Koalas' hospital records analysis in Queensland

Picture is fromhttps://www.wwf.org.uk

Group Number : I

Jie Zhang s4755838

Kaixuan Li s4685450

Danita Prakash s4745511

Raahul Arun s4724492

Introduction

Existing Problems

‘In February 2022, the status of the koala has recently been changed from vulnerable to endangered.’ [1]

Our focus

Analyze the main causes leading koalas to death.

top reasons of koalas’ injury and sickness.

Use different algorithm models to predict death.

Stakeholders

Animal Protection Organization
State Government
Local community

problem solving with data

Getting the data

Where to get the main dataset.

Data cleaning

Describe the data cleaning process

Analysis

Variables correlation
Model chosen and evaluation
Statistical analysis

Storytelling

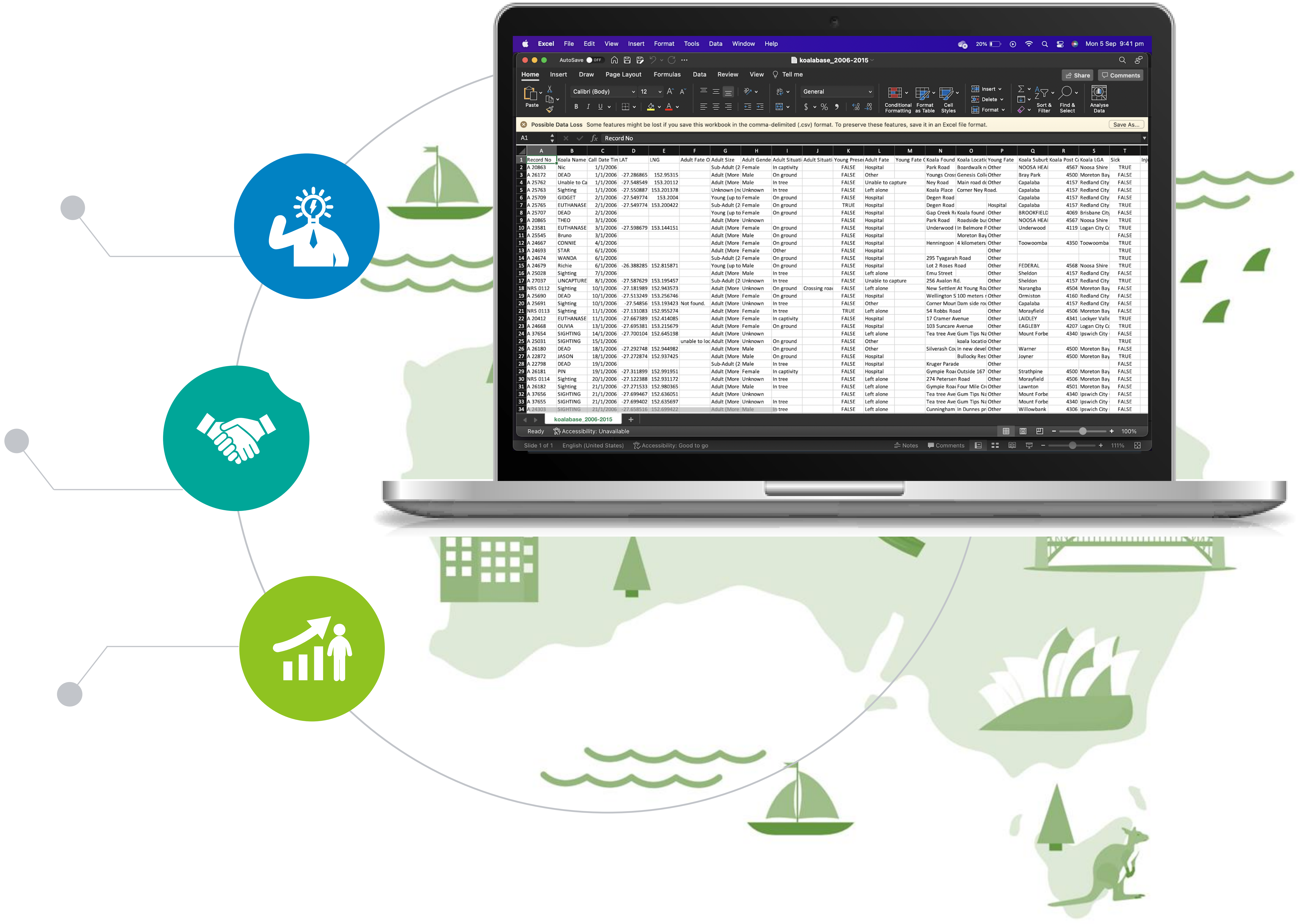
What are the main causes lead koala population decline (death)
Actions taken advice

Getting the data we need

Main Datasets

The main datasets are from the Queensland Government open data portal.
https://www.data.qld.gov.au/dataset/koala-hospital-data
which has records from the year **1996-2022** in hospital, Queensland.

The dataset consists of **56935** rows and **41** columns such as(Record No, Koala Name, Latitude, Longitude, Adult Size etc.)



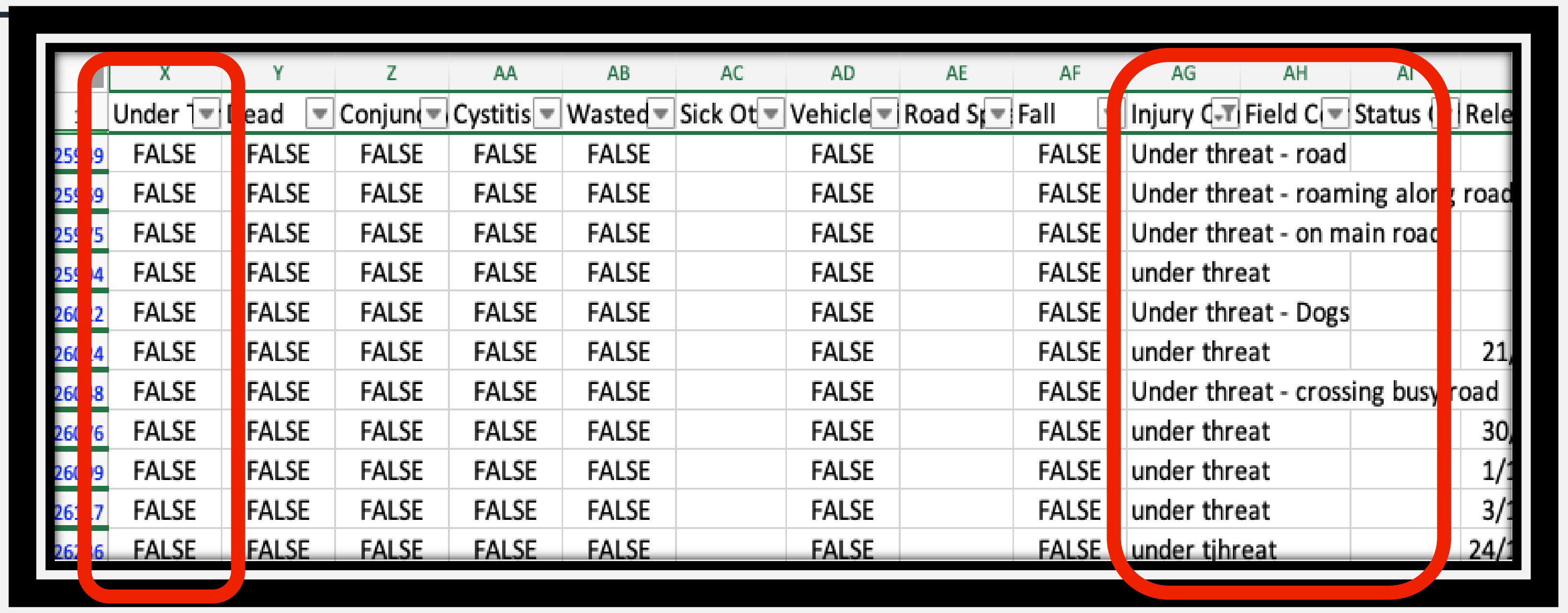
Picture is from https://www.wwf.org.au/what-we-do/species/koala

