	2011	TACC	Tonnes	Actual	Environmen	263	
6316	Yellowfin Tu	2012	TACC	Tonnes	Actual	Environmen	263	
6317	Yellowfin Tu	2013	TACC	Tonnes	Actual	Environmen	263	
6318	Yellowfin Tu	2014	TACC	Tonnes	Actual	Environmen	263	
6319	Yellowfin Tu	2015	TACC	Tonnes	Actual	Environmen	263	
6320	Yellowfin Tu	2016	TACC	Tonnes	Actual	Environmen	263	
6321	Yellowfin Tu	2017	TACC	Tonnes	Actual	Environmen	263	
6322	Yellowfin Tu	2018	TACC	Tonnes	Actual	Environmen	263	
6323								

fish-monetary-stock-account-199

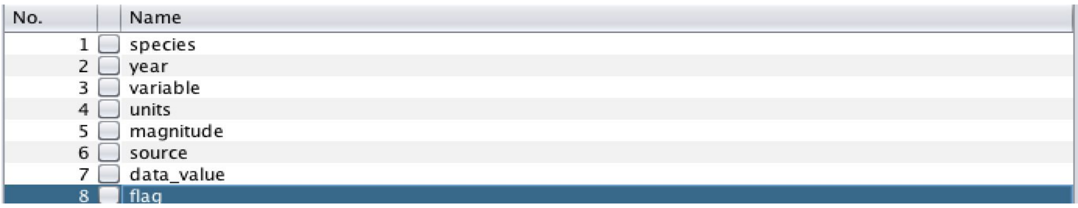


species	year	variable	units	magnitude	source	data_value	flag
Alfonsino & Long-finned Beryx	1996	Asset value	Dollars	Millions	Environmental Accounts	20.3	
Alfonsino & Long-finned Beryx	1997	Asset value	Dollars	Millions	Environmental Accounts	32.6	
Alfonsino & Long-finned Beryx	1998	Asset value	Dollars	Millions	Environmental Accounts	31.2	
Alfonsino & Long-finned Beryx	1999	Asset value	Dollars	Millions	Environmental Accounts	31.4	
Alfonsino & Long-finned Beryx	2000	Asset value	Dollars	Millions	Environmental Accounts	17.8	
Alfonsino & Long-finned Beryx	2001	Asset value	Dollars	Millions	Environmental Accounts	17.8	
Alfonsino & Long-finned Beryx	2002	Asset value	Dollars	Millions	Environmental Accounts	24.2	
Alfonsino & Long-finned Beryx	2003	Asset value	Dollars	Millions	Environmental Accounts	36.1	
Alfonsino & Long-finned Beryx	2004	Asset value	Dollars	Millions	Environmental Accounts	35.5	
Alfonsino & Long-finned Beryx	2005	Asset value	Dollars	Millions	Environmental Accounts	37.5	
Alfonsino & Long-finned Beryx	2006	Asset value	Dollars	Millions	Environmental Accounts	33.9	
Alfonsino & Long-finned Beryx	2007	Asset value	Dollars	Millions	Environmental Accounts	36.6	
Alfonsino & Long-finned Beryx	2008	Asset value	Dollars	Millions	Environmental Accounts	30.6	
Alfonsino & Long-finned Beryx	2009	Asset value	Dollars	Millions	Environmental Accounts	32.3	
Alfonsino & Long-finned Beryx	2010	Asset value	Dollars	Millions	Environmental Accounts	32.5	
Alfonsino & Long-finned Beryx	2011	Asset value	Dollars	Millions	Environmental Accounts	36.6	
Alfonsino & Long-finned Beryx	2012	Asset value	Dollars	Millions	Environmental Accounts	34.6	
Alfonsino & Long-finned Beryx	2013	Asset value	Dollars	Millions	Environmental Accounts	41.3	
Alfonsino & Long-finned Beryx	2014	Asset value	Dollars	Millions	Environmental Accounts	52.2	
Alfonsino & Long-finned Beryx	2015	Asset value	Dollars	Millions	Environmental Accounts	63	
Alfonsino & Long-finned Beryx	2016	Asset value	Dollars	Millions	Environmental Accounts	63.2	R
Alfonsino & Long-finned Beryx	2017	Asset value	Dollars	Millions	Environmental Accounts	80.3	
Alfonsino & Long-finned Beryx	2018	Asset value	Dollars	Millions	Environmental Accounts	66.8	
Alfonsino & Long-finned Beryx	1996	Catch	Tonnes	Actual	Environmental Accounts		
Alfonsino & Long-finned Beryx	1997	Catch	Tonnes	Actual	Environmental Accounts		
Alfonsino & Long-finned Beryx	1998	Catch	Tonnes	Actual	Environmental Accounts		

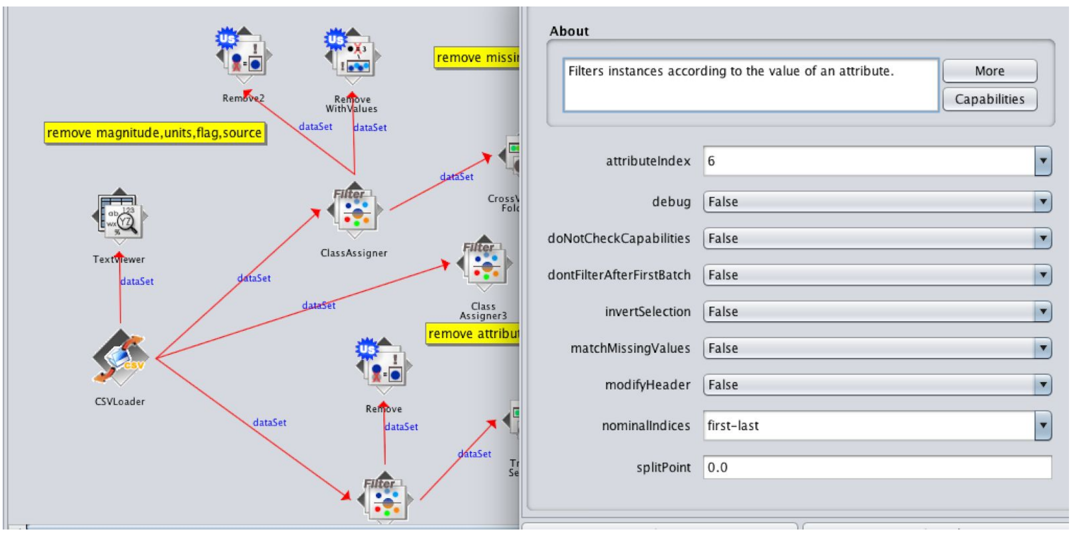
Data Preparation:

And this dataset has a lot of flaws：

1. For missing value ,I choose to remove rows with missing value(data_value is the only attribute may have missing value which atrributeIndex is 6) because the number of data is big enough.(RemoveWithValues)
2. I chose the top ten fish in the amount of catch.
3. I remove asset_value(unit is Dollar) in variable because it is not unified with the other two data units(Catch&TACC tons), and actually I don't need this kind of data.
4. I remove All_species value in variable attribute because it have great impact of result



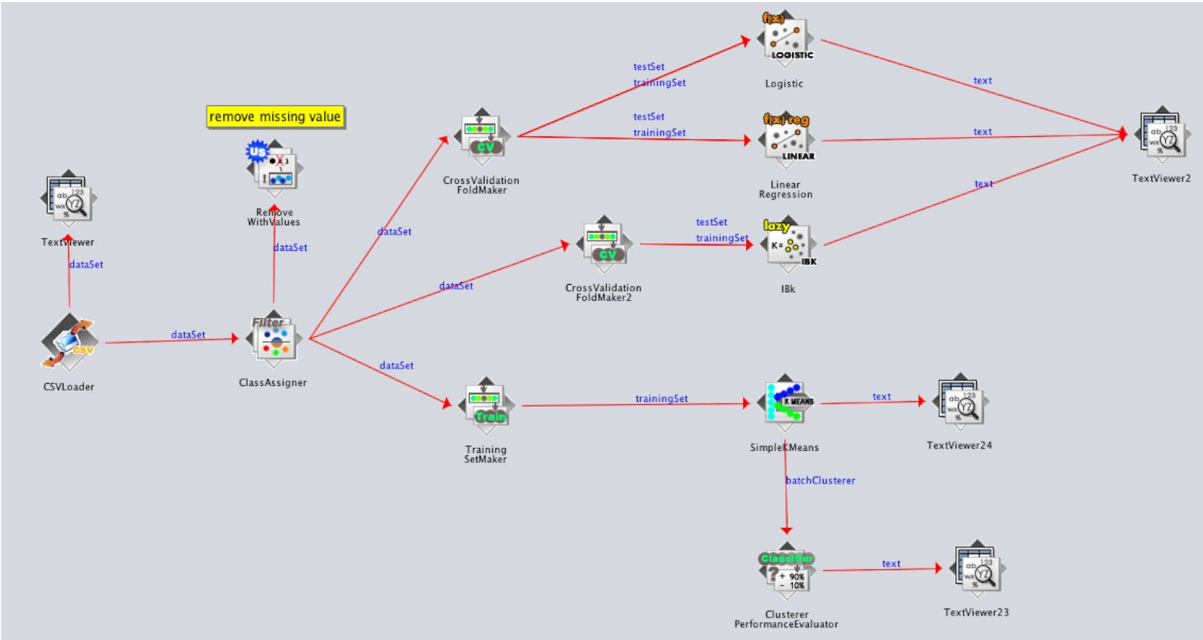
No.	Name
1	species
2	year
3	variable
4	units
5	magnitude
6	source
7	data_value
8	flag



Modeling:

Before modelling, techniques I decide to use are Classification, Cluster and Regression.

Pipeline:



Evaluation:

Reslut of Pipeline:

Linear regression:

```
Linear Regression Model

data_value =

   4168.9504 *  species=Snapper,Silver_Warehou,Orange_Roughy,Ling,Southern_Blue_Whiting,Hoki +
   4549.6825 *  species=Silver_Warehou,Orange_Roughy,Ling,Southern_Blue_Whiting,Hoki +
   8284.08   *  species=Ling,Southern_Blue_Whiting,Hoki +
  18341.682  *  species=Southern_Blue_Whiting,Hoki +
 112705.6704 *  species=Hoki +
   -536.1291 * year +
   4580.8125 * variable=TACC +
1076342.781

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9341
Mean absolute error                   7704.1214
Root mean squared error              16819.2613
Relative absolute error                 26.9878 %
Root relative squared error             35.62   %
Total Number of Instances              384
```

Since the algorithm filters out irrelevant data automatically, the data for the linear regression algorithm now retains only the annual catch (in tons) and the year and type of fish and data_value(how many fish be caught in tons). Because the data value is continuous,so linear regression can be used. The purpose of our experiment is to predict future trends from known data. This satisfies the purpose of linear regression. For my dataset, it is based on the existing annual The amount of fishing to predict the future catch, and finally through the correlation coefficient is not difficult to see the results are very satisfactory, so the linear regression algorithm is very suitable. Linear regression is the ability to describe the relationship between data more accurately with a straight line. This way, when new data appears, it is possible to predict a simple value.

Logistic regression:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         331               55.0749 %
Incorrectly Classified Instances       270               44.9251 %
Kappa statistic                          0.4999
Mean absolute error                      0.1149
Root mean squared error                  0.236
Relative absolute error                 63.9002 %
Root relative squared error             78.7275 %
Total Number of Instances              601
```

Logistic regression is different from linear regression. The essence of logistic regression algorithms is actually a classification algorithm, and it is not intended to predict a certain value. So the result is in the same form as the classification. Logistic regression is used for the classification of discrete variables, that is, the range of its output y is a discrete set, mainly used for class discrimination, and its output value y represents the class belonging to a certain class.

Logistic Regression is mainly used to classify problems. It is often used to predict probabilities. For example, knowing a person's age, weight, height, blood pressure and other information, predicting the probability of suffering from heart disease. The classic LR is used for the two-category problem (only 0, 1 and 2).

Logistic function:

For any x value, the corresponding y value is within the interval (0, 1).

The function formula is: $f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}},$

IBK:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         154               40.1042 %
Incorrectly Classified Instances       230               59.8958 %
Kappa statistic                         -0.2169
Mean absolute error                      0.5982
Root mean squared error                  0.7715
Relative absolute error                121.6966 %
Root relative squared error            155.6247 %
Total Number of Instances              384

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.465    0.683    0.470      0.465    0.468      -0.217   0.429     0.579     TACC
                 0.317    0.535    0.314      0.317    0.315      -0.217   0.429     0.416     Catch
Weighted Avg.    0.401    0.618    0.402      0.401    0.401      -0.217   0.429     0.508

=== Confusion Matrix ===

   a   b   <-- classified as
 101 116 |   a = TACC
 114  53 |   b = Catch
```

Because it is a classification algorithm, the result is correctly classified instances, but in fact, the purpose of the experiment does not have the need to classify the dataset.