Data cleaning process

Incorrect records

Variables were **not recorded,** but comments wrote down the details

Examples	Cleaning Method
494 under threat but records were False	Change Under Threat > TRUE
116 orphan records were not recorded correctly, comments wrote mum dead	Change Orphaned > TRUE



	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	
1	Under T	ead	Conjun	Cystitis	Wasted	Sick Ot	Vehicle	Road Sp	Fall	Injury C	Field C	Status (	Rele
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	Under threat - road			
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	Under threat - roaming alo road			
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	Under threat - on main roa			
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	under threat			
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	Under threat - Dogs			
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	under threat			21
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	Under threat - crossing busy road			
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	under threat			30
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	under threat			1/
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	under threat			3/
[illegible]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE		FALSE	under tjhreat			24/

Imputations

	V	W	X	Y	AI
1	Caused By Dog	Orphan	Under T	Dead	Status Other
37856	FALSE	FALSE	TRUE	FALSE	suspect injury from dog
38086	FALSE	FALSE	FALSE	FALSE	suspect dog attack
48273	FALSE	FALSE	TRUE	FALSE	in same tree for weeks
48287	TRUE	FALSE	FALSE	FALSE	suspect dog attack
48492	FALSE	TRUE	FALSE	FALSE	mum euthanased - dog
48679	TRUE	TRUE	FALSE	FALSE	mum attacked by dogs
48719	FALSE	FALSE	TRUE	FALSE	suspect dog attack
48763	FALSE	FALSE	TRUE	FALSE	suspect dog attack
48775	TRUE	FALSE	FALSE	FALSE	suspect dog attack
48892	FALSE	FALSE	FALSE	FALSE	suspect dog attack
48910	FALSE	FALSE	FALSE	FALSE	suspect dog attack

Blank records

Specific reasons: Caused By Dog/ Orphaned/ Under Threat/ Conjunctivitis/ Cystitis Wasted/ Vehicle Hit/ Fall using **FALSE** or **TRUE** to fill, according to the comments' details

Variables we don't use

In this report, we use limited variables to do analysis, some of the columns we can use in **future study.**
Important variables involved but records don't make sense

Delete

Record no/ koala name/ Post code/ LAT/ LNG/ Adult Fate/ Other Adult situation/ Other Young Fate/ Other Koala/ Found Address/ Koala Location/ Description Road/ Speed Limit/ Release Date/ Release Location/ Release Suburb Release Post Code/ Release LAT/ Release LNG/ Field Comments
sick are blank and Injured are blank **88 records**

Add columns

S	T	U	V	Y
Vehicle Hit	Fall	Attacked by animals	BLIND	Caught in human place
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE

Important lost variables

301 records of attacked by animals (farm: mainly cows or horses) : create a new column Attacked by animals

306 records of blind: create a column Blind

create Cancer 75 records

create Bursitis 75 records

create 444 Caught in human place

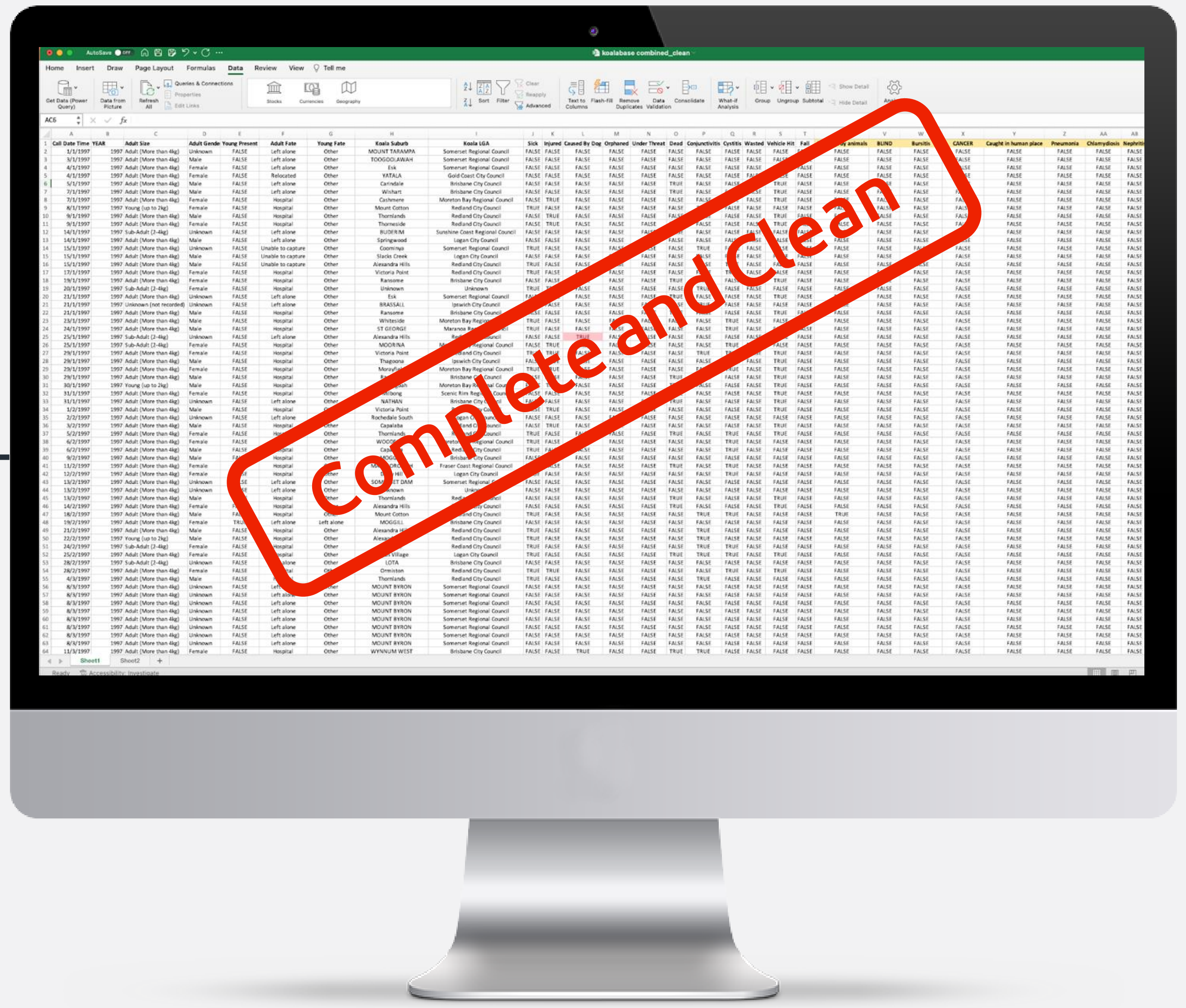
Making the data confess

Combine 3 main datasets

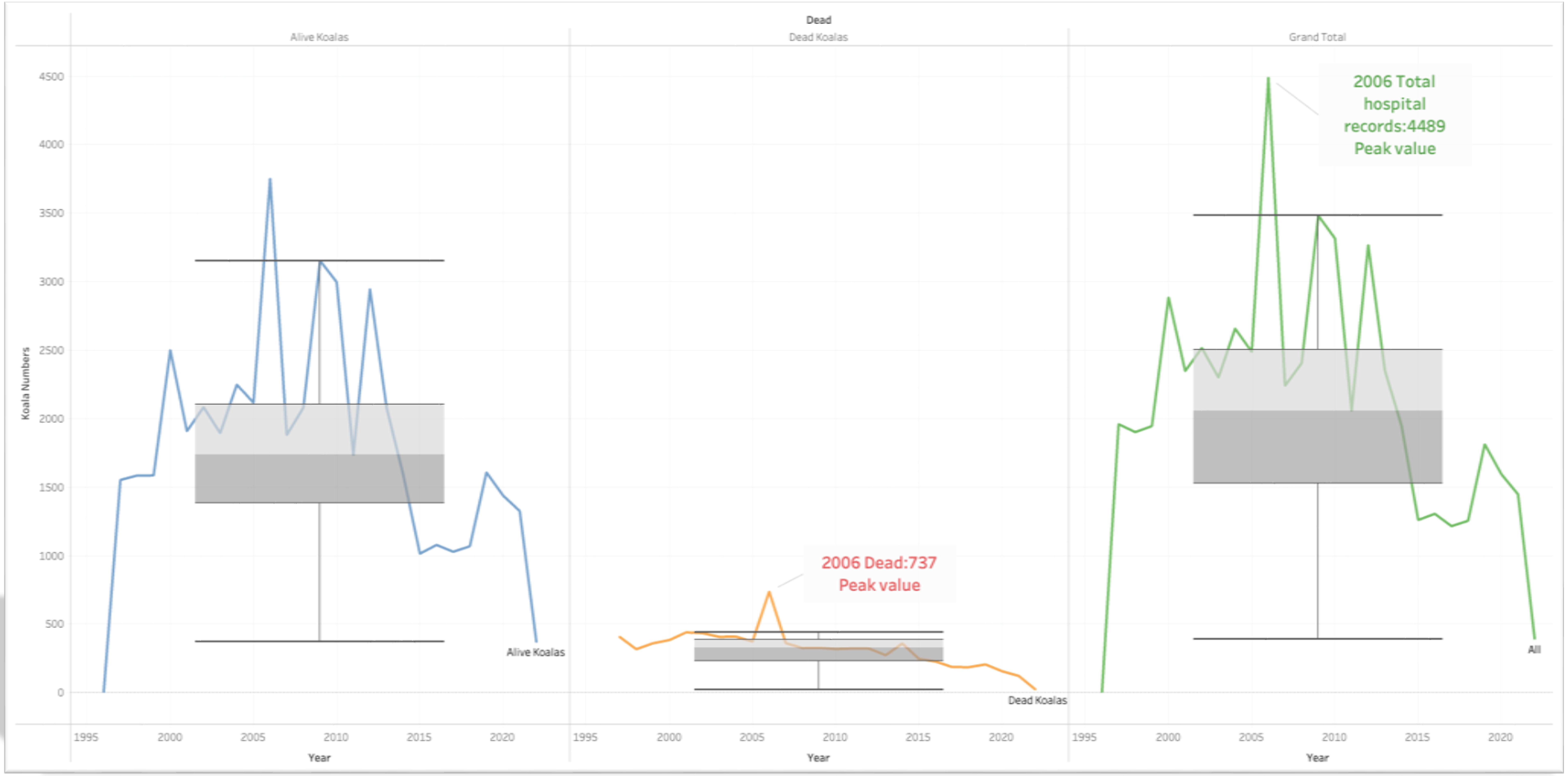
Final main dataset is: 56847 rows 28 columns

Dataset is unique correct and complete

Final dataset



Koala death records yearly trend



☑ Conclusion: 2006 was peak, then the trend is **decreasing**

General data analysis - Injury

S. No	Injured Koalas	Frequency	Proportion
1	Vehicle Hit	11329	19.92%
2	Caused By Dog	4150	7.30%
3	Fall	1045	1.83%
4	Caught in Human Place	443	0.77%
5	Attacked by Animals	301	0.53%
6	Injured	11263	19.81%

Here **top3** common causes of injury are **Vehicle Hit, Caused by Dog** and **Fall.**

Attacked by Animals and Caught in Human Place are less in number.

General data analysis - Sick

S. No	Causes of Sickness	Frequency	Proportion
1	Cystitis	11289	19.85%
2	Wasted	10918	19.20%
3	Conjunctivitis	8365	14.71%
4	Pneumonia	435	0.76%
5	BLIND	306	0.53%
6	Chlamydiosis	152	0.26%
7	CANCER	83	0.14%
8	Bursitis	75	0.13%
9	Nephritis	59	0.10%
10	Sick	20418	35.91%

Here, the **top3** leading causes of sickness are **Cystitis, wasted** and **Conjunctivitis**

While sickness like Cancer, Bursitis and Nephritis are very rare.

Variables correlation analysis

Injury

Sick

☑ Conclusion:

Most common sickness in Koalas are

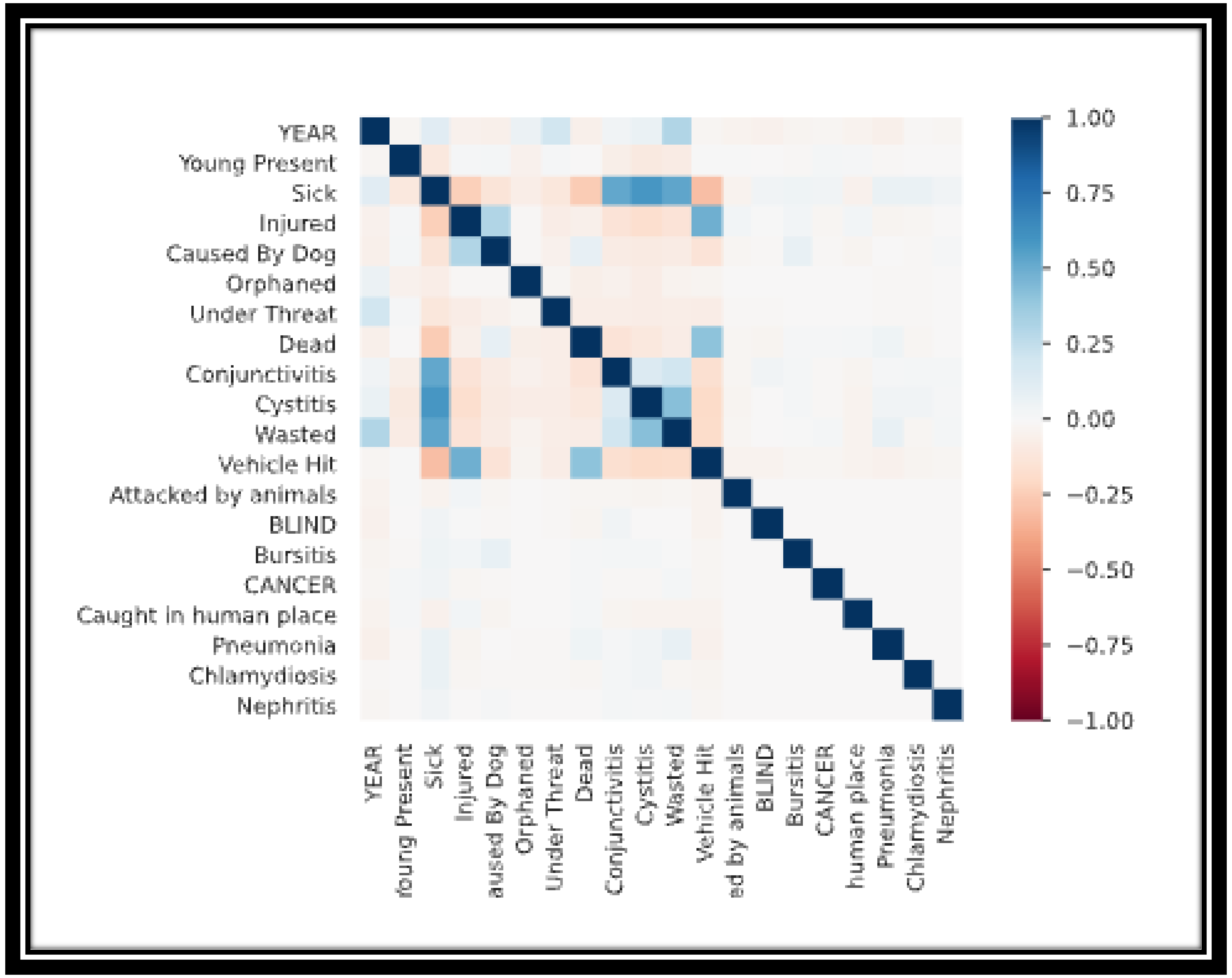
Conjunctivitis, Cystitis, Wasted.

Many of the attacks by dogs were when young were present and most of the injury to koalas were when they were attacked by animals.

Vehicle Hit caused Death

Sick also caused Death

Orphaned is not related to any factors.



Pearson’s correlation heatmap

Define variables

Independent variables

Injury is defined by these variables:
‘Caused By Dog’, ‘Vehicle Hit’, ‘Fall’, ‘Attacked by animals’, ‘Caught in human place’ ,'Under Threat‘ and ’Injured’ itself, in total 7.

Independent variables

Sick is defined by these variables:
'Conjunctivitis’, 'Cystitis’, 'Wasted’, 'Bursitis’, 'CANCER','Pneumonia','Chlamydiosis','BLIND' ,'Nephritis’and 'Sick’ itself, in total 10.

Dependent variable

Dead is the label ’Dead’.