SimplyKmean:

```
Final cluster centroids:
                                                 Cluster#
Attribute                         Full Data             0                     1                     2                     3
                                    (601.0)       (106.0)                (38.0)                (57.0)                (69.0)
==========================================================================================================================================
 species                               Hoki      Blue_Cod         Orange_Roughy           Rock_Lobster                  Hoki
year                              2008.2047     2002.9717             2007.7105             2003.3333             2013.5942
variable                        Asset_value   Asset_value                  TACC                  TACC                 Catch
units                                Tonnes       Dollars                Tonnes                Tonnes                Tonnes
magnitude                            Actual      Millions                Actual                Actual                Actual
source         Environmental_Accounts Environmental_Accounts Environmental_Accounts Environmental_Accounts Environmental_Accounts Environme
data_value                       15979.1083      196.2236            12971.4447             6674.5579             40127.787
flag                                      R             R                     R                     R                     R


Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0      106 ( 18%)
 1       38 (  6%)
 2       57 (  9%)
 3       69 ( 11%)
 4       81 ( 13%)
 5       27 (  4%)
 6       31 (  5%)
 7       46 (  8%)
 8      111 ( 18%)
 9       35 (  6%)
```

I think this experimental cluster is not applicable. The main reason is that the data have a label. We don't want to divide data into different groups.This deviates from our experimental purpose.

Difference between those algorithm:

Linear regression and Logistic regression:

First of all, the above two different regression algorithms are mentioned: linear regression and logistic regression. The ordinary linear regression is mainly used for the prediction of continuous variables. That is, the output y of the linear regression ranges from the whole real interval (y∈R).So it's suitable for my data(data_value)
Logistic regression is used for the classification of discrete variables, that is, the range of its output y is a discrete set, mainly used for class discrimination, and its output value y represents the probability of belonging to a certain class.This is different from my experimental purpose.

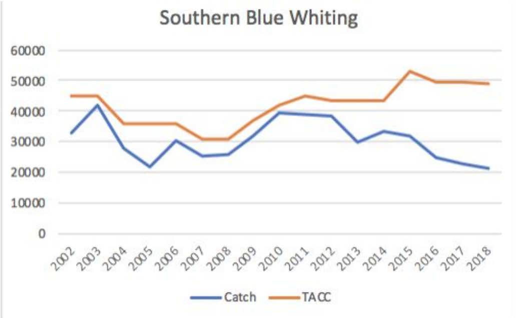
SimpleKmean and IBK:

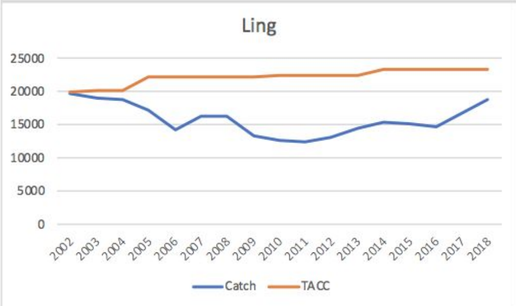
Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. And K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.However, it is not suitable for my purpose.
And IBK is also a classification method ,so IBK is not applicable for the same reason as SimpleKmean above.

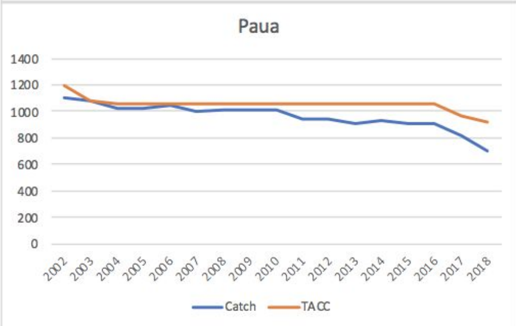
Conclusion:

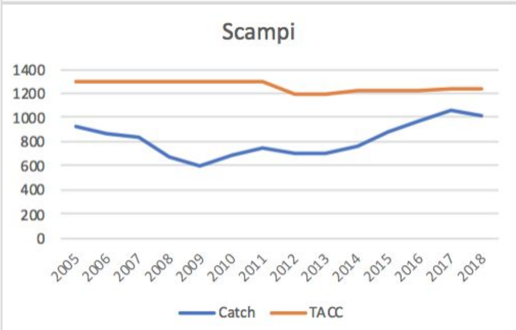
All in all, linear regression have the best performance (coefficient is 0.9341) ,the correlation coefficient is pretty high. It shows attributes have strong relationship. Linear regression is the ability to describe the relationship between data more accurately with a straight line. This way, when new data appears, it is possible to predict a simple value, so it can be said that the number of fishing increases with the year. **But now I don't have enough evidence to show whether we did overfish or not ,the only thing what I can get is the number of fishing is positively related to the year.**

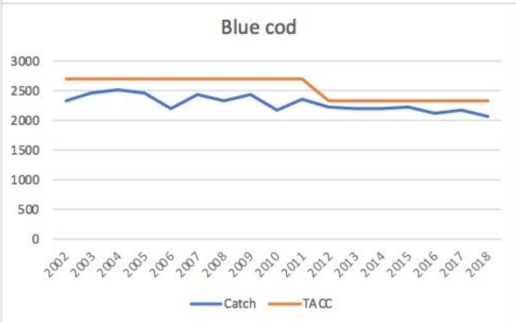
At the same time, I made a diagram between TACC and Catch attributes in order to intuitively discover the relationship between them. And we can see only two catch values of fish exceed TACC.