Pictures are from https://www.wwf.org.au/what-we-do/species/koala

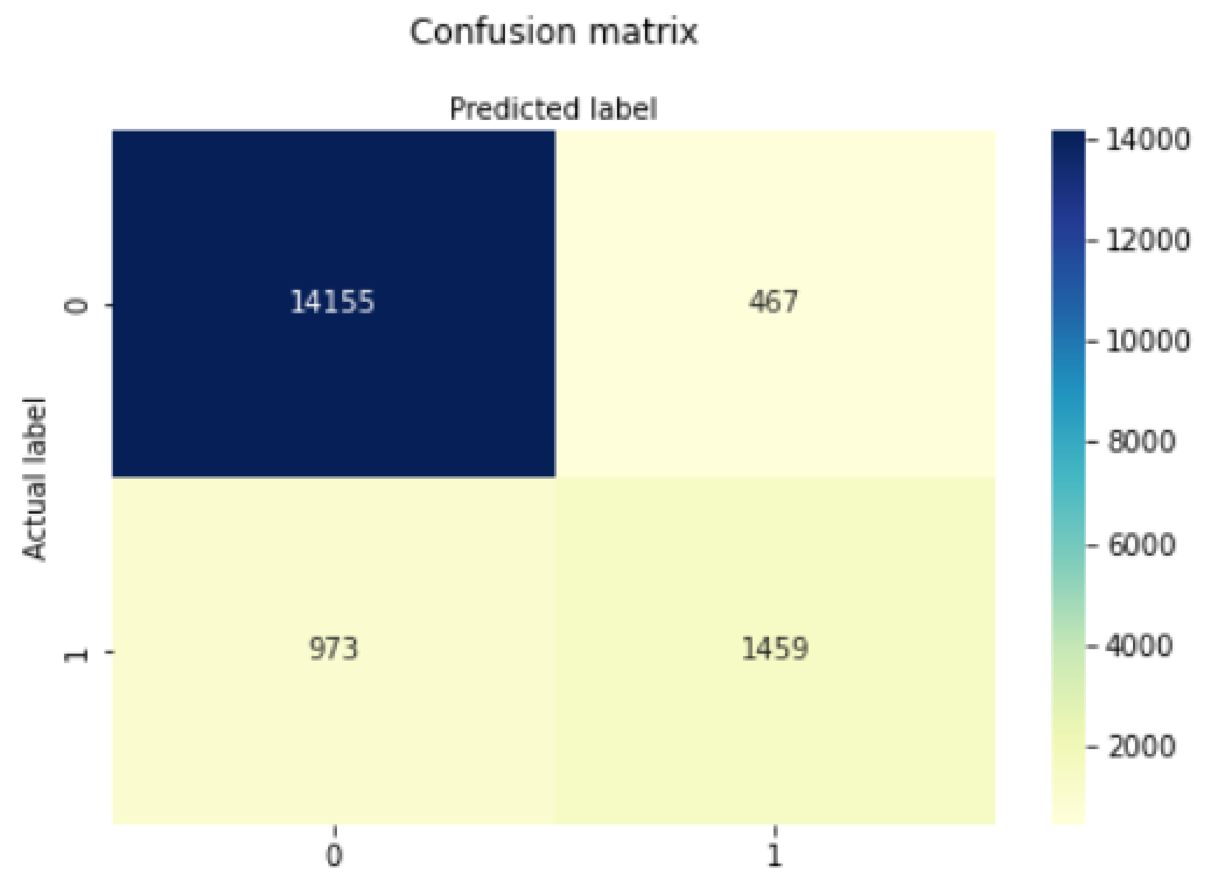
Logistic Regression1-1

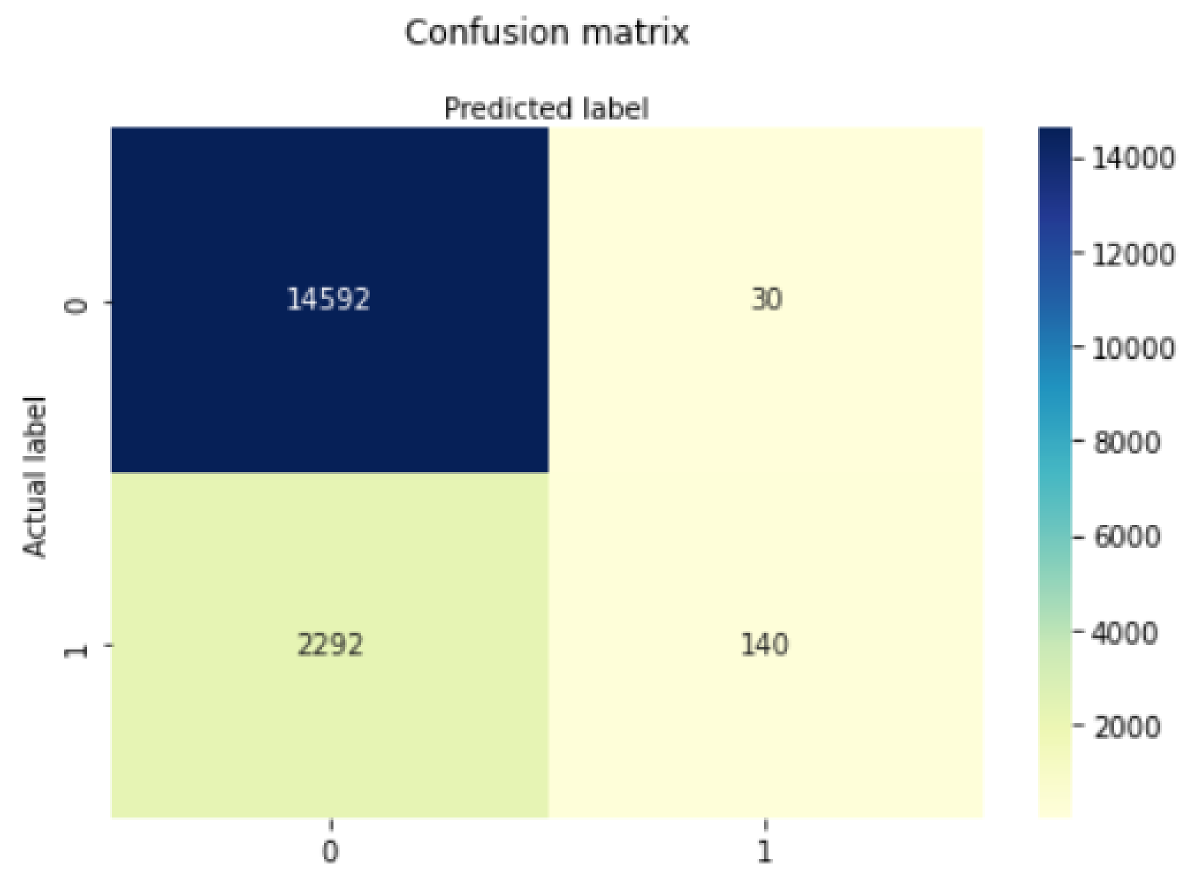
Training data 70%

Testing data 30%

Conclusion: Overall performances are good

Injury (all variables) predict Dead	Sick (all variables) predict Dead






```
array([[14592,    30],
       [ 2292,   140]])
```

	precision	recall	f1-score	support
Alive	0.94	0.97	0.95	14622
Dead	0.76	0.60	0.67	2432
accuracy			0.92	17054
macro avg	0.85	0.78	0.81	17054
weighted avg	0.91	0.92	0.91	17054

```
array([[14155,   467],
       [  973,  1459]])
```

	precision	recall	f1-score	support
Alive	0.86	1.00	0.93	14622
Dead	0.82	0.06	0.11	2432
accuracy			0.86	17054
macro avg	0.84	0.53	0.52	17054
weighted avg	0.86	0.86	0.81	17054

When logistic regression predict Koala's death by using **injury** is more **accurate.**

92% > 86%

While except recall (with dead) in the model, the other **performance** index are **acceptable.**

Model limitation:

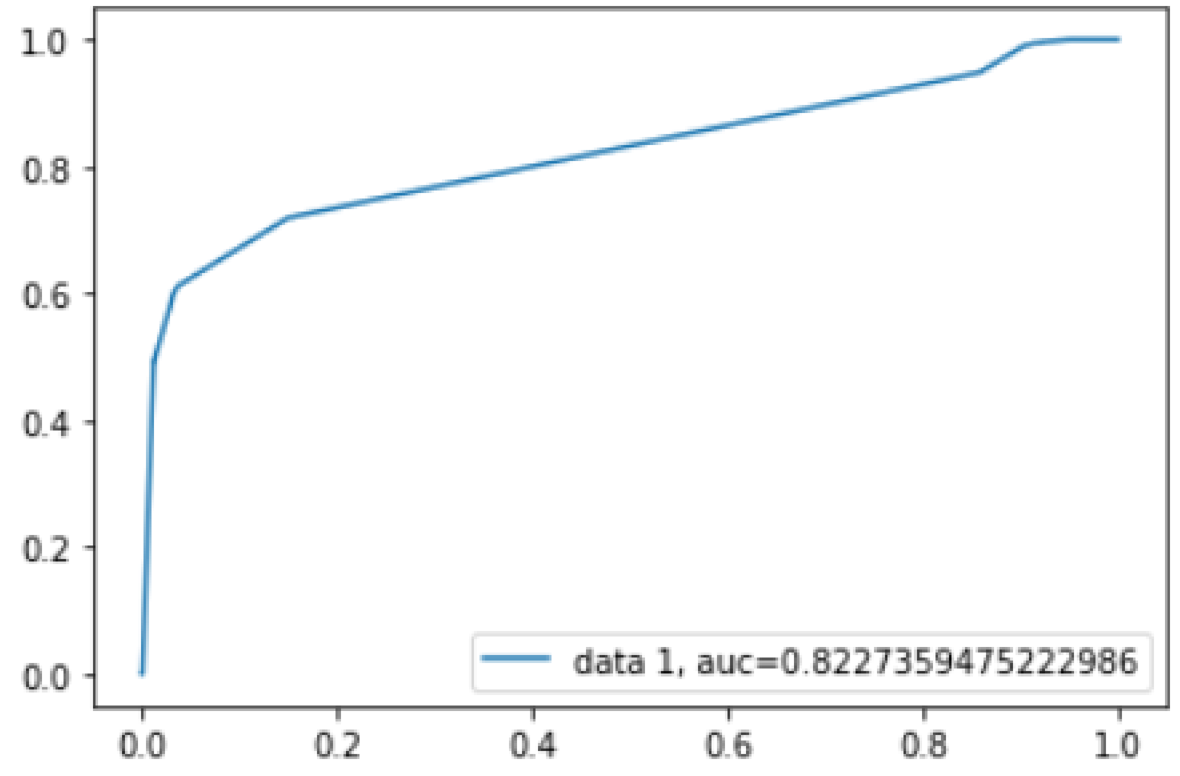
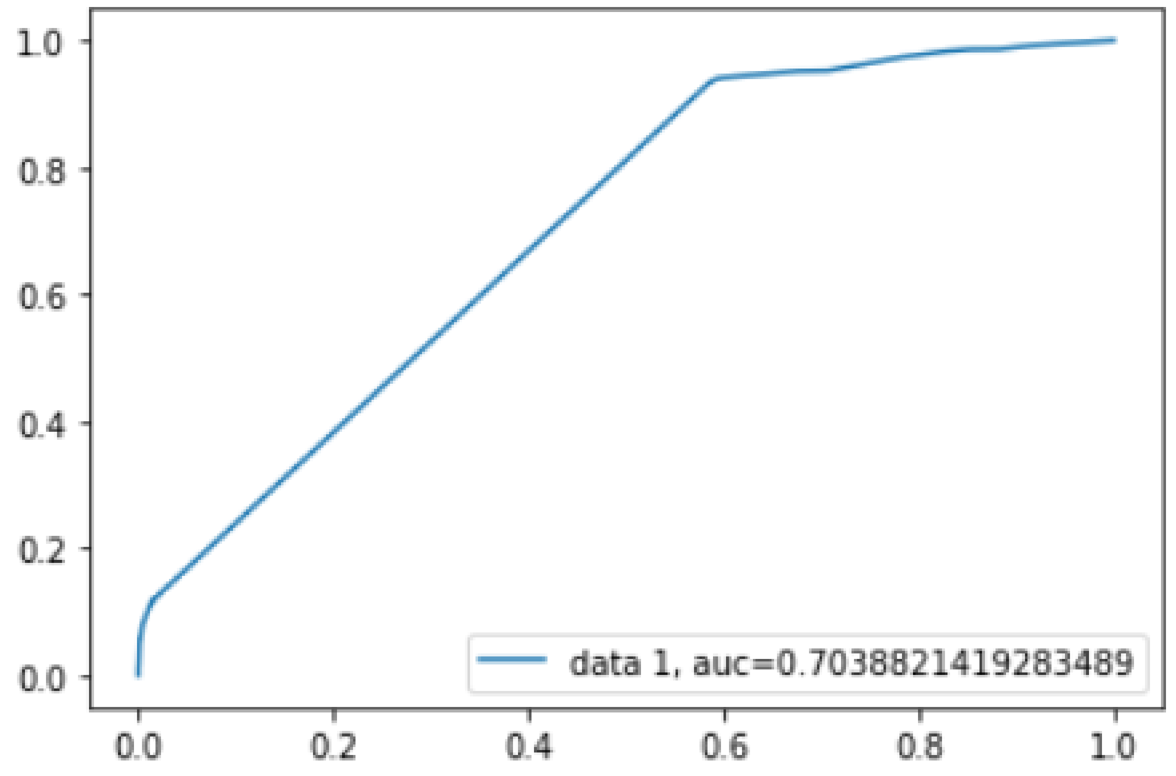
Recall of dead in Sick model is only **6%.**

Recall refers to the percentage of total relevant results correctly classified by your algorithm.

Logistic Regression1-2

Training data 70%

Testing data 30%

Injury (all variables) predict Dead ROC/AUC	Sick (all variables) predict Dead ROC/AUC
	

When we ROC/AUC to justify the performance of the logistic regression model, AUC of **Injury** model is **0.82,**

AUC of **Sick** model is **0.70**

☑ **Conclusion: ROC/AUC are also indicated good performance**

Statistical evaluations

Injury causes and sick causes using logistic regression

Logistic regression Injury causes

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	33989.504[a]	.328	.520

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CausedByDog	3.899	.043	8125.077	1	.000	49.371
	Orphaned	.104	.096	1.176	1	.278	1.110
	VehicleHit	3.772	.035	11778.471	1	.000	43.456
	Fall	3.605	.069	2705.934	1	.000	36.789
	Attackedbyanimals	2.740	.126	469.309	1	.000	15.486
	Caughtinhumanplace	2.521	.110	523.815	1	.000	12.443
	UnderThreat	-1.672	.280	35.660	1	.000	.188
	Constant	-3.413	.029	13719.983	1	.000	.033

a. Variable(s) entered on step 1: CausedByDog, Orphaned, VehicleHit, Fall, Attackedbyanimals, Caughtinhumanplace, UnderThreat.

conclusion

most influential factors of Injury are:
Caused by dog, vehicle hit and fall

most influential factors of sick are:
Chlamydiosis, Conjunctivitis, Bursitis

The two **R-squares** in the table explain the interval for the proportion of variation in the dependent variable that can be explained in this model.

[Cox&Snell R Square, Nagelkerke R Square]

Logistic regression Sick causes

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29198.068[a]	.547	.751

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Conjunctivitis	4.955	.058	7231.718	1	.000	141.832
	Cystitis	4.153	.044	8901.489	1	.000	63.649
	Wasted	3.123	.040	6105.004	1	.000	22.705
	BLIND	1.513	.173	76.377	1	.000	4.541
	Bursitis	4.207	.492	73.056	1	.000	67.130
	CANCER	3.132	.298	110.738	1	.000	22.925
	Pneumonia	1.826	.158	133.928	1	.000	6.212
	Chlamydiosis	5.155	.520	98.239	1	.000	173.230
	Nephritis	3.859	.509	57.557	1	.000	47.423
	Constant	-2.661	.021	16003.872	1	.000	.070

a. Variable(s) entered on step 1: Conjunctivitis, Cystitis, Wasted, BLIND, Bursitis, CANCER, Pneumonia, Chlamydiosis, Nephritis.