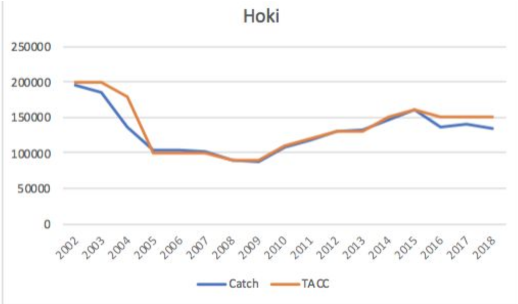


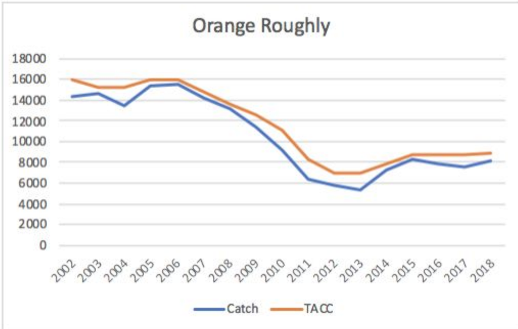


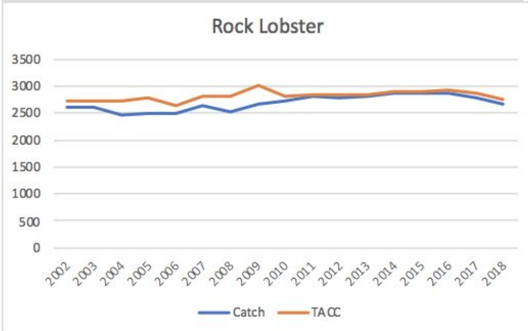


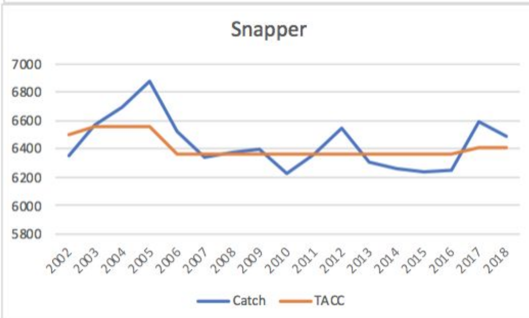


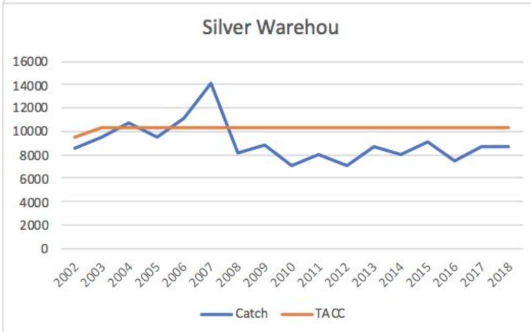












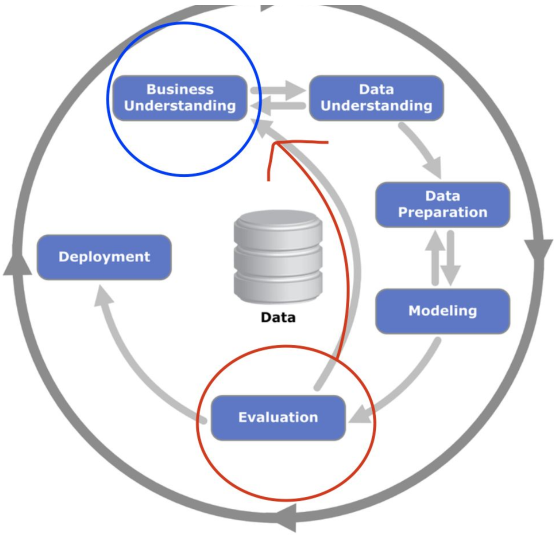
So question in business understanding I can partially prove:

-Business objective: My aim is to prove whether overfishing happened in New Zealand.

-Data mining goal : I want to use data to prove whether people's fishing activities have affected fish in the New Zealand.

For these two question above, based on these diagram above, we did overfish on Snapper and Silver Warehou these two species but we can not ignore other majority of species were not overfished.So we can only say that overfishing happened but did not cause the collapse of fish because Snapper and Silver Warehou fluctuate but both of them stays within a range.

However we can not get definitive conclusion by those results, because my data is not good enough and there is only government regulations - Total allowable commercial catch. There is no data about limitation of biological (ecological balance) catch.So the next step I need to do is restart CRISP-DM and try to find more evidence and more useful dataset.



And when I do CRISP-DM again , I need to keep tracking following question:
and New question in business understanding:

Question 1: Is there any evidence of fish stocks collapsing in NZ waters?

In another aspect, how much human fishing has negatively affect on fish in NZ and whether it is cause fish stocks collapsing. Whether the fishing has caused a decrease of fish ？

Question 2: Whether the fishing exceeds the biological limit?

Because TACC is made by gorvernment and I really coursious about how they decide TACC value.

Question 3: Whether the definition of TACC is related to biological limit of fishing?