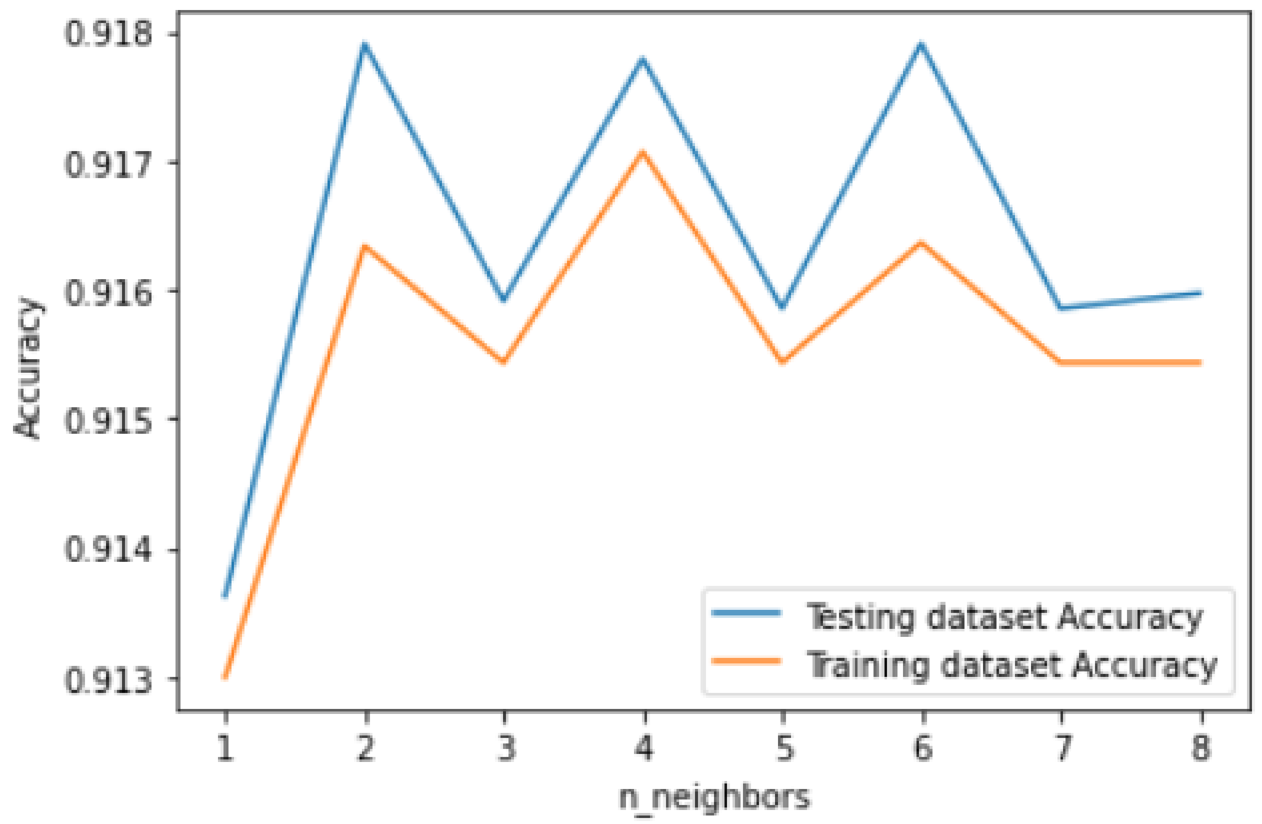
K nearest neighbors

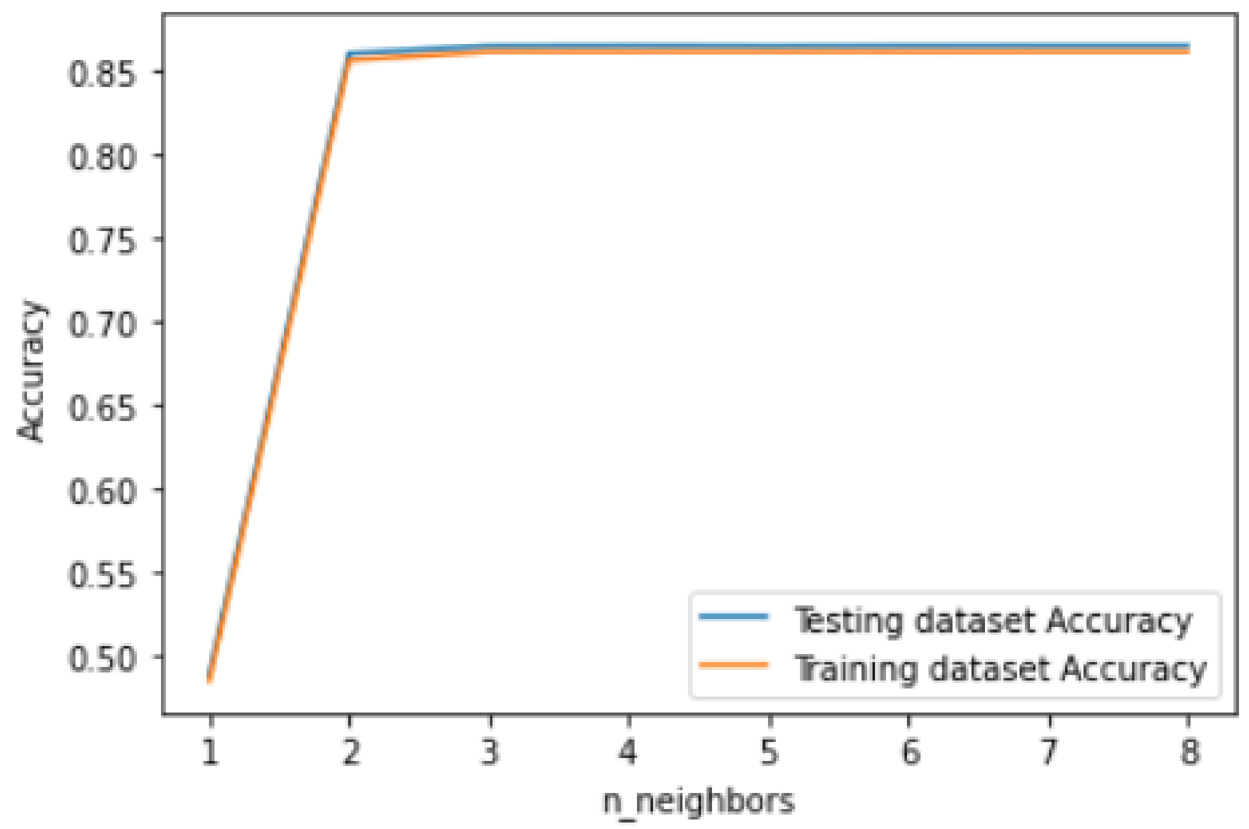
Training data 70%

Testing data 30%

Conclusion: Overall performances are good

Injury (all variables) predict Dead	Sick (all variables) predict Dead






```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, y_train)
# Predict on dataset which model has not seen before
print(knn.predict(X_test))
train_accuracy = knn.score(X_train, y_train)
print(train_accuracy)
test_accuracy= knn.score(X_test, y_test)
print(test_accuracy)
```

✓ 17.4s

```
[0 0 0 ... 0 1 0]
0.9163399678327302
0.9179078222117978
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, y_train)
# Predict on dataset which model has not seen before
print(knn.predict(X_test))
train_accuracy = knn.score(X_train, y_train)
print(train_accuracy)
test_accuracy= knn.score(X_test, y_test)
print(test_accuracy)
```

✓ 14.9s

```
[0 0 0 ... 0 0 0]
0.8563530357860877
0.8603846604902076
```

K nearest neighbors:

How to choose K?

Both the Injury model and Sick model are better when we choose K=2, according to the graph.

Injury model is more accurate than

Sick one, 91.8% > 86%

Both in training and testing data.