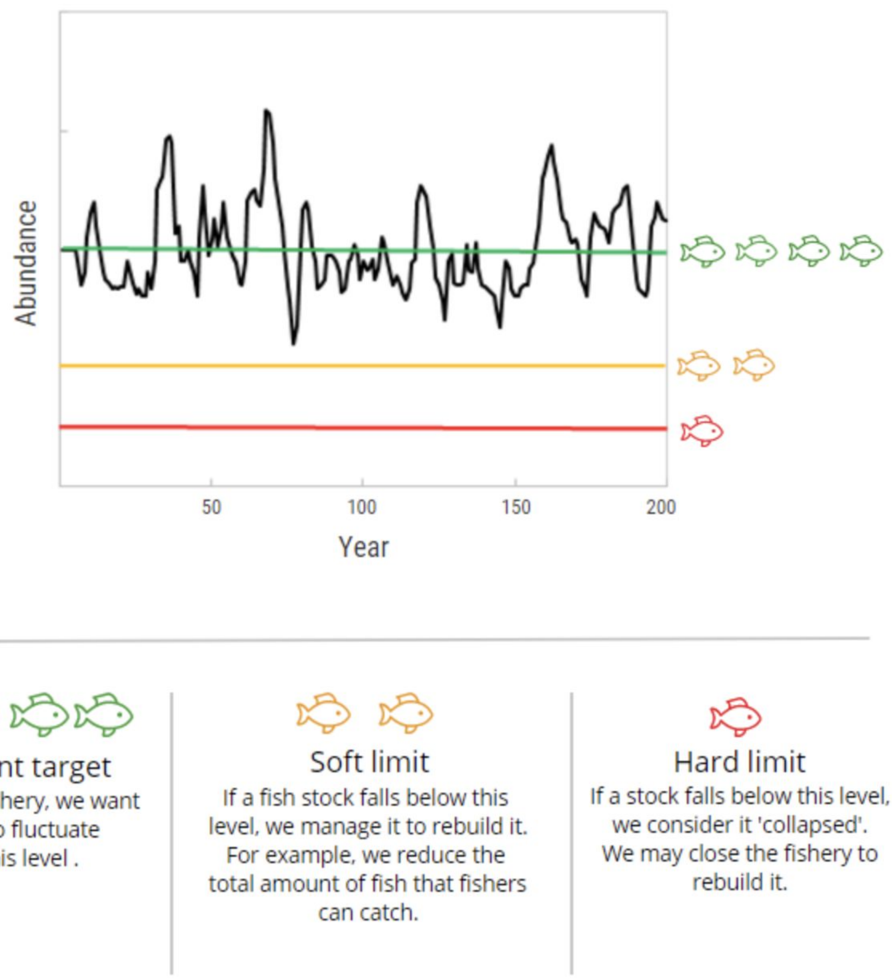
So after I made Question 2 I was thinking is there any possible TACC have relation with limit of fishing.

Part 2 : Completion

The question I chose is Question 3: Whether the definition of TACC is related to biological limit of fishing?

The reason why this is an interesting question is the dataset I used before doesn't contain a reference indicator can be used to compare with my data. So this is a good question and essential question.As mentioned before, a single dataset is not enough to prove whether we overfishing in NZ because TACC is a man-made regulation, I want to know if there is one thing that has been affecting TACC. Therefore, I need an indicator to compare the relationship between my data and this indicator, so I need to introduce another two datasets , which are soft limit of fishing and hard limit of fishing. These two dataset contains different index of limit of fishing in different years.

Explaination of soft limit and hard limit:



Picture from https://www.mpi.govt.nz/growing-and-harvesting/fisheries/fisheries-management/fish-stock-status/

The soft limit dataset I got from:
https://data.mfe.govt.nz/table/53467-performance-of-assessed-fish-stock-in-relation-to-the-soft-limit-200915/data/

year	performance_of_stocks_soft_limit	percent
2009	landings_from_stocks_above_soft_limit	94
2010	landings_from_stocks_above_soft_limit	94.8
2011	landings_from_stocks_above_soft_limit	95.1
2012	landings_from_stocks_above_soft_limit	96.6
2013	landings_from_stocks_above_soft_limit	96.1
2014	landings_from_stocks_above_soft_limit	96.4
2015	landings_from_stocks_above_soft_limit	96.8
2009	stocks_above_soft_limit	81.1
2010	stocks_above_soft_limit	86.7
2011	stocks_above_soft_limit	85
2012	stocks_above_soft_limit	83.2
2013	stocks_above_soft_limit	82
2014	stocks_above_soft_limit	83.6
2015	stocks_above_soft_limit	82.8

The hard limit dataset I got from:
https://data.mfe.govt.nz/table/53469-performance-of-assessed-fish-stock-in-relation-to-the-hard-limit-200915/data/

year	performance_of_stocks_hard_limit	percent
2009	landings_from_stocks_above_hard_limit	99.5
2010	landings_from_stocks_above_hard_limit	99.1
2011	landings_from_stocks_above_hard_limit	97.1
2012	landings_from_stocks_above_hard_limit	99.5
2013	landings_from_stocks_above_hard_limit	99.5
2014	landings_from_stocks_above_hard_limit	99.6
2015	landings_from_stocks_above_hard_limit	99.6
2009	stocks_above_hard_limit	93.9
2010	stocks_above_hard_limit	93.8
2011	stocks_above_hard_limit	93.9
2012	stocks_above_hard_limit	93.9
2013	stocks_above_hard_limit	93.5
2014	stocks_above_hard_limit	94.3
2015	stocks_above_hard_limit	94

Data preparation:

So what I want to do is make a dataset that only retains values of TACC , soft limit and hard limit, then see what relationship between them. (E.g. TACC grows with the growth of soft and hard.).First, I only keep values of landing from stock above hard&soft limit, Because being caught ashore is the real impact on fish. What’s more ,I find value of landing from stock above hard limit vey close to 100% ,which means we have not made a devastating thing.So I decide only use soft limit dataset.

1.My question is Whether the definition of TACC is related to biological limit of fishing? So I remove Catch value from my dataset and focus on TACC and my new data .