Models' comparison

Predicting death is a classification problem, so we tried four different algorithms and compared with the performances of models

Logistic Regression

1. Use Injury and the detailed variables to predict Dead.
2.Use Sick and related variables to predict Dead.

Evaluation index

Confusion matrix
ROC / AUC

KNN

1. Use Injury and the detailed variables to predict Dead.
2.Use Sick and related variables to predict Dead.

Evaluation index

Confusion matrix
K chosen

SVM

Evaluation index

```
# Model Precision: what percentage of positive tuples are lal
print("Precision:",metrics.precision_score(y_test, y_pred))
# Model Recall: what percentage of positive tuples are label
print("Recall:",metrics.recall_score(y_test, y_pred))
✓ 0.2s
Precision: 0.6073619631901841
Recall: 0.081414473684421052
```

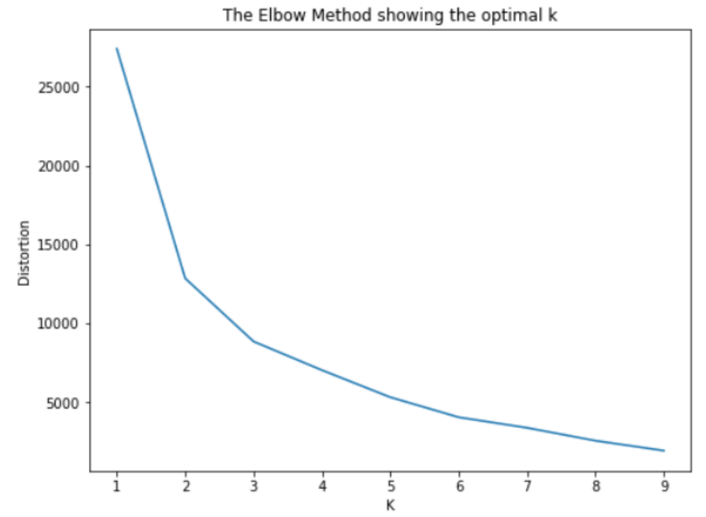
SVM: Precision:61% Recall:8.1%

K means

Evaluation index

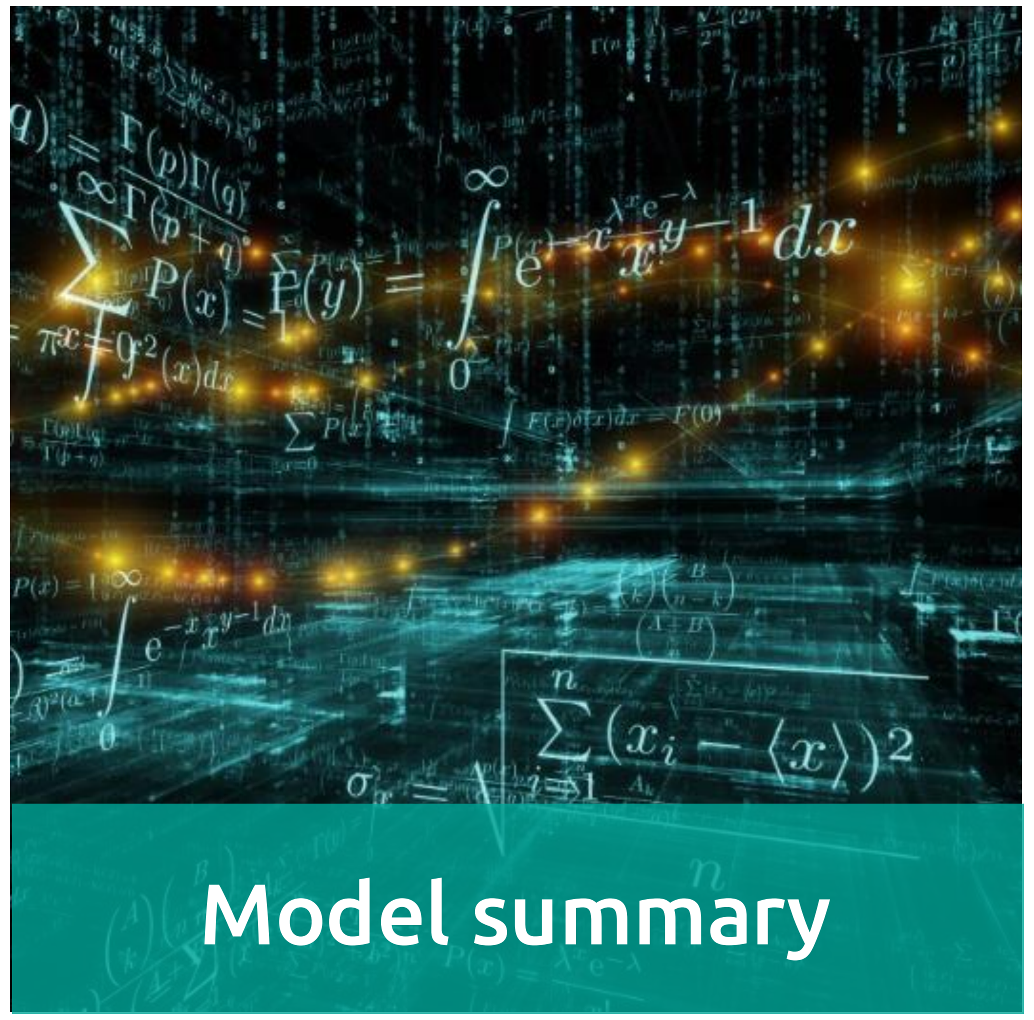
```
score = metrics.accuracy_score(y_test,kmeans.predict(X_test))
print(score)
✓ 0.2s
0.23548727571244282
```

K means Accuracy: 23.5% K=2



☑ **Both logistic Regression and K nearest neighbors (k=2) are good model to choose**

Storytelling and Advice



1. Main causes conclusion

- The main causes lead to koala's death are **Vehicle Hit** and **sick.**
- Most common sickness in Koalas are:

  Conjunctivitis, Cystitis ,Wasted, Chlamydiosis and Bursitis.
- most influential factors of Injury are:

  Caused by dog, vehicle hit and fall.

2. Model summary

- **Logistic regression** and **KNN** are two good performances models to predict death, both in injury and sick models.

3. Actions advice

- Animal Protection Organization: Fundings on koala sickness research.
- State Government: Take actions to protect koalas' home, such as put signs near roads of koalas' habitat.
- Local community: Koalas information telling.

4. Future study

- Using other variables (deleted ones) to do research, such as location clustering.
- Generate deep learning models to predict independent variables.

References and Improvements

No	Trail presentation feedback	Final presentation improvement
1	How does your project solve a problem? Don’t be general. Talk about specifics.	Page 2:Analyze the main causes leading koalas to death. Using different algorithm models to predict death.
2	data collection: you didn't explain.	Page 4: how we get our main datasets
3	5-data confess: predict death from injury and sickness. Why not use the type of injury and sickness? You can use a one-hot encoding of categorical variables (dependent) to predict death.	Page 12: Define variables
4	what are the most frequent causes of death? What types of injury/sickness?	Page 8-9: General data analysis - Injury/Sick Page 11: Vehicle Hit caused Death/Sick also caused Death
5	6-storytelling: The slides on deforestation have nothing to do with your analysis. Your conclusions should be taken from your analysis.	We abandon the deforestation data

[1] https://environment.des.qld.gov.au/wildlife/animals/living-with/koalas/facts

Thank You