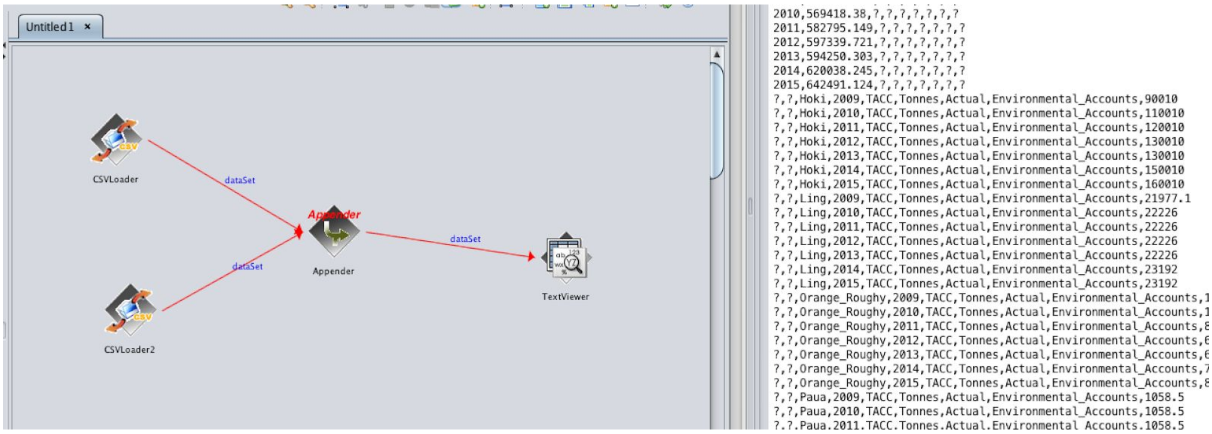
species	year	variable	units	magnitude	source	data_value	performance_of_stocks_soft_limit	percent_of_ls	performance_of_
Hoki	2009	TACC	Tonnes	Actual	Environmenta	90010	landings_from_stocks_above_soft_limit	94	stocks_above_sc
Hoki	2010	TACC	Tonnes	Actual	Environmenta	110010	landings_from_stocks_above_soft_limit	94.8	stocks_above_sc
Hoki	2011	TACC	Tonnes	Actual	Environmenta	120010	landings_from_stocks_above_soft_limit	95.1	stocks_above_sc
Hoki	2012	TACC	Tonnes	Actual	Environmenta	130010	landings_from_stocks_above_soft_limit	96.6	stocks_above_sc
Hoki	2013	TACC	Tonnes	Actual	Environmenta	130010	landings_from_stocks_above_soft_limit	96.1	stocks_above_sc
Hoki	2014	TACC	Tonnes	Actual	Environmenta	150010	landings_from_stocks_above_soft_limit	96.4	stocks_above_sc
Hoki	2015	TACC	Tonnes	Actual	Environmenta	160010	landings_from_stocks_above_soft_limit	96.8	stocks_above_sc
Ling	2009	TACC	Tonnes	Actual	Environmenta	21977.1	landings_from_stocks_above_soft_limit	94	stocks_above_sc
Ling	2010	TACC	Tonnes	Actual	Environmenta	22226	landings_from_stocks_above_soft_limit	94.8	stocks_above_sc
Ling	2011	TACC	Tonnes	Actual	Environmenta	22226	landings_from_stocks_above_soft_limit	95.1	stocks_above_sc
Ling	2012	TACC	Tonnes	Actual	Environmenta	22226	landings_from_stocks_above_soft_limit	96.6	stocks_above_sc
Ling	2013	TACC	Tonnes	Actual	Environmenta	22226	landings_from_stocks_above_soft_limit	96.1	stocks_above_sc
Ling	2014	TACC	Tonnes	Actual	Environmenta	23192	landings_from_stocks_above_soft_limit	96.4	stocks_above_sc
Ling	2015	TACC	Tonnes	Actual	Environmenta	23192	landings_from_stocks_above_soft_limit	96.8	stocks_above_sc
Orange_Rou	2009	TACC	Tonnes	Actual	Environmenta	12532	landings_from_stocks_above_soft_limit	94	stocks_above_sc
Orange_Rou	2010	TACC	Tonnes	Actual	Environmenta	11062	landings_from_stocks_above_soft_limit	94.8	stocks_above_sc
Orange_Rou	2011	TACC	Tonnes	Actual	Environmenta	8221	landings_from_stocks_above_soft_limit	95.1	stocks_above_sc
Orange_Rou	2012	TACC	Tonnes	Actual	Environmenta	6941	landings_from_stocks_above_soft_limit	96.6	stocks_above_sc
Orange_Rou	2013	TACC	Tonnes	Actual	Environmenta	6941	landings_from_stocks_above_soft_limit	96.1	stocks_above_sc
Orange_Rou	2014	TACC	Tonnes	Actual	Environmenta	7841	landings_from_stocks_above_soft_limit	96.4	stocks_above_sc
Orange_Rou	2015	TACC	Tonnes	Actual	Environmenta	8736	landings_from_stocks_above_soft_limit	96.8	stocks_above_sc
Paua	2009	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	94	stocks_above_sc
Paua	2010	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	94.8	stocks_above_sc
Paua	2011	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	95.1	stocks_above_sc
Paua	2012	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	96.6	stocks_above_sc
Paua	2013	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	96.1	stocks_above_sc
Paua	2014	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	96.4	stocks_above_sc
Paua	2015	TACC	Tonnes	Actual	Environmenta	1058.5	landings_from_stocks_above_soft_limit	96.8	stocks_above_sc
Rock_Lobste	2009	TACC	Tonnes	Actual	Environmenta	3021.4	landings_from_stocks_above_soft_limit	94	stocks_above_sc
Rock_Lobste	2010	TACC	Tonnes	Actual	Environmenta	2802.5	landings_from_stocks_above_soft_limit	94.8	stocks_above_sc
Rock_Lobste	2011	TACC	Tonnes	Actual	Environmenta	2847.7	landings_from_stocks_above_soft_limit	95.1	stocks_above_sc
Rock_Lobste	2012	TACC	Tonnes	Actual	Environmenta	2833.1	landings_from_stocks_above_soft_limit	96.6	stocks_above_sc
Rock_Lobste	2013	TACC	Tonnes	Actual	Environmenta	2850.6	landings_from_stocks_above_soft_limit	96.1	stocks_above_sc
Rock_Lobste	2014	TACC	Tonnes	Actual	Environmenta	2895.7	landings_from_stocks_above_soft_limit	96.4	stocks_above_sc
Rock_Lobste	2015	TACC	Tonnes	Actual	Environmenta	2898.2	landings_from_stocks_above_soft_limit	96.8	stocks_above_sc
Scampi	2009	TACC	Tonnes	Actual	Environmenta	1291	landings_from_stocks_above_soft_limit	94	stocks_above_sc
Scampi	2010	TACC	Tonnes	Actual	Environmenta	1291	landings_from_stocks_above_soft_limit	94.8	stocks_above_sc

2. I found that their units are not uniform, the unit of landing_from_stocks_above_hard_limit is percentage but unit of data_value is tons.So I need to transfer one to another, then I use data_value of All_species in my original dataset I removed before multiply the percent then I get how many fish we catch above the soft limit.

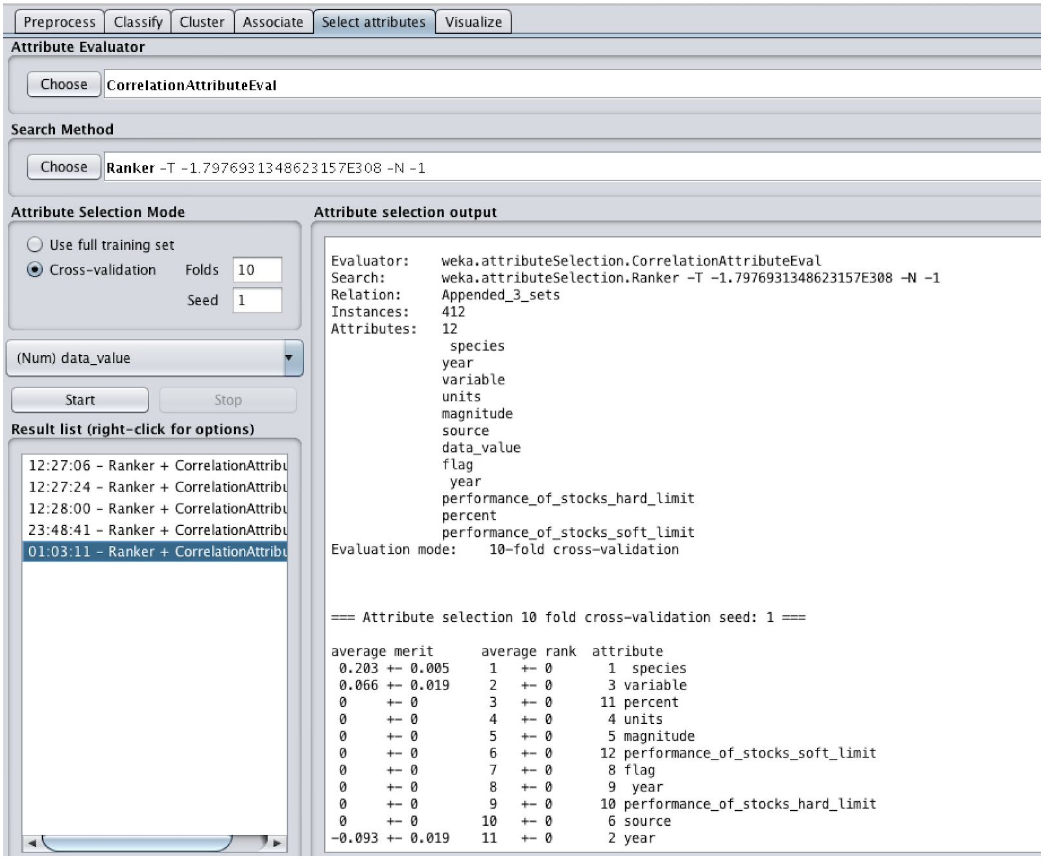
year	landings_from_stocks_above_hard_limit		year	total_value
2009	94	All species	2009	572260.1
2010	94.8	All species	2010	600652.3
2011	95.1	All species	2011	612823.5
2012	96.6	All species	2012	618364.1
2013	96.1	All species	2013	618366.6
2014	96.4	All species	2014	643193.2
2015	96.8	All species	2015	663730.5

year	value_above	
2009	537924.494	572260.1 * 94%
2010	569418.38	600652.3 * 94.8%
2011	582795.149	612823.5 * 95.1%
2012	597339.721	618364.1 * 96.6%
2013	594250.303	618366.6 * 96.1%
2014	620038.245	643193.2 * 96.4%
2015	642491.124	663730.5 * 96.8%

3. Merge: Then I merge my original dataset with soft limit dataset.

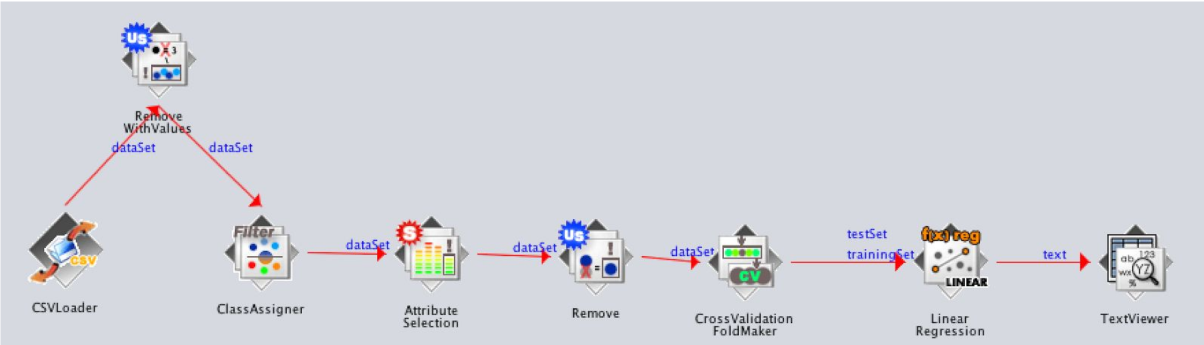


4. Dimensionality reduction:



After I merge two data set we can see source, flag, magnitude ,units ,soft limit ,hard limit, have no relation with data_value. So we can remove these irrelevant attributes for making our result more reliable.

Modelling :



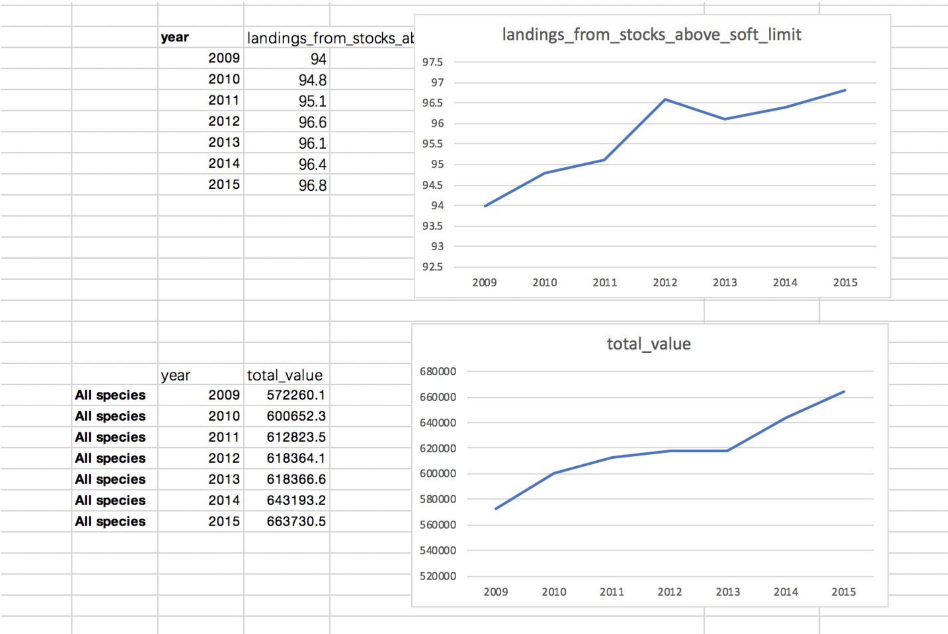
Evaluation:

```
=== Classifier model (full training set) ===


Linear Regression Model

data_value =

   6629.1452 *  species=Snapper,Orange_Roughy,Silver_Warehou,Ling,Southern_Blue_Whiting,Hoki +
  13921.8476 *  species=Ling,Southern_Blue_Whiting,Hoki +
  21401.5571 *  species=Southern_Blue_Whiting,Hoki +
  83284.8571 *  species=Hoki +
      0.0785 * value_above +
 -44548.5919

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9714
Mean absolute error                   4540.5963
Root mean squared error               8955.7443
Relative absolute error                 17.8326 %
Root relative squared error             23.4105 %
Total Number of Instances               70
```

Linear regression have a very good performance (coefficient is 0.9714) ,the correlation coefficient is pretty high. It shows attributes have strong relationship. Linear regression is the ability to describe the relationship between data more accurately with a straight line. This way, when new data appears, it is possible to predict a simple value, so it can be said that the value of TACC have relation with value_above(value of soft limit).

So now I can make sure when government made TACC is partially based on soft limit value.(Question I made before can be answered)

And I draw a line graph below to show the relation between them :



	year	landings_from_stocks_ab
	2009	94
	2010	94.8
	2011	95.1
	2012	96.6
	2013	96.1
	2014	96.4
	2015	96.8

	year	total_value
All species	2009	572260.1
All species	2010	600652.3
All species	2011	612823.5
All species	2012	618364.1
All species	2013	618366.6
All species	2014	643193.2
All species	2015	663730.5

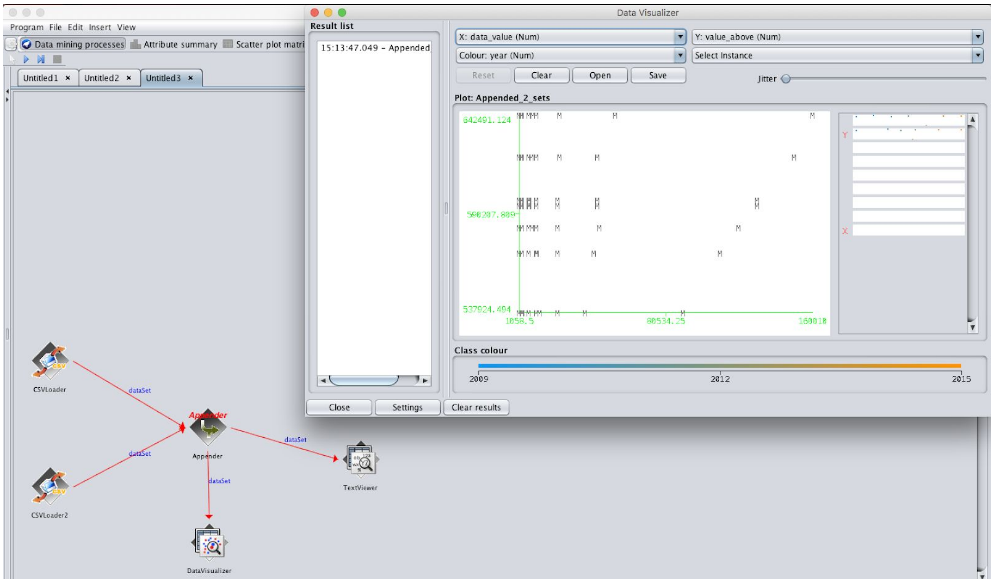
So we can see the relationship more apparently. The more value above soft limit line, the more fish we can catch , in another aspect, if there are some part below the line government will decrease TACC to protect fish.

Conclusion: Soft limit value →TACC →How much we can catch
It is a reasonable procedure to judge how many fish we can catch per year, and this is why overfishing didn't happen in New Zealand.

Part 3 : Challenge

---

graph I got from pipeline:

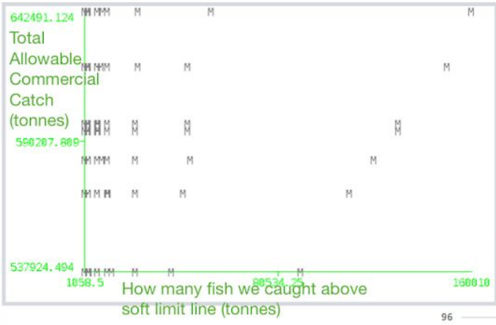


Comp309 A2 Yan Zichu 300476924

This is my poster and I submited it as PDF in submission system.

This line diagrams show TACC and catch value of top 10 species from 2002 to 2018.We can see there are not too much value of catch exceed TACC.

The New Zealand government has effectively controlled the annual catch by properly formulating the TACC so that the fish were not overfished during 2009 to 2015



The point diagram above shows the relation between TACC and the value of how many fish we caught above safe line(soft limit line).So we can find the number of fish we caught is influenced by TACC.

Soft limit value →TACC →How much we can catch

The line graphs below demonstrate All species we caught is based on soft limit value.